Fabian Theis
Andrzej Cichocki
Arie Yeredor
Michael Zibulevsky (Eds.)

LNCS 7191

Latent Variable Analysis and Signal Separation

10th International Conference, LVA/ICA 2012
Tel Aviv, Israel, March 2012
Proceedings

Springer

Lecture Notes in Computer Science 7191

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Fabian Theis Andrzej Cichocki
Arie Yeredor Michael Zibulevsky (Eds.)

Latent Variable Analysis and Signal Separation

10th International Conference, LVA/ICA 2012
Tel Aviv, Israel, March 12-15, 2012
Proceedings

Springer

Volume Editors

Fabian Theis
Helmholtz Center Munich, Institute of Bioinformatics and Systems Biology
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
E-mail: fabian.theis@helmholtz-muenchen.de

Andrzej Cichocki
Riken Brain Science Institute, Laboratory for Advanced Brain Signal Processing
2-1, Hirosawa, Wako-shi, 351-0198 Saitama, Japan
E-mail: a.cichocki@riken.jp

Arie Yeredor
Tel-Aviv University, School of Electrical Engineering
Tel-Aviv 69978, Israel
E-mail: arie@eng.tau.ac.il

Michael Zibulevsky
Technion – Israel Institute of Technology, Department of Computer Science
Haifa 32000, Israel
E-mail: mzib@cs.technion.ac.il

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-28550-9 e-ISBN 978-3-642-28551-6
DOI 10.1007/978-3-642-28551-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012931865

CR Subject Classification (1998): C.3, I.4, I.5, I.6, F.2.2, H.5.5

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the full papers presented at the 10th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA 2012, which was held in Tel Aviv, Israel, during March 12–15, 2012, at the Sheraton Tel-Aviv Hotel and Towers.

The series began nearly 13 years ago under the title of Independent Component Analysis (ICA) workshops (held approximately every 18 months), and has attracted hundreds of participants over the years, continuously broadening its horizons. Starting with the fundamentals of ICA and Blind Source Separation (BSS) in the late 1990s and early 2000, the theme of the series has gradually expanded to include additional forms and models of general mixtures of latent variables, and was therefore re-titled Latent Variable Analysis (LVA) for the previous (9th) LVA/ICA conference in St. Malo (France) in 2010, keeping the acronym ICA as well (at least for a while), for reference to its roots and origins. This volume of Springer's *Lecture Notes on Computer Science* (LNCS) continues the tradition which began in ICA 2004 (held in Granada, Spain), to publish the conference proceedings in this form.

The submissions to LVA/ICA 2012 reflected the diversity of research fields covered by the call for papers, in accordance with the expanded scope of the theme of the series. Topics ranging from theoretical issues such as causality analysis and measures, through novel methods for employing the well-established concepts of sparsity and non-negativity for matrix and tensor factorization, down to a variety of related applications ranging from audio and biomedical signals to precipitation analysis, can all be found among the papers collected in this volume. In addition, LVA/ICA 2012 continued a tradition established in ICA 2009 (in Paraty, Brazil), to host presentations and discussions related to the Signal Separation Evaluation Campaign (SiSEC). SiSEC 2011 consisted of two types of tasks: audio source separation and biomedical data analysis. Several papers associated with submissions to SiSEC 2011 can be found in this volume.

Four world-renowned keynote speakers were invited by the Organizing Committee to present highlights of their recent research:

- Michael Elad (Technion, Israel Institute of Technology, Israel) on "The Analysis Sparse Model—Definition, Pursuit, Dictionary Learning, and Beyond"
- Lieven De Lathauwer (Katholieke Universiteit Leuven, Belgium) on "Block Component Analysis, a New Concept for Blind Source Separation"
- Amnon Shashua (Hebrew University of Jerusalem, Israel) on "The Applications of Tensor Factorization in Inference, Clustering, Graph Theory, Coding and Visual Representations"
- Paris Smaragdis (University of Illinois at Urbana-Champaign, USA) on "From Bases to Exemplars, from Separation to Understanding"

Continuing an initiative introduced at LVA/ICA 2010, the Organizing Committee announced a Late-Breaking / Demo session, for presentation of results and ideas that were not yet fully formalized and evaluated by the full-paper submission deadline, but were sufficiently ripe for presentation at the time of the conference. These included signal separation methods or systems evaluated in SiSEC 2011 but not associated with a full-paper submission to the conference. Submissions to this session were in the form of a "Title + Abstract" only, and are not included in the proceedings.

We received more than 80 full-paper submissions to regular sessions and to special sessions. Each submission of a regular full paper was peer reviewed by at least two members of our Technical Program Committee (TPC) or by competent sub-reviewers assigned by the TPC members. Most papers received three reviews, and some papers received four reviews. Submissions to the special Audio Sessions were peer reviewed by other participants of the same sessions. This volume contains the 20 full papers accepted for oral presentation and 42 full papers accepted for poster presentation, for the regular as well as for the special sessions. In addition, the volume contains the two overview papers of SiSEC 2011, and a paper by Lieven De Lathauwer associated with his keynote talk.

The growing share of audio-processing-related submissions prompted the successful organization of two dedicated special sessions, titled "Real-World Constraints and Opportunities in Audio Source Separation" and "From Audio Source Separation to Multisource Content Analysis." Moreover, the Organizing Committee decided to designate one full-day of the conference as an "Audio-Day," dedicated to the presentation of audio-related papers, including the keynote talk by Paris Smaragdis (but excluding the SiSEC-related contributions, which were presented as part of the SiSEC Special Sessions).

The Organizing Committee would like to extend its warm thanks to those who made LVA/ICA 2012 possible. First and foremost, these are the authors, and, of course, the members of the Program Committee and the sub-reviewers. In addition, we thank the members of the International ICA Steering Committee for their support and advice. We also thank the SiSEC Chairs for their close and fruitful collaboration. We are deeply indebted to the Faculty of Engineering at Bar-Ilan University, and especially to Sharon Gannot and to the Speech and Audio Lab for hosting the Audio Day. The organizing team at Ortra Ltd., and especially Sharon Lapid, were very helpful and always responsive. We also thank Springer and the LNCS team for their continued collaboration, and in particular Frank Holzwarth, Anna Kramer, Christine Reiss and Alfred Hofmann for their help and responsiveness. Finally, we would like to thank our sponsors, The Yitzhak and Chaya Weinstein Research Institute for Signal Processing, Tel Aviv University, Bar-Ilan University, The Technion - Israel Institute of Technology, and the Advanced Communication Center at Tel Aviv University.

January 2012

Andrzej Cichocki
Fabian Theis
Arie Yeredor
Michael Zibulevsky

Organization

Organizing Committee

General Chairs

Arie Yeredor	Tel-Aviv University, Israel
Michael Zibulevsky	Technion - Israel Institute of Technology, Israel

Program Chairs

Andrzej Cichocki	RIKEN Brain Science Institute, Japan
Fabian Theis	Helmholz Center Munich, Germany

SiSEC Evaluation Chairs

Shoko Araki	NTT, Japan
Guido Nolte	Fraunhofer Institute, Germany
Francesco Nesta	Fondazione Bruno Kessler IRST, Italy

Special Sessions

Emmanuel Vincent	INRIA, France
Pierre Leveau	Audionamix, France

Audio Day Arrangements

Sharon Gannot	Bar-Ilan University, Israel

Technical Program Committee

Tülay Adalı (USA)
Shoko Araki (Japan)
Cesar Caiafa (Argentina)
Jonathon Chambers (UK)
Fengyu Cong (Finland)
Sergio Cruces (Spain)
Yannick Deville (France)
Shuxue Ding (Japan)
Cédric Févotte (France)
Rémi Gribonval (France)
Christian Jutten (France)
Juha Karhunen (Finland)
Zbyněk Koldovský (Czeck Republic)
Elmar Lang (Germany)
Yuanqing Li (China)
Danilo Mandic (UK)
Ali Mansour (Australia)
Anke Meyer-Baese (USA)
Anh Huy Phan (Japan)
Mark Plumbley (UK)
Barnabás Póczos (USA)
Saeid Sanei (UK)
Xizhi Shi (China)
Paris Smaragdis (USA)
Petr Tichavský (Czeck Republic)
Ricardo Vigário (Finland)
Vincent Vigneron (France)
Emmanuel Vincent (France)
Vicente Zarzoso (France)
Liqing Zhang (China)
Guoxu Zhou (Japan)

Additional Reviewers

Pablo Aguilera Bonet
Francis Bach
Nancy Bertin
Andrzej Cichocki
Nicolas Dobigeon
Derry Fitzgerald
Jinyu Han
Rodolphe Jenatton
Hong Kook Kim
Matthieu Kowalski
Pierre Leveau
Yanfeng Liang
Morten Mørup
Pejman Mowlaee
Gautham J. Mysore
Alexey Ozerov
Auxiliadora Sarmiento Vega
Hiroshi Sawada
Fabian Theis
Hugo Van Hamme
Arie Yeredor
Michael Zibulevsky

International ICA Steering Committee

Mark Plumbley (UK) (Chair)
Tülay Adalı (USA)
Jean-François Cardoso (France)
Andrzej Cichocki (Japan)
Lieven De Lathauwer (Belgium)
Scott Douglas (USA)
Rémi Gribonval (France)
Christian Jutten (France)
Te-Won Lee (USA)
Shoji Makino (Japan)
Klaus Robert Müller (Germany)
Erkki Oja (Finland)
Paris Smaragdis (USA)
Fabian Theis (Germany)
Ricardo Vigário (Finland)
Emmanuel Vincent (France)
Arie Yeredor (Israel)

Sponsoring Institutions

The Yitzhak and Chaya Weinstein Research Institute for Signal Processing
Tel-Aviv University
The Technion - Israel Institute of Technology
Bar-Ilan University
The Advanced Communication Center

Table of Contents

General LVA/ICA Theory, Methods and Extensions

Sparsity, Sparse Coding and Dictionary Learning

Non-negative and Other Factorizations

Audio Separation and Analysis

SiSEC 2011 Evaluation Campaign

Other Applications

Block Component Analysis, a New Concept for Blind Source Separation

Lieven De Lathauwer

Katholieke Universiteit Leuven Campus Kortrijk,
E. Sabbelaan 53, 8500 Kortrijk, Belgium
Lieven.DeLathauwer@kuleuven-kortrijk.be
http://homes.esat.kuleuven.be/~delathau/

Abstract. The fact that the decomposition of a matrix in a minimal number of rank-1 terms is not unique, leads to a basic indeterminacy in factor analysis. Factors and loadings are only unique under certain assumptions. Working in a multilinear framework has the advantage that the decomposition of a higher-order tensor in a minimal number of rank-1 terms (its Canonical Polyadic Decomposition (CPD)) is unique under mild conditions. We have recently introduced Block Term Decompositions (BTD) of a higher-order tensor. BTDs write a given tensor as a sum of terms that have low multilinear rank, without having to be rank-1. In this paper we explain how BTDs can be used for factor analysis and blind source separation. We discuss links with Canonical Polyadic Analysis (CPA) and Independent Component Analysis (ICA). Different variants of the approach are illustrated with examples.

Keywords: Blind source separation, independent component analysis, canonical polyadic decomposition, block term decomposition, higher-order tensor, multilinear algebra.

1 Algebraic Tools

We start with a few basic definitions from multilinear algebra. These are subsequently used to define two tensor decompositions.

Definition 1. *A* mode-n vector *of an Nth-order tensor $\mathcal{T} = [t_{i_1 i_2 \ldots i_N}]$ is a vector obtained by varying the n-th index and keeping the other indices fixed.*

Definition 2. *The* multilinear rank *of an Nth-order tensor is the N-tuplet consisting of the dimension of the space spanned by the mode-1 vectors, the dimension of the space spanned by the mode-2 vectors, and so on.*

Definition 3. *The* (tensor) outer product $\mathcal{A} \otimes \mathcal{B}$ *of a tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times \ldots \times I_P}$ and a tensor $\mathcal{B} \in \mathbb{C}^{J_1 \times J_2 \times \ldots \times J_Q}$ is the tensor defined by $(\mathcal{A} \otimes \mathcal{B})_{i_1 i_2 \ldots i_P j_1 j_2 \ldots j_Q} = a_{i_1 i_2 \ldots i_P} b_{j_1 j_2 \ldots j_Q}$, for all values of the indices.*

For instance, the outer product $\mathcal{T}$ of three vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ is defined by $t_{ijk} = a_i b_j c_k$.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 1–8, 2012.

Definition 4. *An Nth-order tensor has* rank 1 *iff it equals the outer product of N nonzero vectors.*

Definition 5. *The* rank *of a tensor $\mathcal{T}$ is the minimal number of rank-*1 *tensors that yield $\mathcal{T}$ in a linear combination.*

We can now define a first basic tensor decomposition.

Definition 6. *A* Canonical Polyadic Decomposition (CPD) *of a rank-R tensor $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \times \cdots \times I_N}$ is a decomposition of $\mathcal{T}$ in a sum of R rank-1 terms:*

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \otimes \mathbf{a}_r^{(2)} \otimes \cdots \otimes \mathbf{a}_r^{(N)} \ . \tag{1}$$

The decomposition was for the first time used for data analysis in [3] and [14], where it was called Canonical Decomposition (CANDECOMP) and Parallel Factor Decomposition (PARAFAC), respectively. The term CPD, where "CP" may also stand for "CANDECOMP/PARAFAC", is now becoming more common. An important advantage over the decomposition of a matrix in rank-1 terms, is that CPD of a higher-order tensor is unique under mild conditions, see [11,16,17,20,21] and references therein. (Uniqueness is up to permutation of terms and scaling/counterscaling of factors within a term.) For algorithms, see [11,16,22,23] and references therein.

Consider a third-order tensor $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ that has CPD

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r \ . \tag{2}$$

Define $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \ldots \ \mathbf{a}_R] \in \mathbb{C}^{I_1 \times R}$, $\mathbf{B} = [\mathbf{b}_1 \ \mathbf{b}_2 \ \ldots \ \mathbf{b}_R] \in \mathbb{C}^{I_2 \times R}$ and $\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \ldots \ \mathbf{c}_R] \in \mathbb{C}^{I_3 \times R}$. Eq. (2) is often written as

$$\mathbf{T}_{:,:,i_3} = \mathbf{A} \cdot \operatorname{diag}(c_{i_3 1}, c_{i_3 2}, \ldots, c_{i_3 R}) \cdot \mathbf{B}^T, \qquad 1 \leqslant i_3 \leqslant I_3 \ , \tag{3}$$

in which we use MATLAB colon notation. We see that all slices $\mathbf{T}_{:,:,i_3}$ are linear combinations of the same rank-1 terms $\mathbf{a}_r \mathbf{b}_r^T$, $1 \leqslant r \leqslant R$, where the coefficients are given by the entries of $\mathbf{C}$.

In [8,9,13] we introduced Block Term Decompositions (BTD) of a higher-order tensor. BTDs are a generalization of CPD. A specific case is the following.

Definition 7. *A* decomposition of a tensor $\mathcal{T} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$ in a sum of rank-$(L_r, L_r, 1)$ terms*, $1 \leqslant r \leqslant R$, is a decomposition of $\mathcal{T}$ of the form*

$$\mathcal{T} = \sum_{r=1}^{R} (\mathbf{A}_r \cdot \mathbf{B}_r^T) \otimes \mathbf{c}_r \ , \tag{4}$$

in which each of the matrices $\mathbf{A}_r \in \mathbb{C}^{I_1 \times L_r}$ and $\mathbf{B}_r \in \mathbb{C}^{I_2 \times L_r}$ has linearly independent columns and in which the vectors $\mathbf{c}_r \in \mathbb{C}^{I_3}$ are nonzero, $1 \leqslant r \leqslant R$. We assume that R is minimal.

Conditions under which this decomposition is unique, have been established in [8,9,10]. (Here, uniqueness is up to permutation of terms, scaling/counterscaling of factors within a term, and post-multiplication of $\mathbf{A}_r$ by a square nonsingular matrix $\mathbf{W}_r$ provided $\mathbf{B}_r^T$ is pre-multiplied by $\mathbf{W}_r^{-1}$, $1 \leqslant r \leqslant R$.) Algorithms have been presented in [13,18,19,22]. Note that $(L_r, L_r, 1)$ is the multilinear rank of the r-th term.

Define $\mathbf{A} = [\mathbf{A}_1 \ \mathbf{A}_2 \ \dots \ \mathbf{A}_R] \in \mathbb{C}^{I_1 \times \sum_r L_r}$, $\mathbf{B} = [\mathbf{B}_1 \ \mathbf{B}_2 \ \dots \ \mathbf{B}_R] \in \mathbb{C}^{I_2 \times \sum_r L_r}$ and $\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \dots \ \mathbf{c}_R] \in \mathbb{C}^{I_3 \times R}$. Eq. (4) can also be written as

$$\mathbf{T}_{:,:,i_3} = \mathbf{A} \cdot \operatorname{diag}(c_{i_3 1}\mathbf{I}_{L_1 \times L_1}, c_{i_3 2}\mathbf{I}_{L_2 \times L_2}, \dots, c_{i_3 R}\mathbf{I}_{L_R \times L_R}) \cdot \mathbf{B}^T, \quad 1 \leqslant i_3 \leqslant I_3 \ . \tag{5}$$

All slices $\mathbf{T}_{:,:,i_3}$ are linear combinations of the same rank-L_r matrices $\mathbf{A}_r\mathbf{B}_r^T$, $1 \leqslant r \leqslant R$, where the coefficients are given by the entries of $\mathbf{C}$.

In the next section we explain how CPD and decomposition in rank-$(L_r, L_r, 1)$ terms can be used for blind source separation.

2 Block Component Analysis: The Concept

Factor analysis and blind source separation aim at decomposing a data matrix $\mathbf{X} \in \mathbb{C}^{K \times N}$ into a sum of interpretable rank-1 terms:

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S}^T = \sum_{r=1}^{R} \mathbf{a}_r \mathbf{s}_r^T \ . \tag{6}$$

Here, $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_R] \in \mathbb{C}^{K \times R}$ is the unknown mixing matrix and the columns of $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \dots \ \mathbf{s}_R] \in \mathbb{C}^{N \times R}$ are the unknown sources. (We consider the noiseless case for clarity of exposition.) Since the decomposition of a matrix is not unique, some assumptions need to be made. In Independent Component Analysis (ICA) the most important assumption is that the sources are mutually statistically independent [5,6,15].

If we dispose of a data tensor, then things are simpler, in the sense that the decomposition in rank-1 terms is unique under mild conditions, as mentioned above. This uniqueness makes CPD a powerful tool for data analysis [4,17,21]. ICA can actually be seen as a form of Canonical Polyadic Analysis (CPA). Namely, algebraic methods for ICA typically rely on the CPD of a higher-order cumulant tensor or a third-order tensor in which a set of covariance matrices is stacked. The links are explicitly discussed in [11].

The crucial observation on which Block Component Analysis (BCA) is based, is that also the constraints in CPA are in a certain sense restrictive. Namely, in (3) the matrix slices are decomposed in terms that are rank-1, i.e., they consist of the outer product of two vectors. One could wish to decompose the slices in terms that just have low rank, since the latter enable the modelling of more general phenomena. As explained, this corresponds to the decomposition of a tensor in rank-$(L_r, L_r, 1)$ terms, which is still unique under certain conditions. Probably CPA owes much of its success to rank-1 terms that capture the essence of components that actually are more complex. In such cases it could be of interest to check whether BCA provides more detail.

BCA can be applied to matrix data as well. First, the data need to be tensorized. For instance, one could map the rows of $\mathbf{X}$ to $(I \times J)$ Hankel matrices, with $I + J - 1 = N$, yielding a tensor $\mathcal{X} \in \mathbb{C}^{I \times J \times K}$. Formally, we define:

$$(\mathcal{X})_{ijk} = (\mathbf{X})_{k,i+j-1}, \qquad 1 \leqslant i \leqslant I,\ 1 \leqslant j \leqslant J,\ 1 \leqslant k \leqslant K. \tag{7}$$

Since the mapping is linear, we have:

$$\mathcal{X} = \sum_{r=1}^{R} \mathbf{H}_r \otimes \mathbf{a}_r, \tag{8}$$

in which $\mathbf{H}_r \in \mathbb{C}^{I \times J}$ is the Hankel matrix associated with the r-th source, $1 \leqslant r \leqslant R$. An interesting property is that, for a sufficient number of samples, Hankel matrices associated with exponential polynomials have low rank. Exponential polynomials are functions that can be written as sums and/or products of exponentials, sinusoids and/or polynomials. BCA allows the blind separation of such signals, provided decomposition (8) is unique. Uniqueness conditions guarantee that the components are sufficiently different to allow separation, which in turn implies a bound on the number of components one can deal with. Also, there is a trade-off between complexity, measured by rank L_r, and number of components. For theory underlying the blind separation of exponential polynomials by means of a decomposition in rank-$(L_r, L_r, 1)$ terms, we refer to [10].

Hankelization is just one way to tensorize matrix data. What is essential is that we use a linear transformation that maps the sources to matrices that (approximately) have low rank. Possible alternatives are spectrograms, wavelet representations, etc. For comparison we repeat that in ICA the problem is typically tensorized through the computation of higher-order statistics or sets of second-order statistics.

In the next section we illustrate the principle of BCA by means of examples.

3 Illustration

3.1 Toy Example: Audio

We consider the following sources: $\mathbf{s}_1$ consists of samples 50–80 of the chirp demo signal and $\mathbf{s}_2$ consists of samples 250–280 of the train demo signal in MATLAB (version 7.13). These two signals are shown in Fig. 1. The singular values of the corresponding Hankel matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{R}^{16 \times 16}$ are shown in Fig. 2. We see that $\mathbf{H}_1$ and $\mathbf{H}_2$ can be very well approximated by low-rank matrices. The entries of $\mathbf{A} \in \mathbb{R}^{5 \times 2}$ are drawn from a zero-mean unit-variance Gaussian distribution. Hankelization of $\mathbf{X} \in \mathbb{R}^{5 \times 31}$ yields a tensor $\mathcal{X}^{(H)} \in \mathbb{R}^{16 \times 16 \times 5}$. We also map $\mathbf{X}$ to a tensor $\mathcal{X}^{(W)} \in \mathbb{R}^{40 \times 31 \times 5}$ by means of the biorthogonal spline wavelet 1.3 [7]. This transformation maps every observed time signal to a (scale × time) matrix, where we take the scale values equal to $0.8/(0.05s)$, $1 \leqslant s \leqslant 40$. The singular values of the wavelet representations $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{40\times 31}$ of the two sources are shown in Fig. 2. Tensors $\mathcal{X}^{(H)}$ and $\mathcal{X}^{(W)}$ are decomposed in a sum of a rank-$(L_1, L_1, 1)$ and a rank-$(L_2, L_2, 1)$ term. We conduct a Monte Carlo experiment consisting of 100 runs for different values of L_1 and L_2. The mean and median Signal-to-Interference Ratio (SIR) are shown in Table 1. This table demonstrates that BCA allows one to accurately separate the sources. Moreover, the choice of L_1 and L_2 turns out not to be very critical. The ICA algorithm in [5] yields a mean and median SIR of only 15 dB, due to the fact that in this toy example not enough samples are available to allow the reliable estimation of statistics.

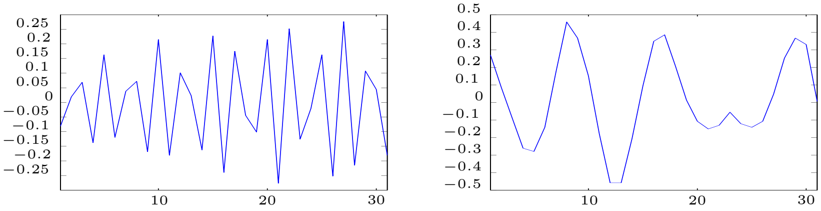


Fig. 1. Chirp (left) and train (right) audio source

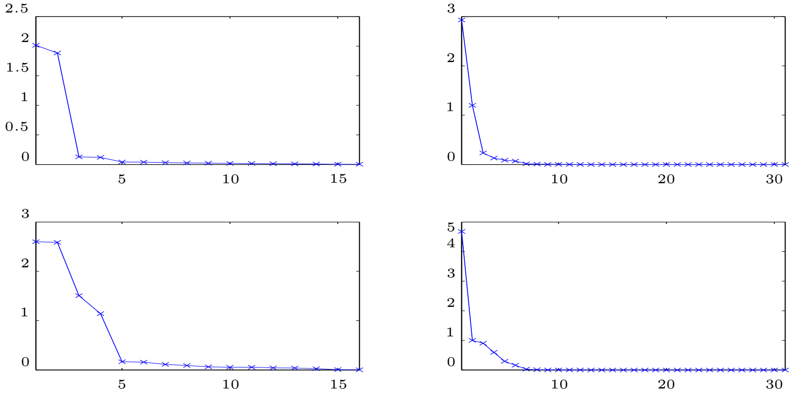


Fig. 2. Left: singular values of Hankel matrices $\mathbf{H}_1$ (top) and $\mathbf{H}_2$ (bottom). Right: singular values of wavelet matrices $\mathbf{W}_1$ (top) and $\mathbf{W}_2$ (bottom).

We next add zero-mean Gaussian noise to the observations and investigate the effect of the Signal-to-Noise Ratio (SNR) on the quality of the separation. We conduct a new Monte Carlo simulation consisting of 100 runs. The value of $L_1 = L_2 = L$ is varied between 1 and 4. The results are shown in Fig. 3. A rank-1 structure turns out to be too simple, at least in the Hankel case.

In the Hankel setting, the signals that correspond to rank-1 matrices are complex exponentials (one frequency, one damping factor). A rank-1 term is

Table 1. Mean (median) SIR [dB] in the noiseless audio example, as a function of L_1 and L_2 (top: Hankel-based BCA, bottom: wavelet-based BCA)

L_1 / L_2	1	2	3	4	5	6	7
1	20 (49)	48 (47)	49 (49)	37 (51)	20 (51)	15 (19)	15 (13)
	43 (43)	41 (41)	19 (41)	19 (41)	16 (16)	14 (13)	13 (12)
2	48 (47)	47 (47)	49 (50)	48 (49)	44 (51)	17 (38)	16 (22)
	41 (41)	46 (46)	47 (49)	48 (52)	33 (33)	17 (18)	14 (17)
3	49 (49)	49 (50)	49 (49)	47 (48)	23 (49)	20 (47)	19 (45)
	19 (41)	47 (49)	46 (46)	27 (41)	25 (32)	18 (33)	13 (12)
4	37 (51)	48 (49)	47 (48)	47 (47)	47 (48)	20 (46)	18 (44)
	19 (41)	48 (52)	27 (41)	52 (52)	16 (21)	47 (48)	13 (12)
5	20 (51)	44 (51)	23 (49)	47 (48)	45 (48)	29 (46)	16 (44)
	16 (16)	33 (33)	25 (32)	16 (21)	13 (13)	28 (35)	12 (12)
6	15 (19)	17 (38)	20 (47)	20 (46)	29 (46)	25 (46)	33 (47)
	14 (13)	17 (18)	18 (33)	47 (48)	28 (35)	46 (47)	17 (25)
7	15 (13)	16 (22)	19 (45)	18 (44)	16 (44)	33 (47)	24 (44)
	13 (12)	14 (17)	13 (12)	13 (12)	12 (12)	17 (25)	17 (20)

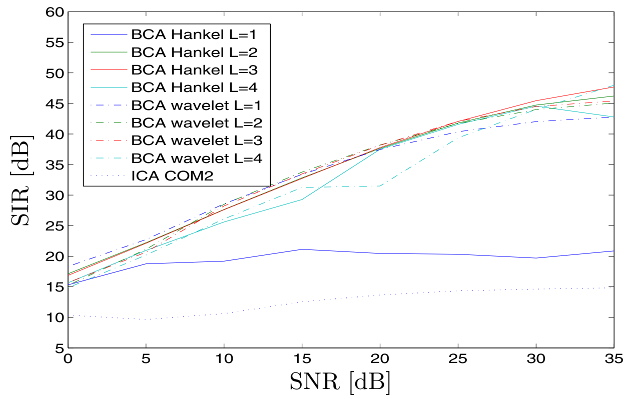


Fig. 3. Mean SIR as a function of SNR in the audio example

sometimes called an atom, since it is a constituent element that cannot be split into smaller parts. In this terminology, CPA consists of splitting a data tensor into atoms. On the other hand, one could say that sounds or melodies, having a certain spectral content, correspond to molecules rather than atoms. BCA is then the separation at the level of molecules.

3.2 Application in Wireless Communication

In spread-spectrum systems that employ an antenna array at the receiver, the received data are naturally represented by the third-order tensor that shows the signal along the temporal, spectral and spatial axis. In [20] it was shown for Direct Sequence - Code Division Multiple Access (DS-CDMA) systems that, in simple propagation scenarios that do not cause Inter-Symbol-Interference (ISI), every user contributes a rank-1 term to the received data. Consequently, in a non-cooperative setting multiple access can be realized through the computation of a CPD. In propagation scenarios that do involve ISI, rank-1 terms are a too restrictive model. It was shown in [12] that, when reflections only take place in the far field of the receive array, multiple access can be realized through the computation of a decomposition in rank-$(L_r, L_r, 1)$ terms. In [18] a more general type of BTD was used to deal with cases where reflections do not only take place in the far field. The same ideas can be applied to other systems with at least triple diversity.

4 Discussion and Conclusion

CPA makes a strong assumption on the components that one looks for, namely, that they are rank-1. In the analysis of text data, web documents, biomedical data, images, ... it is often questionable whether this assumption is satisfied. Low (multilinear) rank may be a better approximation of reality. In this paper we introduced BCA as an analysis technique based on the computation of BTDs. BCA can be used for the analysis of matrix data, after these have been tensorized. To this end, one can compute statistics, like in ICA, but one can also consider Hankel representations, wavelet representations, etc. Deterministic variants of BCA may be useful for the analysis of short data sequences.

BCA is related to Sparse Component Analysis (SCA) [1]. In SCA, the sources are low-dimensional in the sense that they are most often zero. In BCA, the sources have a low intrinsic dimension, characterized by multilinear rank. BCA is also related to compressive sensing [2]. In compressive sensing, low intrinsic dimensionality is used for compact signal representation. In BCA, it is used as the basis for signal separation.

In this paper we limited ourselves to the decomposition in rank-$(L_r, L_r, 1)$ terms. In [8,9,13] more general types of BTD were introduced, which allow a more general analysis.

Acknowledgments. Research supported by: (1) Research Council K.U.Leuven: GOA-MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), CIF1 and STRT1/08/023, (2) F.W.O.: Research Communities ICCoS, ANMMM and MLDM, (3) the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, "Dynamical systems, control and optimization", 2007–2011), (4) EU: ERNSI.

References

1. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. 51(1), 34–81 (2009)
2. Candes, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. on Information Theory 52(2), 489–509 (2006)
3. Carroll, J.D., Chang, J.J.: Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart–Young" decomposition. Psychometrika 35(3), 283–319 (1970)

4. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Wiley & Sons (2009)
5. Comon, P.: Independent Component Analysis, a new concept? Signal Processing 36(3), 287–314 (1994)
6. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press (2010)
7. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61. SIAM (1994)
8. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms — Part I: Lemmas for partitioned matrices. SIAM J. Matrix Anal. Appl. 30, 1022–1032 (2008)
9. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms — Part II: Definitions and uniqueness. SIAM J. Matrix Anal. Appl. 30, 1033–1066 (2008)
10. De Lathauwer, L.: Blind separation of exponential polynomials and the decomposition of a tensor in rank-$(L_r, L_r, 1)$ terms. SIAM J. Matrix Anal. Appl. 32(4), 1451–1474 (2011)
11. De Lathauwer, L.: A short introduction to tensor-based methods for factor analysis and blind source separation. In: Proc. 7th Int. Symp. on Image and Signal Processing and Analysis (ISPA 2011), Dubrovnik, Croatia, September 4–6, pp. 558–563 (2011)
12. De Lathauwer, L., de Baynast, A.: Blind deconvolution of DS-CDMA signals by means of decomposition in rank-$(1, L, L)$ terms. IEEE Trans. on Signal Processing 56(4), 1562–1571 (2008)
13. De Lathauwer, L., Nion, D.: Decompositions of a higher-order tensor in block terms — Part III: Alternating least squares algorithms. SIAM J. Matrix Anal. Appl. 30, 1067–1083 (2008)
14. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics 16 (1970)
15. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2001)
16. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. 51(3), 455–500 (2009)
17. Kroonenberg, P.M.: Applied Multiway Data Analysis. Wiley (2008)
18. Nion, D., De Lathauwer, L.: Block component model based blind DS-CDMA receivers. IEEE Trans. on Signal Processing 56(11), 5567–5579 (2008)
19. Nion, D., De Lathauwer, L.: An enhanced line search scheme for complex-valued tensor decompositions. Application in DS-CDMA. Signal Processing 88(3), 749–755 (2008)
20. Sidiropoulos, N.D., Giannakis, G.B., Bro, R.: Blind PARAFAC receivers for DS-CDMA systems. IEEE Trans. Signal Process. 48(3), 810–823 (2000)
21. Smilde, A., Bro, R., Geladi, P.: Multi-way Analysis with Applications in the Chemical Sciences. John Wiley & Sons, Chichester (2004)
22. Sorber, L., Van Barel, M., De Lathauwer, L.: Optimization-based algorithms for tensor decompositions: Canonical Polyadic Decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms and a new generalization. Tech. rep. 2011-182, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)
23. Tomasi, G., Bro, R.: A comparison of algorithms for fitting the PARAFAC model. Comp. Stat. & Data Anal. 50(7), 1700–1734 (2006)

Partially Linear Estimation with Application to Image Deblurring Using Blurred/Noisy Image Pairs

Tomer Michaeli*, Daniel Sigalov, and Yonina C. Eldar

Technion – Israel Institute of Technology,
Haifa, 32000, Israel
{tomermic,dansigal}@tx.technion.ac.il, yonina@ee.technion.ac.il

Abstract. We address the problem of estimating a random vector X from two sets of measurements Y and Z, such that the estimator is linear in Y. We show that the partially linear minimum mean squared error (PLMMSE) estimator requires knowing only the second-order moments of X and Y, making it of potential interest in various applications. We demonstrate the utility of PLMMSE estimation in recovering a signal, which is sparse in a unitary dictionary, from noisy observations of it and of a filtered version of it. We apply the method to the problem of image enhancement from blurred/noisy image pairs. In this setting the PLMMSE estimator performs better than denoising or deblurring alone, compared to state-of-the-art algorithms. Its performance is slightly worse than joint denoising/deblurring methods, but it runs an order of magnitude faster.

Keywords: Bayesian estimation, minimum mean squared error, linear estimation.

1 Introduction

Bayesian estimation is concerned with the prediction of a random quantity X based on a set of observations Y, which are statistically related to X. It is well known that the estimator minimizing the mean squared error (MSE) is given by the conditional expectation $\hat{X} = \mathbb{E}[X|Y]$. There are various scenarios, however, in which the minimal MSE (MMSE) estimator cannot be used. This can either be due to implementation constraints, because of the fact that no closed form expression for $\mathbb{E}[X|Y]$ exists, or due to lack of complete knowledge of the joint distribution of X and Y. In these cases, one often resorts to linear estimation. The appeal of the linear MMSE (LMMSE) estimator is rooted in the fact that it possesses an easily implementable closed form expression, which merely requires knowledge of the joint first- and second-order moments of X and Y.

For example, the amount of computation required for calculating the MMSE estimate of a jump-Markov Gaussian random process from its noisy version

* This work was supported in part by a Google Research Award.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 9–16, 2012.

grows exponentially in time. By contrast, the LMMSE estimator in this setting possesses a simple recursive implementation, similar to the Kalman filter [1]. A similar problem arises in the area of sparse representations, in which the use of Bernoulli-Gaussian and Laplacian priors is very common. The complexity of calculating the MMSE estimator under the former prior is exponential in the vector's dimension, calling for approximate solutions [2,3]. The MMSE estimator under the latter prior does not possess a closed form expression [4], which has motivated the use of alternative estimation strategies such as the maximum a-posteriori (MAP) method.

In practical situations, the reasons for not using the MMSE estimator may only apply to a subset of the measurements. Then, it may be desirable to construct an estimator that is linear in part of the measurements and nonlinear in the rest. One such scenario arises when estimating a sparsely representable vector X from two sets of measurements Y and Z, one blurred and one noisy. Indeed, as we show in this paper, when working with unitary dictionaries, the MMSE estimate $\mathbb{E}[X|Z]$ from the noisy measurements alone possesses an easy-to-implement closed form solution. However the complexity of computing the MMSE estimate $\mathbb{E}[X|Y,Z]$ from both sets of measurements is exponential. In this setting, the PLMMSE method, which is linear in Y, is computationally cheap and often comes close to the MMSE solution $\mathbb{E}[X|Y,Z]$ in terms of performance.

Partially linear estimation was studied in the statistical literature in the context of regression [5]. In this line of research, it is assumed that the conditional expectation $g(y,z)=\mathbb{E}[X|Y=y,Z=z]$ is linear in y. The goal, then, is to approximate $g(y,z)$ from a set of examples $\{x_i,y_i,z_i\}$ drawn independently from the joint distribution of X, Y and Z. In this paper, our goal is to derive the partially linear MMSE (PLMMSE) estimator. Namely, we do not make any assumptions on the structure of the MMSE estimate $\mathbb{E}[X|Y,Z]$, but rather look for the estimator that minimizes the MSE among all functions $g(Y,Z)$ that are linear in Y.

Due to space limitations, we state here the main results without their proofs, which can be found in [6].

2 Partially Linear Estimation

Suppose that X, Y and Z are random variables (RVs) taking values in $\mathbb{R}^M$, $\mathbb{R}^N$ and $\mathbb{R}^Q$, respectively, such that X is the quantity to be estimated and Y and Z are two sets of measurements thereof. We denote by $\boldsymbol{\Gamma}_{XX}$, $\boldsymbol{\Gamma}_{XY}$, the auto-covariance of X and the cross-covariance of X and Y, respectively.

Our goal is to design a partially linear estimator of X based on Y and Z, which has the form

$$\hat{X}=\boldsymbol{A}Y+b(Z). \tag{1}$$

Here $\boldsymbol{A}$ is a deterministic matrix and $b(z)$ is a vector-valued (Borel measurable) function.

Theorem 1. *The MMSE estimator of the form* (1) *is given by*

$$\hat{X} = \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}W + \mathbb{E}[X|Z], \tag{2}$$

where $W \triangleq Y - \mathbb{E}[Y|Z]$.

Note that (2) is of the form of (1) with $\boldsymbol{A} = \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}$ and $b(Z) = \mathbb{E}[X|Z] - \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}[Y|Z]$. As we show in [6], (2) can be equivalently written as

$$\hat{X} = \left(\boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}\right)\left(\boldsymbol{\Gamma}_{YY} - \boldsymbol{\Gamma}_{\hat{Y}_Z\hat{Y}_Z}\right)^{\dagger}\left(Y - \hat{Y}_Z\right) + \hat{X}_Z, \tag{3}$$

where $\hat{X}_Z \triangleq \mathbb{E}[X|Z]$ and $\hat{Y}_Z \triangleq \mathbb{E}[Y|Z]$. Therefore, all we need to know in order to be able to compute the PLMMSE estimator (2) is the covariance matrix $\boldsymbol{\Gamma}_{XY}$, the conditional expectation $\mathbb{E}[X|Z]$ and the joint distribution of Y and Z.

The intuition behind (2) is similar to that arising in dynamic estimation schemes, such as the Kalman filter. Specifically, we begin by constructing the MMSE estimate $\mathbb{E}[X|Z]$ of X from Z. We then update it with the LMMSE estimate of X based on the *innovation* W of Y with respect to $\mathbb{E}[X|Z]$.

One particularly interesting example is the case where X is observed through two linear systems as

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} \boldsymbol{H} \\ \boldsymbol{G} \end{pmatrix} X + \begin{pmatrix} U \\ V \end{pmatrix}, \tag{4}$$

where U and V are statistically independent. It is easily shown that in this setting, the PLMMSE estimate reduces to

$$\hat{X} = \boldsymbol{A}Y + (\boldsymbol{I} - \boldsymbol{H}\boldsymbol{A})\hat{X}_Z, \tag{5}$$

where $\boldsymbol{I}$ denotes the identity matrix and

$$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^T + \boldsymbol{\Gamma}_{UU})^{\dagger} \tag{6}$$

with $\boldsymbol{C} = \boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{\hat{X}_Z\hat{X}_Z}$.

3 Application to Sparse Approximations

Consider the situation in which X is known to be sparsely representable in a unitary dictionary $\boldsymbol{\Psi} \in \mathbb{R}^{M\times M}$ in the sense that

$$X = \boldsymbol{\Psi}A \tag{7}$$

for some RV A that is sparse with high probability. More concretely, we assume, as in [2,3], a Bernoulli-Gaussian prior, so that the elements of A are given by

$$A_i = S_iB_i, \quad i = 1, \ldots, M, \tag{8}$$

where the RVs $\{B_i\}$ and $\{S_i\}$ are statistically independent, $B_i \sim \mathcal{N}(0, \sigma_{B_i}^2)$ and $\mathbb{P}(S_i = 1) = 1 - \mathbb{P}(S_i = 0) = p_i$.

Assume X is observed through two linear systems, as in (4), where $\boldsymbol{H}$ is an arbitrary matrix, $\boldsymbol{G}$ is an orthogonal matrix satisfying $\boldsymbol{G}^T\boldsymbol{G} = \alpha^2\boldsymbol{I}$ for some $\alpha \neq 0$, and U and V are Gaussian RVs with $\boldsymbol{\Gamma}_{UU} = \sigma_U^2\boldsymbol{I}$ and $\boldsymbol{\Gamma}_{VV} = \sigma_V^2\boldsymbol{I}$. In this case the expression for the MMSE estimate $\mathbb{E}[X|Y,Z]$ comprises 2^M summands [2] rendering its computation prohibitively expensive even for modest values of M. Various approaches have been devised to approximate this solution by a small number of terms (see *e.g.,* [2,3] and references therein).

There are some special cases, however, in which the MMSE estimate possesses a simple structure, which can be implemented efficiently. One such case is when both the channel's response and the dictionary over which X is sparse correspond to orthogonal matrices. As in our setting $\boldsymbol{\Psi}$ is unitary and $\boldsymbol{G}$ is orthogonal, this implies that we can efficiently compute the MMSE estimate $\mathbb{E}[X|Z]$ of X from Z. Therefore, instead of resorting to schemes for approximating $\mathbb{E}[X|Y,Z]$, we can employ the PLMMSE estimator of X based on Y and Z, which, in this situation, possesses the simple closed form expression (5). This approach is particularly effective when the SNR of the observation Y is much worse than that of Z, since the MMSE estimate $\mathbb{E}[X|Y,Z]$ in this case is close to being partially linear in Y. Such a setting is demonstrated in the sequel. We have the following result.

Theorem 2. *The MMSE estimate of X of (7) given Z of (4) is*

$$\mathbb{E}[X|Z] = \boldsymbol{\Psi}\tilde{f}\left(\frac{1}{\alpha}\boldsymbol{\Psi}^T\boldsymbol{G}^T Z\right), \tag{9}$$

where $\tilde{f}(\tilde{z}) = (f(\tilde{z}_1), \ldots, f(\tilde{z}_M))^T$, with

$$f(\tilde{z}_i) = \frac{\frac{\alpha\sigma_{B_i}^2}{\alpha^2\sigma_{B_i}^2+\sigma_V^2}\, p_i\,\mathcal{N}(\tilde{z}_i; 0, \alpha^2\sigma_{B_i}^2 + \sigma_V^2)\,\tilde{z}_i}{p_i\,\mathcal{N}(\tilde{z}_i; 0, \alpha^2\sigma_{B_i}^2 + \sigma_V^2) + (1-p_i)\,\mathcal{N}(\tilde{z}_i; 0, \sigma_V^2)}. \tag{10}$$

Here, $\mathcal{N}(\alpha; \mu, \sigma^2)$ denotes the normal probability density function with mean μ and variance σ^2, evaluated at α.

Therefore, if, *e.g.,* $\boldsymbol{\Psi}$ is a wavelet basis and $\boldsymbol{G} = \boldsymbol{I}$ (so that $\alpha = 1$), then $\mathbb{E}[X|Z]$ can be efficiently computed by taking the wavelet transform of Z (multiplication by $\boldsymbol{\Psi}^T$), applying a scalar shrinkage function on each of the coefficients (namely calculating $f(\tilde{z}_i)$ for the ith coefficient) and applying the inverse wavelet transform (multiplication by $\boldsymbol{\Psi}$) on the result.

Equipped with a closed form expression for $\mathbb{E}[X|Z]$, we can now compute the terms needed for implementing the PLMMSE estimator (5). First, we note that

$$\boldsymbol{\Gamma}_{XX} = \boldsymbol{\Psi}\boldsymbol{\Gamma}_{AA}\boldsymbol{\Psi}^T, \tag{11}$$

where $\boldsymbol{\Gamma}_{AA}$ is a diagonal matrix with $(\boldsymbol{\Gamma}_{AA})_{i,i} = p_i\sigma_{B_i}^2$. Similarly,

$$\boldsymbol{\Gamma}_{\hat{X}_Z\hat{X}_Z} = \boldsymbol{\Psi}\mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z}))\boldsymbol{\Psi}^T, \tag{12}$$

where $\mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z}))$ is a diagonal matrix whose (i,i) element is $\beta_i = \mathbb{V}\mathrm{ar}(f(\tilde{Z}_i))$. This is due to the fact that the elements of $\tilde{Z}$ are statistically independent and the fact that the function $\tilde{f}(\cdot)$ operates element-wise on its argument. Hence, the PLMMSE estimator is given by (5) with $\mathbb{E}[X|Z]$ of (9) and with the matrix

$$\boldsymbol{A} = \boldsymbol{\Psi C \Psi}^T \boldsymbol{H}^T \left(\boldsymbol{H \Psi C \Psi}^T \boldsymbol{H}^T + \sigma_U^2 \boldsymbol{I} \right)^{\dagger}, \tag{13}$$

where here $\boldsymbol{C} = \boldsymbol{\Gamma}_{AA} - \mathbb{C}\text{ov}(\tilde{f}(\tilde{Z})) = \text{diag}(p_1\sigma_{B_1}^2 - \beta_1, \ldots, p_M\sigma_{B_M}^2 - \beta_M)$. Observe that there is generally no closed form expression for the scalars β_i, rendering it necessary to compute them numerically.

An important special case corresponds to the setting in which $p_i = p$ and $\sigma_{B_i}^2 = \sigma_B^2$ for every i. In this situation, we also have that $\beta_i = \beta$ for every i. Furthermore,

$$\boldsymbol{\Gamma}_{XX} = \boldsymbol{\Psi} \left(p\sigma_B^2 \boldsymbol{I} \right) \boldsymbol{\Psi}^T = p\sigma_B^2 \boldsymbol{I} \tag{14}$$

and

$$\boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z} = \boldsymbol{\Psi}(\beta \boldsymbol{I})\boldsymbol{\Psi}^T = \beta \boldsymbol{I}, \tag{15}$$

so that $\boldsymbol{A}$ is simplified to

$$\boldsymbol{A} = (p\sigma_B^2 - \beta)\boldsymbol{H}^T \left((p\sigma_B^2 - \beta)\boldsymbol{H H}^T + \sigma_U^2 \boldsymbol{I} \right)^{\dagger}. \tag{16}$$

As can be seen, here $\boldsymbol{A}$ does not involve multiplication by $\boldsymbol{\Psi}$ or $\boldsymbol{\Psi}^T$. Thus, if $\boldsymbol{H}$ corresponds to a convolution operation, so does $\boldsymbol{A}$, meaning that it can be efficiently applied in the Fourier domain.

3.1 Image Deblurring with Blurred/Noisy Image Pairs

When taking photos in dim light using a hand-held camera, there is a tradeoff between noise and motion blur, which can be controlled by tuning the shutter speed. Using a long exposure time, the image typically comes out blurred due to camera shake. On the other hand, with a short exposure time (and high camera gain), the image is very noisy. In [7] it was demonstrated how a high quality image can be constructed by properly processing two images of the same scene, one blurred and one noisy.

We now show how the PLMMSE approach can be applied in this setting to obtain plausible recoveries at a speed several orders of magnitude faster than any other sparsity-based method. In our setting X, Y and Z correspond, respectively, to the original, blurred (and slightly noisy) and noisy images. Thus, the measurement model is that described by (4), where $\boldsymbol{H}$ corresponds to spatial convolution with some blur kernel, $\boldsymbol{G} = \boldsymbol{I}$, and U and V correspond to white Gaussian noise images with small and large variances respectively. We further assume that the image X is sparse in some orthogonal wavelet basis $\boldsymbol{\Psi}$, such that it can be written as in (7) and (8).

As we have seen, in this setting, the PLMMSE estimator can be computed in two stages. First, we calculate $\hat{X}_Z = \mathbb{E}[X|Z]$ by computing the wavelet transform $\tilde{Z} = \boldsymbol{\Psi}^T Z$, applying the scalar shrinkage function (10) on each wavelet coefficient, and taking the inverse wavelet transform of the result. This stage requires knowledge of the parameters $\{p_i\}$, $\{\sigma^2_{B_i}\}$ and σ^2_V. To this end, we assume that p_i and $\sigma^2_{B_i}$ are the same for wavelets coefficients at the same level. Namely, all wavelet coefficients of Z at level ℓ correspond to independent draws from the Gaussian mixture

$$f_{\tilde{Z}_i}(\tilde{z}) = p^\ell \mathcal{N}(\tilde{z}; 0, \alpha^2\sigma^2_{B^\ell} + \sigma^2_V) + (1-p)\mathcal{N}(\tilde{z}; 0, \sigma^2_V). \tag{17}$$

Consequently, p^ℓ, $\sigma^2_{B^\ell}$ and σ^2_V can be estimated by expectation maximization (EM). In our experiments, we assumed that σ^2_V is known.

In the second stage, the denoised image $\hat{X}_Z$ needs to be combined with the blurred image Y using (5) with $\boldsymbol{A}$ of (13). As discussed in Section 3, this can be carried out very efficiently if $p_i = p$ and $\sigma^2_{B_i} = \sigma^2_B$ for all i. For the sake of efficiency we therefore abandon the assumption that p_i and $\sigma^2_{B_i}$ vary across wavelet levels and assume henceforth that all wavelet coefficients are independent and identically distributed. In this case, $\boldsymbol{A}$ corresponds to the filter

$$A(\omega) = \frac{(\sigma^2_A - \beta)H^*(\omega)}{(\sigma^2_A - \beta)|H(\omega)|^2 + \sigma^2_U}, \tag{18}$$

where $H(\omega)$ is the frequency response of the blur kernel. Consequently, the final PLMMSE estimate corresponds to the inverse Fourier transform of

$$\hat{X}^{\mathrm{F}}_{\mathrm{PLMMSE}}(\omega) = \frac{(\sigma^2_A - \beta)H^*(\omega)Y^{\mathrm{F}}(\omega) + \sigma^2_U \hat{X}^{\mathrm{F}}_Z(\omega)}{(\sigma^2_A - \beta)|H(\omega)|^2 + \sigma^2_U}, \tag{19}$$

where $Y^{\mathrm{F}}(\omega)$ and $\hat{X}^{\mathrm{F}}_Z(\omega)$ denote the Fourier transforms of Y and $\hat{X}_Z$, respectively. In our experiment, we assumed that the blur $H(\omega)$ and noise variance σ^2_U are known. In practice, they can be estimated from Y and Z, as proposed in [7]. This stage also requires knowing the scalars $\sigma^2_A = \mathbb{E}[A^2]$ and $\beta = \mathbb{E}[f^2(\tilde{z})]$, which we estimate as $\widehat{\sigma^2_A} = \frac{1}{M}\sum_{i=1}^M \tilde{z}^2_i - \sigma^2_V$ and $\widehat{\beta} = \frac{1}{M}\sum_{i=1}^M f^2(\tilde{z}_i)$.

Fig. 1 demonstrates our approach on the 512×512 Gold-hill image. In this experiment, the blur corresponded to a Gaussian kernel with standard deviation 3.2. To model a situation in which the noise in Y is due only to quantization errors, we chose $\sigma_U = 1/\sqrt{12} \approx 0.3$ and $\sigma_V = 45$. These parameters correspond to a peak signal to noise ratio (PSNR) of 25.08dB for the blurred image and 15.07dB for the noisy image.

We used the orthogonal Symlet wavelet of order 4 and employed 10 EM iterations to estimate p^ℓ and $\sigma^2_{B^\ell}$ in each wavelet level. The entire process takes 1.1 seconds on a Dual-Core 3GHz computer with un-optimized Matlab code. We note that our approach can be viewed as a smart combination of Wiener filtering for image debluring and wavelet thresholding for image denoising, which are among the simplest and fastest methods available. Consequently, the running time is at least an order of magnitude faster than any other sparsity-based methods (see, *e.g.,* comparisons in [2]).

As can be seen in Fig. 1, the quality of the recoveries corresponding to the denoised image $\hat{X}_Z$ and deblurred image $\hat{X}^{\mathrm{L}}_Y$ is rather poor with respect to the

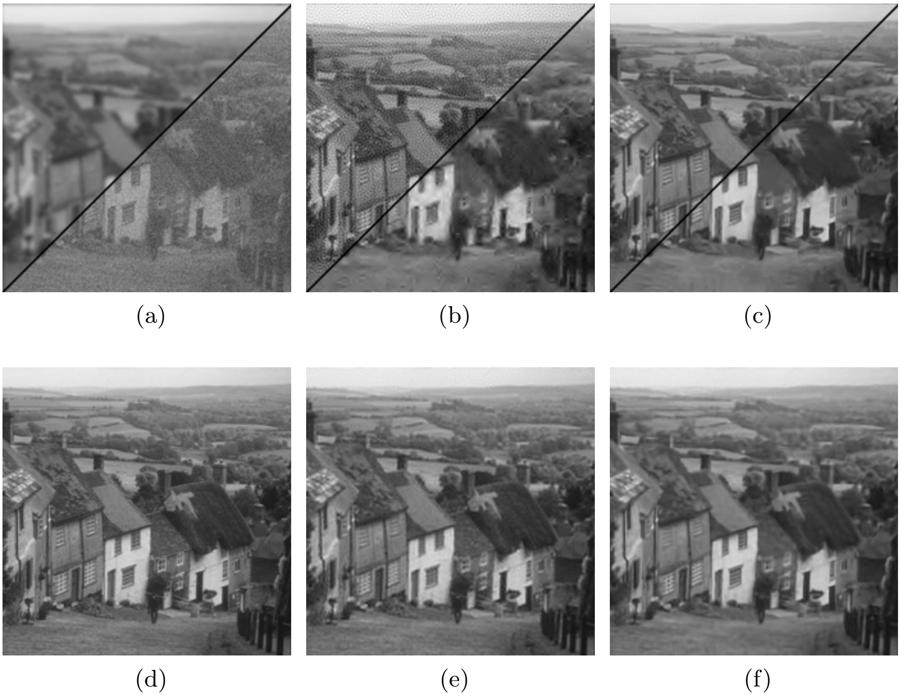


Fig. 1. Debluring with a blurred/noisy image pair using PLMMSE estimation and RD [7]. (a) Blurred image Y (top left) and noisy image Z (bottom-right). (b) LMMSE-deblurred image $\hat{X}_Y^{\mathrm{L}}$ (top-left) and MMSE-denoised image $\hat{X}_Z$ (bottom-right). (c) BM3D-deblurred image (top left) and BM3D-denoised image (bottom-right). (d) Original image X. (e) PLMMSE estimate $\hat{X}_{\mathrm{PLMMSE}}$ from Y and Z. (f) RD recovery.

state-of-the-art BM3D debnoising method [8] and BM3D debluring algorithm [9]. However, the quality of the joint estimate $\hat{X}_{\mathrm{PLMMSE}}$ surpasses each of these techniques. The residual deconvolution (RD) method [7] for joint debluring and denoising outperforms the PLMMSE method in terms of recovery error but the visual differences are not prominent.

A quantitative comparison on several test images is given in Table 1. The PSNR attained by the PLMMSE method is, on average, 0.3dB higher than BM3D debluring, 0.4db higher than BM3D denoising, and 0.8dB lower than RD. In terms of running times, however, our method is, on average, 11 times faster than BM3D deblurring, 16 times faster than BM3D denoising and 18 times faster than RD. Note that RD requires initialization with a denoised version of Z, for which purpose we used the BM3D algorithm. Hence, the running times reported in the last column of Table 1 include the running times of the BM3D denoising method.

Table 1. Performance of deblurring/denoising on several images

	$\hat{X}_Z$	$\hat{X}_Y^{\text{L}}$	BM3D Denoise	BM3D Deblur	PLMMSE	RD
Boat	25.39/0.83	23.45/0.06	27.85/13.52	28.40/10.23	28.05/0.88	29.22/15.31
Lena	26.93/0.73	24.59/0.03	29.47/13.22	30.58/8.90	30.58/0.81	31.37/15.19
Mandrill	21.40/0.64	20.59/0.06	22.72/13.58	21.78/9.57	22.58/0.72	23.30/15.58
Peppers	26.74/0.81	24.89/0.08	29.49/13.14	29.74/8.91	29.80/0.88	31.52/15.03
Mountain	19.23/0.95	17.69/0.09	20.11/15.24	18.45/11.12	20.03/1.05	20.42/17.47
Frog	23.23/0.94	22.35/0.16	24.00/16.07	24.40/13.37	24.69/1.09	24.69/21.14
Gold-hill	25.90/0.69	24.26/0.06	27.52/13.41	28.70/9.54	28.82/1.09	29.09/21.14
Average	24.12/0.81	22.55/0.08	25.88/14.03	26.01/10.23	26.31/0.89	27.09/16.19

4 Conclusion

In this paper, we derived the PLMMSE estimator and showed that it depends only on the joint second-order statistics of X and Y, rendering it applicable in a wide variety of situations. We demonstrated the utility of our approach in sparse signal recovery from a measurement pair. In the context of image enhancement from blurred/noisy image pairs, we showed that PLMMSE estimation performs close to state-of-the-art algorithms while running much faster.

References

1. Costa, O.L.V.: Linear minimum mean square error estimation for discrete-time Markovian jump linear systems. IEEE Trans. Autom. Control 39(8), 1685–1689 (1994)
2. Schniter, P., Potter, L.C., Ziniel, J.: Fast Bayesian matching pursuit. In: Information Theory and Applications Workshop (ITA 2008), pp. 326–333 (2008)
3. Soussen, C., Idier, J., Brie, D., Duan, J.: From bernoulli-gaussian deconvolution to sparse signal restoration. IEEE Trans. Signal Process. (99), 4572–4584 (2010)
4. Girolami, M.: A Variational method for learning sparse and overcomplete representations. Neural Computation 13(11), 2517–2532 (2001)
5. Härdle, W., Liang, H.: Partially linear models. In: Statistical Methods for Biostatistics and Related Fields, pp. 87–103 (2007)
6. Michaeli, T., Sigalov, D., Eldar, Y.: Partially linear estimation with application to sparse signal recovery from measurement pairs. IEEE Trans. Signal Process. (2011) (accepted)
7. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image deblurring with blurred/noisy image pairs. In: ACM SIGGRAPH 2007 Papers, pp. 1–10. ACM (2007)
8. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans. Image Process. 16(8), 2080–2095 (2007)
9. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image restoration by sparse 3d transform-domain collaborative filtering. In: SPIE Electronic Imaging (2008)

Causal Discovery for Linear Non-Gaussian Acyclic Models in the Presence of Latent Gaussian Confounders

Zhitang Chen and Laiwan Chan

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Hong Kong
{ztchen,lwchan}@cse.cuhk.edu.hk

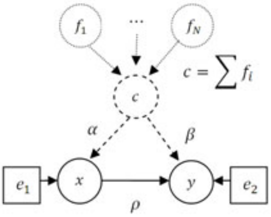
Abstract. LiNGAM has been successfully applied to casual inferences of some real world problems. Nevertheless, basic LiNGAM assumes that there is no latent confounder of the observed variables, which may not hold as the confounding effect is quite common in the real world. Causal discovery for LiNGAM in the presence of latent confounders is a more significant and challenging problem. In this paper, we propose a cumulant-based approach to the pairwise causal discovery for LiNGAM in the presence of latent confounders. The method assumes that the latent confounder is Gaussian distributed and statistically independent of the disturbances. We give a theoretical proof that in the presence of latent Gaussian confounders, the causal direction of the observed variables is identifiable under the mild condition that the disturbances are both super-gaussian or sub-gaussian. Experiments on synthesis data and real world data have been conducted to show the effectiveness of our proposed method.

Keywords: Causal analysis, LiNGAM, latent Gaussian confounder, cumulant-based measure.

1 Introduction

Causal discovery from non-specifically controlled experimental data has received extensive attention in recent years. Many models such as structural equation models (SEMs) and Bayesian networks (BNs) have been proposed to explain the data generating mechanisms and widely applied to social science, econometrics and medical science [1] [2]. However, traditional methods assume the Gaussian disturbances and only employ second order statistics. In general, such methods can only obtain a class of equivalent models [3] and fail to identify the full causal structure without prior knowledge in most cases [4]. Recently, it has been shown that by employing the non-gaussianity of the disturbances, the causal structure can be fully identified. In [3], authors proposed a Linear Non-Gaussian Acyclic Model (LiNGAM) and showed that the full structure can be identified by Independent Component Analysis (ICA) [5] [6]. The Direct-LiNGAM framework was proposed later to avoid iterative searching [7]. The advantages of LiNGAM over conventional methods are: (1) a full and unique causal structure can be identified instead of a class of Markov equivalent models. (2) no prior knowledge of the network structure is needed. (3) Compared to BNs which may require large amount of conditional independent tests, the computational complexity of

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 17–24, 2012.



$$
\begin{aligned}
x &= e_1 + \alpha c \\
y &= \rho x + e_2 + \beta c
\end{aligned} \tag{1}
$$

Fig. 1. Casual pair with latent confounder

LiNGAM is much lower. In spite of the advantages mentioned, basic LiNGAM considers the causally sufficient case where there is no unobserved confounders [3]. However, this assumption may not hold in many real world problems.

In this paper, we will deal with a more challenging problem: causal discovery for causal pairs with latent confounders, which can be represented graphically by figure 1. x and y are observed variables of which we aim to discover the causal direction. e_1 and e_2 are non-gaussian disturbances. c is the unobserved latent confounder which can be regarded as the total effect of many latent factors f_i. ρ, α and β are the corresponding causal strengths. Due to the extra dependence introduced by c, the causal discovery for x and y becomes much more challenging. Previous methods such as BNs, LiNGAM and DirectLiNGAM may obtain misleading results as they does not consider the latent confounders. Recently, Aapo Hyvärinen [8] proposed new measures of causal directions for causal pairs in the scenario of no latent confounders. Inspired by [8], we propose a new cumulant-based measure to discover the causal directions for LiNGAM in the presence of Gaussian Confounders(LiNGAM-GC). The basic idea is the use of a specially designed measure which is immune to the latent Gaussian confounder. We prove that the causal direction can be simply identified by investigating the sign of the proposed cumulant-based measure, i.e. if the measure $\tilde{R}_{xy} > 0$ we can conclude that x causes y. The **advantages** of our proposed cumulant-based measure over the one proposed in [8] are that our measure does not require the explicit estimations of the regression coefficient ρ and more importantly our cumulant-based measure is **immune** to the **latent Gaussian confounder**. In the paper, due to the limit of pages, we mainly deal with causal-effect pairs. However, the algorithm developed in this paper can be easily extended to the model with more than two variables following the similar manner as DirectLiNGAM.

The rest of this paper is organized as follows: In section 2, we briefly introduce some related works concerning latent confounders in recent years. In section 3, we firstly introduce the cumulant-based measure proposed in [8] for pairwise causal discovery. Secondly, we propose our LiNGAM-GC model and a new cumulant-based measure to tackle the causal discovery problem in this model. In section 4, experiments on synthesis data and real world data are conducted to show the effectiveness of our proposed approaches. In section 5, we conclude our paper.

2 Related Works Concerning Latent Confounders

Recently, several papers were published concerning the latent confounder [9] [10] [11]. In [9], authors treated the causal discovery problem for LiNGAM in the presence of confouders as an overcomplete ICA problem and developed algorithms to derive canonical models from observed variables. They used overcomplete ICA algorithms to estimate the whole causal structure. However, overcomplete ICA is an ill-posed and challenging problem which still remains open. There is no reliable and accurate method existing for this problem especially when the dimensionality of the problem is high.

In [10], authors proposed a model called Confounders with Additive Noise (CAN):

$$\begin{aligned} X &= u(T) + N_X \\ Y &= v(T) + N_Y \end{aligned} \tag{2}$$

where X and Y are observed effects; T is the latent confounder; N_X and N_Y are disturbances. T, N_X and N_Y are statistically independent. Authors showed that under certain assumptions, the confounder is identifiable. Note that in the CAN model, there is no direct edge between node X and Y. The variances of N_X and that of N_Y are assumed to be small [10]. However, we are interested in a more general case: there is a direct edge between X and Y; N_X and N_Y are not necessarily small.

In [11], authors proposed a new model called GroupLiNGAM. In this model, latent confounders are allowed but restricted within subsets of the observed variables. They used the Direct-LiNGAM framework by iteratively finding and removing the exogenous subsets of the observed variables from the remaining subsets until the whole causal ordering of the subsets are identified [11]. However, in the GroupLiNGAM model, the confounders are restricted within certain subsets. Furthermore, the causal direction and strength remain unidentified within the subsets yet.

3 Causal Discovery for Causal Pairs with Latent Confounders

3.1 Cumulant-Based Measure by Aapo Hyvärinen

Firstly, we introduce the cumulant-based measure proposed in [8], which lays the foundation for our work. Suppose we have observed two random variables x and y with zero means and generated by equation 1 but with $\alpha = 0$ and $\beta = 0$. e_1 and e_2 are independent non-gaussian distributed disturbances; ρ is the causal strength. Denote by $\hat{x}$ and $\hat{y}$ the normalized x and y with unit variances. Denote by $\hat{\rho}$ the corresponding regression coefficient. It is easy to know that $|\hat{\rho}| < 1$. The cumulant-based measure proposed in [8] is given as below:

$$\tilde{R}_{c4}(\hat{x}, \hat{y}) = sign(kurt(\hat{x}))\tilde{\rho}\hat{E}\{\hat{x}^3\hat{y} - \hat{x}\hat{y}^3\} \tag{3}$$

where $\hat{E}$ means average and $\tilde{\rho}$ is the estimated regression coefficient. Note that the above cumulant-based measure fails to give any decision when the estimated kurtosis of $\hat{x}$ and that of $\hat{y}$ have opposite signs. According theorem 1 in [8], we have the following:

$$\begin{aligned} \hat{x} \rightarrow \hat{y} \Leftrightarrow \tilde{R}_{c4} > 0 \\ \hat{y} \rightarrow \hat{x} \Leftrightarrow \tilde{R}_{c4} < 0 \end{aligned} \tag{4}$$

This cumulant-based measure works quite well in the scenario of no latent confounders. However, the presence of latent confounders is a very common phenomenon in real world problems. This measure may obtain misleading results in real applications. In the following, we consider a new model in presence of latent Gaussian confounders.

3.2 Cause-Effect Pairs in Presence of Latent Gaussian Confounders

Assume x and y are with zero means and generated by equation 1. c is Gaussian distributed. Note that this model is an extension of LiNGAM by allowing the presence of latent Gaussian confounders. We argue that the Gaussianity assumption of the latent confounder is not strong as in real world, there may be a number of latent factors which are the common causes of the observed effects. The confounder we introduce here can be regarded as the total effect of such factors as illustrated graphically by figure 1. According to the **central limit theorem**, the summation of a large number of independent random variables with finite expectations and finite variances tends to be Gaussian distributed. Due to the presence of c, the causal inference becomes problematic: (1) due to the extra dependence introduced by c, the causal direction of x and y can not be simply inferred by testing independence of regressors and regression residues. (2) the causal strength especially the sign of causal strength estimated may be severely biased. To tackle this difficulty, we propose a new cumulant-based measure which is an extension of the cumulant-based measure proposed in [8] . First of all, we investigate two different normalization schemes.

3.3 Normalization to Unit Variance / Unit Absolute Kurtosis

Lemma 1. *Assume that the observed variables x and y are generated according to equation 1 and fulfill $\sigma_2^2 + (2\alpha\rho + \beta)\beta\sigma_c^2 > 0$, where σ_2 and σ_c are the variances of e_2 and c respectively. Denote by $\hat{x}$ and $\hat{y}$ the normalized x and y with unit variances, then the casual strength $\hat{\rho}$ between $\hat{x}$ and $\hat{y}$ has the property of $|\hat{\rho}| < 1$.*

Note that $|\hat{\rho}| < 1$ is a working condition for our cumulant-based measure to be introduced later in this section. We prove that $|\hat{\rho}| < 1$ under assumption of $\sigma_2^2 + (2\alpha\rho + \beta)\beta\sigma_c^2 > 0$, if we normalize x and y to unit variance. However, the assumption may not hold in some real world problems. Below, we propose another normalization method which can guarantee that $|\hat{\rho}| < 1$ under a much weaker assumption. Let $k_x = \sqrt[4]{|kurt(x)|}$ and $k_y = \sqrt[4]{|kurt(y)|}$. We have:

$$k_x^4 = |kurt(x)| = |kurt(e_1 + \alpha c)| \doteq |kurt(e_1)|$$

$$k_y^4 = |kurt(y)| = |kurt(\rho e_1 + e_2 + (\alpha\rho + \beta)c)| \doteq |\rho^4 kurt(e_1) + kurt(e_2)|$$

The above $\doteq$ holds as $kurt(c) = 0$ (Remind that c is Gaussian distributed). Normalizing x and y by $\hat{x} = x/k_x$ and $\hat{y} = y/k_y$, we have the following lemma.

Lemma 2. *Assume the kurtosis of e_1 and the kurtosis of e_2 have the same sign, i.e. e_1 and e_2 are both super-gaussian or sub-gaussian. Denote by $\hat{x}$ and $\hat{y}$ the normalized x and y with unit absolute kurtosis , then the casual strength $\hat{\rho}$ between $\hat{x}$ and $\hat{y}$ has the property of $|\hat{\rho}| < 1$.*

Proof. We skip the proof due to the limit of pages.

3.4 New Cumulant-Based Measure

For convenience, in the rest of this paper, we use notations x, y, c, α, β and ρ but we assume that x and y are standardized (either by normalizing to unit variance or absolute kurtosis). According to lemma 1 or 2, $|\rho| < 1$. Inspired by [8], we give the following cumulant-based measure which can be used to determine the causal direction for LiNGAM-GC. Let:

$$\begin{aligned} C_{xy} &= \hat{E}\{x^3y\} - 3\hat{E}\{xy\}\hat{E}\{x^2\} \\ C_{yx} &= \hat{E}\{xy^3\} - 3\hat{E}\{xy\}\hat{E}\{y^2\} \end{aligned}$$

Define new cumulant-based measure as:

$$\tilde{R}_{xy} = (C_{xy} + C_{yx})(C_{xy} - C_{yx}) \tag{5}$$

$\hat{E}$ means sample average. We have the following theorem:

Theorem 1. *If the causal direction is $x \to y$, we have:*

$$\tilde{R}_{xy} = \rho^2(1+\rho^2)(1-\rho^2)kurt(e_1)^2 \tag{6}$$

where $kurt(e_1) = E\{e_1^4\} - 3E\{e_1^2\}^2$ is the kurtosis of e_1.
If the causal direction is $y \to x$, we have:

$$\tilde{R}_{xy} = \rho^2(1+\rho^2)(\rho^2-1)kurt(e_2)^2 \tag{7}$$

Proof. Consider the fourth-order cumulant

$$C(x,y) = cum(x,x,x,y) = E\{x^3y\} - 3E\{xy\}E\{x^2\} \tag{8}$$

$$\tilde{R}_{xy} = \{C(x,y) + C(y,x)\}\{C(x,y) - C(y,x)\}$$

If $x \to y$, i.e. the observed x and y fulfill the generating mechanism described in equation 1, based on the properties of cumulant [8], we have:

$$\begin{aligned} C(x,y) &= cum(x,x,x,y) = \rho cum(e_1,e_1,e_1,e_1) + \alpha^3(\alpha\rho+\beta)cum(c,c,c,c) \\ &= \rho kurt(e_1) + \alpha^3(\alpha\rho+\beta)kurt(c) \end{aligned}$$

As we assume that the latent confounder c is Gaussian distributed, we have $cum(c,c,c,c) = kurt(c) = 0$ and therefore we have $C(x,y) = \rho kurt(e_1)$.

$$C(y,x) = cum(y,y,y,x) = \rho^3 kurt(e_1) + \alpha(\alpha\rho+\beta)^3 kurt(c) = \rho^3 kurt(e_1)$$

$$\begin{aligned} \tilde{R}_{xy} &= \{\rho kurt(e_1) + \rho^3 kurt(e_1)\}\{\rho kurt(e_1) - \rho^3 kurt(e_1)\} \\ &= \rho^2(1+\rho^2)(1-\rho^2)kurt(e_1)^2 \end{aligned}$$

If $y \to x$, through similar derivation, we have the following:

$$\begin{aligned} \tilde{R}_{xy} &= \{\rho^3 kurt(e_2) + \rho kurt(e_2)\}\{\rho^3 kurt(e_2) - \rho kurt(e_2)\} \\ &= \rho^2(1+\rho^2)(\rho^2-1)kurt(e_2)^2 \end{aligned}$$

According to lemma 1 or 2, we know that $|\rho| < 1$, and therefore we have the following causal inferencing rule:

$$\begin{aligned} x \to y \Leftrightarrow \tilde{R}_{xy} > 0 \\ y \to x \Leftrightarrow \tilde{R}_{xy} < 0 \end{aligned} \tag{9}$$

4 Experiment

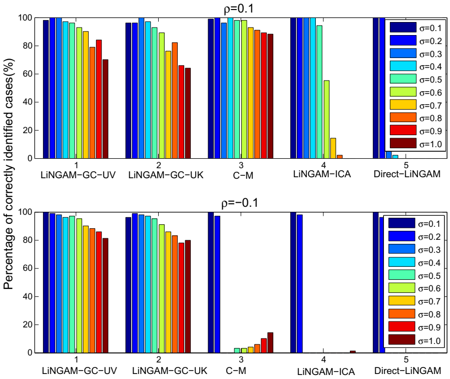
4.1 Synthesis Data

In this experiment, we use our proposed cumulant-base measure with different normalization methods: unit variance (LiNGAM-GC-UV) and unit absolute kurtosis(LiNGAM-GC-UK) , ICA-LiNGAM[1], Direct-LiNGAM[2], Cumulant-based Measure(C-M) [8] to identify the causal direction for causal pairs. The purpose of this experiment is to show that latent confounders can be problematic if they are not considered. We consider the causal pairs generated by equation 1. e_1 and e_2 are generated by $e_i = sign(n_i)|n_i|^2$ and normalized to unit variance, where n_i are standard Gaussian random variable. c is Gaussian distributed with zero mean and standard deviation σ_c [3]. We fix $\alpha = 1.2$ and $\beta = 1.6$. In order to learn how ρ and σ_c^2 affect the accuracies of five algorithms, we conduct a series of experiments as follows: $\{\rho = \pm 0.1, \pm 0.3, \sigma_c = 0.1, 0.2, \cdots, 1.3\}$. Note that the experimental settings guarantee the assumptions of lemma 1 and lemma 2. For each parameter setting $\{\rho, \sigma_c\}$, we randomly generate 100 datasets with sample size of 5000. The percentages of correctly identified datasets for different methods are shown in figure 2 and 3. The experimental results suggest that LiNGAM-GC-UV and LiNGAM-GC-UK have the best performances in different scenarios. Although wrong decisions still occur in the case of small causal strength ρ, it is due to the finite sample size. If the sample size is large enough, both methods are expected to achieve perfect performances. From figure 2 and 3, we also learn that C-M performs very well in the case of positive casual strength but performs badly in the case of negative casual strength . The explanation for this observation is that in the scenario of positive ρ, α and β, the C-M algorithm is immune to the latent Gaussian confounder. However,when the true causal strength $\rho < 0$, the presence of latent confounder may cause C-M algorithm to get wrong estimation of ρ with opposite sign, which in turn leads to the wrong causal direction. This shows that the performance of C-M algorithm depends on whether the effect of the latent confounder is strong enough to flip the sign of the estimated causal strength. The performances of LiNGAM-ICA and Direct-LiNGAM depend on the variance of the confounder. When the variance of the latent confounder is large enough, the performance of both algorithms degenerate dramatically.

[1] http://www.cs.helsinki.fi/group/neuroinf/lingam/

[2] http://www.ar.sanken.osaka-u.ac.jp/~inazumi/dlingam.html

[3] We also conduct the experiment where the confounder c is mildly Non-Gaussian and the result shows that the proposed measure is robust. While in the case of strongly Non-Gaussian confounder, the proposed measure fails to give the correct identification. Due to the limit of pages, we do not present the result here.



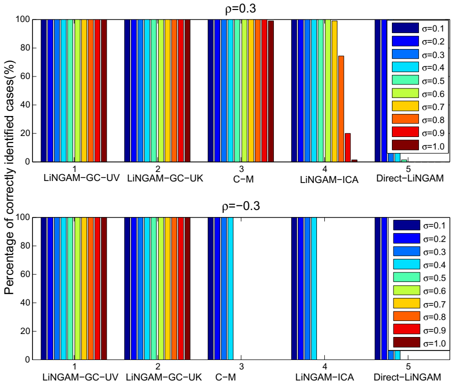


Fig. 2. Results of synthesis data: $\rho = \pm 0.1$

Fig. 3. Results of synthesis data: $\rho = \pm 0.3$

4.2 Real World Data

In order to show the applicability of our proposed methods in real world data, we compare the performances of difference methods in the real-world cause-effect pairs[4]. We select a total of 45 pairs in this dataset[5]. As the computational time of Direct-LiNGAM increases dramatically for large sample size, we use at most 1000 samples for each cause-effect pair. The performances of different algorithms are given in Table 1.

Table 1. Percentage of recovering the true causal direction in 45 real world cause-effect pairs

Algorithm	LiNGAM-GC-UV	LiNGAM-GC-UK	C-M	LiNGAM-ICA	Direct-LiNGAM	IGCI
Accuracy	71.11%	**73.33%**	62.22%	46.67%	55.56%	60%

Table 1 shows that for causal discovery of real world data, our proposed LiNGAM-GC-UV and LiNGAM-GC-UK have the best performances followed by C-M. LiNGAM-GC-UK performs better than LiNGAM-GC-UV possibly due to the milder assumption. IGCI [12] achieves only 60% accuracy mainly due to the fact that it is originally proposed for deterministic causal relations inference. Direct-LiNGAM performs slightly better than random guess while LiNGAM-ICA has the accuracy less than 50%. From this experiment, we learn that by taking into consideration of latent confounders, the causal inference becomes more reliable and accurate.

5 Conclusion

A new Linear Non-Gaussian Acyclic Model in the presence of latent Gaussian confounders is proposed in this paper. By allowing the presence of latent confounders, this

[4] http://webdav.tuebingen.mpg.de/cause-effect/

[5] We make a simple preprocessing of pair #75 to make the relation more linear and use two processed pairs $\{x, \frac{1}{y}\}$ and $\{\frac{1}{x}, y\}$ instead of the original one.

model is expected to give more accurate description of the real world phenomenon. We propose a cumulant-based measure to infer the causal structure of the observed variables for this model. We discuss and prove under what conditions the causal structure is identifiable by our proposed approach. Experimental results show that our algorithms work better on synthesis data and real world cause-effect pairs than the compared methods. The theoretical limit of our proposed method is that its performance is affected by the sample size due to the estimation of higher order cumulant. Future work will focus on developing a more robust measure in the case of small sample size. Using unbiased estimation of cumulant will be an important issue of the future work.

Acknowledgments. The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region, China.

References

1. Pearl, J.: Causality: models, reasoning, and inference. Cambridge Univ. Pr. (2000)
2. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction, and search. The MIT Press (2000)
3. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. The Journal of Machine Learning Research 7, 2003–2030 (2006)
4. Sogawa, Y., Shimizu, S., Hyvärinen, A., Washio, T., Shimamura, T., Imoto, S.: Discovery of Exogenous Variables in Data with More Variables Than Observations. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6352, pp. 67–76. Springer, Heidelberg (2010)
5. Comon, P.: Independent component analysis, a new concept? Signal Processing 36(3), 287–314 (1994)
6. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5), 411–430 (2000)
7. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: Directlingam: A direct method for learning a linear non-gaussian structural equation model. Journal of Machine Learning Research 12, 1225–1248 (2011)
8. Hyvärinen, A.: Pairwise measures of causal direction in linear non-gaussian acyclic models. In: JMLR Workshop and Conference Proceedings (Proc. 2nd Asian Conference on Machine Learning, ACML 2010), vol. 13, pp. 1–16 (2010)
9. Hoyer, P.O., Shimizu, S., Kerminen, A.J.: Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. Arxiv preprint cs/0603038 (2006)
10. Janzing, D., Peters, J., Mooij, J., Schölkopf, B.: Identifying confounders using additive noise models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 249–257. AUAI Press (2009)
11. Kawahara, Y., Bollen, K., Shimizu, S., Washio, T.: Grouplingam: Linear non-gaussian acyclic models for sets of variables. Arxiv preprint arXiv:1006.5041 (2010)
12. Daniušis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., Schölkopf, B.: Inferring deterministic causal relations. In: Proceedings of the Twenty-Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2010), pp. 143–150. AUAI Press, Corvallis (2010)

Alleviating the Influence of Weak Data Asymmetries on Granger-Causal Analyses

Stefan Haufe[1,2], Vadim V. Nikulin[3], and Guido Nolte[4]

[1] Berlin Institute of Technology, Machine Learning
stefan.haufe@tu-berlin.de
[2] Bernstein Focus Neurotechnology, Berlin
[3] Charité University Medicine, Berlin, Neurophysics
[4] Fraunhofer Institute FIRST, Berlin, Intelligent Data Analysis

Abstract. We introduce the concepts of weak and strong asymmetries in multivariate time series in the context of causal modeling. Weak asymmetries are by definition differences in univariate properties of the data, which are not necessarily related to causal relationships between time series. Nevertheless, they might still mislead (in particular Granger-) causal analyses. We propose two general strategies to overcome the negative influence of weak asymmetries in causal modeling. One is to assess the confidence of causal predictions using the antisymmetry-symmetry ratio, while the other one is based on comparing the result of a causal analysis to that of an equivalent analysis of time-reversed data. We demonstrate that Granger Causality applied to the SiSEC challenge on causal analysis of simulated EEG data greatly benefits from our suggestions.

Keywords: Weak/strong asymmetries, ASR, time inversion, Granger Causality, SiSEC challenge.

1 Introduction

Many measures of causal interaction (a. k. a. effective connectivity) are based on the principle that the cause precedes the effect. However, it would be misleading to assume that temporal ordering is necessarily the dominant factor when estimating causal relationship on the basis of the available techniques, such as Granger causality. In fact, methods to estimate causal relations are based on general asymmetries between two (or more) signals out of which the temporal order is just one specific feature. Other asymmetries, like different signal-to-noise ratios, different overall power or spectral details, may in general also affect causal estimates depending on which method is used.

We here propose to distinguish between two different kinds of asymmetries. We call the first type 'strong asymmetries' defined as asymmetries in the relation between two (or more) signals like the temporal ordering. The second type is called 'weak asymmetry' and denotes different univariate properties as given, e. g., by the spectral densities. Weak asymmetries can hence be detected from two signals without estimating any functional relationship between them whereas a strong asymmetry is a property of that functional relationship.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 25–33, 2012.

Altough the concepts presented here (but not the test presented below) could be generalized to other cases in a straight forward way we restrict ourselves in the following to the discussion of stationary and Gaussian distributed data. Let $x_j(t)$ be the signal in channel j at time t. Then the statistical properties are completely defined by the cross-covariance matrices

$$C(p) = \left\langle (\mathbf{x}(t) - \widehat{\mu}_{\mathbf{x}})\,(\mathbf{x}(t-p) - \widehat{\mu}_{\mathbf{x}})^{\top} \right\rangle \ , \tag{1}$$

where $\langle\cdot\rangle$ denotes expectation. The process is now said to contain a strong asymmetry if for some i, j and some p it is found that $C_{i,j}(p) \neq C_{j,i}(p)$, i. e. $C(p)$ is asymmetric for at least one p. The process is said to contain a weak asymmetry if for some i, j and some p it is found that $C_{i,i}(p) \neq C_{j,j}(p)$, i. e. the diagonals are not all equal. Since the power spectrum of the i-th signal is given by the Fourier transform of $C_{i,i}(p)$ the process contains a weak asymmetry if and only if it contains signals with different power spectra.

Methods to detect causality are typically sensitive to both weak and strong asymmetries. Weak asymmetries can be detected more robustly but can also be considered as weaker evidence for causal relations. This can be illustrated if data are instantaneous mixtures of independent sources. In this case all cross-covariances are weighted sums of auto-covariances of the sources. Since auto-covariances are always symmetric functions of the delay p and since generally $C(-p) = C^{\top}(p)$ it follows that $C(p) = C^{\top}(p)$ for mixtures of independent sources [4]. Hence, such mixtures can only contain weak asymmetries but not strong ones.

For methods which are sensitive to both weak and strong asymmetries it is in general difficult to tell on what property of the data an estimate of causal drive is based. However, using empirical estimators of the cross spectra, it is possible to measure the proportions of weak and strong asymmetries in a dataset. In this paper, we demonstrate that a quantity called antisymmetry-symmetry-ratio is a meaningful predictor of the success of the causal estimation for methods that are knowingly affected by weak asymmetries. Moreover, we introduce a procedure based on time inversion, by which it is possible to test whether weak asymmetries are the dominant cause for a given connectivity estimate. We demonstrate that our approaches dramatically reduce the number of wrong predictions of Granger Causality (GC). As a result, GC's performance in the 2011 Signal Separation Evaluation Campaign (SiSEC) challenge on causal analysis of simulated EEG data is significantly improved. Our approaches can be regarded as sanity checks which are applicable in any causal analysis testing temporal delays between driver and receiver.

The paper starts with introducing Granger Causality, the SiSEC challenge dataset and the two novel approaches proposed to improve causal estimations in the Methods section. The Results section confirms that these approaches effectively reduce the number of wrong predictions of Granger Causality on the challenge dataset. In the Discussion section, we elaborate on the applicability of our approaches and draw connections to permutation testing, which is also typically used in conjunction with Granger-causal measures.

2 Methods

2.1 SISEC Challenge Simulated EEG Dataset

To demonstrate our ideas we consider a set of simulated EEG data, which is part of the 2011 Signal Separation Evaluation Campaign. The data consists of 1 000 examples of bivariate data for 6 000 time points. Each example is a superposition of a signal (of interest) and noise. The causally-interacting signals are constructed using a unidirectional bivariate autoregressive (AR) model of order 10 with (otherwise) random AR-parameters and uniformly distributed innovations. The noise is constructed of three independent sources, generated with three univariate AR-models with random parameters and uniformly distributed input, which were instantaneously mixed into the two sensors with a random mixing matrix. The relative strength of noise and signal (i. e. signal-to-noise ratio, SNR) was set randomly. The task of the challenge is to determine the direction of the causal interaction. One point is awarded for every correct prediction, while every wrong prediction causes a penalty of -10 points. If no prediction is given for a dataset, this results in 0 points. The maximum score attainable is 1 000 points, while the minimum score (considering that predictors with less than 50 % accuracy can be improved by sign-flipping) is -4 500 points.

The simulation addresses a conceptual problem of EEG data, namely that the signals of interest are superimposed by mixed noise. However, the actual spectra can be quite different from real EEG data. Volume conduction (i. e., mixing of the signals of interest), which is typically also observed in EEG datasets and poses serious challenges on its own [2], is omitted here in order to facilitate an objective evaluation. We use Matlab code provided by the organizers of the challenge to generate 1 000 new instances of the problem with known directions of causal flow.

2.2 Granger Causality

The *multivariate AR* (MVAR) model is given by

$$\mathbf{x}(t) = \sum_{p=1}^{P} B(p)\mathbf{x}(t-p) + \boldsymbol{\varepsilon}(t) \ , \tag{2}$$

where $B(p)$ are matrices describing the time-delayed influences of $\mathbf{x}(t-\tau)$ on $\mathbf{x}(t)$. Notably, the off-diagonal parts $B_{i,j}(p), i \neq j$ describe time-lagged influences between different time series. Granger Causality [1] involves fitting a multivariate AR model for the full set $\mathbf{x}_{\{1,\ldots,M\}} = \mathbf{x}$, as well as for the reduced set $\mathbf{x}_{\{1,\ldots,M\}\setminus\{i\}}$ of available time series, where $M = 2$ here. Denoting the prediction errors of the full model by $\boldsymbol{\varepsilon}^{\text{full}}$ and those of the reduced model by $\boldsymbol{\varepsilon}^{\setminus i}$, the *Granger score* GC describing the influence of x_i on x_j is defined as the log-ratio of the mean-squared errors (MSE) of the two models with respect to x_j. i. e.,

$$\mathrm{GC}_{i,j} = \log \left(\frac{\sum_{t=P+1}^{T} \left[\varepsilon_j^{\mathrm{full}}(t) \right]^2}{\sum_{t=P+1}^{T} \left[\varepsilon_j^{\setminus i}(t) \right]^2} \right) . \tag{3}$$

This definition, which is based on the ratio of prediction errors, is independent of the scale of the time series x_i and x_j. However, as has been demonstrated in [5], [6], it is influenced by asymmetries in the signal-to-noise ratio.

2.3 Exploiting Statistical Characterics of Non-/interacting Signals for Assessing the Reliability Causal Predictions

Due to additive noise and (in our case) innovation noise introduced by AR modeling, cross-covariances of realistic measurements are never exactly symmetric nor are they exactly antisymmetric. Nevertheless, the amount of symmetric vs. antisymmetric cross-covariance contained in a dataset provides important information about the SNR and hence how difficult the problem of estimating the causal direction is. We propose to use an index called antisymmetry-symmetry ratio (ASR) defined as

$$\mathrm{ASR} = \log \left(\frac{\left\| \left(\widehat{C}(1) - \widehat{C}^{\top}(1), \ldots, \widehat{C}(P) - \widehat{C}^{\top}(P) \right) \right\|_{\mathcal{F}}}{\left\| \left(\widehat{C}(1) + \widehat{C}^{\top}(1), \ldots, \widehat{C}(P) + \widehat{C}^{\top}(P) \right) \right\|_{\mathcal{F}}} \right) \tag{4}$$

for quantifying the confidence in a given causal estimation, where $(A1, \ldots, A_P)$ is the horizontal concatenation of the matrices $A_1, \ldots, A_P$, $A_{\mathcal{F}}$ denotes the Frobenius norm (sum of squared entries) of a matrix and $\widehat{C}(p)$ are empirical estimates of the cross-covariance matrices. The higher the ASR, the lower the proportion of (potentially misguiding) signal parts with symmetric cross-covariance is. Hence, one strategy to avoid false predictions in Granger- (and other) causal analyses is to evaluate only datasets characterized by high ASR.

2.4 A Test for Assessing the Time-Lagged Nature of Interactions

As a second simple test to distinguish weak from strong asymmetries we here suggest to compare the specific result of a causal analysis with the outcome of the method applied on time-reversed signals. This corresponds to the general intuitive idea that when all the signals are reversed in time, the direction of information flow should also reverse. More specifically, if temporal order is crucial to tell a driver from recipient the result can be expected to be reverted if the temporal order is reverted. The mathematical basis for this is the simple observation that the cross-covariance for the time inverted signals, say $\widetilde{C}(p)$, is given as

$$\widetilde{C}(p) = C(-p) = C^{\top}(p) \tag{5}$$

implying that time inversion inverts all strong asymmetries but none of the weak asymmetries. If now a specific measure is essentially identical for original and time inverted signals we conclude that the causal estimate in that specific case is based only on weak asymmetry. To avoid estimation biases introduced by weak asymmetries, one may therefore require that a causality measure delivers significant and *opposing* flows on original and time-reversed signals. Alternatively, one may require that the difference of the results obtained on original and time-reversed signals is significant.

2.5 Experiments

As baselines for the numerical evaluation, we apply Granger Causality as well as the Phase-slope Index (PSI) [5] to all 1 000 datasets and compute the respective score according to the rules of the SiSEC challenge. Granger Causality is calculated using the true model order $P = 10$. The Phase-slope Index is calculated using the authors' implementation[1] in a wide-band on segments of length $N = 100$. For both methods, *net flow*, i. e. the difference between the flows in both directions is assessed. Standard deviations of the methods' results are estimated using the jackknife method. Standardized results with absolute values greater than 2 are considered significantly different from zero. Insignificant results are not reported, i. e. lead to zero points in the evaluation. The whole procedure is repeated 100 times for different realizations of the 1 000 datasets to compute average challenge scores and confidence intervals.

The idea introduced in subsection 2.3 is implemented by ordering the datasets according to their ASR (calculated with $P = 30$), and evaluating the competition score attained when only the first K datasets with highest ASR are analyzed. That is, even significant results might be discarded, if the ASR is low. We consider three additional variants of GC, in which results are reported only if additional restrictions are met. The first variant, 'GC inv both' reports a causal net flow only if it is significant, and if the net flow on time-reversed data points to the opposite direction and is also significant. The variant 'GC inv diff' requires that the difference of the net flows estimated from original and time-reversed data is significantly different from zero. Finally, we compare time inversion to general random permutations of the samples (using the same permutation for all channel) according to the 'difference' approach. The resulting procedure is denoted by 'GC perm diff'.

3 Results

Figure 1 illustrates that interacting signals and mixed independent noise are characterized by different proportions of symmetric and antisymmetric parts in their cross-covariances. The upper-left plot depicts the log-norms of symmetric and antisymmetric cross-covariances of normalized signal and noise time series as a scatter plot, while the upper right plot depicts the respective ASR. In both

[1] http://ml.cs.tu-berlin.de/causality/

plots, signal and noise are highly separable. In the lower left plot, the ASR of the observation is plotted against the signal-to-noise ratio. Apparently, there exists a quasi-linear functional relationship between the two, which is the basis of our idea to use the ASR as an indicator for the difficulty of causal predictions.

Figure 2 summarizes the results of the numerical evaluation of the various causal prediction strategies according to the rules of the SiSEC challenge. For all methods considered, the challenge score is plotted as a function of the number of datasets analyzed (starting from datasets with highest ASR). The scores depicted on the very right hence correspond to the standard situation that all 1 000 datasets are analyzed. The scores obtained by the six contributors of the SiSec challenge are marked by black horizontal bars.

As in previous analyses [6], PSI outperforms Granger Causality having a total score of 593 ± 3 points compared to -438 ± 11 points after evaluation of all 1 000 datasets. However, as the plot also strikingly shows, the inferior performance of GC is a result of a huge number of false predictions predominantly made on data with low ASR. Hence, by avoiding decisions on low-ASR data, GC's score increases dramatically with the maximum of 384 ± 4 points reached if only the 539 datasets with highest ASR are analyzed. Note that this score is not anymore dramatically worse than the score obtained by PSI for the same amount of data, which is 485 ± 1 points. All three alternative variants of Granger Causality perform better than the conventional GC strategy with scores of 353 ± 2 points, 437 ± 5 points and 79 ± 4 points attained for 'GC inv both', 'GC inv diff' and 'GC perm diff', respectively when all datasets are analyzed. Note that this means that both 'GC inv both' and 'GC inv diff' outperform the winning contribution of the SiSec challenge, which achieved a score of 252 points. At the same time, the difference between the score attained when analyzing all 1 000 datasets and the maximal score attained when analyzing fewer datasets is dramatically reduced. This difference is 2 ± 1 points for 'GC inv both', 36 ± 2 points for 'GC inv diff' and 116 ± 4 points for 'GC perm diff', which is much closer to the value of 11 ± 1 points measured for PSI than to the value of 841 ± 10 points measured for conventional GC. Hence, all three proposed variants can be seen as robustifications of conventional GC, which prevent decisions that are solely based on weak asymmetries. Among the three proposed strategies, 'GC inv diff' performs best with scores that are competitive to those attained by PSI, while 'GC perm diff' performs worst. Note that the curve of 'GC inv diff' is located strictly above the curve of 'GC', which means that the additional restriction imposed by the time inversion causes no loss in performance for high-ASR data.

4 Discussion

Our results confirm that the proposed strategies drastically reduce the number of false predictions for methods that are prone to be dominated by weak asymmetries in the data such as Granger Causality. While for conventional Granger Causality the inclusion of the ASR as an additional criterion guiding the prediction is highly benefical, this is less helpful for modified variants that take

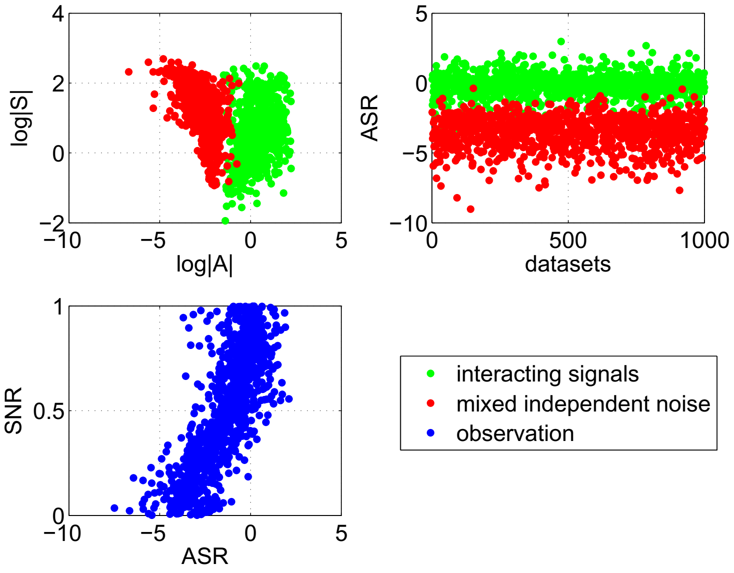


Fig. 1. Upper left: characterization of interacting signals and mixed independent noise by means of the log-norms of the symmetric and antisymmmetric parts of the cross-covariance matrices. Upper right: separation of signal and noise by means of the antisymmetry-symmetry ratio (ASR). Lower left: approximately linear relationship between the ASR of the observations and the signal-to-noise ratio (SNR).

the results obtained on time-reversed (or permuted) data into account. These modifications make GC behave more similarly to PSI, which is itself robust to many weak asymmetries by construction and in particular rather unaffected by dominant symmetric cross-covariances as indicated by low ASR. The choice of the ASR threshold remains an open problem, which is outside the scope of this paper. Empirical strategies to adjust the threshold are, however, conceivable.

Notably, the idea of performing pairwise testing of results obtained on original and time-reversed signals is a special case of permutation testing, as proposed, for example, by [3] in the context of Granger-causal analysis of EEG data using the directed transfer function (DTF). Both approaches have in common that the reordered data shares certain weak asymmetries with the original data, which are likely to cancel out in pairwise comparisons. However, time-reversed data additionally contains strong asymmetries in the opposite direction, which increases the statistical power of the comparison of original and time-reversed data. Consequently, our empirical results indicate that time inversion outperforms permutation testing by far and should be a viable alternative also when using DTF. Interestingly, PSI exactly flips its sign (direction) upon time inversion, for which reason pairwise testing against time-reversed data cannot be used to improve PSI.

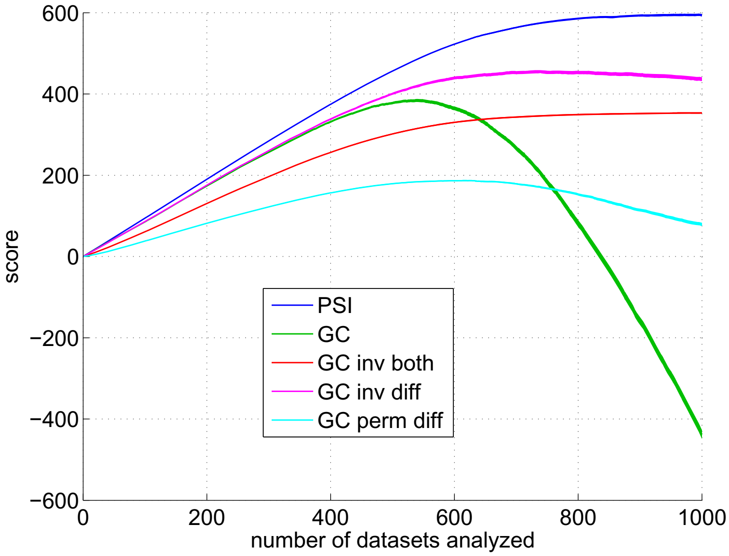


Fig. 2. Score according to the rules of the Signal Separation Evaluation Campaign (SiSEC) 2011 challenge on causal analysis of simulated EEG data as a function of the number of datasets analyzed for the Phase-slope Index (PSI) and different variants of Granger Causality (GC). Confidence intervals are indicated by linewidths. GC: original approach, requiring significant net flow. GC inv both: improved approach, requiring significant net flow and significant opposing net flow on time-reversed data. GC inv diff: improved approach, requiring significantly different net flows on original and time-reversed data. GC perm diff: improved approach, requiring significantly different net flows on original and temporally permuted data. Datasets are ordered by their antisymmetry-symmetry ratio (ASR) to illustrate that the analysis of datasets with low ASR with conventional Granger Causality is error-prone.

5 Conclusion

We proposed two strategies for robustifying Granger-causal analyses, which boost its performance in the SiSEC 2011 challenge.

References

1. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438 (1969)
2. Haufe, S.: Towards EEG source connectivity analysis. PhD thesis, TU Berlin (2011)
3. Kamiński, M., Ding, M., Truccolo, W.A., Bressler, S.L.: Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biol. Cybern. 85, 145–157 (2001)

4. Nolte, G., Meinecke, F.C., Ziehe, A., Müller, K.-R.: Identifying interactions in mixed and noisy complex systems. Phys. Rev. E 73, 051913 (2006)
5. Nolte, G., Ziehe, A., Nikulin, V.V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.R.: Robustly estimating the flow direction of information in complex physical systems. Phys. Rev. Lett. 100, 234101 (2008)
6. Nolte, G., Ziehe, A., Krämer, N., Popescu, F., Müller, K.-R.: Comparison of Granger causality and phase slope index. JMLR W&CP 6, 267–276 (2010)

Online PLCA for Real-Time Semi-supervised Source Separation

Zhiyao Duan[1,⋆], Gautham J. Mysore[2], and Paris Smaragdis[2,3]

[1] EECS Department, Northwestern University
[2] Advanced Technology Labs, Adobe Systems Inc
[3] University of Illinois at Urbana-Champaign

Abstract. Non-negative spectrogram factorization algorithms such as probabilistic latent component analysis (PLCA) have been shown to be quite powerful for source separation. When training data for all of the sources are available, it is trivial to learn their dictionaries beforehand and perform supervised source separation in an online fashion. However, in many real-world scenarios (e.g. speech denoising), training data for one of the sources can be hard to obtain beforehand (e.g. speech). In these cases, we need to perform semi-supervised source separation and learn a dictionary for that source during the separation process. Existing semi-supervised separation approaches are generally offline, i.e. they need to access the entire mixture when updating the dictionary. In this paper, we propose an online approach to adaptively learn this dictionary and separate the mixture over time. This enables us to perform online semi-supervised separation for real-time applications. We demonstrate this approach on real-time speech denoising.

1 Introduction

In recent years, non-negative matrix factorization (NMF) and its probabilistic counterparts such as probabilistic latent component analysis (PLCA) have been widely used for source separation [1]. The basic idea is to represent the magnitude spectrum of each time frame of the mixture signal as a linear combination of dictionary elements from source dictionaries. In the language of PLCA, for a sound mixture of two sources, this can be written as:

$$P_t(f) \approx \sum_{z \in \mathcal{S}_1 \bigcup \mathcal{S}_2} P(f|z)P_t(z) \text{ for } t = 1, \cdots, T \tag{1}$$

where T is the total number of frames; $P_t(f)$ is the normalized magnitude spectrum of the t-th frame of the mixture; $P(f|z)$ for $z \in \mathcal{S}_1$ and $z \in \mathcal{S}_2$ represent the elements (analogous to basis vectors) of the dictionaries of source 1 and source 2 respectively. $P_t(z)$ represents the activation weights of the different dictionary elements at time t. All these distributions are discrete and nonnegative.

[⋆] This work was performed while interning at Adobe Systems Inc.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 34–41, 2012.

Given a mixture spectrogram, we can estimate the dictionary elements and activation weights using the expectation–maximization (EM) algorithm. The source spectra in the t-th frame can then be reconstructed as $\sum_{z\in\mathcal{S}_1} P(f|z)P_t(z)$ and $\sum_{z\in\mathcal{S}_2} P(f|z)P_t(z)$, respectively. This is unfortunately a highly underconstrained problem and rarely leads to useful parameter estimates. One way to address this issue is to perform *supervised* source separation [1], in which we first learn the dictionaries for both sources from their isolated training data. Then in the separation stage, we fix these dictionaries and only the estimate the activation weights, from which we can reconstruct the spectra of each source.

However, in a lot of real-world problems, training data for one source might be hard to obtain beforehand. For example, in the application of speech denoising, we want to separate speech from noise. It is relatively easy to obtain training data for noise, but hard for speech. In these cases, we need to perform *semi-supervised* source separation [1], where we first learn the dictionary for one source (e.g. noise) from its training data beforehand, and then learn the dictionary for the other source (e.g. speech) in addition to the activation weights of both sources from the mixture. Finally, separation can be performed.

For supervised separation, the algorithm in [1] is intrinsically online, since the activation weights in different frames are estimated independently. For semi-supervised separation, however, the algorithm in [1] needs to access the entire mixture to learn the dictionary for the un-pretrained source, hence is offline.

In recent years, researchers have proposed several online NMF algorithms for dictionary learning in different applications (e.g. dictionary learning for image databases [2], document clustering [3], audio reconstruction [4]). The idea is to learn a dictionary to well explain the entire input data, after processing all the inputs, in an online fashion. However, we argue that these algorithms are not suitable for real-time semi-supervised source separation. The reason is that these algorithms only care about the final learned dictionary, after processing all of the input frames. They do not care about the intermediate estimates of the learned dictionary during processing the input frames. Therefore, the dictionary learned after receiving the current frame is not necessarily good enough to explain that frame and to separate it. In fact, processing all of the input frames once is often not enough and it has been shown that cycling over the input data set several times and randomly permuting samples at each cycle [2,3,4] improves the results.

In this paper, we propose an online PLCA algorithm tailored for real-time semi-supervised source separation. We learn the dictionary for the source that does not have training data, from the mixture, and apply it to separate the mixture, in an online fashion. When a new mixture frame comes in, the dictionary is adaptively updated to explain the current frame instead of explaining the entire mixture frames. In this way, we can use a much smaller-sized dictionary compared to the offline PLCA. We show that the performance of the proposed algorithm is almost as good as that of the offline PLCA algorithm (numerically equivalent to offline NMF using KL divergence), but significantly better than an existing online NMF algorithm for this application.

2 Proposed Algorithm

For the real-time source separation problem of two sources, assuming that some isolated training excerpts of source 1 ($\mathcal{S}_1$) are available beforehand and are long enough to capture $\mathcal{S}_1$'s characteristics, we can follow the semi-supervised source separation paradigm presented in Section 1[1]. We first learn a dictionary of $\mathcal{S}_1$ from the spectrogram of its training excerpts beforehand. Then during separation, as each incoming mixture frame arrives, we learn and update the dictionary of $\mathcal{S}_2$ using the proposed online PLCA algorithm, with $\mathcal{S}_1$'s dictionary fixed.

2.1 Online Separation and Dictionary Learning

In order to separate the t-th frame of the mixture signal, we need to decompose its magnitude spectrum using Eq. (1). Here, $P(f|z)$ for $z \in \mathcal{S}_1$ is the pre-learned dictionary of $\mathcal{S}_1$ from training excerpts, and is kept fixed in this decomposition. We need to estimate the dictionary $P(f|z)$ for $z \in \mathcal{S}_2$ and activation weights $P_t(z)$ for all z, such that the decomposition is as accurate as possible, i.e.

$$\underset{P(f|z) \text{ for } z\in\mathcal{S}_2,\ P_t(z) \text{ for all } z}{\arg\min} \quad d_{KL}(P_t(f)||Q_t(f)) \tag{2}$$

where d_{KL} is the KL divergence between two distributions. $P_t(f)$ is the normalized mixture spectrum at time t and $Q_t(f)$ is the reconstructed mixture spectrum i.e. the LHS and RHS of Eq. (1).

However, this is a highly unconstrained problem, since the number of parameters to estimate is much more than the number of equations (i.e. the number of frequency bins in Eq. (1)), even if there is only one element in $\mathcal{S}_2$'s dictionary. A trivial solution that makes the KL divergence in Eq. (2) equal to zero is to use only one dictionary element in $\mathcal{S}_2$, such that the dictionary element is the same as the mixture and the corresponding activation weight equals to one (with all other weights being zero). In practice, this trivial solution is almost always achieved, essentially making the separated source 2, the same as the mixture.

We therefore need to constrain the dictionary of source 2 to avoid this overfitting. We do this by requiring $\mathcal{S}_2$'s dictionary to not only explain $\mathcal{S}_2$'s spectrum in the current frame, but also those in a number of previous frames. We denote this set of frames as $\mathcal{B}$, representing a running buffer. We update $\mathcal{S}_2$'s dictionary in every frame using $\mathcal{B}$. We also set the size of $\mathcal{S}_2$'s dictionary to be much smaller than the size of $\mathcal{B}$. This avoids the overfitting because a compact dictionary will now be used to explain a much larger number of frames.

Clearly these buffer frames need to contain $\mathcal{S}_2$'s spectra, otherwise the dictionary will be incorrectly learned. We will describe how to determine if a mixture frame contains $\mathcal{S}_2$'s spectrum or not in Section 2.2. Suppose we can identify the previous mixture frames that contain $\mathcal{S}_2$'s spectra, we need to decide which ones to include in $\mathcal{B}$. On one hand, $\mathcal{S}_2$'s spectra in the buffer frames need to be different from those in the current frame, so that the learned $\mathcal{S}_2$'s dictionary does

[1] It is straightforward to extend this to N sources if isolated training excerpts for N-1 sources are available.

not overfit the mixture spectra in the current frame. On the other hand, we do not want $\mathcal{S}_2$'s spectra in the buffer frames to be too different from those in the current frame so that we have a more "localized" and compact dictionary. In the real-time source separation problem, it is intuitive to use the L most recent identified mixture frames to balance the tradeoff, as they are not the same as the current frame but tend to be similar. Based on this, the objective becomes:

$$\underset{P(f|z)\ \text{for}\ z\in\mathcal{S}_2,\ P_t(z)\ \text{for all}\ z}{\arg\min} d_{KL}(P_t(f)||Q_t(f)) + \frac{\alpha}{L}\sum_{s\in\mathcal{B}} d_{KL}(P_s(f)||Q_s(f)) \tag{3}$$

where α is the tradeoff between the original objective (good reconstruction of the current frame) and the added constraint (good reconstruction of buffer frames).

With this new objective, we learn $\mathcal{S}_2$'s dictionary and the current frame's activation weights. However, we fix the activation weights of the buffer frames as the values learned when separating them. There are two advantages of fixing them than updating them: First, it makes the algorithm faster. Second, it imposes a heavier constraint on $\mathcal{S}_2$'s dictionary that the newly learned dictionary must not deviate from those learned in the buffer frames too much. We use the EM algorithm to optimize Eq. (3), which is described in Algorithm 1.

Algorithm 1. Single Frame Dictionary Learning

```
Require: $\mathcal{B}$ (buffer frames set), $V_{fs}$ for $s \in \mathcal{B}\bigcup\{t\}$ (normalized magnitude spectra of
    buffer frames and current frame, each frame becomes a probability distribution),
    $P(f|z)$ for $z \in \mathcal{S}_1$ ($\mathcal{S}_1$'s dictionary), $P(f|z)$ for $z \in \mathcal{S}_2$ (initialization of $\mathcal{S}_2$'s dictionary),
    $P_s(z)$ for $s \in \mathcal{B}\bigcup\{t\}$ and $z \in \mathcal{S}_1\bigcup\mathcal{S}_2$ (input activation weights of buffer
    frames and current frame), $\alpha$ (tradeoff between reconstruction of buffer frames and
    current frame), $M$ (number of EM iterations).
for $i = 1$ to $M$ do
    E Step:
        $P_s(z|f) \leftarrow \frac{P_s(z)P(f|z)}{\sum_{z\in\mathcal{S}_1\bigcup\mathcal{S}_2} P_s(z)P(f|z)},\ \text{for}\ s \in \mathcal{B}\bigcup\{t\}. \quad (4)$
    M Step:
        $\phi(f|z) \leftarrow V_{ft}P_t(z|f) + \frac{\alpha}{|\mathcal{B}|}\sum_{s\in\mathcal{B}} V_{fs}P_s(z|f),\ \text{for}\ z \in \mathcal{S}_2, \quad (5)$
        $\phi_t(z) \leftarrow \sum_f V_{ft}P_t(z|f),\ \text{for}\ z \in \mathcal{S}_1\bigcup\mathcal{S}_2. \quad (6)$
        Normalize $\phi(f|z)$ and $\phi_t(z)$ to get $P(f|z)$ and $P_t(z)$ respectively.
end for
return learned dictionary $P(f|z)$ for $z \in \mathcal{S}_2$ and activation weights $P_t(z)$ for
    $z \in \mathcal{S}_1\bigcup\mathcal{S}_2$ of the current frame $t$.
```

2.2 Mixture Frame Classification

The problem that has not been addressed in Section 2.1 is how to determine whether a mixture frame contains $\mathcal{S}_2$'s spectrum or not. For a mixture of two sound sources, we can address this by decomposing the magnitude spectrum of the mixture frame using only the learned $\mathcal{S}_1$'s dictionary as follows:

$$P_t(f) \approx \sum_{z \in \mathcal{S}_1} P(f|z) P_t(z) \tag{7}$$

Since the dictionary is fixed, we learn only the activation weights. If the KL divergence between $P_t(f)$ and the RHS is smaller than a threshold θ_{KL}, it means the mixture spectrum can be well explained by only using $\mathcal{S}_1$'s dictionary, hence $\mathcal{S}_2$'s spectrum is not likely to be present. Otherwise, $\mathcal{S}_1$'s dictionary is not enough to explain the spectrum, hence $\mathcal{S}_2$'s spectrum is likely to be present.

We learn the threshold θ_{KL} by decomposing $\mathcal{S}_1$'s training excerpts again, with its pre-learned dictionary. We calculate the mean and standard deviation of the KL divergences of all the frames, and set the threshold as $\theta_{KL} = mean + std$.

If the current frame is classified as not containing $\mathcal{S}_2$, then we do not include it in the running buffer $\mathcal{B}$. However, just in case there is some amount of $\mathcal{S}_2$ in the frame, we still perform supervised separation on the frame using the pre-learned dictionary of $\mathcal{S}_1$ and the previously updated dictionary of $\mathcal{S}_2$. If the current frame is classified as containing $\mathcal{S}_2$, we run Algorithm 1 on it to update $\mathcal{S}_2$'s dictionary and separate the frame. After separation, we include this frame into the running buffer $\mathcal{B}$ for future use.

2.3 Algorithm Summary

The whole online semi-supervised source separation algorithm is summarized in Algorithm 2. Note that in Line 6 we make a "warm" initialization of $\mathcal{S}_2$'s dictionary using the one learned in the previous frame. This makes Algorithm 1 converge fast, as spectra in successive frames do not often change much.

3 Experiments

We test the proposed online semi-supervised source separation algorithm for real-time speech denoising. The two sources are therefore noise and speech. We learn the dictionary of noise from its training excerpts beforehand[2], and learn and update the dictionary of speech during real-time separation.

We use clean speech files and clean noise files to construct a noisy speech dataset for our experiments. For clean speech files, we use the full speech corpus in the NOIZEUS dataset[3]. This corpus has thirty short English sentences (each about three seconds long) spoken by three female and three male speakers. We concatenate sentences from the same speaker into one long sentence, and therefore obtain six long sentences, each of which is about fifteen seconds long.

For clean noise files, we collected ten different types of noise, including *birds*, *casino*, *cicadas*, *computer keyboard*, *eating chips*, *frogs*, *jungle*, *machine guns*,

[2] Training excerpts for noise is relatively easy to obtain in applications such as teleconferencing, since a few seconds at the beginning in which no one is talking are likely to be long enough to capture the noise characteristics throughout the teleconference.

[3] `http://www.utdallas.edu/~loizou/speech/noizeus/`

Algorithm 2. Online Semi-supervised Source Separation

```
Require: $V_{ft}$ for $t = 1, \cdots, T$ (magnitude spectra of the mixture signal), $P(f|z)$ for $z \in \mathcal{S}_1$ ($\mathcal{S}_1$'s dictionary), $P^{(0)}(f|z)$ for $z \in \mathcal{S}_2$ (random initialization of $\mathcal{S}_2$'s dictionary), $\theta_{KL}$ (threshold to classify a mixture frame), $\mathcal{B}$ (buffer frames set).
for $t = 1$ to $T$ do
    Decompose normalized magnitude spectrum $P_t(f) = \frac{V_{ft}}{\sum_f V_{ft}}$ by Eq. (7).
    if $d_{KL}(P_t(f) || \sum_{z \in \mathcal{S}_1} P(f|z)P_t(z)) < \theta_{KL}$ then
        Supervised separation using $P(f|z)$ for $z \in \mathcal{S}_1$ and $P^{(t-1)}(f|z)$ for $z \in \mathcal{S}_2$ and $P^{(t)}(f|z) \leftarrow P^{(t-1)}(f|z)$.
    else
        Learn $\mathcal{S}_2$'s dictionary $P^{(t)}(f|z)$ for $z \in \mathcal{S}_2$ and activation weights $P_t(z)$ using Algorithm 1, with $P^{(t)}(f|z)$ for $z \in \mathcal{S}_2$ initialized as $P^{(t-1)}(f|z)$.
        Set $\mathcal{S}_2$'s magnitude spectrum as:
            $$V_{ft} \frac{\sum_{z \in \mathcal{S}_2} P^{(t)}(f|z)P_t(z)}{\sum_{z \in \mathcal{S}_1} P(f|z)P_t(z) + \sum_{z \in \mathcal{S}_2} P^{(t)}(f|z)P_t(z)}. \quad (8)$$
        Replace the oldest frame in $\mathcal{B}$ with the $t$-th frame.
    end if
end for
return separated magnitude spectra of the current frame.
```

motorcycles and *ocean*. Each noise file is at least one minute long. The first twenty seconds are used to learn the noise dictionary. The rest are used to construct the noisy speech files.

We generate a noisy speech file by adding a clean speech file and a random portion of a clean noise file with one of the following signal-to-noise ratios (SNR): -10dB, -5dB, 0dB, 5dB and 10dB. By exploring all combinations of speech, noise and SNRs, we generate a total of 300 noisy speech files, each of which is about fifteen seconds long. The sampling rate of all the files is 16kHz.

For comparison, we run offline semi-supervised PLCA [1] (denoted as "PLCA") on this dataset. We segment the mixture into frames of 64ms long and 48ms overlap. We set the speech dictionary size as 20, since we find it is enough to get a perceptually good reconstruction of the clean speech files. We use different sizes of the noise dictionary for different noise types, due to their different characteristics and inherent complexities. We set this value by choosing from $\{1, 2, 5, 10, 20, 50, 100, 200\}$ the size that achieves the best denoising results in the condition of SNR of 0dB. The number of EM iterations is set to 100 as it always converged in that many iterations in our experiments.

We also implement an existing online NMF algorithm [4] (denoted as "O-IS-NMF"), which is designed for audio reconstruction. We apply it to this dataset in the semi-supervised paradigm. We use the same frame sizes and dictionary sizes as PLCA. As suggested in [4], we set the mini-batch parameter β to 1 to avoid inherent delay, and the scaling factor ρ to 1 to match β.

For the proposed algorithm, we use the same frame sizes and noise dictionary sizes as PLCA. We set the buffer size L as 60, which is about one second long.

Since the speech dictionary is only supposed to explain the speech spectra in the current frame and buffer frames, we can use a much smaller size of speech dictionary. We set this value to 7 (opposed to 20 in PLCA), since we find that the average KL divergence in decomposing one second of speech spectra with seven dictionary elements is about the same as that of the average KL divergence in decomposing fifteen seconds of speech spectra with twenty dictionary elements. We choose the tradeoff factor α for each different noise, from the set $\{1, 2, \cdots, 20\}$ as the one that achieves the best denoising results in the condition of SNR of 0dB. We run only 20 EM iterations in processing each frame, which we find almost assures convergence due to the "warm" initialization as described in Section 2.3.

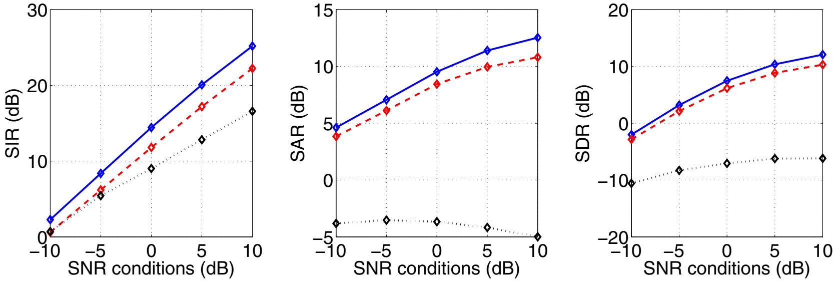


Fig. 1. Average performances on all types of noise of PLCA [1] (blue solid line), O-IS-NMF [4] (black dotted line) and the proposed algorithm (red dash line)

We use the BSS-EVAL metrics [5] to evaluate the separated speech files. Figure 1 shows the average results over all noise types and speakers, for each algorithm and SNR condition. Source-to-interference ratio (SIR) reflects noise suppression, source-to-artifacts ratio (SAR) reflects the artifacts introduced by the separation process, and source-to-distortion ratio (SDR) reflects the overall separation performance. It can be seen that for all the three metrics, the proposed algorithms achieves almost as good of a performance as PLCA. This is a promising result, since the proposed algorithm is an online algorithm and it uses a much smaller speech dictionary than PLCA. The performance of O-IS-NMF is significantly worse than PLCA and the proposed algorithm. As argued in Section 1, we think this algorithm is not suitable for real-time source separation.

Table 1 presents the performances of PLCA and the proposed algorithm for different noise types in the SNR condition of 0dB. The noise-specific parameters for the two algorithms are also presented. It can be seen that for different noise types, the results vary significantly. This is due to the inherent complexity of the noise and whether the training data can cover the noise characteristics or not. For some noise, like *birds*, *cicadas* and *frogs*, the performance of PLCA is significantly better than the proposed algorithm. For other noise like *casino*, *computer keyboard*, *machine guns* and *ocean*, the proposed algorithm achieves similar results to PLCA. The K_n parameter does not change much, except for

Table 1. Performances and noise-specific parameters for different noise types in the SNR condition of 0dB. K_n is the noise dictionary size and α is the tradeoff factor.

	SIR		SAR		SDR			
Noise type	PLCA	Proposed	PLCA	Proposed	PLCA	Proposed	K_n	α
birds	20.0	18.4	10.7	8.9	10.1	8.3	20	14
casino	5.3	7.5	8.6	7.2	3.2	3.9	10	13
cicadas	29.9	18.1	14.8	10.5	14.7	9.7	200	12
computer keyboard	18.5	12.2	8.9	10.2	8.3	7.9	20	3
eating chips	14.0	13.3	8.9	7.0	7.3	5.7	20	13
frogs	11.9	10.9	9.3	7.2	7.1	5.0	10	13
jungle	8.5	5.3	5.6	7.0	3.2	2.5	20	8
machine guns	19.3	16.0	11.8	11.5	10.9	10.0	10	2
motorcycles	10.2	8.0	7.9	7.0	5.6	4.5	10	10
ocean	6.8	7.4	8.8	8.0	4.3	4.3	10	10

the *cicada* noise. The α parameter is usually around 12, with the exception of *computer keyboard* and *machine gun* noise. Since these two noises are pulse-like noise with relatively simple spectra, the optimal α values are much smaller to have a weaker constraint.

The Matlab implementation of the proposed algorithms takes about 25 seconds to denoise each noisy speech file (which is about 15 seconds long), in a modern laptop computer with a 4-core 2.13GHz CPU. It would be easy to make it work in real-time in a C++ implementation or in a more advanced computer.

4 Conclusions

In this paper, we presented an online PLCA algorithm for real-time semi-supervised source separation. For the real-time speech denoising application, we showed that it achieves almost as good results as offline PLCA and significantly better results than an existing online NMF algorithm.

References

1. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 414–421. Springer, Heidelberg (2007)
2. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. J. Machine Learning Research 11, 19–60 (2010)
3. Wang, F., Tan, C., König, A.C., Li, P.: Efficient Document Clustering via Online Nonnegative Matrix Factorizations. In: SDM (2011)
4. Lefèvre, A., Bach, F., Févotte, C.: Online Algorithms for Nonnegative Matrix Factorization with the Itakura-Saito Divergence. In: WASPAA (2011)
5. Vincent, E., Fevotte, C., Gribonval, R.: Performance Measurement in Blind Audio Source Separation. IEEE Trans. on Audio Speech Lang. Process. 14(4), 1462–1469 (2006)

Cramér-Rao Bound for Circular Complex Independent Component Analysis

Benedikt Loesch and Bin Yang

Insitute for Signal Processing and System Theory, University of Stuttgart
{benedikt.loesch,bin.yang}@iss.uni-stuttgart.de

Abstract. Despite an increased interest in complex independent component analysis (ICA) during the last two decades, a closed-form expression for the Cramér-Rao bound (CRB) of the complex ICA problem has not yet been established. In this paper, we fill this gap for the noiseless case and circular sources. The CRB depends on the distributions of the sources only through two characteristic values which can be easily calculated. In addition, we study the CRB for the family of circular complex generalized Gaussian distributions (GGD) in more detail and compare it to simulation results using several ICA estimators.

Keywords: Cramér-Rao bound, Fisher Information, independent component analysis, blind source separation, circular complex distribution.

1 Introduction

Independent Component Analysis (ICA) is a relatively recent signal processing method to extract unobservable source signals or independent components from their observed linear mixtures. We assume a linear square noiseless mixing model

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{x} \in \mathbb{C}^N$ are N linear combinations of the N source signals $\mathbf{s} \in \mathbb{C}^N$. We make the following assumptions:

A1. The mixing matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ is deterministic and invertible.

A2. $\mathbf{s} = [s_1, \cdots, s_N]^T \in \mathbb{C}^N$ are N independent random variables with zero mean and unit variance (after scaling the rows of $\mathbf{A}$ suitably). The probability density functions (pdfs) $p_i(s_i)$ of s_i can be different. We assume the sources to be circular, i.e. $p_i(s_i) = p_i(s_i e^{j\alpha})$ $\forall\alpha \in \mathbb{R}$. Hence $E[s_i^2] = 0$. Furthermore, $p_i(s_i)$ is continuously differentiable with respect to s_i and s_i^* in the sense of Wirtinger derivatives [1] which will be introduced in Sect. 2. The expectations in (15) and (20) exist.

The task of ICA is to demix the signals $\mathbf{x}$ by a demixing matrix $\mathbf{W} \in \mathbb{C}^{N\times N}$

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} \tag{2}$$

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 42–49, 2012.

such that $\mathbf{y}$ is "as close to $\mathbf{s}$" as possible according to some metric. The ideal solution for $\mathbf{W}$ is $\mathbf{A}^{-1}$ neglecting scaling, phase and permutation ambiguity [2].

It is very useful, to have a lower bound for the variance of estimation of $\mathbf{W}$. The Cramér-Rao bound (CRB) provides a lower bound on the covariance matrix of any unbiased estimator of a parameter vector. Although much research in the field of ICA has been undertaken, a closed-form expression for the CRB of the real instantaneous ICA problem has been derived only recently [3, 4]. However, in many practical applications, such as telecommunication or audio processing in frequency domain, the signals are complex. Although many different algorithms for complex ICA have been proposed [5–9], the CRB for this problem has not yet been established. In this paper, we fill this gap by deriving closed-form expressions for the CRB of the vectorized parameter $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}^T)$ and for the CRB of $\boldsymbol{\vartheta} = \mathrm{vec}((\mathbf{WA})^T)$. Due to the intrinsic phase ambiguity in circular complex ICA (cf. A2.: $p_i(s_i) = p_i(s_i e^{j\alpha})\ \forall \alpha \in \mathbb{R}$), we can only derive a CRB with the constraint $[\mathbf{WA}]_{ii} \in \mathbb{R}$. The CRB depends on the distributions of the sources only through two scalars defined in (15) which can be easily calculated.

2 Prerequisites

2.1 Complex Functions and Complex Random Vectors

Define the partial derivative of a complex function $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{u}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + j\mathbf{v}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\alpha} = \Re[\boldsymbol{\theta}]$ as $\partial\mathbf{g}/\partial\boldsymbol{\alpha} = \partial\mathbf{u}/\partial\boldsymbol{\alpha} + j\partial\mathbf{v}/\partial\boldsymbol{\alpha}$ and with respect to $\boldsymbol{\beta} = \Im[\boldsymbol{\theta}]$ as $\partial\mathbf{g}/\partial\boldsymbol{\beta} = \partial\mathbf{u}/\partial\boldsymbol{\beta} + j\partial\mathbf{v}/\partial\boldsymbol{\beta}$. Then the complex partial differential operators $\partial/\partial\boldsymbol{\theta}$ and $\partial/\partial\boldsymbol{\theta}^*$ are defined as

$$\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}} = \frac{1}{2}\left(\frac{\partial\mathbf{g}}{\partial\boldsymbol{\alpha}} - j\frac{\partial\mathbf{g}}{\partial\boldsymbol{\beta}}\right), \quad \frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}^*} = \frac{1}{2}\left(\frac{\partial\mathbf{g}}{\partial\boldsymbol{\alpha}} + j\frac{\partial\mathbf{g}}{\partial\boldsymbol{\beta}}\right). \tag{3}$$

These differential operators have first been introduced for real valued $\mathbf{g}$ by Wirtinger [1]. As long as the real and imaginary part of a complex function $\mathbf{g}$ are real-differentiable, the two Wirtinger derivatives in (3) also exist [10]. The direction of steepest descent of a real function $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{u}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is given by $\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}^*}$ and not $\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}}$ [11]. The complex Jacobian matrix of a complex function $\mathbf{g}$: $\mathbb{C}^M \to \mathbb{C}^N$ is defined as the complex $2N \times 2M$ matrix

$$\mathbf{D}_{\mathbf{g}} = \begin{bmatrix} \frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}} & \frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}^*} \\ \left(\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}^*}\right)^* & \left(\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}}\right)^* \end{bmatrix}, \tag{4}$$

i.e. it is the augmented matrix of $\partial\mathbf{g}/\partial\boldsymbol{\theta}$ and $\partial\mathbf{g}/\partial\boldsymbol{\theta}^*$. The covariance matrix of a complex random vector $\mathbf{x} = \mathbf{x}_R + j\mathbf{x}_I \in \mathbb{C}^N$ is $\mathrm{cov}(\mathbf{x}) = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^H\right]$. The pseudo-covariance matrix of $\mathbf{x}$ is $\mathrm{pcov}(\mathbf{x}) = E\left[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T\right]$.

2.2 Cramér-Rao Bound for a Complex Parameter

We briefly review the CRB for complex parameters (see for example [12]) before we derive the CRB for circular complex ICA. Assume that L observations of

$\mathbf{x}$ are i.i.d. distributed having the pdf $p(\mathbf{x};\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$. The complex Fisher Information Matrix (FIM) of complex parameter $\boldsymbol{\theta}$ is defined as

$$\mathbf{J}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\theta}} & \mathcal{P}_{\boldsymbol{\theta}} \\ \mathcal{P}_{\boldsymbol{\theta}}^* & \mathcal{I}_{\boldsymbol{\theta}}^* \end{bmatrix}, \tag{5}$$

where $\mathcal{I}_{\boldsymbol{\theta}} = E\left[\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta})\{\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta})\}^H\right]$ is called the information matrix and $\mathcal{P}_{\boldsymbol{\theta}} = E\left[\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta})\{\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta})\}^T\right]$ the pseudo-information matrix. Here $\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{2}\left(\nabla_{\boldsymbol{\alpha}} \log p(\mathbf{x};\boldsymbol{\theta}) + j\nabla_{\boldsymbol{\beta}} \log p(\mathbf{x};\boldsymbol{\theta})\right)$ is the column gradient vector of $\log p(\mathbf{x};\boldsymbol{\theta})$, i.e. $[\partial/\partial\theta_1^*, \cdots, \partial/\partial\theta_N^*]^T \log p(\mathbf{x};\boldsymbol{\theta})$.

The inverse of the FIM of $\boldsymbol{\theta}$ gives, under regularity conditions, the CRB of the augmented covariance matrix of an unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and hence

$$\begin{bmatrix} \mathrm{cov}(\hat{\boldsymbol{\theta}}) & \mathrm{pcov}(\hat{\boldsymbol{\theta}}) \\ \mathrm{pcov}(\hat{\boldsymbol{\theta}})^* & \mathrm{cov}(\hat{\boldsymbol{\theta}})^* \end{bmatrix} \geq L^{-1}\mathbf{J}_{\boldsymbol{\theta}}^{-1} = \frac{1}{L}\begin{bmatrix} \mathcal{I}_{\boldsymbol{\theta}} & \mathcal{P}_{\boldsymbol{\theta}} \\ \mathcal{P}_{\boldsymbol{\theta}}^* & \mathcal{I}_{\boldsymbol{\theta}}^* \end{bmatrix}^{-1}. \tag{6}$$

It holds $\mathrm{cov}(\hat{\boldsymbol{\theta}}) \geq L^{-1}(\mathcal{I}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}}\mathcal{I}_{\boldsymbol{\theta}}^{-*}\mathcal{P}_{\boldsymbol{\theta}}^*)^{-1} = L^{-1}\mathbf{R}_{\boldsymbol{\theta}}^{-1}$ with $\mathbf{R}_{\boldsymbol{\theta}} = \mathcal{I}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}}\mathcal{I}_{\boldsymbol{\theta}}^{-*}\mathcal{P}_{\boldsymbol{\theta}}^*$. The CRB for a transformed vector $\boldsymbol{\vartheta} = \mathbf{g}(\boldsymbol{\theta})$ is given by the right-hand-side of

$$\begin{bmatrix} \mathrm{cov}(\hat{\boldsymbol{\vartheta}}) & \mathrm{pcov}(\hat{\boldsymbol{\vartheta}}) \\ \mathrm{pcov}(\hat{\boldsymbol{\vartheta}})^* & \mathrm{cov}(\hat{\boldsymbol{\vartheta}})^* \end{bmatrix} \geq L^{-1}\mathbf{D}_{\mathbf{g}}\mathbf{J}_{\boldsymbol{\theta}}^{-1}\mathbf{D}_{\mathbf{g}}^T. \tag{7}$$

3 Derivation of Cramér-Rao Bound

In ICA, the parameter of interest is the demixing matrix $\mathbf{W}$. We form the parameter vector $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}^T) = [\mathbf{w}_1^T, \cdots, \mathbf{w}_N^T]^T \in \mathbb{C}^{N^2}$, where $\mathbf{w}_i \in \mathbb{C}^N$ are the row vectors of $\mathbf{W}$. The operator $\mathrm{vec}(\cdot)$ stacks the columns of its argument into one long column vector. The pdf of $\mathbf{x} = \mathbf{As}$ is defined as $p(\mathbf{x};\boldsymbol{\theta}) = |\det(\mathbf{W})|^2 \prod_{i=1}^N p_i(\mathbf{w}_i\mathbf{x})$, where $p_i(s_i)$ denotes the pdf of s_i and $\mathbf{W} = \mathbf{A}^{-1}$. By using matrix derivatives, we obtain

$$\frac{\partial}{\partial \mathbf{W}^H} \log p(\mathbf{x};\boldsymbol{\theta}) = \mathbf{A}^* - \mathbf{x}^*\boldsymbol{\varphi}^T(\mathbf{Wx}) = \mathbf{A}^*(\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s}))^* \tag{8}$$

where $\boldsymbol{\varphi}(\mathbf{s}) = [\varphi_1(s_1), \cdots, \varphi_N(s_N)]^T$ and $\varphi_i(s_i) = -\frac{\partial}{\partial s_i^*} \log p_i(s_i)$.

Since $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{W}^T)$, we get $\nabla_{\boldsymbol{\theta}^*} \log p(\mathbf{x};\boldsymbol{\theta}) = \mathrm{vec}\left(\frac{\partial}{\partial \mathbf{W}^H} \log p(\mathbf{x};\boldsymbol{\theta})\right)$ and

$$\mathcal{I}_{\boldsymbol{\theta}} = \left((\mathbf{I}\otimes\mathbf{A})\mathbf{M}_1(\mathbf{I}\otimes\mathbf{A}^H)\right)^*, \quad \mathcal{P}_{\boldsymbol{\theta}} = \left((\mathbf{I}\otimes\mathbf{A})\mathbf{M}_2(\mathbf{I}\otimes\mathbf{A}^T)\right)^*, \tag{9}$$

where $\mathbf{M}_1 = E\left[\mathrm{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}\mathrm{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}^H\right]$, $\mathbf{M}_2 = E\left[\mathrm{vec}\{...\}\mathrm{vec}\{...\}^T\right]$ and $\otimes$ denotes the Kronecker product.

3.1 CRB for G = WA

For simplicity, we first derive the CRB for the transformed parameter $\boldsymbol{\vartheta} = \mathrm{vec}((\mathbf{WA})^T) = (\mathbf{I}\otimes\mathbf{A}^T)\boldsymbol{\theta}$. The covariance of $\hat{\boldsymbol{\vartheta}} = \mathrm{vec}((\hat{\mathbf{W}}\mathbf{A})^T)$ is given by $\mathrm{cov}(\hat{\boldsymbol{\vartheta}}) = (\mathbf{I}\otimes\mathbf{A}^T)\mathrm{cov}(\hat{\boldsymbol{\theta}})(\mathbf{I}\otimes\mathbf{A}^*)$ where $\hat{\boldsymbol{\theta}} = \mathrm{vec}(\hat{\mathbf{W}}^T)$. Hence it holds

$$\mathrm{cov}(\hat{\boldsymbol{\vartheta}}) \geq L^{-1}(\mathbf{I}\otimes\mathbf{A}^T)(\mathcal{I}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}}\mathcal{I}_{\boldsymbol{\theta}}^{-*}\mathcal{P}_{\boldsymbol{\theta}}^*)^{-1}(\mathbf{I}\otimes\mathbf{A}^*) = L^{-1}\mathbf{R}_{\boldsymbol{\vartheta}}^{-1} \tag{10}$$

with $\mathbf{R}_{\boldsymbol{\vartheta}} = (\mathbf{M}_1 - \mathbf{M}_2\mathbf{M}_1^{-*}\mathbf{M}_2^*)^*$.

As shown in the appendix, $\mathbf{R}_{\boldsymbol{\vartheta}} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{\kappa_i \kappa_j - 1}{\kappa_j} \right) \mathbf{L}_{ii} \otimes \mathbf{L}_{jj}$, with $\kappa_i = E\left[|\varphi_i(s_i)|^2 \right]$. $\mathbf{L}_{ii}$ denotes an $N \times N$ matrix with a 1 at the (i,i) position and 0's elsewhere. $\mathbf{R}_{\boldsymbol{\vartheta}}$ is a diagonal matrix of rank $N^2 - N$. The CRB for G_{ij} yields

$$\operatorname{var}(\hat{G}_{ij}) \geq \frac{1}{L} \frac{\kappa_j}{\kappa_i \kappa_j - 1} \quad i \neq j \tag{11}$$

where $\hat{\mathbf{G}} = \hat{\mathbf{W}}\mathbf{A}$. Eq. (11) looks the same as in the real case [3, 4], but in the complex case κ_i is defined using Wirtinger derivatives instead of real derivatives.

Due to the phase ambiguity in circular complex ICA, the Fisher information for the diagonal elements G_{ii} is 0 and hence their CRB does not exist. However, we can constrain G_{ii} to be real and derive the constrained CRB [13] for $\theta = G_{ii}$: The constraint can be formulated as $f(\theta) = \theta - \theta^* = 0$. We then need to calculate $\mathbf{F}(\theta) = \begin{bmatrix} \partial f/\partial\theta & \partial f/\partial\theta^* \\ \partial f^*/\partial\theta & \partial f^*/\partial\theta^* \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ and find an orthonormal 2×1 matrix $\mathbf{U}$ in the null-space of $\mathbf{F}(\theta)$, i.e. $\mathbf{FU} = \mathbf{0}$. We choose $\mathbf{U} = 1/\sqrt{2}\left[1\ 1\right]^T$. The CRB for the constrained parameter $\theta = G_{ii}$ then yields

$$\begin{bmatrix} \operatorname{var}(\theta) & \operatorname{pvar}(\theta) \\ \operatorname{pvar}^*(\theta) & \operatorname{var}(\theta) \end{bmatrix} \geq \frac{1}{L}\mathbf{U}\left(\mathbf{U}^H \begin{bmatrix} \mathcal{I}_\theta & \mathcal{P}_\theta \\ \mathcal{P}_\theta^* & \mathcal{I}_\theta \end{bmatrix} \mathbf{U} \right)^{-1} \mathbf{U}^H = \frac{1}{4L(\eta_i - 1)} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{12}$$

where $\mathcal{I}_\theta = \eta_i - 1 = \mathcal{P}_\theta$ and $\eta_i = E\left[|s_i|^2 |\varphi_i(s_i)|^2 \right]$. The CRB in (12) is valid for a phase-corrected G_{ii} such that $G_{ii} \in \mathbb{R}$. Eq. (12) matches the real case [3, 4], where $\operatorname{var}(\hat{G}_{ij}) \geq L^{-1}(\bar{\eta}_i - 1)^{-1}$ since η_i is defined using Wirtinger derivatives instead of real derivatives and hence for the real case $4(\eta_i - 1) = \bar{\eta}_i - 1$.

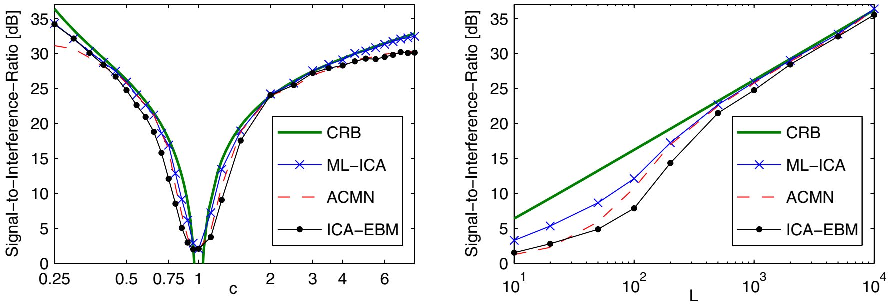
Performance of ICA is often measured using $\hat{\mathbf{G}}$ and hence it can be directly compared to (11), (12). The absolute values of the diagonal elements $|\hat{G}_{ii}|$ should be close to 1. They reflect how well we can estimate the power of each component. The absolute values of the off-diagonal elements $|\hat{G}_{ij}|$ should be close to 0 and reflect how well we can suppress interfering components.

3.2 CRB for W

It holds $\operatorname{vec}(\mathbf{W}^T) = \boldsymbol{\theta} = (\mathbf{I} \otimes \mathbf{A}^T)^{-1}\boldsymbol{\vartheta} = (\mathbf{I} \otimes \mathbf{W}^T)\boldsymbol{\vartheta}$ since $\mathbf{W} = \mathbf{A}^{-1}$. We can estimate the rows of $\mathbf{W}$ only up to an arbitrary phase for each row. We can derive a CRB for the phase-corrected $\mathbf{W}$, for which $\left[\mathbf{WA}\right]_{ii} \in \mathbb{R}$: We use the CRB for the constrained G_{ii} (12) together with the CRB for G_{ij} (11) to form the inverse FIM for the constrained $\mathbf{G}$ as $\mathbf{R}_{\boldsymbol{\vartheta}}^{-1} = \sum_{i=1}^{N} \frac{1}{4(\eta_i - 1)} \mathbf{L}_{ii} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{\kappa_i \kappa_j - 1}{\kappa_j} \right) \mathbf{L}_{ii} \otimes \mathbf{L}_{jj}$. The CRB for constrained $\mathbf{W}$ is then given by $\mathbf{R}_{\boldsymbol{\theta}}^{-1} = (\mathbf{I} \otimes \mathbf{W}^T)\mathbf{R}_{\boldsymbol{\vartheta}}^{-1}(\mathbf{I} \otimes \mathbf{W}^*)$ and $\operatorname{cov}(\hat{\boldsymbol{\theta}}) \geq L^{-1}\mathbf{R}_{\boldsymbol{\theta}}^{-1}$.

4 Results for Generalized Gaussian Distribution (GGD)

A circular complex GGD with zero mean and variance $E[|s|^2] = 1$ is given by the pdf $p(s, s^*) = \frac{c\alpha}{\pi\Gamma(1/c)} \exp\left(-\left[\alpha s s^*\right]^c\right)$ [14], with $\alpha = (\Gamma(2/c))/(\Gamma(1/c))$. $\Gamma(\cdot)$



(a) Varying shape parameter c, $L = 1000$ **(b)** Varying sample size L, $c = 0.5$

Fig. 1. Comparison of performance of three ICA estimators with CRB

denotes the Gamma function. The shape parameter $c > 0$ varies the form of the pdf from super-Gaussian ($c < 1$) to sub-Gaussian ($c > 1$). For $c = 1$, the pdf is Gaussian. By integration in polar coordinates, we find κ, η and β in (15) and (20) as $\kappa = \frac{c^2 \Gamma(2/c)}{\Gamma^2(1/c)}$, $\eta = \beta = c + 1$. For the simulation study, we consider $N = 3$ identically distributed sources with random mixing matrices $\mathbf{A}$ with independent uniform distributions for the real and imaginary parts of each entry (between -1 and 1). We conduct 100 experiments with different $\mathbf{A}$ and different realizations of the source signals and consider the following different ICA estimators: Complex ML-ICA [7], adaptable complex maximization of nongaussianity (ACMN) [9] and complex ICA by entropy bound minimization (ICA-EBM) [8]. We correct for permutation ambiguity and then calculate the signal-to-interference ratio (SIR) averaged over all N sources: $\text{SIR} = \frac{1}{N}\sum_i \left(E\left[|G_{ii}|^2\right] / \sum_{j \neq i} E\left[|G_{ij}|^2\right] \right)$. Fig. 1 (a) compares the SIR given by the CRB with the empirical SIR of the different ICA estimators for varying shape parameter c and a sample size of $L = 1000$. Since all sources are identically distributed, $\text{CRB}(G_{ij}) \to \infty$ and $\text{SIR} \to 0$ for $c \to 1$ (Gaussian). In this case, ICA fails to separate the sources. Clearly, the performance of complex ML-ICA is close to the CRB for a wide range of the shape parameter c. ACMN outperforms ICA-EBM in most cases except for strongly super-Gaussian sources: ACMN uses a GGD model and hence is better suited for separating circular GGD sources. However, ACMN uses prewhitening and then constrains the demixing matrix to be unitary which ICA-EBM does not. Fig. 1 (b) studies the influence of sample size L on ICA performance for $c = 0.5$. Again, complex ML-ICA performs the best as expected. Except for small sample sizes, all algorithms come quite close to the CRB.

5 Conclusion

In this paper, we have derived the CRB for the noiseless ICA problem with circular complex sources. Due to the phase ambiguity in circular complex ICA, the CRB for the diagonal elements of the demixing-mixing-matrix-product $\mathbf{G} = \mathbf{WA}$

does not exist, but a constrained CRB with $G_{ii} \in \mathbb{R}$ can be derived. Simulation results with sources following a circular complex generalized Gaussian distribution have shown that for large enough sample size some ICA estimators can achieve a signal-to-interference ratio close to that given by the CRB.

A Useful Matrix Algebra

Similarly to [4], we make use of some matrix algebra in the derivation of the CRB. We briefly review the required properties here: Let $\mathbf{L}_{ij}$ denote a $N \times N$ matrix with a 1 at the (i, j) position and 0's elsewhere. It is useful to note that

$$\mathbf{A}\mathbf{L}_{ij}\mathbf{A}^T = \mathbf{a}_i\mathbf{a}_j^T, \quad \mathbf{L}_{ij}\mathbf{L}_{kl} = \mathbf{0} \text{ for } j \neq k, \quad \mathbf{L}_{ij}\mathbf{L}_{jl} = \mathbf{L}_{il} \tag{13}$$

where $\otimes$ denotes the Kronecker product. We also note that any $N^2 \times N^2$ block matrix $\mathbf{A}$ can be written using its $N \times N$ diagonal blocks $\mathbf{A}[i, i]$ and $N \times N$ off-diagonal blocks $\mathbf{A}[i, j], i \neq j$ as follows:

$$\mathbf{A} = \sum_{i=1}^{N} \mathbf{L}_{ii} \otimes \mathbf{A}[i, i] + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbf{L}_{ij} \otimes \mathbf{A}[i, j]. \tag{14}$$

B Some Steps in the Derivation of the CRB for G

The derivation of the CRB for $\mathbf{G}$, proceeds in three steps: First, we calculate $\mathbf{M}_1$ and $\mathbf{M}_2$. Then, we obtain $\mathbf{R}_{\boldsymbol{\vartheta}} = (\mathbf{M}_1 - \mathbf{M}_2\mathbf{M}_1^{-*}\mathbf{M}_2^*)^*$ and finally invert $\mathbf{R}_{\boldsymbol{\vartheta}}$.

Using $E[\mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})] = \mathbf{I}$, we can simplify $\mathbf{M}_1$ as

$$\mathbf{M}_1 = E\left[\text{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}\text{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}^H\right] = \boldsymbol{\Omega}_1 - \text{vec}\{\mathbf{I}\}\text{vec}\{\mathbf{I}\}^H,$$

where $\boldsymbol{\Omega}_1 = E\left[\text{vec}\{\mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}\text{vec}\{\mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}^H\right]$ is a $N^2 \times N^2$ block matrix. The (i, i) block $\boldsymbol{\Omega}_1[i, i] = E\left[\mathbf{s}\mathbf{s}^H|\varphi_i(s_i)|^2\right]$ is diagonal since the components of $\mathbf{s}$ are independent and zero mean. The diagonal elements $\boldsymbol{\Omega}_1[i, i]_{(j,j)}$ are given by

$$\boldsymbol{\Omega}_1[i, i]_{(j,j)} = \begin{cases} E\left[|s_i|^2|\varphi_i(s_i)|^2\right] =: \eta_i & i = j \\ E\left[|s_j|^2|\varphi_i(s_i)|^2\right] = E\left[|\varphi_i(s_i)|^2\right] =: \kappa_i & i \neq j \end{cases}. \tag{15}$$

κ_i and η_i are real since $E[g(s)]$ with $g(s) \in \mathbb{R}$ is real. The (i, j) block $\boldsymbol{\Omega}_1[i, j]$ $(i \neq j)$ can be calculated as $\boldsymbol{\Omega}_1[i, j] = E\left[\mathbf{s}\mathbf{s}^H\varphi_i^*(s_i)\varphi_j(s_j)\right]$. It has 1 at entry (i, j) and 0 at entry (j, i), since

$$\boldsymbol{\Omega}_1[i, j]_{(i,j)} = E\left[s_i s_j^* \varphi_i^*(s_i)\varphi_j(s_j)\right] = E\left[s_i\varphi_i^*(s_i)\right]E\left[s_j^*\varphi_j(s_j)\right] = 1, \tag{16}$$

$$\boldsymbol{\Omega}_1[i, j]_{(j,i)} = E\left[s_i^* s_j \varphi_i^*(s_i)\varphi_j(s_j)\right] = E\left[s_i^*\varphi_i^*(s_i)\right]E\left[s_j\varphi_j(s_j)\right] = 0. \tag{17}$$

All other entries of $\boldsymbol{\Omega}_1[i, j]$ are zero since the components of $\mathbf{s}$ are independent and zero mean. Using the matrix algebra from appendix A, we can write $\boldsymbol{\Omega}_1$ as

$$\boldsymbol{\Omega}_1 = \sum_{i=1}^{N} \eta_i \mathbf{L}_{ii} \otimes \mathbf{L}_{ii} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \kappa_i \mathbf{L}_{ii} \otimes \mathbf{L}_{jj} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbf{L}_{ij} \otimes \mathbf{L}_{ij}. \tag{18}$$

Using $\mathrm{vec}\{\mathbf{I}\}\mathrm{vec}\{\mathbf{I}\}^H = \sum_{i=1}^N \sum_{\substack{j=1\\j\neq i}}^N \mathbf{L}_{ij} \otimes \mathbf{L}_{ij} + \sum_{i=1}^N \mathbf{L}_{ii} \otimes \mathbf{L}_{ii}$, we get $\mathbf{M}_1$ as

$$\mathbf{M}_1 = \sum_{i=1}^N (\eta_i - 1)\mathbf{L}_{ii} \otimes \mathbf{L}_{ii} + \sum_{i=1}^N \sum_{\substack{j=1\\j\neq i}}^N \kappa_i \mathbf{L}_{ii} \otimes \mathbf{L}_{jj}. \tag{19}$$

We note that $\mathbf{M}_1$ is a real diagonal matrix.

$\mathbf{M}_2$ can be calculated similarly. It holds:

$$\mathbf{M}_2 = E\left[\mathrm{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}\mathrm{vec}\{\mathbf{I} - \mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}^T\right] = \boldsymbol{\Omega}_2 - \mathrm{vec}\{\mathbf{I}\}\mathrm{vec}\{\mathbf{I}\}^T,$$

where $\boldsymbol{\Omega}_2 = E\left[\mathrm{vec}\{\mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}\mathrm{vec}\{\mathbf{s}\boldsymbol{\varphi}^H(\mathbf{s})\}^T\right]$ is a $N^2 \times N^2$ block matrix. The (i,i) block $\boldsymbol{\Omega}_2[i,i] = E\left[\mathbf{s}\mathbf{s}^T(\varphi_i^*(s_i))^2\right]$ is diagonal since the components of $\mathbf{s}$ are independent and zero mean. The diagonal elements $\boldsymbol{\Omega}_2[i,i]_{(j,j)}$ are given by

$$\boldsymbol{\Omega}_2[i,i]_{(j,j)} = \begin{cases} E\left[s_i^2(\varphi_i^*(s_i))^2\right] =: \beta_i & i = j \\ E\left[s_j^2(\varphi_i^*(s_i)|^2\right] = E\left[s_j^2\right] E\left[(\varphi_i^*(s_i))^2\right] = 0 & i \neq j, \end{cases}. \tag{20}$$

since $E\left[s_j^2\right] = 0$. If s_i is circular, it can be shown that $\beta_i = \eta_i$: For circular $s = s_R + js_I$, $p(-s_R, s_I) = p(s_R, s_I)$, $p(s_R, -s_I) = p(s_R, s_I)$ and $p(s_R, s_I) = g(s_R^2 + s_I^2)$. Let $f(r^2) = f(s_R^2 + s_I^2) = \log p(s_R, s_I)$. It holds

$$\begin{aligned}
\beta &= \frac{1}{4}E\left[\left(s_R^2 + s_I^2\right)\left(\left(\frac{\partial f}{\partial s_R}\right)^2 + \left(\frac{\partial f}{\partial s_I}\right)^2\right)\right], \\
\eta &= \frac{1}{4}E\left[\left(s_R^2 - s_I^2\right)\left(\left(\frac{\partial f}{\partial s_R}\right)^2 - \left(\frac{\partial f}{\partial s_I}\right)^2\right) + 4s_R s_I \left(\frac{\partial f}{\partial s_R}\right)\left(\frac{\partial f}{\partial s_I}\right)\right], \\
4(\eta - \beta) &= -2E\left[s_R^2\left(\frac{\partial f}{\partial s_I}\right)^2 + s_I^2\left(\frac{\partial f}{\partial s_R}\right)^2 - 2s_R s_I\left(\frac{\partial f}{\partial s_R}\right)\left(\frac{\partial f}{\partial s_I}\right)\right] = 0,
\end{aligned}$$

where we used $E\left[s_R s_I\left(\left(\frac{\partial f}{\partial s_R}\right)^2 - \left(\frac{\partial f}{\partial s_I}\right)^2\right)\right] = 0$ and $E\left[\left(s_R^2 - s_I^2\right)\left(\frac{\partial f}{\partial s_R}\right)\left(\frac{\partial f}{\partial s_I}\right)\right] = 0$ in the third line and $\frac{\partial f}{\partial s_R} = 2s_R\frac{\partial f(r^2)}{\partial r^2}$ and $\frac{\partial f}{\partial s_I} = 2s_I\frac{\partial f(r^2)}{\partial r^2}$ in the last line.

The (i,j) block $\boldsymbol{\Omega}_2[i,j]$ $(i \neq j)$ can be calculated as $\boldsymbol{\Omega}_2[i,j] = E\left[\mathbf{s}\mathbf{s}^T\varphi_i^*(s_i)\varphi_j^*(s_j)\right]$. It has 1 at entry (i,j) and (j,i), since

$$\boldsymbol{\Omega}_2[i,j]_{(i,j)} = \boldsymbol{\Omega}_2[i,j]_{(j,i)} = E\left[s_i\varphi_i^*(s_i)\right] E\left[s_j\varphi_j^*(s_j)\right] = 1. \tag{21}$$

All other entries of $\boldsymbol{\Omega}_2[i,j]$ are zero since the components of $\mathbf{s}$ are independent and zero mean. Hence, we can calculate $\mathbf{M}_2 = \boldsymbol{\Omega}_2 - \mathrm{vec}\{\mathbf{I}\}\mathrm{vec}\{\mathbf{I}\}^T$ as

$$\mathbf{M}_2 = \sum_{i=1}^N (\beta_i - 1)\mathbf{L}_{ii} \otimes \mathbf{L}_{ii} + \sum_{i=1}^N \sum_{\substack{j=1\\j\neq i}}^N (\mathbf{L}_{ij} \otimes \mathbf{L}_{ji}). \tag{22}$$

We note that $\mathbf{M}_2$ is a real diagonal matrix.

Since $\mathbf{M}_1$ and $\mathbf{M}_2$ are real matrices, it holds $\mathbf{R}_{\boldsymbol{\vartheta}} = (\mathbf{M}_1 - \mathbf{M}_2\mathbf{M}_1^{-*}\mathbf{M}_2^*)^* = \mathbf{M}_1 - \mathbf{M}_2\mathbf{M}_1^{-1}\mathbf{M}_2$. After some calculations, we get

$$\mathbf{R}_{\boldsymbol{\vartheta}} = \sum_{i=1}^{N} \frac{(\eta_i - 1)^2 - (\beta_i - 1)^2}{\eta_i - 1} \mathbf{L}_{ii} \otimes \mathbf{L}_{ii} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{\kappa_i \kappa_j - 1}{\kappa_j} \right) \mathbf{L}_{ii} \otimes \mathbf{L}_{jj} \quad (23)$$

which simplifies to $\mathbf{R}_{\boldsymbol{\vartheta}} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left(\frac{\kappa_i \kappa_j - 1}{\kappa_j} \right) \mathbf{L}_{ii} \otimes \mathbf{L}_{jj}$ due to $\beta_i = \eta_i$.

References

1. Wirtinger, W.: Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. Mathematische Annalen 97(1), 357–375 (1927)
2. Eriksson, J., Koivunen, V.: Complex random vectors and ICA models: identifiability, uniqueness, and separability. IEEE Transactions on Information Theory 52(3), 1017–1029 (2006)
3. Tichavsky, P., Koldovsky, Z., Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. IEEE Trans. on Sig. Proc. 54(4) (April 2006)
4. Ollila, E., Kim, H.-J., Koivunen, V.: Compact Cramér-Rao bound expression for independent component analysis. IEEE Trans. on Sig. Proc. 56(4) (April 2008)
5. De Lathauwer, L., De Moor, B.: On the blind separation of non-circular sources. In: EUSIPCO 2002, Toulouse, France (September 2002)
6. Douglas, S.C.: Fixed-point algorithms for the blind separation of arbitrary complex-valued non-gaussian signal mixtures. EURASIP J. Appl. Signal Process. 2007(1) (January 2007)
7. Li, H., Adali, T.: Algorithms for complex MLICA and their stability analysis using Wirtinger calculus. IEEE Trans. on Sig. Proc. 58(12), 6156–6167 (2010)
8. Li, X.-L., Adali, T.: Complex independent component analysis by entropy bound minimization. IEEE Transactions on Circuits and Systems I: Regular Papers 57(7), 1417–1430 (2010)
9. Novey, M., Adali, T.: Adaptable nonlinearity for complex maximization of nongaussianity and a fixed-point algorithm. In: Proc. IEEE Workshop on Machine Learning for Signal Processing (September 2006)
10. Remmert, R.: Theory of complex functions. Graduate texts in mathematics. Springer, Heidelberg (1991)
11. Brandwood, D.H.: A complex gradient operator and its application in adaptive array theory. IEE Proc. 130, 11–16 (1983)
12. Ollila, E., Koivunen, V., Eriksson, J.: On the Cramér-Rao bound for the constrained and unconstrained complex parameters. In: Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop, SAM (2008)
13. Jagannatham, A.K., Rao, B.D.: Cramér-Rao lower bound for constrained complex parameters. IEEE Sig. Proc. Letters 11(11) (November 2004)
14. Novey, M., Adali, T., Roy, A.: A complex generalized gaussian distribution – characterization, generation, and estimation. IEEE Trans. on Sig. Proc. 58(3), 1427–1433 (2010)

Complex Non-Orthogonal Joint Diagonalization Based on LU and LQ Decompositions

Ke Wang, Xiao-Feng Gong, and Qiu-Hua Lin

School of Information and Communication Engineering,
Dalian University of Technology, Dalian 116024, China
xfgong@dlut.edu.cn

Abstract. In this paper, we propose a class of complex non-orthogonal joint diagonalization (NOJD) algorithms with successive rotations. The proposed methods consider LU or LQ decompositions of the mixing matrices, and propose to solve the NOJD problem via two successive stages: L-stage and U (or Q)-stage. Moreover, as the manifolds of target matrices in these stages could be appropriately parameterized by a sequence of simple elementary triangular or unitary matrices, which depend on only one or two parameters, the high-dimensional minimization problems could be replaced by a sequence of lower-dimensional ones. As such, the proposed algorithms are of simple closed-form in each iteration, and do not require the target matrices to be Hermitian nor positive definite. Simulations are provided to compare the proposed methods to other complex NOJD methods.

Keywords: Complex non-orthogonal joint diagonalization, Blind source separation, LU, LQ.

1 Introduction

Joint diagonalization (JD) is instrumental in solving many blind source separation (BSS) problems. For example, for an instantaneous linear mixing model $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$, where $\boldsymbol{s}(t)$, $\boldsymbol{A}$, and $\boldsymbol{x}(t)$ are the source, mixing matrix, and observation, respectively, we can calculate fourth-order cumulant [1] or time-varying covariance matrices $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_K$ [2] by assuming source uncorrelation (along with non-stationarity) or independence, that share the following common JD structure:

$$\boldsymbol{C}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^H \tag{1}$$

where $\boldsymbol{D}_k$ is diagonal, $k = 1,\ldots,K$, and superscript 'H' denotes conjugated transpose. JD then seeks an estimate of $\boldsymbol{A}$ by fitting the above common JD structure.

Numerous JD algorithms have been proposed, which can be classified into two categories: the orthogonal and the non-orthogonal ones. The orthogonal JD (OJD) methods, such as Cardoso's Jacobi-like algorithm, often require $\boldsymbol{A}$ to be unitary, and thus prewhitening must be added to orthogonalize the observation to fulfill this requirement. As prewhitening is always inaccurate and the errors introduced in this

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 50–57, 2012.

stage can not be corrected by OJD that follows [3], the non-orthogonal JD (NOJD) which does not require prewhitening has attracted growing attention in the past decade. Generally speaking, NOJD often uses a cost function to measure the fitting of the JD structure and performs minimization of this cost function to update the estimate of $\boldsymbol{A}$ in an iterative manner. To list a few, the weighted least squares (WLS) criterion formulates JD as a set of subspace fitting problems [4-6]. These methods do not necessarily require the target matrices to be Hermitian nor square, yet are sometimes computationally expensive. Information theoretic criterion is used in [7] that allows for super-efficient estimation with positive definite target matrices. The sum of off-diagonal squared norms is also widely used [8, 9], where the problem is the possible convergence to trivial solutions (e.g. singular or zero matrix), especially for gradient based or Newton-type methods [12].

Among the afore-mentioned JD algorithms, those using successive rotations are of a particular kind [1, 9-12]. Instead of optimizing one of the above-mentioned criteria for target matrices over all rows and columns, these methods consider lower-dimensional sub-optimization over two specific row and column indices at each iteration, and repeat the same sub-optimization procedure for all pairs of row and column indices to fulfill NOJD. As the elementary rotation matrix used in each iteration is nonsingular and determined by very few parameters, these methods are often of simple closed-form, and are free of trivial solutions. More exactly, the works in [9, 10] use polar decomposition of the elementary rotation matrix, while the work in [11] considers LU and LQ decompositions. Recently, non-parameterized elementary rotation matrix is considered for NOJD [12]. However, the methods based on parametrized elementary rotation matrices [9-11], which have been proven quite effective in solving NOJD problems, are mostly real-valued. Therefore, it is of great interests as how to extend these methodologies to the complex domain, especially given that complex BSS is more and more encountered in practical problems.

In this paper, we will extend the NOJD algorithms based on successive LU or LQ decompositions [11] to the complex domain. It should be noted that this extension is not trivial as the complex-valued version involves more parameters in the sub-optimization problem for each iteration than the real-valued case. In the rest of the paper, section 2 presents the proposed algorithm, section 3 provides comparisons with other complex NOJD algorithms via simulations, and section 4 concludes this paper.

2 Proposed Algorithms

2.1 Framework for the Proposed Algorithms

For a set of complex-valued matrices $\boldsymbol{C} = \{\boldsymbol{C}_1, ..., \boldsymbol{C}_K\}$ sharing the JD structure as formulated in (1), we seek the estimate for $\boldsymbol{B} \triangleq \boldsymbol{A}^{-1}$ such that $\{\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^H\}_{k=1}^K$ are as diagonal as possible. To solve the above JD problem, we propose to minimize the sum of off-diagonal squared norms for the estimation of $\boldsymbol{B}$ as follows:

$$\boldsymbol{B} = \arg\min_{\boldsymbol{B}} \sum\nolimits_{k=1}^{K} \operatorname{off}(\boldsymbol{B}\boldsymbol{C}_k\boldsymbol{B}^H) \qquad (2)$$

where $\mathrm{off}(\boldsymbol{P}) \triangleq \sum_{1\le i\ne j\le N} |p_{i,j}|^2$ for $\boldsymbol{P} \in \mathrm{C}^{N\times N}$. Moreover, we can reasonably assume that $\boldsymbol{B}$ is with unit determinant so that it could be factorized as follows:

$$\boldsymbol{B} = \boldsymbol{L}\boldsymbol{V} \tag{3}$$

where $\boldsymbol{L} \in \mathrm{C}^{N\times N}$ is a lower-triangular matrix with ones at its diagonal, $\boldsymbol{V} \in \mathrm{C}^{N\times N}$ is upper-triangular if (3) corresponds to LU factorization, and is unitary if (3) denotes LQ decomposition. Since any complex non-singular square matrix admits these two decompositions, it is reasonable to consider the unmixing matrix $\boldsymbol{B} = \boldsymbol{L}\boldsymbol{V}$ as in (3). As such, (2) could be solved in the following alternating manner:

$$\tilde{\boldsymbol{V}} = \arg\min_{\boldsymbol{V}} \sum\nolimits_{k=1}^{K} \mathrm{off}(\boldsymbol{V}\boldsymbol{C}_k\boldsymbol{V}^H) \tag{4.a}$$

$$\tilde{\boldsymbol{L}} = \arg\min_{\boldsymbol{L}} \sum\nolimits_{k=1}^{K} \mathrm{off}(\boldsymbol{L}\boldsymbol{C}'_k\boldsymbol{L}^H) \tag{4.b}$$

where $\boldsymbol{C}'_k = \tilde{\boldsymbol{V}}\boldsymbol{C}_k\tilde{\boldsymbol{V}}^H$, and $\tilde{\boldsymbol{B}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{V}}$ is the estimate of $\boldsymbol{B}$.

In succesive rotation based methods, the minimization problems in (4) are solved by repeating the following scheme for all possible index pairs (i, j), $1 \le i < j \le N$, and iterate until convergence:

$$\boldsymbol{C}_{k,new} = \boldsymbol{T}_{(i,j)}\boldsymbol{C}_{k,old}\boldsymbol{T}^H_{(i,j)}, \quad \boldsymbol{V}_{new} = \boldsymbol{T}_{(i,j)}\boldsymbol{V}_{old} \tag{5.a}$$

$$\boldsymbol{C}'_{k,new} = \boldsymbol{T}'_{(i,j)}\boldsymbol{C}'_{k,old}\boldsymbol{T}'^H_{(i,j)}, \quad \boldsymbol{L}_{new} = \boldsymbol{T}'_{(i,j)}\boldsymbol{L}_{old} \tag{5.b}$$

where $\boldsymbol{C}_{k,new}$, $\boldsymbol{C}'_{k,new}$, $\boldsymbol{V}_{new}$ and $\boldsymbol{L}_{new}$ denote the updates of $\boldsymbol{C}_k$, $\boldsymbol{C}'_k$, $\boldsymbol{V}$ and $\boldsymbol{L}$ in the current iteration, and $\boldsymbol{C}_{k,old}$, $\boldsymbol{C}'_{k,old}$, $\boldsymbol{V}_{old}$ and $\boldsymbol{L}_{old}$ are the results obtained in the previous iteration, $k = 1,...,K$. $\boldsymbol{T}_{(i,j)}$ and $\boldsymbol{T}'_{(i,j)}$ are elementary rotation matrices for problems (4.a) and (4.b), respectively, which equal the identity matrix except the entries indexed (i,i), (i, j), (j,i), and (j, j). The goal is then to find optimal $\boldsymbol{T}_{(i,j)}$ and $\boldsymbol{T}'_{(i,j)}$ in each iteration to solve (4.a) and (4.b), respectively. Noting that the elementary rotation matrices $\boldsymbol{T}_{(i,j)}$ and $\boldsymbol{T}'_{(i,j)}$ are determined by only one or two parameters (as will be shown later), the higher-dimensional optimization problem in (2) could be reduced to a sequence of one- or two-dimentional simple sub-problems.

2.2 Schemes to Find Optimal Elementary Rotation Matrices

Firstly, we consider the LU decomposition of $\boldsymbol{B}$ so that the matrix $\boldsymbol{V}$ in (3) is an upper-triangular matrix $\boldsymbol{U} \in \mathrm{C}^{N\times N}$, and the JD problem could be solved via two alternating stages termed as U-stage and L-stage, respectively, as indicated in (4) ($\boldsymbol{V}$ in (4.a) is replaced by $\boldsymbol{U}$ for LU decomposition). We break the minimization problem in the U-stage into a sequence of sub-problems via (5.a) with elementary rotation matrix $\boldsymbol{T}_{(i,j)}$ equal to the identity matrix except the $(i, j)th$ upper entry $\alpha_{i,j}$. As such, for index pair (i, j), we note that $\boldsymbol{T}_{(i,j)}\boldsymbol{C}_{k,old}\boldsymbol{T}^H_{(i,j)}$ only impacts the ith row and column of $\boldsymbol{C}_{k,old}$, $k = 1,2,...,K$. As a result, the minimization of $\sum_{k=1}^{K} \mathrm{off}(\boldsymbol{T}^H_{(i,j)}\boldsymbol{C}_{k,old}\boldsymbol{T}_{(i,j)})$ amounts to minimizing the sum-of squared norms of the off-diagonal elements in the ith row and column of $\boldsymbol{C}_{k,new}$:

$$\xi_{i,j} \triangleq \sum_{k=1}^{K} \sum_{p=1,p\neq i}^{N} \left[\left| C_{k,new}(i,p) \right|^2 + \left| C_{k,new}(p,i) \right|^2 \right] \tag{6}$$

Noting $C_{k,new}(i,p) = \alpha_{i,j} C_{k,old}(j,p) + C_{k,old}(i,p)$, and $C_{k,new}(p,i) = C_{k,old}(p,i) + \alpha_{i,j}^{*} C_{k,old}(p,j)$, (6) could be rewritten as follows after a few manipulations:

$$\begin{aligned} \xi_{i,j} = \textstyle\sum_{k=1}^{K} \sum_{p=1,p\neq i}^{N} \{ & (C_{k,old}(j,p) C_{k,old}^{*}(j,p) + C_{k,old}(p,j) C_{k,old}^{*}(p,j)) \alpha_{i,j} \alpha_{i,j}^{*} \\ & + [C_{k,old}(i,p) C_{k.old}^{*}(j,p) + C_{k.old}(p,j) C_{k.old}^{*}(p,i)] \alpha_{i,j}^{*} \\ & + [C_{k.old}(j,p) C_{k,old}^{*}(i,p) + C_{k.old}(p,i) C_{k.old}^{*}(p,j)] \alpha_{i,j} \\ & + [C_{k.old}(i,p) C_{k.old}^{*}(i,p) + C_{k.old}(p,i) C_{k,old}^{*}(p,i)] \} \end{aligned} \tag{7}$$

As such, the optimal parameter $\alpha_{i,j}$ could be obtained by setting the derivative of $\xi_{i,j}$ with regards to $\alpha_{i,j}^{*}$ to zero, which yields the following:

$$\alpha_{i,j} = -\frac{\sum_{k=1}^{K} \sum_{p=1,p\neq i}^{N} [C_{k,old}(i,p) C_{k,old}^{*}(j,p) + C_{k,old}(p,j) C_{k,old}^{*}(p,i)]}{\sum_{k=1}^{K} \sum_{p=1,p\neq i}^{N} [C_{k,old}(j,p) C_{k,old}^{*}(j,p) + C_{k,old}(p,j) C_{k,old}^{*}(p,j)]} \tag{8}$$

The minimization problem in the L-stage could be solved similarly to the U-stage, with the only exception that the iterations are repeated for $1 \le j < i \le N$, and one U-stage and L-stage make up a sweep. As a result, the JD problem is tackled by alternating the U-stage and L-stage until convergence in the LU based method. The LU-based method for complex NOJD is termed as LUCJD for short.

In the second proposed algorithm, we consider the LQ decomposition of $\boldsymbol{B}$, and thus the matrix $\boldsymbol{V}$ in (3) is a unitary matrix $\boldsymbol{Q} \in \mathbb{C}^{N\times N}$. Similarly to LUCJD, JD is herein solved by alternating Q-stage and L-stage as indicated in (4) ($\boldsymbol{V}$ in (4.a) is replaced by $\boldsymbol{Q}$ for LQ decomposition). Moreover, noting that the L-stage could be handled similarly as that in LUCJD, and the Q-stage involves an OJD problem, which could be actually tackled by Cardoso's Jacobi-like method [1]. As a result, the JD problem is solved by alternating the Q-stage that adopts Jacobi-like algorithm, and the L-stage that uses the scheme proposed in LUCJD, until convergence is reached. The LQ-based algorithm is termed as LQCJD for clarity.

2.3 Remarks and Summarization

We have some implementation remarks on the proposed LUCJD and LQCJD:

Remark 1: The proposed algorithms do not require the target matrices be Hermitian, and therefore could be used for solving BSS problems that involve JD of non-Hermitian complex matrices (e.g. time-lagged covariance matrices or fourth-order cumulant matrices for complex-valued signals).

Remark 2: There are several termination criteria for the proposed algorithms. For example, we could monitor the value of sum of off-diagonal squared norms, and stop the iterations when the decrease in it is smaller than a preset threshold. In this paper, we observe the change of $\left\| \boldsymbol{LV} - \boldsymbol{I}_N \right\|_F^2$ ($\boldsymbol{V}$ is upper-trianguler matrix $\boldsymbol{U}$ in LUCJD, and is unitary matrix $\boldsymbol{Q}$ in LQCJD, $\boldsymbol{I}_N$ is an $N \times N$ identity matrix) between two succesive sweeps and stop the iterations if it's smaller than a threshold.

We summarize the proposed algorithms in Table 1:

Table 1. Summarization of the proposed algorithm

```
• Input: A set of $N \times N$ square matrices $\boldsymbol{C}_1, \boldsymbol{C}_2, \cdots, \boldsymbol{C}_K$, and a threshold $\tau$
• Output: The estimated unmixing matrix $\boldsymbol{B}$
• Implementation:
  $\boldsymbol{B} \leftarrow \boldsymbol{I}_N$, $\gamma_{old} \leftarrow 0$, $\zeta \leftarrow \tau + 1$.
  while $\zeta \geq \tau$ do
      The U-stage or Q-stage: $\boldsymbol{V} \leftarrow \boldsymbol{I}_N$
      for all $1 \leq i < j \leq N$ do
          - For U-stage in LUCJD: obtain optimal elementary upper-triangular matrix $\boldsymbol{T}_{(i,j)}$
            with its $(i, j)$th element determined by (8)
            For Q-stage in LQCJD: obtain optimal elementary unitary matrix $\boldsymbol{T}_{(i,j)}$ via the Jacobi-like algorithm [1]
          - Update matrices: $\boldsymbol{V} \leftarrow \boldsymbol{T}_{(i,j)}\boldsymbol{V}$, $\boldsymbol{C}_k \leftarrow \boldsymbol{T}_{(i,j)}\boldsymbol{C}_k\boldsymbol{T}_{(i,j)}^H$, $k = 1, \cdots, K$
      end for
      The L-stage: $\boldsymbol{L} \leftarrow \boldsymbol{I}_N$
      for all $1 \leq j < i \leq N$ do
          - obtain optimal elementary lower-triangular matrix $\boldsymbol{T}'_{(i,j)}$ with its $(i, j)$th element determined by (8)
          - Update matrices: $\boldsymbol{L} \leftarrow \boldsymbol{T}'_{(i,j)}\boldsymbol{L}$, $\boldsymbol{C}_k \leftarrow \boldsymbol{T}'_{(i,j)}\boldsymbol{C}_k\boldsymbol{T}'^H_{(i,j)}$ $(k = 1, \cdots, K)$
      end for
      $\boldsymbol{B} \leftarrow \boldsymbol{LVB}$, $\gamma_{new} \leftarrow \|\boldsymbol{LV} - \boldsymbol{I}_N\|_F^2$, $\zeta \leftarrow |\gamma_{new} - \gamma_{old}|$, $\gamma_{old} \leftarrow \gamma_{new}$
  end while
```

3 Simulation Results

We provide simulations to compare the proposed LUCJD and LQCJD with X. Guo's nonparametric Jacobi transformation based JD (JTJD)[12], Li's fast approximate JD (FAJD) [5], Tichavsky and Yeredor's uniformly weighted exhaustive diagonalization by Gaussian iteration (UWEDGE) [6]. We generate the target matrices as:

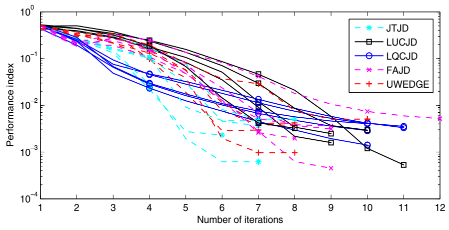
$$\boldsymbol{C}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^H + \sqrt{\sigma}\boldsymbol{N}_k \quad (k = 1, \ldots, K) \tag{9}$$

where $\boldsymbol{D}_k \in \mathbb{C}^{N \times N}$ is a diagonal matrix with its diagonal elements normally distributed with zero mean. $\boldsymbol{A} \in \mathbb{C}^{N \times N}$ and $\boldsymbol{N}_k \in \mathbb{C}^{N \times N}$ are the mixing matrix and the noise term which are randomly generated from normal distribution with zero mean. σ is the noise level. We note herein that the above target matrices are neither Hermitian nor positive definite. In addition, all the compared methods are initialized with identity matrix, and uniform weights are used for FAJD and UWEDGE. The performance index (PI) [3] is used to measure the performance of the algorithms:

$$\mathrm{Index}(\boldsymbol{P}) = [2N(N-1)]^{-1}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{N} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N} \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1\right)\right] \tag{10}$$

where $\boldsymbol{P} = \tilde{\boldsymbol{B}}\boldsymbol{A}$, with $\tilde{\boldsymbol{B}} \in \mathbb{C}^{N \times N}$ being the estimate of the unmixing matrix.

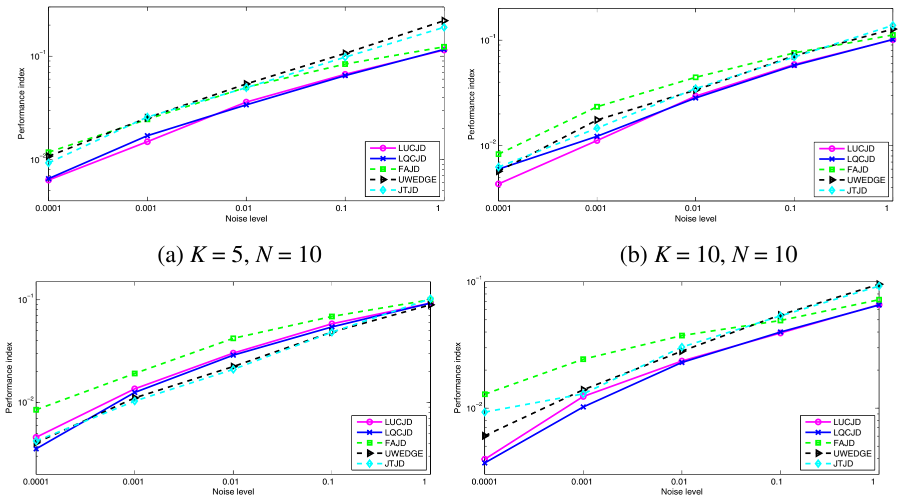
In the first simulation, we compare the convergence speed of the considered algorithms. We fix the number and the size of target matrices as $K = 10$ and $N = 10$, respectively, and plot in Fig. 1 the PI curves from 5 independent runs against the number of iterations for 2 different noise levels : (a) $\sigma = 0.0001$; (b) $\sigma = 1$. From Fig. 1. (a) we see that LUCJD and LQCJD are of almost equal convergence speed as FAJD, which is slightly slower than JTJD and UWEDGE, when the noise is at a low level of $\sigma = 0.0001$. However, when the noise level increases to $\sigma = 1$, we note from Fig. 1. (b) that the number of iterations for JTJD, UWEDGE, and FAJD significantly increase as well, while LUCJD and LQCJD are still able to yield robust converging behavior, with LQCJD being the fastest one among all the compared algorithms.



(a) $\sigma = 0.0001$ (b) $\sigma = 1$

Fig. 1. Performance index against number of iterations

In the second simulation, we compare the estimation accuracy of the mixing matrix at different noise levels. We let the noise level σ vary from 0.0001 to 1 and plot in Fig. 2 the PI curves obtained from 200 independent runs versus σ for the following 4 cases: (a) $K = 5$, $N = 10$; (b) $K = 10$, $N = 10$; (c) $K = 30$, $N = 10$; (d) $K = 10$, $N = 20$.



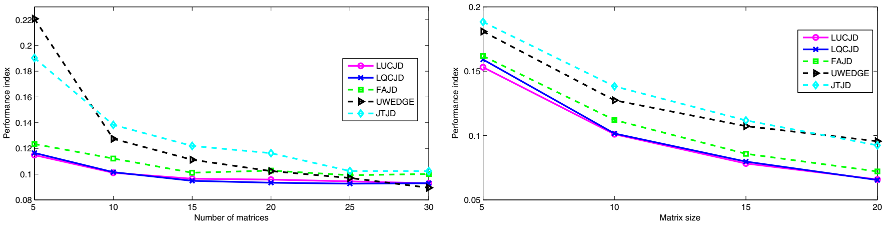
(a) $K = 5$, $N = 10$ (b) $K = 10$, $N = 10$

(c) $K = 30$, $N = 10$ (d) $K = 10$, $N = 20$

Fig. 2. Performance index versus noise level for different numbers and sizes of target matrices

We could see that LUCJD and LQCJD provide almost equal performance in the considered scenarios. It is also demonstrated in Fig. 2. (a) to Fig. 2. (c) that LUCJD and LQCJD outperform JTJD, FAJD, and UWEDGE for small number of target matrices, yet slightly underperform their competitors when the number of target matrices increases. This shows that the proposed algorithms are more robust to noise than JTJD, FAJD, and UWEDGE in difficult situations where only a small number of target matrices are available. In addition, we observe in Fig. 2. (d) that LUCJD and LQCJD offer better estimation precision than their competitors.

In the third simulation, we test the performance of the compared algorithms against the number and size of target matrices for a fixed noise level $\sigma = 1$. We fix the size of target matrices to N = 10, and plot the PI curves of the compared algorithms versus the number of target matrices in Fig. 3. (a). Then we fix the number of target matrices to K = 10, and plot the PI curves against the matrix size in Fig. 3. (b). The shown statistics are obtained from 200 independent runs. From Fig. 3. (a) we see that the performance of LUCJD and LQCJD are very stable when the number of target matrices varies. In particular, we note that LUCJD and LQCJD outperform the other compared algorithms clearly for small number of target matrices, while only slightly underperforms UWEDGE when K = 30. This observation again illustrates the advantage of the proposed methods in difficult scenarios where only a small number of target matrices are available. In addition, from Fig. 3. (b) we note that increasing the matrix size from 5 to 20 can slightly improve the performance of all the compared algorithms, with LUCJD and LQCJD outperforming the other methods.



(a) $N = 10$, $K = 5 \sim 30$

(b) $K = 10$, $N = 5 \sim 20$

Fig. 3. Performance index versus number and size of target matrices with noise level $\sigma = 1$

4 Conclusion

In this paper, we proposed a class of complex non-orthogonal joint diagonalization (NOJD) algorithms based on LU and LQ decompositions. The proposed algorithms (termed as LUCJD and LQCJD, respectively) tackled the NOJD problem by using a sequence of simple parameterized elementary rotation matrices, and thus are of simple closed-form in each iteration and free of trivial solutions. In addition, the proposed algorithms do not require the target matrices to be Hermitian nor positive definite. Simulations are provided to compare the performance of the proposed algorithms with some other complex NOJD algorithms. The results show that the proposed LUCJD and LQCJD are of stable convergence at different noise levels, and LQCJD converges faster than the other compared methods at high noise levels. With regards to the estimation accuracy, LUCJD and LQCJD could provide superior performance for different noise levels and matrix sizes over the other compared methods, especially in case of small number of target matrices. The above simulation results infer that the proposed methods may be preferable in difficult situations where the noise level is high, and only a few target matrices are available.

Acknowledgments. This work is supported by the Fundamental Research Fund for Central Universities of China, and by National Natural Science Foundation of China under grants 60971097, 61072098, and 61105008.

References

1. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. Radar and Signal Processing. IEE Proceedings -F 140, 362–370 (1993)
2. Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E.: A blind source separation technique using second-order statistics. IEEE Transactions on Signal Processing 45, 434–444 (1997)
3. Souloumiac, A.: Joint diagonalization: is non-orthogonal always preferable to orthogonal. In: 3rd International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Dutch Antilles, pp. 305–308 (2009)
4. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Transactions on Signal Processing 50, 1545–1553 (2002)
5. Li, X.-L., Zhang, X.-D.: Nonorthogonal joint diagonalization free of degenerate solution. IEEE Transactions on Signal Processing 55, 1803–1814 (2007)
6. Tichavsky, P., Yeredor, A.: Fast approximate joint diagonalization incorporating weight matrices. IEEE Transactions on Signal Processing 57, 878–891 (2009)
7. Pham, D.-T., Serviére, C., Boumaraf, H.: Blind separation of speech mixtures based on nonstationarity. In: 7th International Symposium on Signal Processing and Its Applications, France, pp. 73–76 (2003)
8. Hori, G.: Joint diagonalization and matrix differential equations. In: 1999 International Symposium on Nonlinear Theory and its Applications, Hawaii, pp. 675–678 (1999)
9. Souloumiac, A.: Nonorthogonal joint diagonalization by combining givens and hyperbolic rotations. IEEE Transactions on Signal Processing 57, 2222–2231 (2009)
10. Luciani, X., Albera, L.: Joint Eigenvalue Decomposition Using Polar Matrix Factorization. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 555–562. Springer, Heidelberg (2010)
11. Afsari, B.: Simple LU and QR Based Non-Orthogonal Matrix Joint Diagonalization. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 1–7. Springer, Heidelberg (2006)
12. Guo, X.-J., Zhu, S.-H., Miron, S., Brie, D.: Approximate joint diagonalization by nonorthogonal nonparametric jacobi transformations. In: 35th International Conference on Acoustics, Speech, and Signal Processing, Dallas, pp. 3774–3777 (2010)

Exact and Approximate Quantum Independent Component Analysis for Qubit Uncoupling

Yannick Deville[1] and Alain Deville[2]

[1] Université de Toulouse, UPS-CNRS-OMP, Institut de Recherche en Astrophysique
et Planétologie, 14 Av. Edouard Belin, 31400 Toulouse, France
yannick.deville@irap.omp.eu

[2] Aix-Marseille Univ, IM2NP, Campus Scientifique Saint-Jérôme,
F-13997 Marseille, France
alain.deville@univ-provence.fr

Abstract. As a tool for solving the Blind Quantum Source Separation problem introduced in our previous papers, we here propose the concept of Quantum Independent Component Analysis (QICA). Starting from quantum bits (qubits) with cylindrical-symmetry Heisenberg coupling, quantum-to-classical conversion yields an original nonlinear mixing model, which leads us to develop QICA methods dedicated to this model. Our first method consists in minimizing the mutual information of the outputs of our nonlinear separating system. It is attractive because it yields an exact solution, without any spurious points thanks to the (Q)ICA separability of the considered model. The second proposed method is a simpler approximation of the first one. It is based on a truncated expansion of differential entropy (or negentropy), derived from the Edgeworth expansion of probability density functions.

Keywords: blind quantum source separation, quantum independent component analysis, nonlinear mixing model, mutual information, Edgeworth expansion, qubit, cylindrical-symmetry Heisenberg coupling.

1 Introduction

Source Separation (SS) is an Information Processing (IP) problem, which consists in retrieving a set of unknown source "signals" (time series, images...) from a set of observations, which are mixtures of these source signals. In particular, the *Blind* Source Separation (BSS) configuration corresponds to the case when the parameter values of the considered mixing model are unknown. On the contrary, these values are known in the non-blind case, which therefore reduces to the inversion of a known mixing model. The BSS field emerged in the 1980s and then yielded major developments, e.g. reported in the handbook [2]. Until recently, all these investigations were performed in a classical, i.e. non-quantum, framework. Independently from BSS, another field within the overall IP domain rapidly developed in the last decades, i.e. Quantum Information Processing (QIP). It is described in detail in [10], and its main features are summarized in [4],[7],[8].

We recently bridged the gap between the classical BSS and QIP fields, by introducing a new field, i.e. Quantum Source Separation (QSS), first proposed

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 58–65, 2012.

in our paper [4] and then described in more detail in particular in [7]. The QSS problem consists in restoring the information contained in individual *quantum* source signals, only starting from quantum mixtures (in SS terms) of these signals. This gives rise to three possible approaches:

1. In the classical-processing approach [4], [7], one first converts the mixed quantum data into classical ones by means of measurements, and then processes the measured data with classical (i.e, again, non-quantum) methods. We showed that original processing methods must then be developed, because the nonlinear mixing model thus encountered has not previously been addressed in the classical (B)SS literature.
2. Quantum-processing methods [7] keep the quantum nature of the mixtures and process them by means of quantum circuits in order to retrieve the quantum sources.
3. Hybrid methods [8] combine the above two approaches, by first partly processing the quantum mixtures with quantum circuits, then converting the resulting quantum data into classical ones by means of measurements, and eventually processing the measured data with classical methods.

In this paper, we only consider the first approach to QSS, based on classical-processing methods, which are the only easily implementable ones nowadays, since the practical design of quantum circuits is only an emerging field. As in the classical SS framework, these QSS methods give rise to two configurations, i.e. the blind and non-blind ones. We here consider the most complex configuration, i.e. the blind one, which requires us to estimate the value(s) of the mixing model parameter(s). In our papers [4], [7], we only described a very basic method for performing this estimation. That method is based on the first-order moment of a measured signal and has the drawback of setting constraints on some source statistics. We therefore here aim at developing much more powerful methods for performing the considered Blind Quantum Source Separation (BQSS) task.

In the classical framework, several classes of methods were proposed for solving the BSS problem, the most popular of them being Independent Component Analysis (ICA). Similarly, as a tool for solving the BQSS problem, we here develop what we will call "Quantum Independent Component Analysis (QICA) methods", in the sense: Independent Component Analysis methods for data which initially have a quantum nature (these data are here converted into classical ones and then processed by classical means). More precisely, we will first describe a method which performs exact QICA, and then an associated approximation. Before this, we now define the considered mixing and separating models.

2 Mixing Model

In the QIP field, "qubits" (i.e. quantum bits) are used instead of classical bits for performing computations. A qubit, with index i, has a quantum state expressed as follows (for a pure state):

$$|\psi_i >= \alpha_i|+ > +\beta_i|- > \tag{1}$$

where $|+>$ and $|->$ are basis vectors, whereas α_i and β_i are two complex-valued coefficients such that

$$|\alpha_i|^2 + |\beta_i|^2 = 1. \tag{2}$$

In [4], [7], we considered the situation when the two qubits respectively associated with two spins of a physical system are separately initialized with states defined by (1) and then get "mixed" (in the SS sense), due to the undesired coupling effect which exists in the considered system (Heisenberg coupling in our case). We proposed an approach which consists in repeatedly initializing the two qubits according to (1) and later measuring spin components associated with the system composed of these two coupled qubits. We showed that this yields four possible measured values, with respective probabilities p_1, p_2, p_3 and p_4. In [7], we derived the expressions of these probabilities with respect to the polar representation of the qubit parameters α_i and β_i, which reads

$$\alpha_i = r_i e^{\mathrm{i}\theta_i} \qquad \beta_i = q_i e^{\mathrm{i}\phi_i} \qquad\qquad \forall i \in \{1, 2\} \tag{3}$$

with $0 \leq r_i \leq 1$, $q_i = \sqrt{1 - r_i^2}$ due to (2), and $\mathrm{i} = (-1)^{\frac{1}{2}}$. The above probabilities may then be expressed as follows:

$$p_1 = r_1^2 r_2^2 \tag{4}$$

$$\begin{aligned} p_2 = {} & r_1^2(1 - r_2^2)(1 - v^2) + (1 - r_1^2) r_2^2 v^2 \\ & -2r_1 r_2 \sqrt{1 - r_1^2}\sqrt{1 - r_2^2}\sqrt{1 - v^2}\, v \sin \Delta_I \end{aligned} \tag{5}$$

$$p_4 = (1 - r_1^2)(1 - r_2^2) \tag{6}$$

where

$$\Delta_I = (\phi_2 - \phi_1) - (\theta_2 - \theta_1) \tag{7}$$

and v is a parameter, defined in [7], which is such that $0 \leq v^2 \leq 1$, and whose value is unknown in most configurations (this corresponds to the *blind* version of this QSS problem). Note that probability p_3 is not considered in this investigation, since it is redundant with the above three ones: we always have

$$p_1 + p_2 + p_3 + p_4 = 1. \tag{8}$$

Eq. (4)-(6) form the nonlinear "mixing model" (in SS terms) of this investigation. The observations involved in this model are the probabilities p_1, p_2 and p_4 measured (in fact, estimated, using repeated qubit initializations [7]) for each choice of the initial states of the qubits. Using standard SS notations, the observation vector is therefore $x = [x_1, x_2, x_3]^T$, where ^{T} stands for transpose and

$$x_1 = p_1, \qquad x_2 = p_2, \qquad x_3 = p_4. \tag{9}$$

The source vector to be retrieved from these observations is $s = [s_1, s_2, s_3]^T$ with $s_1 = r_1, s_2 = r_2$ and $s_3 = \Delta_I$ (the parameters q_i are then obtained as $q_i = \sqrt{1 - r_i^2}$; the four phase parameters in (3) cannot be individually extracted from their combination Δ_I; only two phases have a physical meaning [8]). In the blind configuration considered in this paper, retrieving the sources first requires one to estimate the only unknown mixing parameter of this model, i.e. v.

3 Separating System

In [7], we showed that the above mixing model is invertible (with respect to the considered domain of source values), for any fixed v such that $0 < v^2 < 1$, provided the source values meet the following conditions:

$$0 < r_1 < \tfrac{1}{2} < r_2 < 1 \tag{10}$$

$$-\tfrac{\pi}{2} \leq \Delta_I \leq \tfrac{\pi}{2}. \tag{11}$$

The separating system that we proposed for retrieving (estimates of) the sources by combining the observations then yields an output vector $y = [y_1, y_2, y_3]^T$ which reads

$$y_1 = \sqrt{\frac{1}{2}\left[(1 + p_1 - p_4) - \sqrt{(1 + p_1 - p_4)^2 - 4p_1}\right]} \tag{12}$$

$$y_2 = \sqrt{\frac{1}{2}\left[(1 + p_1 - p_4) + \sqrt{(1 + p_1 - p_4)^2 - 4p_1}\right]} \tag{13}$$

$$y_3 = \mathrm{Arcsin}\left[\frac{y_1^2(1 - y_2^2)(1 - \hat{v}^2) + (1 - y_1^2)y_2^2\hat{v}^2 - p_2}{2y_1y_2\sqrt{1 - y_1^2}\sqrt{1 - y_2^2}\sqrt{1 - \hat{v}^2}\hat{v}}\right] \tag{14}$$

where $\hat{v}$ is the estimate of v used in the separating system. The outputs y_1, y_2 and y_3 respectively restore the sources $s_1 = r_1, s_2 = r_2$ and $s_3 = \Delta_I$.

4 Exact QICA

We here consider the case when each source signal of our BQSS problem is continuous-valued, stochastic, identically distributed (i.d) and all source signals are mutually statistically independent. The observations and separating system outputs are then also stochastic and i.d. We therefore consider the random variables (RVs) defined by all these signals at a single time, and we denote Y_i the RVs thus associated with the outputs of the separating system.

In these conditions, we consider the "global model" (from the source signals s_i to their estimates y_i) obtained by combining the mixing model (4)-(6) and the separating model (12)-(14). We call it "the Heisenberg global model". In [5], we briefly showed that it is "ICA separable". Generally speaking, an arbitrary (memoryless) global model is said to be ICA separable, in the above conditions and for sources having given probability density functions (pdf), if it meets the following property: if the output RVs of the separating system are mutually statistically independent, then they are equal to the source RVs, up to some acceptable indeterminacies which depend on the considered model (e.g., only one sign indeterminacy for the Heisenberg global model, as detailed hereafter). In [5], we first studied the ICA separability of a very general class of global models. We then briefly focused on the Heisenberg global model, and we proved that it is ICA separable. We here aim at proceeding much further in the investigation of this

BQSS problem, by deriving qubit separation methods based on this separability property. This property is indeed very attractive, because it ensures that, by adapting $\hat{v}$ so that the output RVs Y_i become statistically independent, it is guaranteed thay they become equal to the source RVs (up to the indeterminacies of the Heisenberg model). To derive a practical QICA method from this property, we need to define a quantity which detects when the output RVs are independent. A well-known quantity which meets this constraint is the mutual information of these RVs, denoted $I(Y)$, where $Y = [Y_1, Y_2, Y_3]^T$: $I(Y)$ is null when the RVs Y_i are independent and positive otherwise. A separation criterion for the Heisenberg global model therefore consists in adapting $\hat{v}$ so as to minimize (and thus cancel) a function, therefore called "the cost function", defined as $I(Y)$. The above ICA separability property means that $I(Y)$ has no global spurious points, i.e. it reaches its global minimum value only when source separation is achieved, up to the indeterminacies of the model. From the general analysis provided in [5], one may derive that these indeterminacies here reduce to a sign indeterminacy for y_3: when the output RVs are independent, we have $y_3 = \pm s_3$. The other two output signals yield no indeterminacies, i.e. they are equal to the corresponding source signals.

The cost function thus obtained may be expressed as

$$I(Y) = \left(\sum_{i=1}^{3} h(Y_i) \right) - h(Y). \tag{15}$$

In this expression, each term $h(Y_i)$ is the differential entropy of the RV Y_i, which may be expressed as

$$h(Y_i) = -E\{\ln f_{Y_i}(Y_i)\} \tag{16}$$

where $f_{Y_i}(.)$ is the pdf of Y_i and $E\{.\}$ stands for expectation. Similarly, $h(Y)$ is the joint differential entropy of all RVs Y_i, which reads

$$h(Y) = -E\{\ln f_Y(Y)\} \tag{17}$$

where $f_Y(.)$ is the joint pdf of all RVs Y_i.

Moreover, we here use the following general property. Let us consider an arbitrary random vector X with dimension N, to which an arbitrary invertible transform ϕ is applied. We thus get the random vector Y with dimension N, defined as: $Y = \phi(X)$. This transform has the following effect on joint differential entropy [6]:

$$h(Y) = h(X) + E\{\ln |J_\phi(X)|\} \tag{18}$$

where $J_\phi(x)$ is the Jacobian of the transform $y = \phi(x)$, i.e. the determinant of the Jacobian matrix of ϕ. Each element with indices (i, j) of this matrix is equal to $\frac{\partial \phi_i(x)}{\partial x_j}$, where $\phi_i = y_i$ is the ith component of the vector function ϕ and x_j is its jth argument. This property here applies to the output joint differential entropy defined in (17), and the transform ϕ here consists of the separating model defined by (9) and (12)-(14). Eq. (12) and (13) show that y_1 and y_2 do not depend on x_2. Therefore, $J_\phi(x)$ here reduces to

$$J_\phi(x) = J_1 J_2 \tag{19}$$

where

$$J_1 = \frac{\partial y_3}{\partial x_2} \tag{20}$$

$$= -\operatorname{sgn}(\hat{v})\left\{4y_1^2y_2^2(1-y_1^2)(1-y_2^2)(1-\hat{v}^2)\hat{v}^2 \right.$$
$$\left. -[y_1^2(1-y_2^2)(1-\hat{v}^2)+(1-y_1^2)y_2^2\hat{v}^2-x_2]^2\right\}^{-\frac{1}{2}} \tag{21}$$

$$J_2 = \frac{\partial y_1}{\partial x_3}\frac{\partial y_2}{\partial x_1} - \frac{\partial y_1}{\partial x_1}\frac{\partial y_2}{\partial x_3} \tag{22}$$

$$= \frac{1}{4y_1y_2\sqrt{(1+x_1-x_3)^2-4x_1}}. \tag{23}$$

Combining (15) and (18), the considered cost function becomes

$$I(Y) = \left(\sum_{i=1}^{3} h(Y_i)\right) - h(X) - E\{\ln|J_\phi(X)|\}. \tag{24}$$

Its term $h(X)$ does not depend on the separating system parameter $\hat{v}$ to be optimized, but only on the fixed available observations. Besides, (12) and (13) show that the outputs y_1 and y_2, and therefore the differential entropies $h(Y_1)$ and $h(Y_2)$ also do not depend on $\hat{v}$. Therefore, minimizing $I(Y)$ with respect to $\hat{v}$ is equivalent to minimizing the following cost function:

$$C_2(Y) = h(Y_3) - E\{\ln|J_\phi(X)|\}. \tag{25}$$

5 Approximate QICA

The exact QICA criterion developed in the previous section involves the pdf of a separating system output. It therefore requires one to estimate this pdf (or its derivative, used in some optimization algorithms), which is cumbersome. An alternative approach consists in deriving an approximation of this pdf, which yields an associated approximate QICA criterion. We now investigate this approach.

A method for defining an approximation of a pdf, and then of the associated differential entropy or negentropy, consists in using the Edgeworth expansion, which is e.g. detailed in [9]. This approach may be summarized as follows. The considered pdf of an RV U is expressed as the product of a reference pdf, here selected as a Gaussian RV G with the same mean and variance as U, and of a factor expressed as a series (see its explicit expression e.g. in [9]). This then makes it possible to express the negentropy of U, i.e.

$$\mathcal{J}(U) = h(G) - h(U), \tag{26}$$

as a series. Then truncating that series to a given order provides a corresponding approximation of that negentropy.

That approach was used and detailed by Comon in [1], but only for a standardized (i.e. zero-mean and unit-variance) RV. This was motivated by the fact that, for the linear instantaneous mixing model considered in [1]: 1) the zero-mean versions of the observations are linked according to the same model to the zero-mean versions of the source signals, so that one can restrict oneself to zero-mean signals and 2) this model does not fix the scales of the estimated sources, so that one can decide to only consider unit-variance outputs without loss of generality. Comon thus obtained the following approximation of negentropy:

$$\mathcal{J}(U) \simeq \frac{1}{12}\mathrm{cum}_3(U)^2+\frac{1}{48}\mathrm{cum}_4(U)^2+\frac{7}{48}\mathrm{cum}_3(U)^4-\frac{1}{8}\mathrm{cum}_3(U)^2\mathrm{cum}_4(U) \quad (27)$$

where $\mathrm{cum}_i(U)$ is the ith-order cumulant of the standardized RV U.

On the contrary, we here have to consider unstandardized RVs, because our nonlinear mixing model does not yield the above-defined translation property and scale indeterminacy. We therefore aim at determining a (neg)entropy approximation for an unstandardized RV. This could be done by starting from the pdf expansion provided in [9] for an arbitrary RV and then developing the above-defined procedure for deriving the corresponding negentropy expansion. However, these computations would be complicated and may be avoided as follows, by taking advantage of the results already obtained by Comon for standardized RVs. Since we eventually aim at deriving an approximation of the differential entropy of Y_3 involved in (25), we introduce the standardized version of Y_3, defined as:

$$\tilde{Y_3} = \frac{Y_3 - E\{Y_3\}}{\sigma_{Y_3}} \quad (28)$$

where σ_{Y_3} is the standard deviation of Y_3. Then, thanks to the properties of differential entropy [3], we have

$$h(Y_3) = h(\tilde{Y_3}) + \ln \sigma_{Y_3}. \quad (29)$$

$h(\tilde{Y_3})$ may then be expressed with respect to the negentropy of $\tilde{Y_3}$ by using (26), and the fact that the differential entropy of a standardized Gaussian RV is equal to $[\ln(2\pi) + 1]/2$, as may be computed directly or derived from [1]. Applying (27) to the RV defined in (28), and using the translation and scaling properties of cumulants, one eventually gets

$$\begin{aligned} h(Y_3) \simeq \frac{\ln(2\pi)+1}{2} + \ln\sigma_{Y_3} - \frac{1}{12\sigma_{Y_3}^6}\mathrm{cum}_3(Y_3)^2 - \frac{1}{48\sigma_{Y_3}^8}\mathrm{cum}_4(Y_3)^2 \\ -\frac{7}{48\sigma_{Y_3}^{12}}\mathrm{cum}_3(Y_3)^4 + \frac{1}{8\sigma_{Y_3}^{10}}\mathrm{cum}_3(Y_3)^2\mathrm{cum}_4(Y_3). \quad (30) \end{aligned}$$

Inserting the latter expression in (25) yields the approximate cost function $C_3(Y)$ to be minimized. Using the standard cumulant-vs-moment expressions, $C_3(Y)$ may be rewritten as a combination of expectations of explicitly defined RVs. Practical estimators of this cost function may then be derived. Note that, contrary to the initial cost function $C_2(Y)$, it is not guaranteed at this stage that the *approximate* function $C_3(Y)$ obtained here reaches its global minimum exactly and only when source separation is achieved. The analysis of this topic is beyond the space allocated to this paper. Note also that the constant term $[\ln(2\pi)+1]/2$ due to the standardized Gaussian may be removed from $C_3(Y)$, since it has no influence on the minimization of $C_3(Y)$.

6 Extensions and Conclusions

In this paper, we introduced the Quantum Independent Component Analysis (QICA) concept, and we proposed two resulting criteria for performing the separation of coupled qubits, once they have been converted into classical data. Various optimization algorithms may be derived from these criteria. This e.g. includes standard gradient-based approaches. In addition, a straightforward and relatively cheap algorithm for reaching the *global* minimum of the considered cost functions here consists in performing a sweep over the *single* (bounded) tunable parameter of our separating system, and in computing corresponding sample estimates of the above-defined cost functions. We plan to assess the performance of that approach. However, actual data may hardly be presently obtained for performing such tests, since the implementation of QIP systems is only an emerging topic. Therefore, we will first develop a software simulation of coupled qubits.

References

1. Comon, P.: Independent Component Analysis, a new concept? Signal Processing 36, 287–314 (1994)
2. Comon, P., Jutten, C. (eds.): Handbook of blind source separation. Independent component analysis and applications. Academic Press, Oxford (2010)
3. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley, NY (1991)
4. Deville, Y., Deville, A.: Blind Separation of Quantum States: Estimating Two Qubits from an Isotropic Heisenberg Spin Coupling Model. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 706–713. Springer, Heidelberg (2007)
5. Deville, Y.: ICA Separability of Nonlinear Models with References: General Properties and Application to Heisenberg-Coupled Quantum States (Qubits). In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 173–180. Springer, Heidelberg (2010)
6. Deville, Y.: Traitement du signal: signaux temporels et spatiotemporels - Analyse des signaux, théorie de l'information, traitement d'antenne, séparation aveugle de sources, Ellipses Editions Marketing, Paris (2011) ISBN 978-2-7298-7079-9
7. Deville, Y., Deville, A.: Classical-processing and quantum-processing signal separation methods for qubit uncoupling. Quantum Information Processing (to appear), doi: 10.1007/s11128-011-0273-7
8. Deville, Y., Deville, A.: A quantum/classical-processing signal separation method for two qubits with cylindrical-symmetry Heisenberg coupling. In: Deloumeaux, P., et al. (eds.) Information Theory: New Research, ch.5, pp. 185–217. Nova Science Publishers, Hauppauge (to appear) ISBN: 978-1-62100-325-0
9. Kendall, M., Stuart, A.: The advanced theory of statistics, vol. 1. Charles Griffin, London & High Wycombe (1977)
10. Nielsen, M.A., Chuang, I.L.: Quantum computation and quantum information. Cambridge University Press, Cambridge (2000)

A Matrix Joint Diagonalization Approach for Complex Independent Vector Analysis

Hao Shen and Martin Kleinsteuber

Department of Electrical Engineering and Information Technology
Technische Universität München, Germany
{hao.shen,kleinsteuber}@tum.de

Abstract. Independent Vector Analysis (IVA) is a special form of Independent Component Analysis (ICA) in terms of group signals. Most IVA algorithms are developed via optimizing certain contrast functions. The main difficulty of these contrast function based approaches lies in estimating the unknown distribution of sources. On the other hand, tensorial approaches are efficient and richly available to the standard ICA problem, but unfortunately have not been explored considerably for IVA. In this paper, we propose a matrix joint diagonalization approach to solve the complex IVA problem. A conjugate gradient algorithm on an appropriate manifold setting is developed and investigated by several numerical experiments.

Keywords: Complex blind source separation, independent vector analysis, complex oblique projective manifold, conjugate gradient algorithm.

1 Introduction

Nowadays, Independent Component Analysis (ICA) has become a standard statistical tool for solving the Blind Source Separation (BSS) problem, which aims to recover signals from only the mixed observations without knowing the *a priori* information of both the source signals and the mixing process. It is known that application of the standard ICA model is often limited since it requires mutual statistical independence between all individual components. In many real applications, however, there are often groups of signals of interest, where components from different groups are *mutually statistically independent*, while *mutual statistical dependence* is still allowed between components in the same group. Such problems can be tackled by a technique now referred to as Multidimensional Independent Component Analysis (MICA) [1], or Independent Subspace Analysis (ISA) [2].

A special form of ISA arises in solving the BSS problem with convolutive mixtures [3]. After transferring the convolutive observations into frequency domain via short-time Fourier transforms, the convolutive BSS problem ends up with a collection of instantaneous complex BSS problems in each frequency bin. After solving the sub-problems individually, the final stage faces the challenge of aligning all statistically dependent component from different groups, which is

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 66–73, 2012.

referred to as the *permutation* problem. To overcome or avoid the permutation problem, a relatively new approach, namely, Independent Vector Analysis (IVA), is developed, cf. [4]. Many IVA algorithms are developed via optimizing certain contrast functions, cf. [5,6]. The main difficulty of these contrast function based approaches lies in estimating the unknown distribution of sources, which usually require a large number of observations [7].

On the other hand, tensorial approaches are efficient and richly available to the standard ICA problem, but unfortunately have not been explored considerably for IVA. In this paper, by assuming that the cross correlation matrices between different signal groups do not vanish, we propose a matrix joint diagonalization approach to solve the complex IVA problem. After adapting the so-called the complex oblique projective (COP) manifold, an appropriate setting for the standard instantaneous complex ICA problem [8], to the current scenario, we develop an efficient conjugate gradient (CG) based IVA algorithm.

The paper is organized as follows. Section 2 introduces briefly the linear complex IVA problem and recall some basic concepts of the COP manifold required for developing a CG algorithm. In Section 3, we develop an intrinsic conjugate gradient IVA algorithm. Finally in Section 4, performance of our proposed approach in terms of separation quality is investigated by several experiments.

2 Problem Descriptions and Prelimiaries

Let us start with some notations and definitions. In this work, we denote by $(\cdot)^{\mathsf{T}}$ the matrix transpose, $(\cdot)^{\mathsf{H}}$ the Hermitian transpose, $\overline{(\cdot)}$ the complex conjugate of entries of a matrix, and by $Gl(m)$ the set of all $m \times m$ invertible complex matrices.

2.1 Complex Independent Vector Analysis

Given k instantaneous complex linear Independent Component Analysis (ICA) problems

$$w_i(t) = A_i s_i(t), \qquad \text{for } i = 1, \ldots, k, \tag{1}$$

where $s_i(t) = [s_{i1}(t), \ldots, s_{im}(t)]^{\mathsf{T}} \in \mathbb{C}^m$ be a group of m mutually statistically independent complex signals, $A_i \in Gl(m)$ is the mixing matrix, and $w_i(t) = [w_{i1}(t), \ldots, w_{im}(t)]^{\mathsf{T}} \in \mathbb{C}^m$ presents m corresponding observed linear mixtures of $s_i(t)$. One critical assumption of IVA is that signals in all sub-problems are statistically aligned, i.e., all the j-th sources from different sub-problems, i.e. $\{s_{ij}(t)\}_{i=1}^{k}$, are mutually statistically dependent.

As the standard ICA model, we assume without loss of generality that sources $s(t)$ have zero mean and unit variance, i.e.,

$$\mathbb{E}[s_i(t)] = 0, \qquad \text{and} \qquad \operatorname{cov}(s_i) := \mathbb{E}[s_i(t) s_i^{\mathsf{H}}(t)] = I_m, \tag{2}$$

where $\mathbb{E}[\cdot]$ denotes the expectation over time index t, and I_m is the $m \times m$ identity matrix. The expression $\operatorname{cov}(s_i)$ is referred to as the *complex covariance matrix* of the sources $s_i(t)$.

The task of IVA is to find a set of demixing matrices $\{X_i\}_{i=1}^k \subset Gl(m)$ via

$$y_i(t) = X_i^{\mathsf{H}} w_i(t), \tag{3}$$

for $i = 1, \ldots, k$, such that

(1) All k sub-ICA problems are solved, and
(2) The statistical alignment between groups is restored, i.e., the estimated j-th signals $\{y_{ij}(t)\}_{i=1}^k$ are mutually statistically dependent.

The main idea of this work is to exploit the cross correlation matrices between groups of observations, defined as

$$\operatorname{cor}(w_i, w_j) := \mathbb{E}[w_i(t) w_j^{\mathsf{H}}(t)] = A_i \underbrace{\mathbb{E}[s_i(t) s_j^{\mathsf{H}}(t)]}_{=:\operatorname{cor}(s_i, s_j)} A_j^{\mathsf{H}}. \tag{4}$$

Similarly, pseudo cross correlation matrices can also be generated directly, i.e.

$$\operatorname{pcor}(w_i, w_j) := \mathbb{E}[w_i(t) w_j^{\mathsf{T}}(t)] = A_i \operatorname{pcor}(s_i, s_j) A_j^{\mathsf{T}}. \tag{5}$$

In this work, we assume that cross correlations between sources in all groups do not vanish. With a further assumption on sources being nonstationary, i.e. both (pseduo) cross correlation matrices of $s(t)$, and consequently, $w(t)$ as well, are time-varying, we arrive at a problem of jointly diagonalizing two sets of cross correlation and pseudo cross correlation matrices at different time intervals.

To summarize, we are interested in solving the following problem. For a complex IVA problem with k sub-problems, we construct cross correlation and pseudo cross correlation matrices at n time intervals, i.e. for all $i, j = 1, \ldots, k$ and $r = 1, \ldots, n$, a set of Hermitian positive matrices $\{C_{ij}^{(r)}\}_{i<j}$ and a set of complex symmetric matrices $\{R_{ij}^{(r)}\}_{i<j}$. The task is to find a set of matrices $\{X_i\}_{i=1}^k \subset Gl(m)$ such that

$$X_i^{\mathsf{H}} C_{ij}^{(r)} X_j \qquad \text{and} \qquad X_i^{\mathsf{H}} R_{ij}^{(r)} \overline{X}_j, \tag{6}$$

for all $i < j$ and $r = 1, \ldots, n$, are simultaneously diagonalized, or approximately simultaneously diagonalized subject to certain diagonality measure. Note that, the above problem is similar to the simultaneous SVD formulation proposed in [9], whereas in our current setting, the transforms $\{X_i\}$ are not restricted to be *unitary*.

2.2 Complex Oblique Projective Manifold

To make the paper self-contained, in this section, we briefly recall some concepts of the complex oblique projective manifold, and naturally extend it to the product manifold of k copies.

Recall the definition of the $(m-1)$-dimensional complex projective space $\mathbb{CP}^{m-1}$ as

$$\mathbb{CP}^{m-1} := \left\{ P \in \mathbb{C}^{m \times m} \,\middle|\, P^{\mathsf{H}} = P, P^2 = P, \operatorname{tr}(P) = 1 \right\}, \tag{7}$$

i.e. the set of all $(m-1)$-dimensional rank-one Hermitian projectors. Then, the COP manifold, denote by $\mathcal{Q}(m, \mathbb{C})$, is defined as

$$\mathcal{Q}(m, \mathbb{C}) := \left\{ (P_1, \ldots, P_m) \,\middle|\, P_i \in \mathbb{CP}^{m-1}, \det\left(\sum_{i=1}^{m} P_i\right) > 0 \right\}. \tag{8}$$

As $\mathcal{Q}(m, \mathbb{C})$ is an open and dense Riemannian submanifold of the m-times product of $\mathbb{CP}^{m-1}$ with the Euclidean product metric, i.e.

$$\overline{\mathcal{Q}(m, \mathbb{C})} = \underbrace{\mathbb{CP}^{m-1} \times \ldots \times \mathbb{CP}^{m-1}}_{m-\text{times}} =: \left(\mathbb{CP}^{m-1}\right)^m, \tag{9}$$

where $\overline{\mathcal{Q}(m, \mathbb{C})}$ denotes the closure of $\mathcal{Q}(m, \mathbb{C})$, the tangent spaces, the geodesics, and the parallel transport for $\mathcal{Q}(m, \mathbb{C})$ and $(\mathbb{CP}^{m-1})^m$ coincide locally.

Let us denote by

$$\mathfrak{u}(m) := \left\{ \Omega \in \mathbb{C}^{m \times m} \,\middle|\, \Omega = -\Omega^{\mathsf{H}} \right\} \tag{10}$$

the set of skew-Hermitian matrices. Then, given any $\Upsilon = (P_1, \ldots, P_m) \in \mathcal{Q}(m, \mathbb{C})$, the tangent space of $\mathcal{Q}(m, \mathbb{C})$ at Υ is defined as

$$T_\Upsilon \mathcal{Q}(m, \mathbb{C}) \cong T_{P_1}\mathbb{CP}^{m-1} \times \ldots \times T_{P_m}\mathbb{CP}^{m-1}, \tag{11}$$

where $T_{P_i}\mathbb{CP}^{m-1}$ denotes the tangent space of $\mathbb{CP}^{m-1}$ at $P_i \in \mathbb{CP}^{m-1}$, i.e.

$$T_P\mathbb{CP}^{m-1} := \{ [P, \Omega] \mid \Omega \in \mathfrak{u}(m) \} \tag{12}$$

with matrix commutator $[A, B] := AB - BA$.

Let $\Phi = (\phi_1, \ldots, \phi_m) \in T_\Upsilon \mathcal{Q}(m, \mathbb{C})$ with $\phi_i \in T_{P_i}\mathbb{CP}^{m-1}$ for all $i = 1, \ldots, m$, a Riemannian product metric on $T_\Upsilon \mathcal{Q}(m, \mathbb{C})$ is constructed as

$$G \colon T_\Upsilon \mathcal{Q}(m, \mathbb{C}) \times T_\Upsilon \mathcal{Q}(m, \mathbb{C}) \to \mathbb{R}, \qquad G(\Phi, \Psi) := \sum_{i=1}^{m} \Re \operatorname{tr}(\phi_i \cdot \psi_i), \tag{13}$$

where $\Re Z$ is the real part of a complex number Z. The geodesic through $\Upsilon \in \mathcal{Q}(m, \mathbb{C})$ in direction $\Phi \in T_\Upsilon \mathcal{Q}(m, \mathbb{C})$ is given by

$$\gamma_{\Upsilon,\Phi} \colon \mathbb{R} \to \mathcal{Q}(m, \mathbb{C}), \qquad \gamma_{\Upsilon,\Phi}(t) := \left(\gamma_{P_1,\phi_1}(t), \ldots, \gamma_{P_m,\phi_m}(t)\right), \tag{14}$$

where $\gamma_{P,\phi}$ defines the geodesic through $P \in \mathbb{CP}^{m-1}$ in direction $\phi \in T_P\mathbb{CP}^{m-1}$

$$\gamma_{P,\phi} \colon \mathbb{R} \to \mathbb{CP}^{m-1}, \qquad \gamma_{P,\phi}(t) := \mathsf{e}^{t[\phi,P]} P \mathsf{e}^{-t[\phi,P]}. \tag{15}$$

Here, $\mathsf{e}^{(\cdot)}$ denotes the matrix exponential. Then, the parallel transport of $\Psi \in T_\Upsilon \mathcal{Q}(m, \mathbb{C})$ with respect to the Levi-Civita connection along the geodesic $\gamma_{\Upsilon,\Phi}(t)$ is

$$\tau_{\Upsilon,\Phi}(\Psi) := \left(\tau_{P_1,\phi_1}(\psi_1), \ldots, \tau_{P_m,\phi_m}(\psi_m)\right) \tag{16}$$

with $\tau_{P,\phi}$ being the parallel transport of $\psi \in T_P\mathbb{CP}^{m-1}$ with respect to the Levi-Civita connection along the geodesic $\gamma_{P,\phi}(t)$

$$\tau_{P,\phi}(\psi) = \mathsf{e}^{[\phi,P]}\psi\mathsf{e}^{-[\phi,P]}. \tag{17}$$

Trivially, by considering the complex IVA demixing model in (3), a product of k copies of the COP manifold, i.e. $\mathcal{Q}^k(m,\mathbb{C}) := \mathcal{Q}(m,\mathbb{C}) \times \ldots \times \mathcal{Q}(m,\mathbb{C})$, is an appropriate manifold setting to the complex IVA problem. The tangent spaces, the geodesics, and the parallel transport of $\mathcal{Q}^k(m,\mathbb{C})$ follow directly from the product manifold structure. We refer to [10] for further insights of the topic.

3 A CG Algorithm for Simultaneous Non-unitary SVD

Recall the definition of the *complex oblique manifold* as

$$\mathcal{O}(m,\mathbb{C}) := \left\{X \in Gl(m,\mathbb{C}) \,\middle|\, \operatorname{ddiag}(X^{\mathsf{H}}X) = I_m \right\}, \tag{18}$$

where $\operatorname{ddiag}(Z)$ forms a diagonal matrix, whose diagonal entries are just those of Z, and denote by $\mathcal{O}^k(m,\mathbb{C}) := \mathcal{O}(m,\mathbb{C}) \times \ldots \times \mathcal{O}(m,\mathbb{C})$ the product manifold of k copies of $\mathcal{O}(m,\mathbb{C})$. Then the off-norm cost function, a popular diagonality measure of matrices, is straightforwardly adapted to the current setting as

$$\begin{aligned} &f\colon \mathcal{O}^k(m,\mathbb{C}) \to \mathbb{R},\\ &f(X_1,\ldots,X_k) := \sum_{i<j}^{k}\sum_{r=1}^{n} \tfrac{1}{2}\left\|\operatorname{off}(X_i^{\mathsf{H}}C_{ij}^{(r)}X_j)\right\|_F^2 + \tfrac{1}{2}\left\|\operatorname{off}(X_i^{\mathsf{H}}R_{ij}^{(r)}\overline{X}_j)\right\|_F^2, \end{aligned} \tag{19}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrices. A direct calculation gives

$$\begin{aligned} &f(X_1,\ldots,X_k)\\ =&\sum_{i<j}^{k}\sum_{p\neq q}^{m}\sum_{r=1}^{n} x_{ip}^{\mathsf{H}}C_{ij}^{(r)}x_{jq}(x_{ip}^{\mathsf{H}}C_{ij}^{(r)}x_{jq})^{\mathsf{H}} + x_{ip}^{\mathsf{H}}R_{ij}^{(r)}\overline{x}_{jq}(x_{ip}^{\mathsf{H}}R_{ij}^{(r)}\overline{x}_{jq})^{\mathsf{H}}\\ =&\sum_{i<j}^{k}\sum_{p\neq q}^{m}\sum_{r=1}^{n} \operatorname{tr} x_{ip}x_{ip}^{\mathsf{H}}C_{ij}^{(r)}x_{jq}x_{jq}^{\mathsf{H}}C_{ij}^{(r)\mathsf{H}} + \operatorname{tr} x_{ip}x_{ip}^{\mathsf{H}}R_{ij}^{(r)}\left(x_{jp}x_{jp}^{\mathsf{H}}\right)^{\mathsf{T}}R_{ij}^{(r)\mathsf{H}}. \end{aligned} \tag{20}$$

Clearly, the function f induces the following function $\widetilde{f}$ on $\mathcal{Q}^k(m,\mathbb{C})$

$$\begin{aligned} &\widetilde{f}\colon \mathcal{Q}^k(m,\mathbb{C}) \to \mathbb{R},\\ &\widetilde{f}(\varUpsilon_1,\ldots,\varUpsilon_k) := \sum_{i<j}^{k}\sum_{p\neq q}^{m}\sum_{r=1}^{n} \operatorname{tr} P_{ip}C_{ij}^{(r)}P_{jq}C_{ij}^{(r)\mathsf{H}} + \operatorname{tr} P_{ip}R_{ij}^{(r)}P_{jq}^{\mathsf{T}}R_{ij}^{(r)\mathsf{H}}. \end{aligned} \tag{21}$$

Computing the first derivative of $\widetilde{f}$ at $(\varUpsilon_1, \ldots, \varUpsilon_k) \in \mathcal{Q}^k(m, \mathbb{C})$ in direction $(\varPhi_1, \ldots, \varPhi_k) \in T_{(\varUpsilon_1, \ldots, \varUpsilon_k)} \mathcal{Q}^k(m, \mathbb{C})$ gives

$$\begin{aligned} &\mathrm{D}\,\widetilde{f}(\varUpsilon_1, \ldots, \varUpsilon_k)(\varPhi_1, \ldots, \varPhi_k) \\ =&\sum_{i<j}^{k} \sum_{p \neq q}^{m} \sum_{r=1}^{n} \operatorname{tr} \phi_{ip} C_{ij}^{(r)} P_{jq} C_{ij}^{(r)\mathsf{H}} + \operatorname{tr} P_{ip} C_{ij}^{(r)} \phi_{jq} C_{ij}^{(r)\mathsf{H}} + \\ &\qquad + \operatorname{tr} \phi_{ip} R_{ij}^{(r)} P_{jq}^{\mathsf{T}} R_{ij}^{(r)\mathsf{H}} + \operatorname{tr} P_{ip} R_{ij}^{(r)} \phi_{jq}^{\mathsf{T}} R_{ij}^{(r)\mathsf{H}}. \end{aligned} \tag{22}$$

Then, the Riemannian gradient of $\widetilde{f}$ at $(\varUpsilon_1, \ldots, \varUpsilon_k) \in \mathcal{Q}^k(m, \mathbb{C})$, i.e. $(\varPhi_1, \ldots, \varPhi_k) := \nabla_{\widetilde{f}}(\varUpsilon_1, \ldots, \varUpsilon_k) \in T_{(\varUpsilon_1, \ldots, \varUpsilon_k)} \mathcal{Q}^k(m, \mathbb{C})$, is computed, for each element $\phi_{ip} \in T_{P_{ip}} \mathbb{CP}^{m-1}$, as

$$\begin{aligned} \phi_{ip} = \Bigg[P_{ip}, \Bigg[P_{ip}, &\sum_{j>i}^{k} \sum_{p \neq q}^{m} \sum_{r=1}^{n} C_{ij}^{(r)} P_{iq} C_{ij}^{(r)\mathsf{H}} + R_{ij}^{(r)} P_{iq}^{\mathsf{T}} R_{ij}^{(r)\mathsf{H}} \\ &+ \sum_{j<i}^{k} \sum_{p \neq q}^{m} \sum_{r=1}^{n} C_{ij}^{(r)\mathsf{H}} P_{iq} C_{ij}^{(r)} + R_{ij}^{(r)\mathsf{T}} P_{iq}^{\mathsf{T}} \overline{R_{ij}^{(r)}} \Bigg] \Bigg]. \end{aligned} \tag{23}$$

Straightforwardly, a conjugate gradient algorithm for minimizing the function $\widetilde{f}$ as defined in (21) follows. Due to the complexity of our algorithm and the space limit, the algorithm is sketched briefly as follows. We refer to [8,11] and references therein for detailed descriptions.

Algorithm 1. *A conjugate gradient IVA algorithm*

```
Step 1: Given an initial guess $(\varUpsilon_1^{(0)}, \ldots, \varUpsilon_k^{(0)}) \in \mathcal{Q}^k(m, \mathbb{C})$ and set $i = 0$.
Step 2: Set $i = i + 1$, let $(\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)}) = (\varUpsilon_1^{(i-1)}, \ldots, \varUpsilon_k^{(i-1)})$, and compute
            $(\varPhi_1^{(1)}, \ldots, \varPhi_k^{(1)}) = (\varPsi_1^{(1)}, \ldots, \varPsi_k^{(1)}) = -\nabla_{\widetilde{f}}(\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)})$.
Step 3: For $j = 1, \ldots, 2km(m-1) - 1$:
        $(i)$ Update $(\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)}) \leftarrow \gamma_{(\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)}), (\varPhi_1^{(i)}, \ldots, \varPhi_k^{(i)})}(\lambda^*)$, where
                $\lambda^* = \operatorname*{argmin}_{\lambda \in \mathbb{R}} \widetilde{f} \circ \gamma_{(\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)}), (\varPhi_1^{(i)}, \ldots, \varPhi_k^{(i)})}(\lambda)$;
        $(ii)$ Compute $(\varPsi_1^{(j+1)}, \ldots, \varPsi_k^{(j+1)}) = -\nabla_{\widetilde{f}}\left((\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)})\right)$;
        $(iii)$ Update conjugate search directions $(\varPhi_1^{(j+1)}, \ldots, \varPhi_k^{(j+1)})$.
Step 4: If $\left\|(\varUpsilon_1^{(i+1)}, \ldots, \varUpsilon_k^{(i+1)}) - (\varUpsilon_1^{(i)}, \ldots, \varUpsilon_k^{(i)})\right\|$ is small enough, stop.
        Otherwise, go to Step 2.
```

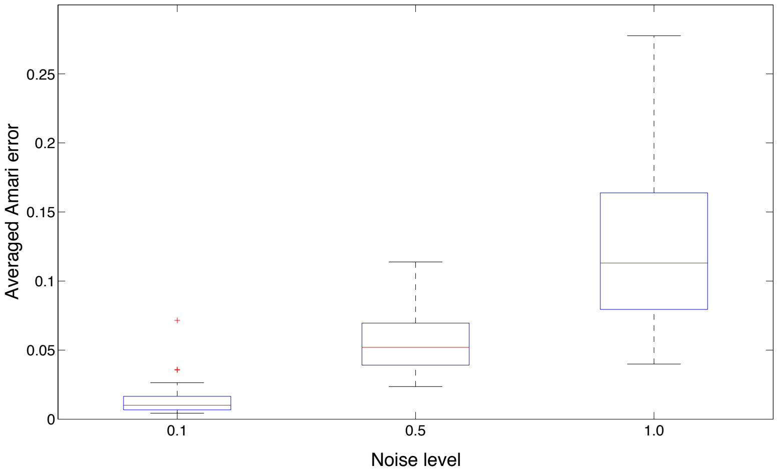


Fig. 1. Separation performance of the proposed CG algorithm

4 Numerical Experiments

In our experiment, we investigate performance of our method in terms of separation quality. Separation performance is measured by the averaged Amari error. Generally, the smaller the Amari error, the better the separation.

The task of our experiment is to jointly diagonalize a set of Hermitian positive definite matrices $\{C_{ij}^{(r)}\}_{i<j}$ and a set of complex symmetric matrices $\{R_{ij}^{(r)}\}_{i<j}$, which are constructed by

$$C_{ij}^{(r)} = A_i \varLambda_{ij}^{(r)} A_j^{\mathsf{H}} + \varepsilon E_H \qquad \text{and} \qquad R_{ij}^{(r)} = A_i \widehat{\varLambda}_{ij}^{(r)} A_j^{\mathsf{T}} + \varepsilon E_S \tag{24}$$

where $A_i \in Gl(m)$ is randomly picked, both real and imaginary parts of the diagonal entries of $\varLambda_{ij}^{(r)}$ and $\widehat{\varLambda}_{ij}^{(r)}$ are drawn from a uniform distribution on the interval $(0, 10)$, matrices $E_H \in \mathbb{C}^{m \times m}$ and $E_S \in \mathbb{C}^{m \times m}$ are a Hermitian and a complex symmetric matrix, respectively, whose real and imaginary parts are generated from a uniform distribution on the unit interval $(-0.5, 0.5)$, representing additive stationary noises, and $\varepsilon \in \mathbb{R}$ is the noise level.

We set $m = 3$, $k = 3$, $n = 3, \varepsilon \in \{0.1, 0.5, 1.0\}$, and run 50 tests. The quartile based boxplot of averaged Amari errors of our proposed algorithm against three different noise levels are drawn in Figure 1. Our CG algorithm demonstrates its correspondingly delaying performance with the increasing noise levels.

Acknowledgments. This work has been supported in parts by the German DFG funded Cluster of Excellence CoTeSys - Cognition for Technical Systems.

References

1. Cardoso, J.F.: Multidimensional independent component analysis. In: Proceedings of the 23rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998), Seattle, WA, USA, pp. 1941–1944 (1998)
2. Hyvärinen, A., Hoyer, P.O.: Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. Neural Computation 12(7), 1705–1720 (2000)
3. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Transactions on Speech and Audio Processing 11(2), 109–116 (2003)
4. Lee, I., Kim, T., Lee, T.W.: Independent vector analysis for convolutive blind speech separation. In: Makino, S., Lee, T.W., Sawada, H. (eds.) Blind Speech Separation. Signals and Communication Technology, pp. 169–192. Springer, Netherlands (2007)
5. Kim, T.: Real-time independent vector analysis for convolutive blind source separation. IEEE Transactions on Circuits and Systems Part I 57(7), 1431–1438 (2010)
6. Hao, J., Lee, I., Lee, T.W., Sejnowski, T.J.: Independent vector analysis for source separation using a mixture of gaussians prior. Neural Computation 22(6), 1646–1673 (2010)
7. Bermejo, S.: Finite sample effects in higher order statistics contrast functions for sequential blind source separation. IEEE Signal Processing Letters 12(6), 481–484 (2005)
8. Shen, H., Kleinsteuber, M.: Complex Blind Source Separation via Simultaneous Strong Uncorrelating Transform. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 287–294. Springer, Heidelberg (2010)
9. Maehara, T., Murota, K.: Simultaneous singular value decomposition. Linear Algebra and its Applications 435(1), 106–116 (2011)
10. Spivak, M.: A Comprehensive Introduction to Differential Geometry, 3rd edn., vol. 1 – 5. Publish or Perish, Inc. (1999)
11. Kleinsteuber, M., Hüper, K.: An intrinsic CG algorithm for computing dominant subspaces. In: Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), pp. IV1405–IV1408 (2007)

Algebraic Solutions to Complex Blind Source Separation

Hao Shen and Martin Kleinsteuber

Department of Electrical Engineering and Information Technology
Technische Universität München, Germany
{hao.shen,kleinsteuber}@tum.de

Abstract. The linear BSS problem can be solved under certain conditions via a joint diagonalization approach of only two matrices. Algebraic solutions, i.e. solutions that only involve eigenvalue decompositions or singular value decompositions, are of particular interest as efficient eigensolvers exist. Success of these methods depends significantly on particular properties of the sources, such as non-stationarity, non-whiteness, non-Gaussianity, and non-circularity. In this work, we propose alternative algebraic solutions to solve the complex BSS problem, which generalize the existing approaches. For example, applicability of SUT is limited to the positive definiteness of the covariance matrix, whereas our approach allows to exploit alternative information, such as autocorrelation and pseudo-autocorrelation, to solve the complex BBS problem.

Keywords: Complex blind source separation, second order statistics, (generalized) eigenvalue decomposition, Takagi factorization.

1 Introduction

Since the pioneering work on Independent Component Analysis (ICA) [1], the problem has attracted enormous attentions from various communities, and many efficient algorithms have been developed, cf. [2]. Despite the major interest in developing numerical iterative algorithms, cf. [3], a relatively small fraction of attention has been focused on the development of algebraic solutions, i.e. solutions that only involve eigenvalue decompositions or singular value decompositions. Although the algebraic approaches are in general less powerful and less robust to noise and estimation errors than their iterative counterparts, cf. [4,5], these methods are of particular interest, as they provide not only general solvability conditions for successful BSS, but also simple, efficient solutions based on various powerful eigensolvers.

By exploiting particular properties of the sources, such as *non-stationarity*, *non-whiteness*, *non-Gaussianity*, and *non-circularity*, the linear BSS problem can be solved by jointly diagonalizing two matrices, namely, the covariance matrix of the observations and an additional matrix, which reflects the assumptions. A corresponding algebraic solution, named Strong Uncorrelating Transform (SUT), has been developed in [6]. It employs one step of Eigenvalue Decomposition

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 74–81, 2012.

(EVD) of a positive definite Hermitian matrix, followed by one Takagi factorization, which is a special form of Singular Value Decomposition (SVD) of a complex symmetric matrix. We refer to [7,8] for more details on powerful EVD/SVD methods. Meanwhile, the other three assumptions lead to a unified approach of Generalized Eigenvalue Decomposition (GEVD), cf. [4], which simply involves two steps of EVD.

Success of algebraic solutions is known to be limited to their pre-selected assumptions. For example, the AMUSE algorithm is only capable to separate non-white signals with distinct autocorrelation coefficients, cf. [9], while the SUT approach fails when dealing with non-circular sources with indistinct circularity coefficients [6]. The present work completes the puzzle of algebraic solutions to the linear BSS problem. We consider all potential combinations of the aforementioned four properties of signals. Alternative algebraic solutions are developed for the cases when the existing approaches fail. In particular, we propose a generalization of the popular SUT algorithm, by eliminating the involvement of the covariance matrix and relaxing the constraint that one matrix needs to be positive definite.

This paper is organized as follows. In Section 2, we briefly introduce the complex valued linear BSS problem, and review second order statistics based approaches. Section 3 presents the main contribution of this work. Finally in Section 4, performance of the proposed algebraic BSS solution is investigated and compared with other algebraic approaches.

2 Complex BSS and Second-Order Statistics

Let us start with some notations and definitions. In this work, we denote by $(\cdot)^{\mathsf{T}}$ the matrix transpose, $(\cdot)^{\mathsf{H}}$ the Hermitian transpose, $(\cdot)^*$ the complex conjugate, and, $|\cdot|$, $\Re z$ and $\Im z$ the modulus $|z| = \sqrt{zz^*}$, the real part and the imaginary part of $z \in \mathbb{C}$ respectively. Furthermore, we denote by $Gl(m)$, $U(m)$ and $O(m)$, the set of all $m \times m$ *invertible*, *unitary* and *complex orthogonal* matrices, respectively.

2.1 Complex Linear BSS Model

Let $s(t) = [s_1(t), \ldots, s_m(t)]^{\mathsf{T}} \in \mathbb{C}^m$ be an m-dimensional vector representing the time series of m statistically independent complex signals. The noise-free instantaneous linear complex BSS model is given by

$$w(t) = As(t), \tag{1}$$

where $A \in \mathbb{C}^{m \times m}$ is the mixing matrix of full rank and $w(t) = [w_1(t), \ldots, w_m(t)]^{\mathsf{T}} \in \mathbb{C}^m$ presents m observed linear mixtures of $s(t)$. Without loss of generality, we assume that the sources $s(t)$ have zero mean, cf. [3], i.e. $\mathbb{E}[s(t)] = 0$, where $\mathbb{E}[\cdot]$ denotes the expectation over the time index t.

The task of the linear complex BSS problem (1) is to recover the source signals $s(t)$ by estimating the mixing matrix A or its inverse A^{-1} based only on the observations $w(t)$ via the demixing model

$$y(t) = X^{\mathsf{H}} w(t), \tag{2}$$

where $X^{\mathsf{H}} \in \mathbb{C}^{m \times m}$ is the demixing matrix, an estimation of A^{-1}, and $y(t) \in \mathbb{C}^m$ represents the corresponding extracted signals.

2.2 Second-Order Statistics Based Algebraic Solutions

Given the mixing model (2), the covariance matrix of the observations $w(t)$ over the time t is computed as

$$C_w := \mathbb{E}[w(t)w^{\mathsf{H}}(t)] = A \underbrace{\mathbb{E}[s(t)s^{\mathsf{H}}(t)]}_{=:C_s} A^{\mathsf{H}}, \tag{3}$$

where the covariance matrix of the sources C_s is diagonal and nonnegative following the statistical independence assumption. When the source signals are assumed to be nonstationary or time varying, the demixing matrix is expected to be identifiable via a joint diagonalization of two covariance matrices within different time intervals [10].

When source signals are stationary but non-white, i.e. with non-zero autocorrelations, the second order statistics in the form of autocorrelations for time lag $\tau > 0$ is often used, i.e.

$$\widetilde{C}_w(\tau) := \mathbb{E}[w(t)w^{\mathsf{H}}(t-\tau)] = A\widetilde{C}_s(\tau)A^{\mathsf{H}}. \tag{4}$$

Note that although the autocorrelation matrix of the sources is still diagonal, it needs not necessarily to be real. In other words, the autocorrelation matrix of the observations is not Hermitian in general. Similarly, the demixing matrix is expected to be identified via a joint diagonalization of one covariance matrix and one autocorrelation matrix within a non-zero time lag [11].

Furthermore, it is well known that for complex valued signals, there are certain properties that are not shared with their real valued counterparts, and that can be employed for complex BSS. Namely, besides the standard covariance matrix (3), a similar statistical quantity of complex valued signals, known as *pseudo-covariance matrix*, can be defined as

$$R_w := \mathbb{E}[w(t)w^{\mathsf{T}}(t)] = AR_s(t)A^{\mathsf{T}}. \tag{5}$$

The works in [6,12] have shown that, when the sources are all non-circular with distinct circularity coefficients, the demixing matrix can be successfully identified by jointly diagonalizing both the covariance and pseudo-covariance matrix. The resulting algebraic solution, namely SUT, provides a simple answer to the complex BSS problem. However, in order to separate non-circular signals with same circularity coefficients, one has to either utilize numerical iterative algorithms or employ some additional information. Recent work in [13] proposes to utilize the *pseudo-autocorrelation matrix* of signals, i.e.

$$\widetilde{R}_w(\tau) := \mathbb{E}[w(t)w^{\mathsf{T}}(t-\tau)] = A\widetilde{R}_s(\tau)A^{\mathsf{T}}, \tag{6}$$

and develops a numerical iterative algorithm to solve the linear BSS problem. Note that both the pseudo-covariance and pseudo-autocorrelation matrix are complex symmetric.

3 Algebraic Solutions to Complex BSS Problem

It is interesting to notice that existing algebraic solutions are only provided for the situations which combine the covariance matrix (3) and one of the other three quantities. The main contribution of this work is to consider all possible mixtures of second order statistics in developing algebraic solution to linear BSS. Thus, the problem studied in this work can be summarized as follows. Let two matrices be generated as

$$C_1 := A\Omega_1 A^{\dagger_1} \qquad \text{and} \qquad C_2 := A\Omega_2 A^{\dagger_2} \tag{7}$$

where $A \in Gl(m)$ and $\Omega_i = \operatorname{diag}(\omega_{i1}, \ldots, \omega_{im}) \in Gl(m)$ are unknown, and $(\cdot)^{\dagger_i}$ denotes either the Hermitian transpose or the matrix transpose. It is important to notice that the model (7) allows mixtures of both Hermitian congruence and matrix congruence. Then, the task is to find a matrix $X \in Gl(m)$, as an estimation of $A^{-\mathsf{H}}$, such that C_1 and C_2 are simultaneously diagonalized via

$$X^{\mathsf{H}} C_1 (X^{\mathsf{H}})^{\dagger_1} \qquad \text{and} \qquad X^{\mathsf{H}} C_2 (X^{\mathsf{H}})^{\dagger_2}. \tag{8}$$

3.1 Two Hermitian or Two Complex Symmetric

Algebraic solutions dealing with two matrices constructed via the Hermitian congruence have been studied in [4]. As the cases with the matrix congruence can be treated in the same way, in this subsection, we only briefly recap the results in [4] in a unified form.

Let two matrices $C_1, C_2 \in Gl(m)$ be constructed by

$$C_1 = A\Omega_1 A^{\dagger} \qquad \text{and} \qquad C_2 = A\Omega_2 A^{\dagger}, \tag{9}$$

where $(\cdot)^{\dagger}$ denotes either the Hermitian transpose or the matrix transpose. Now let us assume that one of the matrices, say C_2, is invertible. Then we compute

$$C_1 C_2^{-1} = A\Omega_1 A^{\dagger} \left(A\Omega_2 A^{\dagger}\right)^{-1} = A\Omega_1 \Omega_2^{-1} A^{-1}, \tag{10}$$

which gives the eigendecomposition of $C_1 C_2^{-1}$. It then follows directly that, if the eigenvalues of $C_1 C_2^{-1}$ are distinct, i.e. the diagonal entries of $\Omega_1 = \operatorname{diag}(\omega_{11}, \ldots, \omega_{1m})$ and of $\Omega_2 = \operatorname{diag}(\omega_{21}, \ldots, \omega_{2m})$ satisfy

$$\frac{\omega_{1i}}{\omega_{2i}} \neq \frac{\omega_{1j}}{\omega_{2j}} \tag{11}$$

for all pairs (i, j) with $i \neq j$, then the mixing matrix A is identifiable up to a column-wise scaling and permutation, and can be computed by an EVD of $C_1 C_2^{-1}$.

3.2 One Hermitian and One Complex Symmetric

In this subsection, we develop an algebraic solution to the situation with one matrix being constructed via Hermitian congruence and the other via matrix congruence. Hence, let us assume that two matrices $C_1, C_2 \in Gl(m)$ are constructed by

$$C_1 = A\Omega_1 A^{\mathsf{H}} \qquad \text{and} \qquad C_2 = A\Omega_2 A^{\mathsf{T}}. \tag{12}$$

Here, Ω_1 refers to either the covariance matrix (3), or more general, the autocorrelation matrix (4) of the sources, and Ω_2 corresponds to their pseudo-counterparts (5) and (6). We emphasize that neither C_1 nor C_2 needs to be positive definite.

It is clear that SUT does not apply to this situation, as the matrix C_1 is neither Hermitian nor positive definite in general. Nevertheless, the construction of our approach is largely inspired by the derivation of SUT. Namely, SUT aims to transfer C_1, restricted to be the covariance matrix, into the identity matrix, and simultaneously bring C_2 into a real diagonal matrix. In our case, we propose to take the opposite direction, i.e. to transfer C_2 into the identity matrix, and C_1 into a diagonal matrix.

Lemma 1. *Let $C_1 \in Gl(m)$ and $C_2 \in Gl(m)$ be constructed as in (12), and let $C_2 = U\Sigma U^{\mathsf{T}}$ be the Takagi factorization of C_2. Then,*

(i) the matrix $\widetilde{C}_1 := \Sigma^{-1/2}U^{\mathsf{H}}C_1 U\Sigma^{-1/2}$ admits a matrix factorization of the form $\widetilde{C}_1 = V\Lambda V^{\mathsf{H}}$, where $V \in O(m)$ and Λ is diagonal;

(ii) the transformation $X := U\Sigma^{-1/2}V^*$ brings C_2 into the identity matrix and C_1 into a diagonal matrix via $X^{\mathsf{H}}C_1X$ and $X^{\mathsf{H}}C_2X^*$.

Proof. (i) Recall the construction of C_2 as in (12), we have

$$A\Omega_2 A^{\mathsf{T}} = U\Sigma U^{\mathsf{T}}. \tag{13}$$

As diagonal entries of Σ are all positive, Equation (13) is equivalent to

$$\Sigma^{-1/2}U^{\mathsf{H}}A\Omega_2 A^{\mathsf{T}}U^*\Sigma^{-1/2} = I_m. \tag{14}$$

By substituting $\Omega_2 = (\Omega_2^{1/2})^2$ into the above equation, it can be seen that $V := \Sigma^{-1/2}U^{\mathsf{H}}A\Omega_2^{1/2}$ is complex orthogonal. With $A = U\Sigma^{1/2}V\Omega_2^{-1/2}$, Equation (12) yields

$$C_1 = A\Omega_1 A^{\mathsf{H}} = U\Sigma^{1/2}V \underbrace{\Omega_2^{-1/2}\Omega_1\Omega_2^{-\mathsf{H}/2}}_{=:\Lambda} V^{\mathsf{H}}\Sigma^{1/2}U^{\mathsf{H}}, \tag{15}$$

where Λ is diagonal. Then, Equation (15) is equivalent to

$$\Sigma^{-1/2}U^{\mathsf{H}}C_1U\Sigma^{-1/2} = V\Lambda V^{\mathsf{H}}. \tag{16}$$

(ii) It is straightforward to verify that

$$X^{\mathsf{H}}C_1X = V^{\mathsf{T}}\Sigma^{-1/2}U^{\mathsf{H}}C_1U\Sigma^{-1/2}V^* = \Lambda, \tag{17}$$

and

$$X^{\mathsf{H}}C_2X^* = V^{\mathsf{T}}\Sigma^{-1/2}U^{\mathsf{H}}C_2U^*\Sigma^{-1/2}V = I_m. \tag{18}$$

Hence the lemma follows. ∎

As the complex symmetric matrix C_2 reflects the pseudo second order statistics of complex signals, we name the matrix X *Pseudo-Uncorrelating Transform (PUT)* in referring its connection to SUT. A small computation shows that the matrix V consists of eigenvectors of $\widetilde{C}_1\widetilde{C}_1^{\mathsf{T}}$, as

$$\widetilde{C}_1\widetilde{C}_1^{\mathsf{T}} = V\varLambda V^{\mathsf{H}}V^*\varLambda V^{\mathsf{T}} = V\varLambda^2 V^{\mathsf{T}}. \tag{19}$$

Thus, if W is a matrix such that $\widetilde{C}_1\widetilde{C}_1^{\mathsf{T}} = W\varLambda' W^{-1}$ and if the eigenvalues $\varLambda'$ are pairwise distinct, it follows by the uniqueness of the EVD, that $V = W(W^{\mathsf{T}}W)^{-1/2}DP$, where P is a permutation and D is diagonal with entries being ±1. Then, the PUT algorithm is summarized as

Algorithm 1. *Pseudo-Uncorrelating Transform (PUT)*

```
Step 1: Construct $C_1, C_2$ from the observations $w(t)$, where $C_1$ and $C_2$ are
        constructed via Hermitian congruence and matrix congruence,
        respectively;
Step 2: Compute the Takagi factorization of $C_2 = U\varSigma U^{\mathsf{T}}$;
Step 3: Let $\widetilde{C}_1 := \varSigma^{-1/2}U^{\mathsf{H}}C_1U\varSigma^{-1/2}$, compute EVD of $\widetilde{C}_1\widetilde{C}_1^{\mathsf{T}} = W\varLambda W^{-1}$;
Step 4: Compute $V = W(W^{\mathsf{T}}W)^{-1/2}$;
Step 5: Compute the PUT matrix $X = U\varSigma^{-1/2}V^*$;
```

Finally, we characterize the applicability of PUT as an effective ICA technique. In the context of BSS, we refer to entries of $\varLambda$ as defined in (15), i.e. $\{\omega_{1i}/|\omega_{2i}|\}$ as the *pseudospectrum* of the sources.

Theorem 1. *Let the source $s(t)$ in the ICA model in* (1) *have pseudospectra $\varLambda$ defined in* (15) *with $\varLambda^2$ being pairwise distinct, then a pseudo-uncorrelating transform of the mixture $w(t)$ is a demixing matrix.*

Proof. Recall Equation (19), as an eigenvalue decomposition, the matrix V is determined up to a permutation and a columnwise sign difference, if the eigenvalues of $\widetilde{C}_1\widetilde{C}_1^{\mathsf{T}}$, i.e. $\varLambda^2$, are pairwise distinct. Let $\widehat{V} := VDP$, where $D = \operatorname{diag}(d_1, \ldots, d_m)$ with $d_i = \pm 1$ and P is a permutation matrix. Then $\widehat{X} := U\varSigma^{-1/2}\widehat{V}^*$ is a PUT matrix. A direct computation shows

$$\begin{aligned} \widehat{X}^{\mathsf{H}}A &= (U\varSigma^{-1/2}\widehat{V}^*)^{\mathsf{H}}U\varSigma^{1/2}V\varOmega_2^{-1/2} \\ &= PDV^{\mathsf{T}}\varSigma^{-1/2}U^{\mathsf{H}}U\varSigma^{1/2}V\varOmega_2^{-1/2} \\ &= PD\varOmega_2^{-\mathsf{H}/2}. \end{aligned} \tag{20}$$

Namely, the PUT matrix $\widehat{X}$ is an estimation of $A^{-\mathsf{H}}$ up to a permutation and a columnwise scaling. Then the result follows. ■

Remark 1. It is interesting to know that, when the matrix C_1 is Hermitian and positive definite, i.e. $\omega_{1i} > 0$ for all $i = 1, \ldots, m$. Then the pseudospectra $\varLambda$ as in (15) are simply the reciprocal of the circularity coefficients of sources. Our result coincides with the identifiability condition of SUT, cf. theorem 2 in [6].

The second observation is that SUT of an arbitrary pair of one positive definite Hermitian and one complex symmetric matrix does always exist, cf. [14]. While PUT does not hold in general for an arbitrary pair of Hermitian and complex symmetric matrix. However, existence of SUT implies the applicability of PUT on an arbitrary pair of positive definite Hermitian and complex symmetric matrix. In other words, PUT can be considered as a generalization of SUT.

Corollary 1. *For an arbitrary pair of one Hermitian positive definite and one nonsingular complex symmetric matrix, a PUT matrix always exists.*

4 Numerical Experiments

In this section, we investigate separation performance of several algebraic BSS solutions. In particular, we denote by *EVD1, EVD2, EVD3* three eigendecomposition based approaches employing non-stationarity (two covariance mateices), non-whiteness (one covariance matrix and one autocorrelation matrix), and non-circularity (one autocorrelation matrix and one pseudo-autocorrelation matrix), respectively. Separation performance is measured by the normalized Amari error proposed in [15]. It is important to notice that estimations of the demixing matrix X from different methods might differ in column-wise scaling. Thus, in order to compare the methods, we normalize all columns of each estimated X. Generally speaking, the smaller the Amari error, the better the separation.

The task of our experiment is to five stationary, non-white, non-circular (with identical circularity coefficients) sources, which are constructed as, for $k = 1, \ldots, 5$,

$$\begin{cases} \Re s_k(t) = \mathcal{N}_{(0,1)}(t) + \sin(\frac{\pi}{100k} t), \\ \Im s_k(t) = \mathcal{N}_{(0,2)}(t) + \cos(\frac{\pi}{100k} t), \end{cases} \tag{21}$$

where $\mathcal{N}_{(0,1)}(t)$ denotes a sample drawn from a standard normal distribution. We run the experiment for 100 times, and the quartile based boxplot of Amari

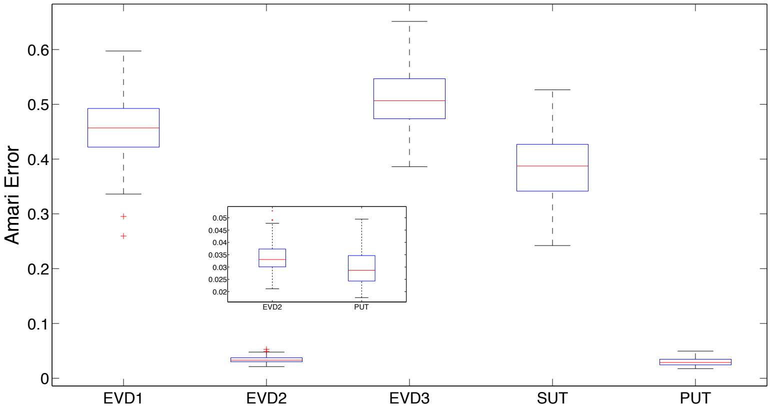


Fig. 1. Separation performance of algebraic solutions

errors for each method are drawn in Fig. 1. For this particular dataset, it is obvious that algebraic approaches based on nonstationarity and noncircularity, i.e. *EVD1, EVD3, SUT*, fail the task. Whereas both *PUT* and *EVD2* succeed in achieving good separations. With a closer look to the result in the zoomed-in window, *PUT* approach outperforms the *EVD2* slightly.

Acknowledgments. This work has been supported in parts by the German DFG funded Cluster of Excellence CoTeSys - Cognition for Technical Systems.

References

1. Comon, P.: Independent component analysis, a new concept? Signal Processing 36(3), 287–314 (1994)
2. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press Inc. (2010)
3. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
4. Parra, L., Sajda, P.: Blind source separation via generalized eigenvalue decomposition. The Journal of Machine Learning Research 4(7-8), 1261–1269 (2004)
5. Shen, H., Kleinsteuber, M.: Complex Blind Source Separation via Simultaneous Strong Uncorrelating Transform. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 287–294. Springer, Heidelberg (2010)
6. Eriksson, J., Koivunen, V.: Complex-valued ICA using second order statistics. In: Proceedings of the 14th IEEE International Workshop on MLSP, pp. 183–191 (2004)
7. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, New York (1985)
8. Kleinsteuber, M.: A sort-Jacobi algorithm for semisimple Lie algebras. Linear Algebra and its Applications 430(1), 155–173 (2009)
9. Tong, L., Liu, R.W., Soon, V.C., Huang, Y.F.: Indeterminacy and identifiability of blind identification. IEEE Transactions on Circuits and Systems 38(5), 499–509 (1991)
10. Pham, D.T., Cardoso, J.F.: Blind separation of instantaneous mixtures of nonstationary sources. IEEE Transactions on Signal Processing 49(9), 1837–1848 (2001)
11. Molgedey, L., Schuster, H.G.: Separation of a mixture of independent signals using time delayed correlations. Physical Review Letters 72(23), 3634–3637 (1994)
12. De Lathauwer, L., de Moor, B.: On the blind separation of non-circular sources. In: Proceedings of the 11th EUSIPCO, pp. 99–102 (2002)
13. Li, X.L., Adalı, T.: Blind separation of noncircular correlated sources using Gaussian entropy rate. IEEE Transactions on Signal Processing 59(6), 2969–2975 (2011)
14. Benedetti, R., Cragnolini, P.: On simultaneous diagonalization of one Hermitian and one symmetric form. Linear Algebra and its Applications 57, 215–226 (1984)
15. Amari, S.I., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 757–763. The MIT Press (1996)

On the Separation Performance of the Strong Uncorrelating Transformation When Applied to Generalized Covariance and Pseudo-covariance Matrices

Arie Yeredor

School of Electrical Engineering, Tel-Aviv University, Israel
arie@eng.tau.ac.il

Abstract. Traditionally, the strong uncorrelating transformation (SUT) is applied to the zero-lag sample autocovariance and pseudo-autocovariance matrices of the observed mixtures for separating complex-valued stationary sources. The performance of the SUT in that context has been recently analyzed. In this work we extend the analysis to the case where the SUT is applied to "generalized" covariance and pseudo-covariance matrices - which are prescribed by an arbitrary symmetric, positive definite matrix, termed an "association matrix". The analysis applies not only to stationary sources, but also to sources with arbitrary complex-valued temporal covariance and pseudo-covariance. As we show, the use of generalized covariance and pseudo-covariance matrices for the SUT entails a potential for significant improvement in the resulting separation performance, as we also demonstrate in simulation.

1 Introduction and Model Assumptions

We address the use of the Strong Uncorrelating Transformation (SUT) for blind separation of complex-valued sources. Classically, the SUT is applied to the zero-lag sample-covariance and sample-pseudo-covariance matrices of the observed mixtures, yielding an estimate of the demixing matrix. The separation performance of the SUT in this context (in terms of the Interference to Source Ratio (ISR)) was recently analyzed in [1] for the case of wide-sense stationary sources.

It is, however, possible to apply the powerful tool of SUT to other second-order statistics matrices of the sources, other than the zero-lag covariance and pseudo-covariance. It would be interesting to explore whether (and if so, when) using the SUT with alternative matrix-pairs can yield improved separation performance relative to its classical use with zero-lag covariance and pseudo-covariance. Therefore, our objective in this work is to derive expressions for the resulting ISR when the SUT is applied to more general ("generalized") covariance and pseudo-covariance matrices.

We address the static, linear, square-invertible and noiseless mixture model $\boldsymbol{X} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}}$ (the reason for the tilde notation will become clear in the sequel), in which $\widetilde{\boldsymbol{S}} \overset{\triangle}{=} [\widetilde{\boldsymbol{s}}_1 \ \widetilde{\boldsymbol{s}}_2 \ \cdots \ \widetilde{\boldsymbol{s}}_K]^T$ is a $K \times N$ matrix containing the K unobserved

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 82–90, 2012.

source signals (each of length N) as its rows; $\widetilde{\boldsymbol{A}}$ is the unknown $K \times K$ mixing matrix (assumed to be nonsingular); and $\boldsymbol{X} \triangleq [\boldsymbol{x}_1 \; \boldsymbol{x}_2 \; \cdots \; \boldsymbol{x}_K]^T$ is the $K \times N$ matrix of K observed mixtures.

To define the "generalized" covariance and pseudo-covariance matrices, let $\boldsymbol{P}$ denote some arbitrary $N \times N$ real-valued symmetric positive-definite matrix, which we term an "association-matrix", and let the sample *generalized covariance* and *generalized pseudo-covariance* matrices $\widehat{\boldsymbol{R}}$ and $\overline{\widehat{\boldsymbol{R}}}$ be given by

$$\widehat{\boldsymbol{R}}_x = \tfrac{1}{N}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^H \qquad\qquad \overline{\widehat{\boldsymbol{R}}}_x = \tfrac{1}{N}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^T \tag{1}$$

(respectively). Evidently, $\widehat{\boldsymbol{R}}_x$ is always Hermitian and $\overline{\widehat{\boldsymbol{R}}}_x$ is always symmetric. Different types of generalized covariance and pseudo-covariance matrices can be attained by different selection of the association matrix, for example:

- If $\boldsymbol{P}$ is taken as the $N \times N$ identity matrix, then $\widehat{\boldsymbol{R}}$ and $\overline{\widehat{\boldsymbol{R}}}$ coincide with the "standard" sample covariance and pseudo-covariance, taken uniformly over the entire observation interval;
- If $\boldsymbol{P}$ is diagonal (with positive, possibly different elements along its diagonal), then $\widehat{\boldsymbol{R}}$ and $\overline{\widehat{\boldsymbol{R}}}$ are temporally-weighted sample covariance and pseudo-covariance, with weighting prescribed by the diagonal values of $\boldsymbol{P}$.
- If $\boldsymbol{P}$ is a Toeplitz matrix then $\widehat{\boldsymbol{R}}$ (resp., $\overline{\widehat{\boldsymbol{R}}}$) is a linear combination of sample correlations (pseudo-correlations) matrices at different lags, as prescribed by the values along the diagonals of $\boldsymbol{P}$.

For our subsequent performance analysis we need to introduce assumptions on the statistical properties of the sources. In addition to the standard ICA assumption, that the sources are zero-mean mutually independent stochastic processes, we shall only need to quantify second-order properties of the sources. We shall not make any particular structural assumption, such as stationarity, but merely denote (for $k = 1, \ldots, K$) by

$$\widetilde{\boldsymbol{C}}_k = E[\widetilde{\boldsymbol{s}}_k \widetilde{\boldsymbol{s}}_k^H], \qquad\qquad \overline{\widetilde{\boldsymbol{C}}}_k = E[\widetilde{\boldsymbol{s}}_k \widetilde{\boldsymbol{s}}_k^T] \tag{2}$$

the complex-valued $N \times N$ temporal covariance and pseudo-covariance matrices of the sources. Note that we do not assume any particular structure of $\widetilde{\boldsymbol{C}}_k$ and / or of $\overline{\widetilde{\boldsymbol{C}}}_k$, and do not assume knowledge of these matrices for separation, but only for the performance analysis. No further information on the distributions of the sources is needed for our small-errors analysis.

2 Normalization Model and the SUT

Due to the inherent scale and phase ambiguity in complex-valued ICA (equivalent to the scale and sign ambiguity in the real-valued case), we assume a scaling and phase correction convention as follows. Given a selected association matrix $\boldsymbol{P}$, denote the following (for each source, $k=1,\dots,K$):

$$\mu_k \overset{\triangle}{=} \tfrac{1}{N}E[\widetilde{\boldsymbol{s}}_k^T\boldsymbol{P}\widetilde{\boldsymbol{s}}_k^*] = \tfrac{1}{N}\operatorname{Tr}\{\boldsymbol{P}\widetilde{\boldsymbol{C}}_k^*\} \quad \rho_k e^{j\phi_k} \overset{\triangle}{=} \tfrac{1}{N}E[\widetilde{\boldsymbol{s}}_k^T\boldsymbol{P}\widetilde{\boldsymbol{s}}_k] = \tfrac{1}{N}\operatorname{Tr}\{\boldsymbol{P}\overline{\widetilde{\boldsymbol{C}}}_k\}, \quad (3)$$

where the superscript * denotes complex-conjugation, $j=\sqrt{-1}$, and all the parameters μ_k, ρ_k and ϕ_k are real-valued. In addition, μ_k is positive (due to the positive definiteness of $\boldsymbol{P}$), and ρ_k is non-negative. Now define (for $k=1,\dots,K$) the normalization factors $\eta_k \overset{\triangle}{=} \sqrt{\mu_k e^{j\phi_k}}$. If we further define "normalized sources" as $\boldsymbol{s}_k \overset{\triangle}{=} \widetilde{\boldsymbol{s}}_k/\eta_k$ and a new mixing-matrix $\boldsymbol{A}=\widetilde{\boldsymbol{A}}\cdot\operatorname{Diag}\{\eta_1,\dots,\eta_K\}$, then we can describe the observed mixtures as (scaled, rotated) mixtures of the "normalized" version of the same sources: $\boldsymbol{X}=\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{S}}=\boldsymbol{A}\boldsymbol{S}$, where $\boldsymbol{S}\overset{\triangle}{=}[\boldsymbol{s}_1 \ \cdots \ \boldsymbol{s}_K]^T$ is the $K\times N$ matrix of normalized sources, each satisfying

$$\tfrac{1}{N}E[\boldsymbol{s}_k^T\boldsymbol{P}\boldsymbol{s}_k^*]=1 \qquad\qquad \tfrac{1}{N}E[\boldsymbol{s}_k^T\boldsymbol{P}\boldsymbol{s}_k]=\frac{\rho_k}{\mu_k}\overset{\triangle}{=}\kappa_k, \quad (4)$$

where $\kappa_k \geq 0$ is the generalized *circularity coefficient* (e.g., [2,3]) of the k-th source with respect to $\boldsymbol{P}$. We assume that the sources have distinct generalized circularity coefficients, and are ordered in a descending order of these coefficients (i.e., $\kappa_1>\kappa_2>\cdots>\kappa_K$). In addition, we shall denote by $\boldsymbol{K}\in\mathbb{R}^{K\times K}$ the diagonal matrix holding $\kappa_1,...,\kappa_K$ along its main diagonal (this matrix is sometimes called the *circularity spectrum* [2] matrix). The source separation goal is now to separate the normalized sources, which is equivalent, under the conventional scaling and phase ambiguities, to separation of the original sources.
In addition, we have (for $k=1,\dots,K$)

$$\boldsymbol{C}_k \overset{\triangle}{=} E[\boldsymbol{s}\boldsymbol{s}^H]=\frac{1}{\mu_k}\widetilde{\boldsymbol{C}}_k \qquad\qquad \overline{\boldsymbol{C}}_k \overset{\triangle}{=} E[\boldsymbol{s}\boldsymbol{s}^T]=\frac{e^{-j\phi_k}}{\mu_k}\overline{\widetilde{\boldsymbol{C}}}_k \quad (5)$$

As mentioned above, we consider a separation scheme in which the SUT is applied to the matrices $\widehat{\boldsymbol{R}}_x$ and $\widehat{\overline{\boldsymbol{R}}}_x$. The SUT finds a matrix $\widehat{\boldsymbol{B}}$, such that $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{R}}_x\widehat{\boldsymbol{B}}^H=\boldsymbol{I}$ and $\widehat{\boldsymbol{B}}\widehat{\overline{\boldsymbol{R}}}_x\widehat{\boldsymbol{B}}^T$ is a diagonal matrix which we shall denote $\widehat{\boldsymbol{K}}$, since it serves as an estimate of the true circularity spectrum matrix $\boldsymbol{K}$. $\widehat{\boldsymbol{B}}$ serves as an estimate of the separation matrix $\boldsymbol{B}=\boldsymbol{A}^{-1}$.

The computation of the SUT proceeds as follows (see, e.g., [4,5,6,3]). A whitening transformation $\widehat{\boldsymbol{W}}$ of $\widehat{\boldsymbol{R}}$ is found first: using the eigenvalues decomposition $\widehat{\boldsymbol{R}}=\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Phi}}^H$ (where $\widehat{\boldsymbol{\Phi}}$ is unitary and $\widehat{\boldsymbol{\Lambda}}$ is diagonal and positive), $\widehat{\boldsymbol{W}}=\widehat{\boldsymbol{\Lambda}}^{-1/2}\widehat{\boldsymbol{\Phi}}^H$ provides $\widehat{\boldsymbol{B}}$ up to multiplication by a unitary matrix $\boldsymbol{U}$, which can be extracted from the Singular Values Decomposition (SVD) of the matrix $\widehat{\boldsymbol{W}}\widehat{\overline{\boldsymbol{R}}}\widehat{\boldsymbol{W}}^T$. Indeed, denoting the SVD of this matrix as $\widehat{\boldsymbol{W}}\widehat{\overline{\boldsymbol{R}}}\widehat{\boldsymbol{W}}^T=\widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^H$ (where $\widehat{\boldsymbol{U}}$ and $\widehat{\boldsymbol{V}}$ are unitary and $\widehat{\boldsymbol{\Sigma}}$ is diagonal non-negative), the desired SUT matrix is given by $\widehat{\boldsymbol{B}}=\widehat{\boldsymbol{U}}^H\widehat{\boldsymbol{W}}$. We note in passing, that since $\widehat{\boldsymbol{W}}\widehat{\overline{\boldsymbol{R}}}\widehat{\boldsymbol{W}}^T$ is symmetric, we also have $\boldsymbol{U}=\boldsymbol{V}^*$, so the SVD essentially yields the Takagi factorization (see, e.g., [7]) in this case.

3 Performance Analysis

Any reasonable performance measure would be based on the overall mixing-unmixing matrix, $\boldsymbol{T} \overset{\triangle}{=} \widehat{\boldsymbol{B}}\boldsymbol{A}$, which would, under the normalization assumptions in Section 2, ideally be the Identity matrix. An important feature of SUT-based separation is the property of equivariance with respect to the mixing matrix $\boldsymbol{A}$ (closely related to the more general property of equivariance of the *generalized* uncorrelating transformation [8]): Given a particular realization of the sources, the resulting overall mixing-unmixing matrix $\boldsymbol{T}$ does not depend on $\boldsymbol{A}$. This appealing property can be shown in a way similar to the derivation in [1], by showing that $\boldsymbol{T}(\boldsymbol{X}) = \widehat{\boldsymbol{B}}(\boldsymbol{X})\boldsymbol{A}$ is actually $\boldsymbol{T}(\boldsymbol{S})$ - namely the SUT matrix of the **sources'** generalized covariance and pseudo-covariance matrices. Recalling (e.g., [2,3]) that if the generalized circularity coefficients are distinct then the SUT is unique, we conclude that $\boldsymbol{T}(\boldsymbol{AS}) = \boldsymbol{T}(\boldsymbol{S})$, regardless of the value of $\boldsymbol{A}$.

Thus, any performance measure which is based on $\boldsymbol{T}(\boldsymbol{X})$ will be independent of $\boldsymbol{A}$. One such popular measure is the ISR matrix, a $K \times K$ matrix in which the (k,ℓ)-th element $(k \neq \ell)$, defined as

$$\mathrm{ISR}_{k,\ell} = E\left[\left|\frac{T[k,\ell](\boldsymbol{X})}{T[k,k](\boldsymbol{X})}\right|^2\right]\cdot\frac{\mathrm{Tr}\{\boldsymbol{C}_\ell\}}{\mathrm{Tr}\{\boldsymbol{C}_k\}} \approx E[\,|T[k,\ell](\boldsymbol{X})|^2]\cdot\frac{\mathrm{Tr}\{\boldsymbol{C}_\ell\}}{\mathrm{Tr}\{\boldsymbol{C}_k\}}, \tag{6}$$

(where $\mathrm{Tr}\{\cdot\}$ denotes the trace operator) represents the mean relative residual energy of the ℓ-th source in the reconstruction of the k-th source. Note that the approximation in (6) is due to the small-errors assumption, which, combined with the scaling convention, enables to assume $T[k,k](\boldsymbol{X}) \approx 1$.

To proceed with our small-errors analysis, assume now that $\boldsymbol{T}(\boldsymbol{S}) = \boldsymbol{I} + \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \in \mathbb{C}^{K\times K}$ is a matrix with small elements representing the (small) deviation of $\boldsymbol{T}(\boldsymbol{S})$ from its ideal value of $\boldsymbol{I}$. We therefore have, from the SUT,

$$(\boldsymbol{I}+\boldsymbol{\Theta})\widehat{\boldsymbol{R}}_s(\boldsymbol{I}+\boldsymbol{\Theta}^H) = \boldsymbol{I}, \qquad (\boldsymbol{I}+\boldsymbol{\Theta})\widehat{\overline{\boldsymbol{R}}}_s(\boldsymbol{I}+\boldsymbol{\Theta}^T) = \widehat{\boldsymbol{K}}. \tag{7}$$

We can also express $\widehat{\boldsymbol{R}}_s = \boldsymbol{I} + \boldsymbol{\mathcal{E}}$ and $\widehat{\overline{\boldsymbol{R}}}_s = \boldsymbol{K} + \overline{\boldsymbol{\mathcal{E}}}$, where (under the small-errors analysis) $\boldsymbol{\mathcal{E}}, \overline{\boldsymbol{\mathcal{E}}} \in \mathbb{C}^{K\times K}$ are also matrices with small elements, representing the (small) deviations of the sources' sample covariance and pseudo-covariance (resp.) from their true values. This leads to a description of $\boldsymbol{\Theta}$ in terms of these deviations as follows:

$$\begin{aligned}(\boldsymbol{I}+\boldsymbol{\Theta})(\boldsymbol{I}+\boldsymbol{\mathcal{E}})(\boldsymbol{I}+\boldsymbol{\Theta}^H) = \boldsymbol{I} \;&\Rightarrow\; \boldsymbol{I}+\boldsymbol{\Theta}+\boldsymbol{\Theta}^H+\boldsymbol{\mathcal{E}} \approx \boldsymbol{I}\\ (\boldsymbol{I}+\boldsymbol{\Theta})(\boldsymbol{K}+\overline{\boldsymbol{\mathcal{E}}})(\boldsymbol{I}+\boldsymbol{\Theta}^T) = \widehat{\boldsymbol{K}} \;&\Rightarrow\; \boldsymbol{K}+\boldsymbol{\Theta}\boldsymbol{K}+\boldsymbol{K}\boldsymbol{\Theta}^T+\overline{\boldsymbol{\mathcal{E}}} \approx \widehat{\boldsymbol{K}},\end{aligned} \tag{8}$$

where we have neglected terms that are quadratic or higher in the elements of the small-valued matrices $\boldsymbol{\Theta}$, $\boldsymbol{\mathcal{E}}$ and $\overline{\boldsymbol{\mathcal{E}}}$. This leads, in turn, to

$$\boldsymbol{\Theta}+\boldsymbol{\Theta}^H = -\boldsymbol{\mathcal{E}}, \qquad \boldsymbol{\Theta}\boldsymbol{K}+\boldsymbol{K}\boldsymbol{\Theta}^T = (\widehat{\boldsymbol{K}}-\boldsymbol{K})-\overline{\boldsymbol{\mathcal{E}}}. \tag{9}$$

Recalling that both $\boldsymbol{K}$ and $\widehat{\boldsymbol{K}}$ are diagonal, we obtain for each off-diagonal term:

$$\begin{aligned} \Theta[k,\ell] + \Theta^*[\ell,k] &= -\mathcal{E}[k,\ell] \\ \kappa_\ell \Theta[k,\ell] + \kappa_k \Theta[\ell,k] &= -\overline{\mathcal{E}}[k,\ell], \end{aligned} \tag{10}$$

where $\Theta[k,\ell]$ denotes the (k,ℓ)-th element of $\boldsymbol{\Theta}$, etc. In matrix form we get:

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ \kappa_\ell & \kappa_k & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & \kappa_\ell & \kappa_k \end{bmatrix}}_{\triangleq \boldsymbol{H}_{k\ell}} \underbrace{\begin{bmatrix} \Theta_R[k,\ell] \\ \Theta_R[\ell,k] \\ \Theta_I[k,\ell] \\ \Theta_I[\ell,k] \end{bmatrix}}_{\triangleq \boldsymbol{\theta}_{k\ell}} = - \underbrace{\begin{bmatrix} \mathcal{E}_R[k,\ell] \\ \overline{\mathcal{E}}_R[k,\ell] \\ \mathcal{E}_I[k,\ell] \\ \overline{\mathcal{E}}_I[k,\ell] \end{bmatrix}}_{\triangleq \boldsymbol{\varepsilon}_{k\ell}} = -\frac{1}{2} \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ -j & j & 0 & 0 \\ 0 & 0 & -j & j \end{bmatrix}}_{\triangleq \boldsymbol{J}} \underbrace{\begin{bmatrix} \mathcal{E}[k,\ell] \\ \mathcal{E}^*[k,\ell] \\ \overline{\mathcal{E}}[k,\ell] \\ \overline{\mathcal{E}}^*[k,\ell] \end{bmatrix}}_{\triangleq \boldsymbol{\epsilon}_{k\ell}} \tag{11}$$

where $\Theta_R[k,\ell]$ and $\Theta_I[k,\ell]$ denote the real and imaginary parts (resp.) of $\Theta[k,\ell]$, with similar notations for elements of $\boldsymbol{\mathcal{E}}$ and $\overline{\boldsymbol{\mathcal{E}}}$. Observe now, that the (k,ℓ)-th elements of $\boldsymbol{\mathcal{E}}$ and $\overline{\boldsymbol{\mathcal{E}}}$ are (for $k \neq \ell$) simply the off-diagonal elements of the sample generalized covariance and pseudo-covariance matrices $\widehat{\boldsymbol{R}}$ and $\widehat{\overline{\boldsymbol{R}}}$ (resp.) of the sources, and therefore $\mathcal{E}[k,\ell] = \frac{1}{N}\boldsymbol{s}_k \boldsymbol{P} \boldsymbol{s}_\ell^*$ and $\overline{\mathcal{E}}[k,\ell] = \frac{1}{N}\boldsymbol{s}_k \boldsymbol{P} \boldsymbol{s}_\ell$. As such, these are obviously zero-mean random variables (since the sources are mutually uncorrelated), which means (from (11)) that the off-diagonal elements of $\boldsymbol{\Theta}$ are also zero-mean (under the small-errors assumption). In order to obtain the variances of the off-diagonal elements of $\boldsymbol{\Theta}$ (which by (6) are the respective elements of the ISR matrix, up to scaling normalization), we first need the covariance matrix of the vector $\boldsymbol{\epsilon}_{k\ell} \triangleq \left[\mathcal{E}[k,\ell] \;\; \mathcal{E}^*[k,\ell] \;\; \overline{\mathcal{E}}[k,\ell] \;\; \overline{\mathcal{E}}^*[k,\ell]\right]^T$. To this end, we note the following joint moments (for all $k \neq \ell$):

$$\begin{aligned} E\left[\mathcal{E}[k,\ell]\mathcal{E}^*[k,\ell]\right] &= \frac{1}{N^2} \sum_{p,q,m,n=1}^{N} E\left[s_k[p]P[p,q]s_\ell^*[q]s_k^*[m]P[m,n]s_\ell[n]\right] \\ &= \frac{1}{N^2} \sum_{p,q,n,m=1}^{N} P[p,q]P[m,n]E\left[s_k[p]s_k^*[m]s_\ell[n]s_\ell^*[q]\right] \\ &= \frac{1}{N^2} \sum_{p,q,n,m=1}^{N} C_k[p,m]P[m,n]C_\ell[n,q]P[q,p] = \mathrm{Tr}\{\boldsymbol{C}_k \boldsymbol{P} \boldsymbol{C}_\ell \boldsymbol{P}\}, \end{aligned} \tag{12}$$

where we have used the statistical independence of the sources, as well as the symmetry of $\boldsymbol{P}$. Similarly, it is straightforward (although somewhat tedious) to verify that the entire covariance matrix of $\boldsymbol{\epsilon}_{k\ell}$ can be expressed as

$$E[\boldsymbol{\epsilon}_{k\ell}\boldsymbol{\epsilon}_{k\ell}^H] = \frac{1}{N^2} \begin{bmatrix} \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\} \\ \mathrm{Tr}\{\overline{\boldsymbol{C}}_k^*\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k^*\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\} \\ \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\} \\ \mathrm{Tr}\{\overline{\boldsymbol{C}}_k^*\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\overline{\boldsymbol{C}}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\overline{\boldsymbol{C}}_k^*\boldsymbol{P}\overline{\boldsymbol{C}}_\ell^*\boldsymbol{P}\} & \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\} \end{bmatrix}. \tag{13}$$

For convenience, we may obtain more compact expressions for the subsequent derivation by defining:

$$\alpha_{k\ell} \triangleq \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\} = \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} \quad \overline{\alpha}_{kl} \triangleq \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}\} = \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}\}$$
$$\beta_{k\ell} \triangleq \mathrm{Tr}\{\boldsymbol{C}_k^*\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\} = \mathrm{Tr}\{\boldsymbol{C}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\}. \tag{14}$$

For the first two identities we used the conjugate symmetry of $\boldsymbol{P}\boldsymbol{C}_\ell\boldsymbol{P}$ and of $\boldsymbol{P}\boldsymbol{C}_\ell^*\boldsymbol{P}$, combined with the property that if $\boldsymbol{F}$ and $\boldsymbol{G}$ are Hermitian, then $\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{G}\}$ is real-valued; For the third identity we used the symmetry of $\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}$, combined with the property that if $\boldsymbol{F}$ is Hermitian and $\boldsymbol{G}$ is symmetric, then $\mathrm{Tr}\{\boldsymbol{F}\boldsymbol{G}\} = \mathrm{Tr}\{\boldsymbol{F}^*\boldsymbol{G}\}$. We also define

$$\gamma_{k\ell} \triangleq \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell^*\boldsymbol{P}\} \qquad \gamma_{k\ell} \triangleq \mathrm{Tr}\{\overline{\boldsymbol{C}}_k\boldsymbol{P}\overline{\boldsymbol{C}}_\ell\boldsymbol{P}\}. \tag{15}$$

Using these terms, we may express the covariance matrix (13) as

$$E[\boldsymbol{\epsilon}_{k\ell}\boldsymbol{\epsilon}_{k\ell}^H] = \frac{1}{N^2}\begin{bmatrix} \alpha_{k\ell} & \gamma_{k\ell} & \beta_{k\ell}^* & \beta_{\ell k} \\ \gamma_{k\ell}^* & \alpha_{k\ell} & \beta_{\ell k}^* & \beta_{k\ell} \\ \beta_{k\ell} & \beta_{\ell k} & \overline{\alpha}_{kl} & \overline{\gamma}_{kl} \\ \beta_{\ell k}^* & \beta_{k\ell}^* & \overline{\gamma}_{k\ell}^* & \overline{\alpha}_{kl} \end{bmatrix}. \tag{16}$$

We now proceed to obtain the covariance matrix of the (real-valued) vector $\boldsymbol{\varepsilon}_{k\ell} = \frac{1}{2}\boldsymbol{J}\boldsymbol{\epsilon}_{k\ell}$, given by

$$\begin{aligned}\boldsymbol{\Psi}_{k\ell} &\triangleq E[\boldsymbol{\varepsilon}_{k\ell}\boldsymbol{\varepsilon}_{k\ell}^T] = \frac{1}{4}\cdot\boldsymbol{J}E[\boldsymbol{\epsilon}_{k\ell}\boldsymbol{\epsilon}_{k\ell}^H]\boldsymbol{J}^H \\ &= \frac{1}{2N^2}\begin{bmatrix} \mathcal{R}\{\alpha_{k\ell}+\gamma_{k\ell}\} & \mathcal{R}\{\beta_{k\ell}+\beta_{\ell k}\} & \mathcal{I}\{\gamma_{k\ell}\} & \mathcal{I}\{\beta_{k\ell}+\beta_{\ell k}\} \\ \mathcal{R}\{\beta_{k\ell}+\beta_{\ell k}\} & \mathcal{R}\{\overline{\alpha}_{kl}+\overline{\gamma}_{kl}\} & \mathcal{I}\{\beta_{\ell k}-\beta_{k\ell}\} & \mathcal{I}\{\overline{\gamma}_{kl}\} \\ \mathcal{I}\{\gamma_{k\ell}\} & \mathcal{I}\{\beta_{\ell k}-\beta_{k\ell}\} & \mathcal{R}\{\alpha_{k\ell}-\gamma_{k\ell}\} & \mathcal{R}\{\beta_{k\ell}-\beta_{\ell k}\} \\ \mathcal{I}\{\beta_{k\ell}+\beta_{\ell k}\} & \mathcal{I}\{\overline{\gamma}_{kl}\} & \mathcal{R}\{\beta_{k\ell}-\beta_{\ell k}\} & \mathcal{R}\{\overline{\alpha}_{kl}-\overline{\gamma}_{kl}\} \end{bmatrix},\end{aligned} \tag{17}$$

where $\mathcal{R}\{\cdot\}$ and $\mathcal{I}\{\cdot\}$ denote the Real and Imaginary parts (resp.). Note that the $\alpha_{k\ell}$ and $\overline{\alpha}_{kl}$ coefficients are always real-valued.

The last step is to obtain the covariance matrix of $\boldsymbol{\theta}_{k\ell} = \boldsymbol{H}_{k\ell}^{-1}\boldsymbol{\varepsilon}_{k\ell}$, evidently given by $\boldsymbol{C}_{k,\ell} \triangleq E[\boldsymbol{\theta}_{k,\ell}\boldsymbol{\theta}_{k,\ell}^T] = \boldsymbol{H}_{k\ell}^{-1}\boldsymbol{\Psi}_{k\ell}\boldsymbol{H}_{k\ell}^{-T}$. Note, however, that for obtaining the (k,ℓ)-th element of the ISR matrix,

$$\mathrm{ISR}_{k,\ell} = E[|\Theta[k,\ell]|^2] = E[\Theta_R^2[k,\ell] + \Theta_I^2[k,\ell]] \tag{18}$$

we only need $\mathrm{var}\{\Theta_R[k,\ell]\} = \boldsymbol{C}_{k,\ell}[1,1]$ and $\mathrm{var}\{\Theta_I[k,\ell]\} = \boldsymbol{C}_{k,\ell}[3,3]$. Noting further that $\boldsymbol{H}_{k,\ell}$ is a block-diagonal matrix (with two 2×2 blocks), we can identify these elements from the $(1,1)$ elements of the respective 2×2 blocks:

$$\begin{aligned}\boldsymbol{C}_{k,\ell}[1:2,1:2] &= \begin{bmatrix} 1 & 1 \\ \kappa_\ell & \kappa_k \end{bmatrix}^{-1} \cdot \boldsymbol{\Psi}_{k\ell}[1:2,1:2] \cdot \begin{bmatrix} 1 & \kappa_\ell \\ 1 & \kappa_k \end{bmatrix}^{-1} \\ &= \frac{1}{2N^2(\kappa_k-\kappa_\ell)^2}\begin{bmatrix} \kappa_k & -1 \\ -\kappa_\ell & 1 \end{bmatrix} \cdot \mathcal{R}\left\{\begin{bmatrix} \alpha_{k\ell}+\gamma_{k\ell} & \beta_{k\ell}+\beta_{\ell k} \\ \beta_{k\ell}+\beta_{\ell k} & \overline{\alpha}_{kl}+\overline{\gamma}_{kl} \end{bmatrix}\right\} \cdot \begin{bmatrix} \kappa_k & -\kappa_\ell \\ -1 & 1 \end{bmatrix}\end{aligned} \tag{19}$$

$$\begin{aligned}\boldsymbol{C}_{k,\ell}[3:4,3:4] &= \begin{bmatrix} 1 & -1 \\ \kappa_\ell & \kappa_k \end{bmatrix}^{-1} \cdot \boldsymbol{\Psi}_{k\ell}[3:4,3:4] \cdot \begin{bmatrix} 1 & \kappa_\ell \\ -1 & \kappa_k \end{bmatrix}^{-1} \\ &= \frac{1}{2N^2(\kappa_k+\kappa_\ell)^2} \begin{bmatrix} \kappa_k & 1 \\ -\kappa_\ell & 1 \end{bmatrix} \cdot \mathcal{R}\left\{ \begin{bmatrix} \alpha_{k\ell}-\gamma_{k\ell} & \beta_{k\ell}-\beta_{\ell k} \\ \beta_{k\ell}-\beta_{\ell k} & \overline{\alpha}_{kl}-\overline{\gamma}_{kl} \end{bmatrix} \right\} \cdot \begin{bmatrix} \kappa_k & -\kappa_\ell \\ 1 & 1 \end{bmatrix}. \end{aligned} \tag{20}$$

Taking the $(1,1)$ element of each of these two matrices, we obtain (resp.):

$$E[\Theta_R^2[k,\ell]] = \mathcal{R}\left\{ \frac{\kappa_k^2(\alpha_{k\ell}+\gamma_{k\ell}) - 2\kappa_k(\beta_{k\ell}+\beta_{\ell k}) + (\overline{\alpha}_{kl}+\overline{\gamma}_{kl})}{2N^2(\kappa_k-\kappa_\ell)^2} \right\} \tag{21}$$

$$E[\Theta_I^2[k,\ell]] = \mathcal{R}\left\{ \frac{\kappa_k^2(\alpha_{k\ell}-\gamma_{k\ell}) + 2\kappa_k(\beta_{k\ell}-\beta_{\ell k}) + (\overline{\alpha}_{kl}-\overline{\gamma}_{kl})}{2N^2(\kappa_k+\kappa_\ell)^2} \right\}. \tag{22}$$

Finally, the asymptotic expression for each $\mathrm{ISR}_{k,\ell}$ is given by the sum of these two expressions ((21) and (22)), normalized by the ratio $\mathrm{Tr}\{\boldsymbol{C}_\ell\}/\,\mathrm{Tr}\{\boldsymbol{C}_k\}$.
We note some important properties of the ISR expression:

- Invariance with respect to the other sources: $\mathrm{ISR}_{k,\ell}$ depends only on the statistics of the k-th and ℓ-th sources, and not on other sources. Note, however, that this property only holds under our small-errors assumption, as a direct result of the approximation made in (8);
- Invariance with respect to the distributions of the sources: The ISR depends only on the temporal SOS of the (relevant) sources, and is independent of their higher-order temporal moments or particular distributions. Note that this property, too, is only valid under the small-errors approximation.
- Non-identifiability condition: If $\kappa_k = \kappa_\ell$, then the resulting $\mathrm{ISR}_{k,\ell}$ and $\mathrm{ISR}_{\ell,k}$ are infinite, meaning that two sources with the same generalized circularity coefficient with respect to $\boldsymbol{P}$ cannot be separated by the SUT of the respective generalized covariance and pseudo-covariance alone.

4 Simulation

To demonstrate the validity of our analytic derivations, as well as the potential performance gain in using generalized covariance and pseudo-covariance matrices, we present the following simulation results. We mixed $K = 3$ stationary sources, generated as follows. Each source $s_k[n]$ was $N = 500$ samples long, obtained as a filtered version of an iid zero-mean complex Gaussian noise source $w_k[n]$, with circularity coefficients 0.9, 0.8, and 0.7 for $w_1[n]$, $w_2[n]$ and $w_3[n]$ (resp.). Thus, each source was generated as $s_k[n] = \sum_{m=0}^{3} h_k[m] w_k[n-m]$ (for $k = 1, 2, 3$). The finite impulse-response (FIR) filters (all of order 3) were structured so as to have the following sets of zeros (in the z-plane): For $h_1[m]$: $\{0.8+0.8\jmath, 1-0.2\jmath, 2.6\jmath\}$; for $h_2[m]$: $\{-0.9, 1.5+0.9\jmath, 1.3+0.6\jmath\}$; and for $h_3[m]$: $\{1.3\jmath, -0.9-0.1\jmath, 0.6-0.6\}$. The sources were mixed by random complex-valued mixing matrices with elements randomly and independently drawn from a zero-mean unit-variance complex Gaussian distribution. The demixing matrix was estimated twice: First, as the SUT of the sample zero-lag covariance and pseudo-covariance matrices (corresponding to $\boldsymbol{P} = \boldsymbol{P}_1 = \boldsymbol{I}$), and then as the SUT of the sample generalized covariance and pseudo-covariance matrices obtained with $\boldsymbol{P} = \boldsymbol{P}_2$, structured as a 500×500 symmetric Toeplitz matrix with the generating (first row) vector $[1 \;\; 0.1 \;\; 0 \;\; -0.4 \;\; \mathbf{0}^T]$. The resulting empirical ISR values (averaged over 5000 independent trials) are presented in Table 1 below for both separation schemes, together with the analytically predicted values (in parentheses). A close match between the empirical and analytically predicted values (up to about 1dB) is observed, as well as significant performance differences (e.g., in $\mathrm{ISR}_{2,1}$) when using different association matrices.

Table 1. Empirical and (in parentheses) theoretically predicted ISR values [dB]

ISR[dB]	s_1	s_2	s_3
$s_1 : \boldsymbol{P} = \boldsymbol{P}_1$		−18.7(−17.6)	−25.8(−25.9)
$s_1 : \boldsymbol{P} = \boldsymbol{P}_2$		−23.0(−23.7)	−26.2(−26.6)
$s_2 : \boldsymbol{P} = \boldsymbol{P}_1$	−17.8(−17.4)		−20.0(−21.6)
$s_2 : \boldsymbol{P} = \boldsymbol{P}_2$	−23.6(−24.5)		−18.3(−18.9)
$s_3 : \boldsymbol{P} = \boldsymbol{P}_1$	−24.7(−24.7)	−19.2(−20.3)	
$s_3 : \boldsymbol{P} = \boldsymbol{P}_2$	−25.7(−26.1)	−17.8(−18.2)	

5 Conclusion

Using a small-errors analysis, we derived expressions for the resulting ISR in SUT-based separation using the observations' generalized covariance and pseudo-covariance matrices. The results depend only on the sources' (complex-valued) SOS and on the association matrix. Theoretically, the analytic expressions can also serve for optimizing the selection of an association matrix, whenever the sources' SOS are known.

References

1. Yeredor, A.: Performance analysis of the strong uncorrelating transformation in blind separation of complex-valued sources. To Appear in IEEE Transactions on Signal Processing (2012)
2. Eriksson, J., Koivunen, V.: Complex random vectors and ICA models: Identifiability, uniqueness, and separability. IEEE Transactions on Information Theory 52(3), 1017–1029 (2006)
3. Schreier, P., Scharf, L.: Statistical Signal Processing of Complex-valued Data. Cambridge University Press (2010)
4. Benedetti, R., Cragnolini, P.: On simultaneous diagonalization of one Hermitian and one symmetric form. Linear Algebra and its Applications 57, 215–226 (1984)
5. De Lathauwer, L., De Moor, B.: On the blind separation of non-circular sources. In: XIth European Signal Processing Conf. (EUSIPCO 2002), pp. 99–102 (September 2002)

6. Eriksson, J., Koivunen, V.: Complex-valued ICA using second order statistics. In: Proceedings of the 2004 IEEE Machine Learning for Signal Processing Workshop XIV, pp. 183–191 (2004)
7. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press (1985)
8. Ollila, E., Koivunen, V.: Complex ICA using generalized uncorrelating transform. Signal Processing 89(4), 365–377 (2009)

A Canonical Correlation Analysis Based Method for Improving BSS of Two Related Data Sets

Juha Karhunen, Tele Hao, and Jarkko Ylipaavalniemi

Dept. of Information and Computer Science, Aalto University, School of Science,
P.O. Box 15400, FI-00076 Aalto, Espoo, Finland
firstname.lastname@aalto.fi
http://ics.tkk.fi/en/

Abstract. We consider an extension of ICA and BSS for separating mutually dependent and independent components from two related data sets. We propose a new method which first uses canonical correlation analysis for detecting subspaces of independent and dependent components. Different ICA and BSS methods can after this be used for final separation of these components. Our method has a sound theoretical basis, and it is straightforward to implement and computationally not demanding. Experimental results on synthetic and real-world fMRI data sets demonstrate its good performance.

1 Introduction

Various independent component analysis (ICA) and blind source separation (BSS) methods [1, 2] are nowadays well-known techniques for blind extraction of useful information from single vector-valued data $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(N_x)]$ with many applications. The data model used in the basic linear ICA is simply

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} s_i(t)\mathbf{a}_i \tag{1}$$

Thus each data vector $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T$ is expressed as a linear combination of independent components or source signals $s_i(t)$, collected respectively to the source vector $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_n(t)]^T$. For simplicity, we first assume that both $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are zero mean n-vectors, and that the mixing matrix $\mathbf{A}$ is a full-rank constant $n \times n$ matrix with column vectors $\mathbf{a}_i$, $i = 1, 2, \ldots, n$.

In standard linear ICA, the index t which usually denotes time or sample index is not important, because the order of the data vectors $\mathbf{x}(t)$ can be arbitrary. This holds if they are samples from some multivariate statistical distribution. However, the data vectors $\mathbf{x}(t)$ have often important underlying temporal structure. Alternative BSS methods have been developed for utilizing such temporal information. They usually utilize either temporal autocorrelations directly or smoothly changing nonstationarity of variances. The assumptions and applications domains of these three major categories of methods based on the simple model (1) vary somewhat [1, 2].

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 91–98, 2012.

The most widely used standard ICA method is currently FastICA [1, 3] due to its efficient implementation and fast convergence which makes it applicable to higher dimensional problems, too. From the many methods using temporal autocorrelations, we have used the TDSEP method [4] which performs usually well. Some attempts have been made to combine different types of BSS methods so that they would be able to separate wider classes of source signals. In [5], an approximate method called UbiBSS is developed which tries to utilize higher-order statistics, temporal autocorrelations, and nonstationarity of variances. We have used its Matlab code [6] in our experiments.

ICA and BSS have been generalized into many directions from the simple linear noiseless model (1) [1, 2]. We consider a generalization in which one tries to find out mutually dependent and independent components from two different but related data sets $\mathbf{X}$ and $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N_y)]$. Data vectors $\mathbf{y}(t)$ have dimension m which can be different from dimension n of the data vectors $\mathbf{x}(t)$ in $\mathbf{X}$, but they obey a similar basic linear model

$$\mathbf{y}(t) = \mathbf{B}\mathbf{r}(t) = \sum_{i=1}^{m} r_i(t)\mathbf{b}_i \tag{2}$$

in which $\mathbf{r}(t)$ is m-vector and $\mathbf{B}$ $m \times m$ matrix.

This generalization of ICA and BSS has not been studied as much as several others, but some related work can be found in [7–9, 11–13]. In most of these methods the data model is more rectrictive than ours, assuming that in the data sets $\mathbf{X}$ and $\mathbf{Y}$ there exist pairs of sources which are mutually dependent, but these sources are independent of all the other sources in $\mathbf{X}$ and $\mathbf{Y}$. In particular, canonical correlation analysis (CCA) or its extension to multiple data sets is applied in [11, 13], but in a different way than we do. Due to space limitations, we do not discuss these related works in more detail here.

2 Our Method

We apply canonical correlation analysis (CCA) to find the subspaces of dependent and independent sources in the two related data sets. CCA [14] is an old statistical technique which measures the linear relationships between two multidimensional datasets $\mathbf{X}$ and $\mathbf{Y}$ using their autocovariances and cross-covariances. CCA finds two bases, one for both $\mathbf{X}$ and $\mathbf{Y}$, in which the cross-correlation matrix between the data sets $\mathbf{X}$ and $\mathbf{Y}$ becomes diagonal and the correlations of the diagonal are maximized.

In CCA, the dimensions of the data vectors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ can be different, but they are assumed to have zero means. The canonical correlations and the respective basis vectors can be computed by solving a generalized eigenvalue problem as discussed in [14]. This solution simplifies considerably if the data vectors $\mathbf{x}$ and $\mathbf{y}$ are prewhitened [1]. It turns out that the basis vectors of CCA can then be determined from the singular value decomposition (SVD) of the cross-covariance matrix $\mathbf{C_{xy}} = \mathrm{E}\{\mathbf{x}\mathbf{y}^T\}$ of $\mathbf{x}$ and $\mathbf{y}$:

$$\mathbf{C}_{\mathbf{xy}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{L} \rho_i \mathbf{u}_i \mathbf{v}_i^T \tag{3}$$

Note that the SVD of $\mathbf{C}_{\mathbf{yx}} = \mathrm{E}\{\mathbf{y}\mathbf{x}^T\} = \mathbf{C}_{\mathbf{xy}}^T$ is quite similar and is obtained by transposing both sides of Eq. (3). There $\mathbf{U}$ and $\mathbf{V}$ are two orthogonal square matrices ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$) containing as their column vectors the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_j$. In our case, these singular vectors are the basis vectors providing canonical correlations. In general, the dimensionalities of the matrices $\mathbf{U}$ and $\mathbf{V}$ and consequently the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are different corresponding to different dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$. The pseudodiagonal matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{4}$$

consists of a diagonal matrix $\mathbf{D}$ containing the non-zero singular values appended with zero matrices so that the matrix $\boldsymbol{\Sigma}$ is compatible with the different dimensions of $\mathbf{x}$ and $\mathbf{y}$. These non-zero singular values are just the non-zero canonical correlations. If the cross-covariance matrix $\mathbf{C}_{\mathbf{xy}}$ has full rank, their number L is the smaller one of the dimensions of the data vectors $\mathbf{x}$ and $\mathbf{y}$.

We first make the data vectors $\mathbf{x} \in \mathbf{X}$ zero mean if necessary. These data vectors are whitened separately:

$$\mathbf{v}_{\mathbf{x}} = \mathbf{V}_{\mathbf{x}}\mathbf{x}, \qquad \mathbf{v}_{\mathbf{y}} = \mathbf{V}_{\mathbf{y}}\mathbf{y} \tag{5}$$

We use standard principal component analysis (PCA) for whitening as discussed in [1]. After this we estimate the cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}}$ of the whitened data vectors $\mathbf{v}_{\mathbf{x}}$ and $\mathbf{v}_{\mathbf{y}}$ in standard manner:

$$\widehat{\mathbf{C}}_{\mathbf{v_x v_y}} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{v}_{\mathbf{x}}(t)\mathbf{v}_{\mathbf{y}}^T(t) \tag{6}$$

There N is the smaller of the numbers N_x and N_y of the data vectors in the two data sets $\mathbf{X}$ and $\mathbf{Y}$, respectively.

We then perform the SVD of the estimated cross-covariance matrix $\widehat{\mathbf{C}}_{\mathbf{v_x v_y}}$ quite similarly as for $\mathbf{C}_{\mathbf{xy}}$ in (3). After inspecting the magnitudes of the singular values in the pseudodiagonal matrix $\boldsymbol{\Sigma}$, we divide the matrices $\mathbf{U}$ and $\mathbf{V}$ of singular vectors into two submatrices:

$$\mathbf{U} = [\mathbf{U}_1\ \mathbf{U}_2], \qquad \mathbf{V} = [\mathbf{V}_1\ \mathbf{V}_2] \tag{7}$$

There $\mathbf{U}_1$ and $\mathbf{V}_1$ correspond to dependent components for which the respective singular values are larger than 0.5, and $\mathbf{U}_2$ and $\mathbf{V}_2$ to the independent components for which the respective singular values are small. The data are then mapped using these submatrices onto subspaces corresponding to the dependent and independent components by computing

$$\mathbf{U}_1^T\mathbf{X}, \quad \mathbf{U}_2^T\mathbf{X}, \quad \mathbf{V}_1^T\mathbf{Y}, \quad \mathbf{V}_2^T\mathbf{Y} \tag{8}$$

where $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(N_x)]$ and $\mathbf{Y} = [\mathbf{y}(1), \dots, \mathbf{y}(N_y)]$. It should be noted that contrary to the customary use of SVD we include in the submatrices $\mathbf{U}_2$ and $\mathbf{V}_2$ also the singular vectors corresponding to small or even zero singular values for being able to separate all the sources in $\mathbf{X}$ and $\mathbf{Y}$. We are not aware that CCA would have used in this way in ICA and BSS previously.

Sometimes CCA alone used in this way is sufficient for coarse separation of sources, but in most cases CCA at least makes clear progress towards separation, providing signal-to-noise ratios of a few decibels. The preliminary separation results of CCA can often be improved by applying to the four mapped data sets defined in (8) some suitable ICA or BSS method. In principle at least it is possible to apply any kind of postprocessing here.

The somewhat surprising result than CCA alone can provide coarse separation can be justfied heuristically as follows. First, let us denote the separating matrices after the whitening step in (5) by $\mathbf{W}_\mathbf{x}^T$ for $\mathbf{v}_\mathbf{x}$ and respectively by $\mathbf{W}_\mathbf{y}^T$ for $\mathbf{v}_\mathbf{y}$. A basic result in the theory of ICA and BSS [1] is that after whitening the separating matrices $\mathbf{W}_\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ become orthogonal: $\mathbf{W}_\mathbf{x}^T\mathbf{W}_\mathbf{x} = \mathbf{I}$, $\mathbf{W}_\mathbf{y}^T\mathbf{W}_\mathbf{y}$ $= \mathbf{I}$. Thus

$$\widehat{\mathbf{s}} = \mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x}\mathbf{x} = \mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x}\mathbf{A}\mathbf{s} = \mathbf{s} \tag{9}$$

where we have for simplicity assumed that the estimated sources $\widehat{\mathbf{s}}$ appear in the same order as the original sources $\mathbf{s}$. Assuming that there are as many linearly independent mixtures $\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ as sources $\mathbf{s}$, so that the mixing matrix $\mathbf{A}$ is a full-rank square matrix, we get from (9) by setting $\widehat{\mathbf{s}} = \mathbf{s}$

$$\mathbf{A} = (\mathbf{W}_\mathbf{x}^T\mathbf{V}_\mathbf{x})^{-1} = \mathbf{V}_\mathbf{x}^{-1}\mathbf{W}_\mathbf{x} \tag{10}$$

due to the orthogonality of the matrix $\mathbf{W}_\mathbf{x}$. Quite similarly, we get for the another mixing matrix $\mathbf{B}$ in (2) the equivalent result $\mathbf{B} = \mathbf{V}_\mathbf{y}^{-1}\mathbf{W}_\mathbf{y}$.

Consider now the cross-covariance matrix after whitening. It is

$$\mathbf{C}_{\mathbf{v_x v_y}} = \mathrm{E}\{\mathbf{v}_\mathbf{x}\mathbf{v}_\mathbf{y}^T\} = \mathbf{V}_\mathbf{x}\mathrm{E}\{\mathbf{x}\mathbf{y}\}\mathbf{V}_\mathbf{y}^T = \mathbf{V}_\mathbf{x}\mathbf{A}\mathbf{Q}\mathbf{B}^T\mathbf{V}_\mathbf{y}^T \tag{11}$$

Here the matrix $\mathbf{Q} = \mathrm{E}\{\mathbf{s}\mathbf{r}^T\}$ is a diagonal matrix, if the sources signals in the source vectors $\mathbf{s}$ and $\mathbf{r}$ are pairwise dependent but otherwise independent of each other. Inserting $\mathbf{A} = \mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{W}_\mathbf{x}$ and $\mathbf{B} = \mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{W}_\mathbf{y}$ into (11) yields finally

$$\mathbf{C}_{\mathbf{v_x v_y}} = \mathbf{W}_\mathbf{x}\mathbf{Q}\mathbf{W}_\mathbf{y}^T \tag{12}$$

But this is exactly the same type of expansion as the SVD of the whitened cross-covariance matrix $\mathbf{C}_{\mathbf{v_x v_y}}$ in (3), because the matrices $\mathbf{W}_\mathbf{x}$ and $\mathbf{W}_\mathbf{y}$ are orthogonal matrices and $\mathbf{Q}$ is a diagonal matrix. Thus on the assumptions made above the SVD of the whitened cross-covariance matrix provides a solution that has the same structure as the separating solution. Even though we cannot from this result directly deduce that the SVD of the whitened cross-covariance matrix (that is, CCA) would provide a separating solution, this seems to hold in simple cases at least as shown by our experiments in the next section.

Another justification is that CCA, or SVD of whitened data vectors, uses second-order statistics (cross-covariances) only for separation, while standard

ICA algorithms such as FastICA use for separation higher-order statistics only after the data has been normalized with respect to their second-order statistics by whitening them. Our method combines both types of statistics. Our experimental results demonstrate that this often provide better results than using solely second-order or higher-statistics for separation. Dividing the separation problem into subproblems using the matrices in (8) may also help. Probably solving two lower dimensional subproblems is easier than solving a higher dimensional separation problem.

3 Experimental Results

We have successfully tested our method with synthetical data sets, with data sets in which real-world sources have been mixed synthetically, and with real-world robot and fMRI (functional magnetic resonance imaging) data. Due to space limitations, we can show some quite selected results only here. More experimental results can be found in [16].

Consider first a set of 6 synthetical stochastic sources which have been purposedly designed so that they are very difficult to separate for most ICA and BSS methods. They are defined in the Matlab code [6] of the UniBSS method and explained in the respective paper [5]. Standard ICA methods based on non-Gaussianity should be able to separate only the two first sources. Methods based on temporal statistics should not able to separate any of them. Method utilizing smoothly changing variances are able to separate only the fifth and sixth source. Only the approximative UniBSS method [5] which utilizes all these properties is able to separate all these 6 sources.

We mixed the first three sources and the fifth one to form the first data set **X**, and the second, third, fourth and sixth source to the second data set **Y**. Thus in these data sets there are two completely dependent sources, while the remaining two sources in them are statistically independent of all the other sources.

Table 1. Signal-to-noise ratios (dB) of different methods for the source signals 1-4 in the first data set **X**

Method	Source 1	Source 2	Source 3	Source 4
CCA	10.3	9.9	10.1	10.3
FastICA	22.5	14.1	9.4	10.6
TDSEP	10.0	30.5	10.0	27.5
UniBSS	33.9	40.7	27.6	28.5
CCA + FastICA	29.3	20.0	21.0	29.4
CCA + TDSEP	30.7	37.9	34.8	30.2
CCA + UniBSS	33.7	48.4	39.2	32.7
Method in [9]	25.7	9.8	9.4	23.1
Method in [13]	12.5	11.4	11.3	13.2

Table 2. Signal-to-noise ratios (dB) of different methods for the source signals 5-8 in the second data set **Y**

Method	Source 5	Source 6	Source 7	Source 8
CCA	9.9	10.1	10.5	10.5
FastICA	9.5	4.6	4.2	5.2
TDSEP	9.7	26.4	9.8	28.8
UniBSS	37.1	27.0	28.6	29.0
CCA + FastICA	21.1	21.9	13.1	13.2
CCA + TDSEP	37.9	34.8	31.6	33.1
CCA + UniBSS	49.4	39.2	31.0	33.0
Method in [9]	9.8	9.4	9.5	9.5
Method in [13]	11.4	11.3	3.6	3.9

We used 5000 data vectors and source signal values ($t = 1, 2, \ldots, 5000$) for providing enough data to the UniBSS method [5]. The other tested methods, CCA, FastICA, TDSEP and their combinations require less samples, especially CCA. We computed the average signal-to-noise ratios of the estimated sources over 100 random realizations of the sources and the data sets **X** and **Y** because the results vary for single realizations. In each realization, the elements of the 4 × 4 mixing matrices were Gaussian random numbers.

We not only tried our CCA based method and its combinations applying either FastICA, TDSEP, or UniBSS for post-processing to achieve better separation, but also compared it with two methods introduced by other authors for the same problem. The first compared method introduced in [9] assumes that the dependent sources in the two data sets are active simultaneously. The second compared method [13] uses multiset canonical correlation analysis. Theoretically its results should coincide with plain CCA for two data sets but in practice this may not hold due to problems such as deflationary nature of the algorithm mentioned in a later paper [12].

The separation results for the four sources 1-4 contained in the first data set **X** are shown in Table 1, and for the 4 sources in the other data set **Y** in Table 2. For clarity, we have numbered these sources from 5 to 8. We set (somewhat arbitrarily) the threshold of successful separation to 10 dB based on visual inspection. Tables 1 and 2 show that CCA alone yields fairly similar separation results for all the 8 sources which already lie at our separation threshold. FastICA can separate clearly the two first sources but fails for the three last sources. The TDSEP method separates well four sources, the other sources lie at the separation threshold. The UniBSS method separates well all the sources. The results are qualitatively similar if the dependent and independent sources are selected otherwise from the 6 original sources.

Combining CCA with post-processing with FastICA, TDSEP, or UniBSS methods improves the results for all these methods, so that also FastICA and TDSEP can now separate well all the sources in this difficult separation problem.

The methods introduced in [9] and [13] provide clearly lower signal-to-noise ratios, failing for some sources. Using CCA combined with FastICA or TDESP methods is in practice often preferable over using the UniBSS method. The UniBSS method requires much more samples for reliable results. It may already converge to a separating solution but then deviates again farther away, and this can happen several times. The UniBSS method also requires different types of nonlinearities for sub-Gaussian and super-Gaussian sources. The FastICA and TDSEP methods don't suffer from this limitation.

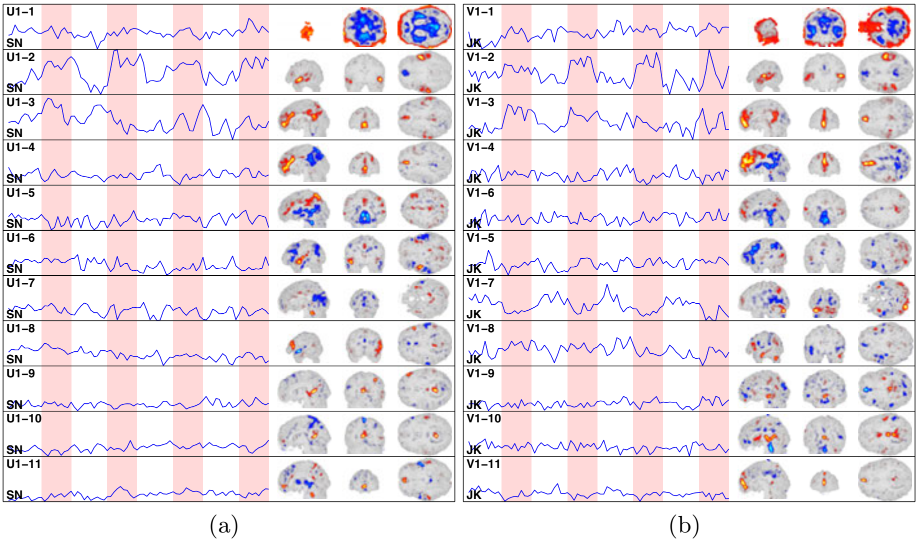


Fig. 1. Experimental results with fMRI data. Each row shows one of the 11 separated components. The activation time-course with the stimulation blocks for reference, shown on the left, and the corresponding spatial pattern on three coincident slices, on the right. Components from (a) the first and (b) the second dataset.

The usefulness of the method was tested with data from a functional magnetic resonance imaging (fMRI) study [10], where it is described in more detail. We used the measurements of two healthy adults while they were listening to spoken safety instructions in 30 s intervals, interleaved with 30 s resting periods. In these experiments we used slow feature analysis (SFA) [15] for post-processing the results given by CCA, because it gave better results than FastICA.

Fig. 1 shows the results of applying our method to the two datasets and separating 11 components from the dependent subspaces **U1** and **V1**. The consistency of the components across the subjects is quite good. The first component shows a global hemodynamic contrast, that may also be related to artifacts originating from smoothing the data in the standard preprocessing. The activity of the second component is focused on the primary auditory cortices. The third and fourth components show both positively and negatively task-related activity around the anterior and posterior cingulate gyrus. These first results are promising and in good agreement with the the ones reported in [10]. Future tasks are extension of the method to multiple datasets for interpreting the found components more thoroughly, and a more extensive comparison with existing ICA and BSS methods using real-world data.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
2. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press (2010)
3. Hyvärinen, A., et al.: The FastICA package for Matlab. Helsinki Univ. of Technology, Espoo (2005), http://www.cis.hut.fi/projects/ica/fastica/
4. Ziehe, A., Müller, K.-R.: TDSEP - an efficient algorithm for blind source separation using time structure. In: Proc. of the Int. Conf. on Artificial Neural Networks (ICANN 1998), Skövde, Sweden (1998)
5. Hyvärinen, A.: A unifying model for blind separation of independent sources. Signal Processing 85(7), 1419–1427 (2005)
6. Hyvärinen, A.: Basic Matlab code for the unifying model for BSS. Univ. of Helsinki, Finland (2003–2006), http://www.cs.helsinki.fi/u/ahyvarin/code/UniBSS.m
7. Akaho, S., Kiuchi, Y., Umeyama, S.: MICA: Multidimensional independent component analysis. In: Proc. of the 1999 Int. Joint Conf. on Neural Networks (IJCNN 1999), pp. 927–932. IEEE Press, Washington, DC (1999)
8. Van Hulle, M.: Constrained subspace ICA based on mutual information optimization directly. Neural Computation 20(4), 964–973 (2008)
9. Gutmann, M.U., Hyvärinen, A.: Extracting Coactivated Features from Multiple Data Sets. In: Honkela, T. (ed.) ICANN 2011, Part I. LNCS, vol. 6791, pp. 323–330. Springer, Heidelberg (2011)
10. Ylipaavalniemi, J., et al.: Analyzing consistency of independent components: An fMRI illustration. NeuroImage 39, 169–180 (2008)
11. Ylipaavalniemi, J., et al.: Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. NeuroImage 48, 176–185 (2009)
12. Anderson, M., Li, X.-L., Adalı, T.: Nonorthogonal Independent Vector Analysis Using Multivariate Gaussian Model. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 354–361. Springer, Heidelberg (2010)
13. Li, Y.-Q., et al.: Joint blind source separation by multiset canonical correlation analysis. IEEE Trans. on Signal Processing 57(10), 3918–3928 (2009)
14. Rencher, A.: Methods of Multivariate Analysis, 2nd edn. Wiley (2002)
15. Wiskott, L., Sejnowski, T.: Slow feature analysis: unsupervised learning of invariances. Neural Computation 14, 715–770 (2002)
16. Karhunen, J., Hao, T.: Finding dependent and independent components from two related data sets. In: Proc. of 2011 Int. J. Conf. on Neural Networks (IJCNN 2011), San Jose, CA, USA, pp. 457–466 (July-August 2011)

A Probability-Based Combination Method for Unsupervised Clustering with Application to Blind Source Separation

Julian Mathias Becker, Martin Spiertz, and Volker Gnann

Institut für Nachrichtentechnik, RWTH Aachen University,
52056 Aachen, Germany
{becker,gnann}@ient.rwth-aachen.de
http://www.ient.rwth-aachen.de

Abstract. Unsupervised clustering algorithms can be combined to improve the robustness and the quality of the results, e.g. in blind source separation. Before combining the results of these clustering methods the corresponding clusters have to be aligned, but usually it is not known which clusters of the employed methods correspond to each other. In this paper, we present a method to avoid this correspondence problem using probability theory. We also present an application of our method in blind source separation. Our approach is better expandable than other state-of-the-art separation algorithms while leading to slightly better results.

1 Introduction

The idea of combining the results of multiple clustering methods has been presented in [1],[2],[3]. For clustering of data, a number of approaches may be applied, usually leading to different results. The intention of combining multiple clustering approaches is to improve the results by using the strengths of all the methods. Unfortunately, in blind clustering methods, the correspondences of the clusters are unknown. When combining methods this correspondence problem has to be solved, see [3]. Fred and Jain propose to use a measure of similarity between patterns [2] to circumvent this problem. We propose a similiar method, extending the approach by using probabilities, which opens another way of clustering the combined results.

An application where multiple clustering methods can be used is blind source separation (BSS). BSS tries to separate the original sources out of a mixed audio signal and can be used as a preprocessing step for many audio processing tasks such as remixing, instrument recognition, or automatic music transcription. Many state-of-the-art algorithms use non-negative tensor factorization (NTF) or non-negative matrix factorization (NMF) to factorize single notes out of the mixture. While [4] and [5] propose extensions to the NTF to factorize complete melodies, [6] presents an approach, where the single notes are being clustered to melodies. In [7], different clustering methods are proposed, one using spectral and the other one temporal features.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 99–106, 2012.

In this paper we propose an approach for combining multiple clustering methods, which is presented in Section 2. In Section 3 we use our algorithm to combine the different clustering methods for BSS as proposed in [7] and analyze the performance in comparison to other approaches. Finally in Section 5, a conclusion is given.

2 The Proposed Combination Algorithm

In this section we present the proposed combination strategy for unsupervised clustering methods. We assume a testset of I data items that have to be grouped into C clusters. In the following items will be named i_m with $1 \leq m \leq I$ and clusters c with $1 \leq c \leq C$. We furthermore assume a number of V different clustering methods, which all return probabilities ${}^{v}p_c(i_m)$ that item i_m belongs to cluster c with $1 \leq v \leq V$. Thus, every method returns a matrix of size $I \times C$. Combining these matrices is not possible because it is not known which clusters of the different methods correspond to each other. So, before combining the matrices it is first necessary to estimate the correspondences of the clusters and to align the columns. This step may induce errors if the correspondences are not estimated correctly. This issue motivates our proposed algorithm.

2.1 The Basic Idea

Instead of evaluating the probabilitiy $p_c(i_m)$ that item i_m belongs to cluster c, we propose to calculate the probability $p(i_m, i_n)$ that the items i_m and i_n belong to the same cluster. This means, for every clustering method v we calculate a matrix

$$\mathbf{Q}_v = \begin{pmatrix} 1 & {}^{v}p(i_1, i_2) & \cdots & {}^{v}p(i_1, i_I) \\ {}^{v}p(i_2, i_1) & 1 & \cdots & {}^{v}p(i_2, i_I) \\ \vdots & \vdots & \ddots & \vdots \\ {}^{v}p(i_I, i_1) & {}^{v}p(i_I, i_2) & \cdots & 1 \end{pmatrix} \tag{1}$$

where the entries ${}^{v}p(i_m, i_n)$ are calculated as

$${}^{v}p(i_m, i_n) = \sum_{k=1}^{C} {}^{v}p_k(i_m) \cdot {}^{v}p_k(i_n) \tag{2}$$

for each clustering method v, leading to V matrices $\mathbf{Q}_v$ of size $I \times I$.
These matrices $\mathbf{Q}_v$ can now be combined without having to be aligned. One possibility to combine the matrices is taking the mean values of the entries ${}^{v}p(i_m, i_n)$ over all v. This leads to a matrix $\mathbf{Q}_{\text{av}}$ with the average probabilities $p_{\text{av}}(i_m, i_n)$

$$\mathbf{Q}_{\text{av}} = \begin{pmatrix} 1 & p_{\text{av}}(i_1, i_2) & \cdots & p_{\text{av}}(i_1, i_I) \\ p_{\text{av}}(i_2, i_1) & 1 & \cdots & p_{\text{av}}(i_2, i_I) \\ \vdots & \vdots & \ddots & \vdots \\ p_{\text{av}}(i_I, i_1) & p_{\text{av}}(i_I, i_2) & \cdots & 1 \end{pmatrix} \tag{3}$$

where the entries $p_{\text{av}}(i_m, i_n)$ are calculated as

$$p_{\text{av}}(i_m, i_n) = \frac{\sum_{v=1}^{V} {}^{v}p(i_m, i_n)}{V}. \tag{4}$$

The indices m and n corresponding to the maximum value in $\mathbf{Q}_{\text{av}}$ denote the items that are most probable to belong to the same cluster.

Other combinations of the matrices are possible. For example instead of taking the mean value, the different methods could be weighted, for example depending on how good they perform. A weighted combination will be used in Section 3.2.

2.2 The Clustering Algorithm

The matrix $\mathbf{Q}_{\text{av}}$ can now be used for clustering. In our clustering algorithm, items will be iteratively grouped together. These groups of items will be named ${}^{q}Z_r$ where q is the current iteration step and $1 \leq r \leq R$ indicates all existing groups. Every group contains at least one item.

Groups can also be interpreted as events in a probability meaning. Every group represents the event, that the items in this group belong to the same cluster.

We define the following notations:

Term	Meaning
$p({}^{q}Z_r)$	Probability that all items that are grouped together in group ${}^{q}Z_r$ belong to the same cluster
$p(\{{}^{q}Z_r, {}^{q}Z_s\})$	Probability that all items that are grouped together in the groups ${}^{q}Z_r$ and ${}^{q}Z_s$ belong to the same cluster
$p({}^{q}Z_r \cap {}^{q}Z_s)$	Probability that the events ${}^{q}Z_r$ and ${}^{q}Z_s$ both occur
${}^{q}Z$	Unites all events ${}^{q}Z_1, {}^{q}Z_2, \ldots, {}^{q}Z_R$. This means the notation $p({}^{q}Z)$ describes the same probability as $p({}^{q}Z_1 \cap {}^{q}Z_2 \cap \ldots \cap {}^{q}Z_R)$
${}^{q+1}Z_r = \{{}^{q}Z_s, {}^{q}Z_t\}$	Indicates, that in the iteration step $q+1$, all items that were grouped together in ${}^{q}Z_s$ and ${}^{q}Z_t$ are merged in group ${}^{q+1}Z_r$

For our clustering algorithm we will need to calculate a matrix similiar to the matrix $\mathbf{Q}_{\text{av}}$ in Eq. (3) in every iteration step. This matrix is denoted ${}^{q}\tilde{\mathbf{Q}}_{\text{av}}$ and will be called probability matrix in the following. The entries at index m, n are now defined as

$$p_{\text{av}}(\{{}^{q}Z_m, {}^{q}Z_n\}|{}^{q}Z) = \frac{\sum_{v=1}^{V} {}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\}|{}^{q}Z)}{V}. \tag{5}$$

We assume that the events ${}^{q}Z_r$ and ${}^{q}Z_s$ are independent, if $r \neq s$. In this case the probabilitiy $p({}^{q}Z_r \cap {}^{q}Z_s)$ reduces to

$$p({}^{q}Z_r \cap {}^{q}Z_s) = p({}^{q}Z_r) \cdot p({}^{q}Z_s). \tag{6}$$

Considering the fact that ${}^{q}Z_s$ is a subset of $\{{}^{q}Z_r, {}^{q}Z_s\}$ it is obvious that

$$p(\{{}^{q}Z_r, {}^{q}Z_s\} \cap {}^{q}Z_s) = p(\{{}^{q}Z_r, {}^{q}Z_s\}). \tag{7}$$

Using the definition of conditional probability, the probability ${}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\}|{}^{q}Z)$ in Eq. (5) can be written as

$$ {}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\}|{}^{q}Z) = \frac{{}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\} \cap {}^{q}Z)}{{}^{v}p({}^{q}Z)}. \tag{8} $$

With Equations 6 and 7 this term reduces to

$$ {}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\}|{}^{q}Z) = \frac{{}^{v}p(\{{}^{q}Z_m, {}^{q}Z_n\})}{{}^{v}p({}^{q}Z_m) \cdot {}^{v}p({}^{q}Z_n)}. \tag{9} $$

This group-based definition of ${}^{q}\tilde{\mathbf{Q}}_{\text{av}}$ allows us to group items together iteratively. For the special case that every group ${}^{q}Z_r$ contains exactly one item, the matrix ${}^{q}\tilde{\mathbf{Q}}_{\text{av}}$ is identical to $\mathbf{Q}_{\text{av}}$ in Eq. (3).

We suggest the following iterative algorithm for combined clustering:

```
Initialize ${}^{0}Z_r = \{i_r\} \, \forall \, r = 1, 2, \ldots, I$
$q = 0$, $q_{\max} = I - C$
**while** $q < q_{\max}$ **do**
    Calculate ${}^{q}\mathbf{Q}_{\text{av}}$ (Eq. (3) and (5))
    $m, n = \underset{\tilde{m},\tilde{n}}{\operatorname{argmax}} \, {}^{q}\mathbf{Q}_{\text{av}}(\tilde{m}, \tilde{n}), \ \tilde{m} < \tilde{n}$
    ${}^{q+1}Z_r = \begin{cases} {}^{q}Z_r & \text{for } r < m \\ \{{}^{q}Z_m, {}^{q}Z_n\} & \text{for } r = m \\ {}^{q}Z_r & \text{for } m < r < n \\ {}^{q}Z_{r+1} & \text{for } r \geq n \end{cases}$
    $q = q + 1$
**end while**
Return ${}^{q_{\max}}Z_r$
```

The C remaining groups ${}^{q_{\max}}Z_r$ are the result of the clustering. Every group can be interpreted as one cluster. All items that belong to this group are assigned to this cluster.

3 Application to Blind Source Separation

In the following the algorithm is applied to BSS. In [7], an approach for BSS was presented which uses two clustering methods. The methods use spectral and temporal information, respectively. The combination of both methods was done by hard-decision. However, it seems reasonable to assume, that even if one of the methods is more reliable, the other method still contains information that could improve the clustering. Hence a soft-decision combination could improve the results.

3.1 Hard-Decision Approach

More detailed information about the hard-decision approach can be found in [7] and [6].

In the following we assume $x(n)$ to be an additive mixture of M monaural sources $s_m(n)$, with n being the time index. In the following we present the signal flow of the algorithm.

- First, the short-time Fourier transform (STFT) of $x(n)$ is taken. The resulting complex-valued spectrogram $\underline{\mathbf{X}}$ is of size $K \times T$ with frequency-bins $1 \leq k \leq K$ and time-bins $1 \leq t \leq T$. In the following only the absolute values, $\mathbf{X} = |\underline{\mathbf{X}}|$, are used.
- In the next step, $\mathbf{X}$ is factorized by the NMF. This results in a separation of I sound events. The NMF outputs the two matrices $\mathbf{B}$ of size $K \times I$ and $\mathbf{G}$ of size $T \times I$. These matrices approximate $\mathbf{X}$ by

$$\mathbf{X}(k,t) \approx \sum_{j=1}^{I} \mathbf{B}(k,j)\mathbf{G}(t,j). \tag{10}$$

  The j-th column of $\mathbf{B}$ corresponds to the spectrum and the j-th column of $\mathbf{G}$ represents the temporal envelope of sound event σ_j. For multichannel signals, the NTF can be used instead of the NMF.
- Signal synthesis is done as described in [6]. The spectrogram $\underline{\mathbf{Y}}_j(k,t)$ corresponding to the estimated time domain signal $y_j(n)$ of sound event σ_j is calculated as:

$$\underline{\mathbf{Y}}_j(k,t) = \underline{\mathbf{X}}(k,t) \cdot \frac{\mathbf{B}(k,j)\mathbf{G}(t,j)}{\sum_{z=1}^{I} \mathbf{B}(k,z)\mathbf{G}(t,z)}. \tag{11}$$

  The output signals $y_j(n)$ are estimated by applying the inverse STFT to $\underline{\mathbf{Y}}_j$.
- The I sound events are clustered into M clusters. A vector $\mathbf{a}$ with I elements is defined, with $1 \leq \mathbf{a}(j) \leq M$, $\mathbf{a}(j) \in \mathbb{N}$. The entries $\mathbf{a}(j)$ of this vector specify the cluster, to which cluster the sound event σ_j is assigned. Clustering is done using the NMF as proposed in [6].
  Features $\mathbf{F_B}$ and $\mathbf{F_G}$ are calculated from the matrices $\mathbf{B}$ and $\mathbf{G}$ using the source-filter model theory for frequency and time domain [7]. The features are independently factorized by an NMF, which gives an approximation

$$\mathbf{F}_{\{\mathbf{B}|\mathbf{G}\}}(k,j) \approx \sum_{m=1}^{M} \mathbf{W}_{\{\mathbf{B}|\mathbf{G}\}}(k,m)\mathbf{V}_{\{\mathbf{B}|\mathbf{G}\}}(j,m). \tag{12}$$

  The index $\{\mathbf{B}|\mathbf{G}\}$ denotes, that either the matrices calculated from $\mathbf{B}$ or from $\mathbf{G}$ are used. While the m-th column of $\mathbf{W}$ corresponds to the m-th cluster center, the m-th column of $\mathbf{V}$ corresponds to the m-th connectivity values. Therefore the clustering vector $\mathbf{a}$ is defined as

$$\mathbf{a}(j) = \underset{m}{\operatorname{argmax}} \mathbf{V}(j,m). \tag{13}$$

  Hence, we get one vector $\mathbf{a_B}(j)$ from the spectral clustering and one vector $\mathbf{a_G}(j)$ from the temporal clustering.

- The decision, which clustering vector to be used is based on the *number of note instances* μ_m of the mixture. This value μ_m is calculated for each column of $\mathbf{G}$ by subtracting the mean value of the corresponding column and counting the zero crossings from negative to positive values. The final value μ_{av} is estimated as the mean value over all μ_m. The clustering vector $\mathbf{a}_{final}(j)$ that is applied for the final clustering is $\mathbf{a}_{\mathbf{B}}(j)$ if $\mu_{\text{av}} \leq \vartheta$, or else $\mathbf{a}_{\mathbf{G}}(j)$, with ϑ being a predefined threshold. In [7], a value of $\vartheta = 1.6$ is proposed.

3.2 Extension to Soft-Decision

Instead of using a hard-decision we propose a soft-decision combination of the clustering methods, using the algorithm presented in Section 2. The correspondences between the given problem and the proposed algorithm are as follows:

Clusters. The clusters for the algorithm are the different estimated sources of the mixture signal. The number of clusters C in the algorithm is therefore M.

Items. The items i_m of the algorithm are the I separated sound events σ_j.

Methods. The clustering methods that have to be combined are the spectral and the temporal clustering methods.

Probabilities. Our combination algorithm requires probabilities instead of hard mappings. We define

$$p_m(\sigma_j) = \frac{\mathbf{V}(j,m)}{\sum_{k=1}^{M} \mathbf{V}(j,k)}. \tag{14}$$

This definition leads to a matrix of probabilities of size $I \times M$ for every clustering method, which can be used as input for the proposed algorithm.

Weightings. Instead of the hard-decision a soft-decision is made by weighting the probability matrices of the different methods before combining them. The probability matrix corresponding to the spectral clustering is weigthed with $w_{\mathbf{B}}$ with

$$w_{\mathbf{B}} = \begin{cases} 1 & \text{if } \mu_{\text{av}} \leq b_l \\ \frac{b_u - n_{\text{av}}}{b_u - b_l} & \text{if } b_l < \mu_{\text{av}} \leq b_u \\ 0 & \text{if } \mu_{\text{av}} > b_u \end{cases}, \tag{15}$$

where b_l and b_u denote a lower and an upper bound. These parameters have to be determined by experiment. The probability matrix corresponding to the temporal clustering is weigthed with $1 - w_{\mathbf{B}}$. For the special case $b_l = b_u$ this transforms to the hard-decision criterion with threshold b_l.

4 Experimental Results

We compare our soft-decision approach with the hard-decision approach in [7]. We also compare our results with the results of [4].

For performance measurement we use the measures SDR, SIR and SAR as proposed in [8]. The test set 1 that we use for comparison with [7] is a set of 1770 two-source mixtures, mixed from 60 monaural recordings, which is identical to test set $\mathcal{A}$ in [7]. Test set 2 are the 25 mixtures used in [4]. This dataset mainly contains very harmonic mixtures.

For fair comparison, we use exactly the same parameters for our algorithm as are used in [7]. In [7] a value of $\vartheta = 1.6$ is used as threshold for the hard-decision. Therefore we chose the values for the upper and the lower bound of the weights for the soft-decision symmetrical around this value. Experiments show, that values of $b_l = 0.8$ and $b_u = 2.4$ lead to good results.

The mean values over SDR, SIR and SAR for test set 1 are shown in Table 1. It can be observed that the soft-decision approach performs slightly better than the approach of [7] for all of the three measures.

The mean values over SDR, SIR and SAR for test set 2 are shown in Table 2. We compare our results with the results of the hard-decision approach [7] and with the results of [4]. Compared to the algorithm in [4], our algorithm leads to lower distortion by artifacts (SAR) but to higher distortion by interferences (SIR). Compared to the hard-decision approach [7] our algorithm leads to slightly better results for all of the three measures. In [7] it is shown that for test set 2, spectral clustering leads to much better results than temporal clustering, which can be explained by the high harmonicity of the sources. However, our results show, that even for such harmonic mixtures the results can be improved by also using temporal information.

Table 1. Results for SDR, SIR and SAR in dB for test set 1

Test set 1	SDR	SIR	SAR
hard-decision [7]	7.20	12.92	13.62
soft-decision	7.23	13.07	13.83

Table 2. Results for SDR, SIR and SAR in dB for test set 2

Test set 2	SDR	SIR	SAR
[4]	9.01	24.91	9.52
hard-decision [7]	9.80	15.05	15.91
soft-decision	9.91	15.21	16.27

It should be noted that besides the slightly better results compared to the hard-decision approach, presented in Table 1 and Table 2, the proposed combination algorithm holds the advantage of beeing easily expandable by appending other matrices of clustering results to the input. It can be assumed that by including more methods, the results could be improved further. This possibility of using more clustering methods is not given in the hard-decision approach.

5 Conclusion

In this paper we present a new way of combining different clustering methods based on probability theory. We calculate the probabilities that different items belong to the same cluster, which makes it possible to combine different methods without having to solve the correspondence problem. We introduce a method of clustering the combined values by iteratively grouping together items that most probably belong to the same cluster.

We use the presented approach to extend the BSS-algorithm proposed in [7] by using a soft-decision combination. We show that this extension leads to slightly better separation results. Furthermore, our approach has the advantage of being easily expandable, using more clustering methods.

References

1. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. 3, 583–617 (2003)
2. Boulis, C., Ostendorf, M.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005)
3. Boulis, C., Ostendorf, M.: Combining Multiple Clustering Systems. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 63–74. Springer, Heidelberg (2004)
4. FitzGerald, D., Cranitch, M., Coyle, E.: Extended nonnegative tensor factorisation models for musical sound source separation. In: Computational Intelligence and Neuroscience (2008)
5. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 550–563 (2010), http://www.irisa.fr/metiss/ozerov/demos.html
6. Spiertz, M., Gnann, V.: Source-filter based clustering for monaural blind source separation. In: Proc. of International Conference on Digital Audio Effects DAFx, Como, Italy (2009)
7. Spiertz, M., Gnann, V.: Note clustering based on 2d source-filter modeling for underdetermined blind source separation. In: Proceedings of the AES 42nd International Conference on Semantic Audio, Ilmenau, Germany (July 2011)
8. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (2006)

Charrelation Matrix Based ICA*

Alon Slapak and Arie Yeredor

School of Electrical Engineering, Tel-Aviv University

Abstract. Charrelation matrices are a generalization of the covariance matrix, encompassing statistical information beyond second order while maintaining a convenient 2-dimensional structure. In the context of ICA, charrelation matrices-based separation was recently shown to potentially attain superior performance over commonly used methods. However, this approach is strongly dependent on proper selection of the parameters (termed *processing-points*) which parameterize the charrelation matrices. In this work we derive a data-driven criterion for proper selection of the set of processing-points. The proposed criterion uses the available mixtures samples to quantify the resulting separation errors' covariance matrix in terms of the processing points. Minimizing the trace of this matrix with respect to the processing points enables to optimize (asymptotically) the selection of these points, thereby yielding better separation results than other methods, as we demonstrate in simulation.

1 Introduction

In the framework of instantaneous ICA, consider the model $\boldsymbol{x}[n] = \mathbf{A}\boldsymbol{s}[n]$, $1 \leq n \leq N$, where $\boldsymbol{x}[n] \in \mathbb{R}^K$ is a multivariate observations signal acquired from K sensors, $\boldsymbol{s}[n] \in \mathbb{R}^K$ is a multivariate source signal originates from mutual statistical independent sources, and $\mathbf{A} \in \mathbb{R}^{K\times K}$ is an unknown invertible mixing matrix. The goal is to obtain an estimate $\widehat{\mathbf{B}}$ of the demixing matrix, $\mathbf{B} \overset{\triangle}{=} \mathbf{A}^{-1}$, which in turn provides an estimate of the sources via $\hat{\boldsymbol{s}}[n] = \widehat{\mathbf{B}}\boldsymbol{x}[n]$. This work addresses weighted approximate joint diagonalization (AJD) of matrix-form statistics (frequently termed *target matrices*) having the appealing property of being strictly diagonal for random vectors with independent components, such as in JADE [1], SOBI [2], BGL [3] and (more recently) WITCHESS [4]. The target-matrices in this work are the sample-charrelation matrices (see Sect. 2). While the potential advantages of charrelation matrix-based ICA were demonstrated in WITCHESS [4], no method for proper (let alone optimal) selection of the processing-points, at which the charrelation matrices are evaluated, has been proposed or developed. The objective of this work is to provide a method for their selection, which would be completely data-driven and would not require prior statistical information on the sources or on the mixing matrix.

We begin with a brief overview of charrelation matrices and their relevant properties in the next section, followed by derivation of the algorithm and simulation results in the following sections.

* This research was supported by the Israeli Science Foundation (grant No. 1255/08).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 107–114, 2012.

2 Charrelation Matrices

In this section, the definition of the *charrelation*[1] *matrix* is introduced. The derivation follows the traditional methodology used in the literature for the covariance matrix.

Definition 1. ***(charmean)***
Given a random vector $\boldsymbol{x} \in \mathbb{R}^K$, and a function $\boldsymbol{g}(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^L$, the charmean *of $\boldsymbol{g}(\boldsymbol{x}) \in \mathbb{R}^L$ with respect to (w.r.t.) $\boldsymbol{x}$ at an arbitrary* processing-point $\boldsymbol{\tau} \in \mathbb{R}^K$ *is defined as:*

$$\boldsymbol{\eta}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x});\boldsymbol{\tau}\right] \triangleq \frac{E\left[\boldsymbol{g}\left(\boldsymbol{x}\right)\exp\{\boldsymbol{x}^T\boldsymbol{\tau}\}\right]}{E\left[\exp\{\boldsymbol{x}^T\boldsymbol{\tau}\}\right]} \in \mathbb{R}^L$$

whenever both means (taken w.r.t. $\boldsymbol{x}$) exist.

The charmean shares many properties with the conventional expectation operator (e.g., linearity in $g(\boldsymbol{x})$, separability in the case of statistical independence), and for $\boldsymbol{\tau} = \mathbf{0}$ both operators coincide.

Definition 2. ***(cross-charrelation and charrelation matrices)***
Given a random vector $\boldsymbol{x} \in \mathbb{R}^K$ and functions $\boldsymbol{g}_1(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^{L_1}$ and $\boldsymbol{g}_2(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^{L_2}$, the cross-charrelation matrix *between $\boldsymbol{g}_1(\boldsymbol{x})$ and $\boldsymbol{g}_2(\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$ at an arbitrary processing-point $\boldsymbol{\tau} \in \mathbb{R}^K$ is defined as:*

$$\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{g}_1(\boldsymbol{x}),\boldsymbol{g}_2(\boldsymbol{x});\boldsymbol{\tau}\right] \triangleq \boldsymbol{\eta}_{\boldsymbol{x}}\left[\boldsymbol{g}_1(\boldsymbol{x})\boldsymbol{g}_2^T(\boldsymbol{x});\boldsymbol{\tau}\right] - \boldsymbol{\eta}_{\boldsymbol{x}}\left[\boldsymbol{g}_1(\boldsymbol{x});\boldsymbol{\tau}\right]\boldsymbol{\eta}_{\boldsymbol{x}}^T\left[\boldsymbol{g}_2(\boldsymbol{x});\boldsymbol{\tau}\right] \in \mathbb{R}^{L_1 \times L_2}$$

whenever all the charmeans involved exist. Similarly, for $\boldsymbol{g}(\cdot) : \mathbb{R}^K \rightarrow \mathbb{R}^L$, the charrelation matrix *of $g(\boldsymbol{x})$ (w.r.t. $\boldsymbol{x}$, at $\boldsymbol{\tau}$) is simply defined as the cross-charrelation between $\boldsymbol{g}(\boldsymbol{x})$ and itself, namely $\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x});\boldsymbol{\tau}\right] \overset{\triangle}{=} \boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x}),\boldsymbol{g}(\boldsymbol{x});\boldsymbol{\tau}\right]$.*

The charrelation matrix is a symmetric, positive semi-definite matrix, sharing many properties with the conventional covariance matrix. Both the cross-charrelation and charrelation matrices coincide with the cross-covariance and covariance matrices (resp.) for $\boldsymbol{\tau} = \mathbf{0}$. For $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$, the charrelation matrix coincides with the Hessian (at $\boldsymbol{\tau}$) of the second generalized characteristic function of $\boldsymbol{x}$, namely $\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{x};\boldsymbol{\tau}\right] = \frac{\partial^2}{\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T}\log E\left[\exp\{\boldsymbol{x}^T\boldsymbol{\tau}\}\right]$.

The following additional properties are relatively straightforward to derive, and would be useful in our subsequent derivations:

Properties 1. ***(Charrelation matrix)***

1. ***Linear transformations:*** *If $\mathbf{C} \in \mathbb{R}^{L \times K}$ and $\boldsymbol{c} \in \mathbb{R}^L$ are some constant matrix and vector, and $\boldsymbol{y} = \mathbf{C}\boldsymbol{x} + \boldsymbol{c} \in \mathbb{R}^L$, then*

$$\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{y};\boldsymbol{\tau}\right] = \mathbf{C}\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{x};\boldsymbol{\tau}\right]\mathbf{C}^T \in \mathbb{R}^{L \times L}$$
$$\boldsymbol{\Psi}_{\boldsymbol{y}}\left[\boldsymbol{y};\boldsymbol{\tau}\right] = \mathbf{C}\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{x};\mathbf{C}^T\boldsymbol{\tau}\right]\mathbf{C}^T \in \mathbb{R}^{L \times L}.$$

[1] Pronounced “car-relation”, reflecting the relation to the characteristic function.

2. ***Independence:*** *If $\boldsymbol{x} \in \mathbb{R}^K$ can be partitioned into two statistically independent groups $\boldsymbol{x}_1 \in \mathbb{R}^{K_1}, \boldsymbol{x}_2 \in \mathbb{R}^{K_2}$ with $K_1 + K_2 = K$, then $\boldsymbol{\Psi}_{\boldsymbol{x}}[\boldsymbol{x}; \boldsymbol{\tau}] \in \mathbb{R}^{K \times K}$ is block-diagonal (with the respective partition) for all $\boldsymbol{\tau} \in \mathbb{R}^K$ at which it exists.*

Convenient estimates of the charmean and the charrelation matrix are obtained from the *sample-charmean* and the *sample-charrelation* (resp.):

$$\widehat{\boldsymbol{\eta}}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\tau}\right] = \frac{\sum\limits_{n=1}^{N} \boldsymbol{g}\left(\boldsymbol{x}\left[n\right]\right) \exp\{\boldsymbol{x}^T\left[n\right] \boldsymbol{\tau}\}}{\sum\limits_{n=1}^{N} \exp\{\boldsymbol{x}^T\left[n\right] \boldsymbol{\tau}\}} \in \mathbb{R}^L \tag{1}$$

$$\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\tau}\right] = \widehat{\boldsymbol{\eta}}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x})\boldsymbol{g}^T(\boldsymbol{x}); \boldsymbol{\tau}\right] - \widehat{\boldsymbol{\eta}}_{\boldsymbol{x}}\left[\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\tau}\right] \widehat{\boldsymbol{\eta}}_{\boldsymbol{x}}^T\left[\boldsymbol{g}(\boldsymbol{x}); \boldsymbol{\tau}\right] \in \mathbb{R}^{L \times L} \tag{2}$$

Though biased in general, both estimates are asymptotically unbiased and consistent. To simplify the exposition, we shall from now on use the notations $\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau})$ and $\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau})$ as shorthand for $\boldsymbol{\Psi}_{\boldsymbol{x}}\left[\boldsymbol{x}; \boldsymbol{\tau}\right]$ and $\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}\left[\boldsymbol{x}; \boldsymbol{\tau}\right]$.

3 Derivation of the Separation Scheme

We begin by deriving the *model function*, which is a functional relationship between the demixing matrix and the observations' charrelation matrices, evaluated at M arbitrary processing-points $\boldsymbol{\tau}_1, ..., \boldsymbol{\tau}_M$. From Property 1.1 above, and from the relation $\boldsymbol{x} = \mathbf{A}\boldsymbol{s}$, we have $\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}) = \mathbf{A}\boldsymbol{\Psi}_{\boldsymbol{s}}(\mathbf{A}^T\boldsymbol{\tau})\mathbf{A}^T$, and thus $\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau})\mathbf{B}^T = \boldsymbol{\Psi}_{\boldsymbol{s}}(\mathbf{A}^T\boldsymbol{\tau})$. Furthermore, from Property 1.2, the charrelation matrix $\boldsymbol{\Psi}_{\boldsymbol{s}}(\mathbf{A}^T\boldsymbol{\tau})$ of the random vector $\boldsymbol{s}$ with mutually independent elements is strictly diagonal. Consequently, a conceivable model function may be:

$$\begin{aligned} &\mathbf{H}\left(\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_1), \ldots, \boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_M); \mathbf{B}\right) = \\ &\qquad \left[\mathbf{Off}(\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_1)\mathbf{B}^T) \cdots \mathbf{Off}(\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_M)\mathbf{B}^T)\right] = \mathbf{0} \in \mathbb{R}^{K \times KM} \end{aligned} \tag{3}$$

where the $\mathbf{Off}(\cdot)$ operator nullifies the main diagonal of its argument matrix. Since all matrices of the form $\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau})\mathbf{B}^T$ are symmetric, considerable reduction may be achieved by rearranging (3) to refer only to the matrix elements above the main diagonal as follows.

First, consider the following bijective indices-transformation:
$\gamma\left(k, \ell, m\right) \triangleq \left(\left(k-1\right)\left(K - \frac{1}{2}k\right) - k + \ell - 1\right) M + m$, where $1 \leq k \leq K-1$, $k < \ell \leq K$, $1 \leq m \leq M$.

This index transformation is used for rearranging all upper off-diagonal elements of the target matrices in a single vector, which is sometimes called the *Off-DIagonal Terms (ODIT)* vector in this context (see, e.g., [4]). Note that $\gamma\left(k, \ell, m\right)$ is in the range $[1 \ldots G]$ where $G = \frac{1}{2}K\left(K-1\right)M$. Thus, define the vectorizing operator $\mathbf{odit}(\cdot)$ which transforms a set of M matrices, each $K \times K$, into a $G \times 1$ vector, such that

$$\left[\mathbf{odit}(\mathbf{D}_1, \ldots, \mathbf{D}_M)\right]_{\gamma(k,\ell,m)} \triangleq \left[\mathbf{D}_m\right]_{k,\ell} \quad 1 \leq k < \ell \leq K \tag{4}$$

where the subscript(s) outside the brackets denote(s) the element index.

The resulting vector-form model function can then be formulated as:

$$\boldsymbol{h}\left(\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)=\mathbf{odit}\left(\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_1)\mathbf{B}^T,\cdots,\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{x}}(\boldsymbol{\tau}_M)\mathbf{B}^T\right)=\mathbf{0}\in\mathbb{R}^G \tag{5}$$

The ensuing weighted least squares (WLS) estimate of $\mathbf{B}$ is obtained by substituting the true (unknown) charrelation matrices in (5) with their sample-means, and seeking the matrix $\widehat{\mathbf{B}}$ which minimizes the weighted norm of the model function subject to some scaling constraint on $\widehat{\mathbf{B}}$ (to avoid the trivial minimizer $\widehat{\mathbf{B}}=\mathbf{0}$).

$$\min_{\mathbf{B}}\left\{\boldsymbol{h}^T\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)\mathbf{W}\boldsymbol{h}\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)\right\}\Rightarrow\widehat{\mathbf{B}} \tag{6}$$

Moreover, under a small-errors assumption it is well-known that the optimal weight matrix $\mathbf{W}$ (in sense of the resulting estimation-error covariance) is the inverse of the covariance matrix of $\boldsymbol{h}\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)$.

Since the model function (5) is non-linear in $\mathbf{B}$, we resort to a Gauss-Newton-based iterative solution for minimizing the WLS criterion. We employ a simple scaling constraint of all-ones elements along the main diagonal[2] of $\widehat{\mathbf{B}}$, which leaves $K(K-1)$ free parameters. To this end, we define another bijective indices-transformation, $\beta(k_1,k_2)\triangleq(K-1)(k_2-1)+k_1-\frac{1}{2}(1+\text{sign}(k_1-k_2))$ where $1\le k_1\neq k_2\le K$ and $\text{sign}(k)$ equals 1 if $k>0$ and -1 if $k<0$ (note that $\beta(k_1,k_2)$ is in the range $[1\ldots K(K-1)]$). We now define the vectorizing operator $\overline{\mathbf{vec}}(\cdot)$ which transforms a $K\times K$ matrix into a $K(K-1)\times 1$ vector, such that

$$\left[\overline{\mathbf{vec}}(\mathbf{D})\right]_{\beta(k_1,k_2)}\triangleq[\mathbf{D}]_{k_1,k_2}\quad 1\le k_1\neq k_2\le K. \tag{7}$$

The iterative updates take the form

$$\overline{\mathbf{vec}}(\widehat{\mathbf{B}}^{[j+1]})=\overline{\mathbf{vec}}(\widehat{\mathbf{B}}^{[j]})-\left(\mathbf{J}_{\boldsymbol{h}}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}^{-1}\mathbf{J}_{\boldsymbol{h}}\right)^{-1}\mathbf{J}_{\boldsymbol{h}}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}^{-1}\boldsymbol{h}\in\mathbb{R}^{K(K-1)} \tag{8}$$

All the elements in the last term depend on $\widehat{\mathbf{B}}^{[j]},\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_M$ which were omitted for brevity of the exposition, $\mathbf{J}_{\boldsymbol{h}}\in\mathbb{R}^{G\times K(K-1)}$ is the derivative of the model function w.r.t. the off-diagonal elements of $\mathbf{B}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}\in\mathbb{R}^{G\times G}$ is the estimated covariance matrix of the model function - both are given by the following propositions:

Proposition 1. *The derivative of the model function w.r.t.* $\overline{\mathbf{vec}}(\mathbf{B})$, *denoted* $\mathbf{J}_{\boldsymbol{h}}\triangleq\frac{\partial\boldsymbol{h}\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)}{\partial\overline{\mathbf{vec}}(\mathbf{B})}\in\mathbb{R}^{G\times K(K-1)}$ *is:*

$$\left[\mathbf{J}_{\boldsymbol{h}}\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m);\mathbf{B}\right)\right]_{\gamma(k,\ell,m),\beta(k_1,k_2)}=\begin{cases}\left[\mathbf{B}\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{\ell,k_2} & k=k_1\\ \left[\mathbf{B}\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{k,k_2} & \ell=k_1\\ 0 & \text{o.w.}\end{cases} \tag{9}$$

[2] This scaling constraint is acceptable due to the inherent scaling ambiguity in ICA, except in rare cases where $\mathbf{B}$ has a zero element on its main diagonal.

Proof. Substituting (4), (7) and (5) in (9), yields:

$$\frac{\partial\left[\boldsymbol{h}\left(\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_1),\ldots,\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_M);\mathbf{B}\right)\right]_{\gamma(k,\ell,m)}}{\partial\left[\mathbf{B}\right]_{k_1,k_2}}=\frac{\partial\left[\mathbf{B}\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\mathbf{B}^T\right]_{k,\ell}}{\partial[\mathbf{B}]_{k_1,k_2}}=$$

$$=\frac{\sum_{i=1}^{K}\sum_{j=1}^{K}[\mathbf{B}]_{k,i}[\mathbf{B}]_{\ell,j}\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{i,j}}{\partial[\mathbf{B}]_{k_1,k_2}}=\begin{cases}\sum_{j=1}^{K}[\mathbf{B}]_{\ell,j}\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{k_2,j} & k=k_1\\ \sum_{i=1}^{K}[\mathbf{B}]_{k,i}\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{i,k_2} & \ell=k_1\\ 0 & \text{o.w.}\end{cases}$$

and using the symmetry of the charrelation matrix, the result is immediate. □

Proposition 2. *The elements of the covariance matrix of the model function* $\boldsymbol{\Sigma}_{\boldsymbol{h}}\in\mathbb{R}^{G\times G}$*, are (for* $1\leq k<\ell\leq K,\quad 1\leq p<q\leq K,\quad 1\leq m,n\leq M$*):*

$$[\boldsymbol{\Sigma}_{\boldsymbol{h}}]_{\gamma(k,\ell,m),\gamma(p,q,n)}=$$
$$=\sum_{i,j,u,v=1}^{K}[\mathbf{B}]_{k,i}[\mathbf{B}]_{\ell,j}[\mathbf{B}]_{p,u}[\mathbf{B}]_{q,v}\operatorname{Cov}\left(\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{i,j},\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_n)\right]_{u,v}\right)$$

where an expression for the inner cross-covariance is given in the Appendix.

Proof.

$$\operatorname{Cov}\left([\boldsymbol{h}]_{\gamma(k,\ell,m)},[\boldsymbol{h}]_{\gamma(p,q,n)}\right)=\operatorname{Cov}\left(\left[\mathbf{B}\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\mathbf{B}^T\right]_{k,\ell},\left[\mathbf{B}\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_n)\mathbf{B}^T\right]_{p,q}\right)=$$

$$=\operatorname{Cov}\left(\sum_{i,j=1}^{K}[\mathbf{B}]_{k,i}\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{i,j}\left[\mathbf{B}^T\right]_{j,\ell},\sum_{u,v=1}^{K}[\mathbf{B}]_{p,u}\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_n)\right]_{u,v}\left[\mathbf{B}^T\right]_{v,q}\right)$$

and the result is immediate from the linearity of the covariance. □

Although $\mathbf{B}$ is unknown in practice, a reasonable estimate of this covariance matrix can be obtained by plugging in some initial estimate $\widehat{\mathbf{B}}^{[0]}$ of $\mathbf{B}$ (attained with any consistent ICA method).

In addition, under a small errors assumption the covariance of the estimated vectorized set of off-diagonal elements of $\mathbf{B}$ is $\boldsymbol{\Sigma}_{\overline{\mathbf{vec}}(\widehat{\mathbf{B}})}=\left(\mathbf{J}_{\boldsymbol{h}}^T\boldsymbol{\Sigma}_{\boldsymbol{h}}^{-1}\mathbf{J}_{\boldsymbol{h}}\right)^{-1}\in\mathbb{R}^{K(K-1)\times K(K-1)}$. While this matrix is unknown, it can be closely estimated for any set of processing-points $\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_M$, using the expressions above, given a preliminary (consistent) estimate $\widehat{\mathbf{B}}^{[0]}$ of $\mathbf{B}$ and the sample-characteristic function, sample-charmeans and sample-charrelations - as prescribed in the Appendix. These closed-form expressions provide the ability to predict the resulting separation performance for any set of selected processing points, based on statistics obtained directly from the observed mixtures alone (no prior knowledge is required). This important feature enables the selection of processing-points, e.g. by finding the set of processing points which minimizes the trace of this (estimated, or predicted) covariance matrix, denoted as $\widehat{\mathbf{\Sigma}}_{\overline{\mathbf{vec}}(\widehat{\mathbf{B}})}\left(\widehat{\mathbf{B}}^{[0]};\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_M\right)$.

We are now ready to summarize our algorithm for charrelation matrices based ICA, including selection of the processing points.

CHARRICA Algorithm:
Input: mixture signals $\boldsymbol{x}[n]$, $1 \leq n \leq N$

```
1.  $\widehat{\mathbf{B}}^{[0]} \leftarrow$ initial estimate of the model parameters provided by any consistent algorithm, rescaled (by rows) to have all-ones along the main diagonal.
2.  Find $\{\boldsymbol{\tau}_m^o\}_{m=1}^M = \underset{\{\boldsymbol{\tau}_m\}_{m=1}^M}{Argmin}\, Tr\left\{\widehat{\mathbf{\Sigma}}_{\overline{\mathbf{vec}}(\widehat{\mathbf{B}})}\left(\widehat{\mathbf{B}}^{[0]};\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_M\right)\right\}$
3.  Estimate the sample-charrelation matrices $\left\{\widehat{\mathbf{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m^o)\right\}_{m=1}^M$
4.  $j \leftarrow 0$
5.  Repeat 6-10 until convergence
6.      Estimate $\widehat{\mathbf{\Sigma}}_{\boldsymbol{h}}\left(\widehat{\mathbf{B}}^{[j]},\boldsymbol{\tau}_1^o,\ldots,\boldsymbol{\tau}_M^o\right)$ using Proposition 2 and the Appendix
7.      Repeat 8-10 until convergence
8.          Compute $\mathbf{J}_{\boldsymbol{h}}$ using Proposition 1
9.          $\overline{\mathbf{vec}}\left(\widehat{\mathbf{B}}^{[j+1]}\right) = \overline{\mathbf{vec}}\left(\widehat{\mathbf{B}}^{[j]}\right) - \left(\mathbf{J}_{\boldsymbol{h}}^T\widehat{\mathbf{\Sigma}}_{\boldsymbol{h}}^{-1}\mathbf{J}_{\boldsymbol{h}}\right)^{-1}\mathbf{J}_{\boldsymbol{h}}^T\widehat{\mathbf{\Sigma}}_{\boldsymbol{h}}^{-1}\boldsymbol{h}$
10.         $j \leftarrow j+1$
```

Output: $\widehat{\mathbf{B}}^{[j]}$.

Since convergence in lines 5 and 7 is usually obtained after 2-3 iterations, the complexity of the CHARRICA algorithm is mainly determined by the covariance matrix estimation in line 6. This is quite a demanding task with complexity of $O(K^8M^2N)$. The second challenging task is the optimization in line 2, which also involves the calculation of the covariance matrix of the model function. The explicit complexity of the optimization is highly dependent on the optimization method.

4 Results

To capture the effectiveness of the proposed algorithm, two simple scenarios where considered, in which each of $K = 3$ sources were drawn independently from the same distribution. In Fig. 1, the attained mean interference to source ratio (ISR) vs. the number of samples N is shown for two distributions: zero-mean unit-variance uniform distribution and a mixture of two equally probable Gaussians $N(\pm 1, 0.04)$. The sources were mixed by a random mixing matrix $\mathbf{A} \in \mathbb{R}^{3\times 3}$, generated with *iid* standard Gaussian elements, followed by normalizing each row to unit norm. First, the ISR matrix was obtained as the empirical average (over 100 independent trials) of the squared elements of the overall mixing-unmixing matrices $(\widehat{\mathbf{B}} \cdot \mathbf{A})$ of each trial, where $\widehat{\mathbf{B}}$ is the estimated unmixing matrix. Second, the overall ISR was calculated as the average

of the off-diagonal terms of the ISR matrix. Five separation algorithms were compared: (1) JADE [1]; (2) FastICA [5]; (3) EFICA [6]; (4) RADICAL [7]; and (5) CHARRICA as proposed in Section 3 with $M = 3$ target matrices. The first four algorithms were performed using the available codes on the internet with the default parameters. In both scenarios, the ISR decreases logarithmically with the number of samples N. For the Uniform distribution, the CHARRICA outperforms JADE, EFICA and RADICAL by about 7 dB, and FastICA by about 10 dB. For the Gaussian Mixture distribution, CHARRICA outperforms JADE, EFICA and RADICAL by about 3 dB, and FastICA by about 6 dB.
Comments: in all of the simulations that were performed the Gauss-Newton algorithm has reached a local minimum. Also, using only the diagonal of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}$ (the covariance matrix of the model function with null off-diagonal elements) provided almost the same performance as with the full matrix.

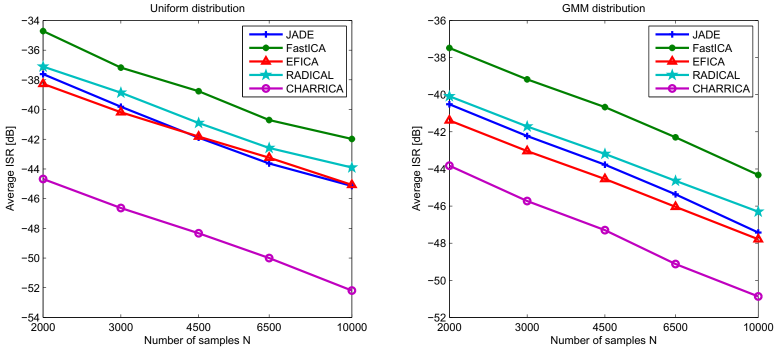


Fig. 1. Performance (ISR [dB] vs. the number of samples N) attained for a Uniform distribution (left) and for a Gaussian Mixture distribution (right), averaged along 100 independent trials, $K = 3$ sources.

5 Discussion

The objective of this work was to demonstrate the potential of the properly-selected charrelation matrices for superior performance over other target matrices in ICA application. The CHARRICA algorithm indeed gives better results compared with known methods such as JADE, FastICA and EFICA. The improvement is achieved, however, at the cost of increased computational complexity which is a considerable drawback of the algorithm.

Using charrelation matrix-based ICA, together with an optimal weighting scheme and a closed-form algorithm for a proper selection of the processing-points, outperforms conventional methods. However, further research should be done to improve the relativity high complexity of the proposed algorithm.

A Appendix

The cross-covariance of two sample-charrelation matrices is [4]:

$$\mathrm{Cov}\left(\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_m)\right]_{i,j},\left[\widehat{\boldsymbol{\Psi}}_{\boldsymbol{x}}(\boldsymbol{\tau}_n)\right]_{u,v}\right)=\frac{1}{N}\frac{\phi_{\boldsymbol{x}}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)}{\phi_{\boldsymbol{x}}\left(\boldsymbol{\tau}_m\right)\phi_{\boldsymbol{x}}\left(\boldsymbol{\tau}_n\right)}\cdot\rho_{ijuv}(\boldsymbol{\tau}_m,\boldsymbol{\tau}_n)\in\mathbb{R}$$

where $\phi_{\boldsymbol{x}}(\boldsymbol{\tau})\triangleq E\left[\exp\{\boldsymbol{x}^T\boldsymbol{\tau}\}\right]$ is the generalized characteristic function (GCF) of $\boldsymbol{x}$ at the processing-point $\boldsymbol{\tau}$, and:

$$\rho_{ijuv}(\boldsymbol{\tau}_m,\boldsymbol{\tau}_n)\triangleq\left(\begin{array}{l}
\Psi_{ij,uv}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)- \\
-\eta_i\left(\boldsymbol{\tau}_m\right)\Psi_{j,uv}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_j\left(\boldsymbol{\tau}_m\right)\Psi_{i,uv}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)- \\
-\eta_u\left(\boldsymbol{\tau}_n\right)\Psi_{ij,v}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_v\left(\boldsymbol{\tau}_n\right)\Psi_{ij,u}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)+ \\
+\eta_i\left(\boldsymbol{\tau}_m\right)\eta_u\left(\boldsymbol{\tau}_n\right)\Psi_{j,v}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)+ \\
+\eta_j\left(\boldsymbol{\tau}_m\right)\eta_u\left(\boldsymbol{\tau}_n\right)\Psi_{i,v}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)+ \\
+\eta_i\left(\boldsymbol{\tau}_m\right)\eta_v\left(\boldsymbol{\tau}_n\right)\Psi_{j,u}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)+ \\
+\eta_j\left(\boldsymbol{\tau}_m\right)\eta_v\left(\boldsymbol{\tau}_n\right)\Psi_{i,u}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)+ \\
+\left(\begin{array}{l}\Psi_{i,j}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\Psi_{i,j}\left(\boldsymbol{\tau}_m\right)+ \\ +(\eta_i\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_i\left(\boldsymbol{\tau}_m\right))(\eta_j\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_j\left(\boldsymbol{\tau}_m\right))\end{array}\right)\cdot \\
\cdot\left(\begin{array}{l}\Psi_{u,v}\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\Psi_{u,v}\left(\boldsymbol{\tau}_n\right)+ \\ +(\eta_u\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_u\left(\boldsymbol{\tau}_n\right))(\eta_v\left(\boldsymbol{\tau}_m+\boldsymbol{\tau}_n\right)-\eta_v\left(\boldsymbol{\tau}_n\right))\end{array}\right)
\end{array}\right)\in\mathbb{R}$$

For shorthand, the subscript $\boldsymbol{x}$ and the operand(s) were omitted and only the indices are left, so that, e.g., $\eta_i(\boldsymbol{\tau}_m)\triangleq\eta_{\boldsymbol{x}}[x_i;\boldsymbol{\tau}_m]$, $\Psi_{i,j}(\boldsymbol{\tau}_m)\triangleq\Psi_{\boldsymbol{x}}[x_i,x_j;\boldsymbol{\tau}_m]$, $\Psi_{j,uv}(\boldsymbol{\tau}_m)\triangleq\Psi_{\boldsymbol{x}}[x_j,x_ux_v;\boldsymbol{\tau}_m]$ and $\Psi_{ij,uv}(\boldsymbol{\tau}_m)\triangleq\Psi_{\boldsymbol{x}}[x_ix_j,x_ux_v;\boldsymbol{\tau}_m]$.

References

[1] Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. IEE Proc.-F. 140(6), 362–370 (1993)

[2] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E.: A blind source separation technique using second-order statistics. IEEE Trans. Signal Processing 45, 434–444 (1997)

[3] Pham, D.-T., Cardoso, J.-F.: Blind separation of instantaneous mixtures of non stationary sources. IEEE Trans. on Signal Processing 49(9), 1837–1848 (2001)

[4] Slapak, A., Yeredor, A.: Weighting for more. Enhancing Characteristic-Function Based ICA with Asymptotically Optimal Weighting. Sig. Process. 91(8), 2016–2027 (2011)

[5] Hyvřinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural Comp. 9(7), 1483–1492 (1997)

[6] Koldovský, Z., Tichavský, P., Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. IEEE Trans. on Neural Networks 17(5), 1265–1277 (2006)

[7] Learned-Miller, E.G., Fisher III, J.W.: ICA using spacings estimates of entropy. J. Mach. Learn. Res. 4(7-8), 1271–1295 (2004)

Contrast Functions for Independent Subspace Analysis

Jason A. Palmer and Scott Makeig

Swartz Center for Computational Neuroscience
University of California San Diego, La Jolla, CA 92093
{jason,scott}@sccn.ucsd.edu

Abstract. We consider the Independent Subspace Analysis problem from the point of view of contrast functions, showing that contrast functions are able to partially solve the ISA problem. That is, basic ICA can solve the ISA problem up to within-subspace separation/analysis. We define sub- and super-Gaussian subspaces and extend to ISA a previous result on freedom of ICA from local optima. We also consider new types of dependent densities that satisfy or violate the entropy power inequality (EPI) condition.

1 Introduction

The mutual information minimization approach to blind source separation has proved very effective at separating linear mixtures of independent, nongaussian sources [1]. This approach is equivalent to a maximum likelihood approach in which the source density models are adapted as well [2]. In general, however, some "sources" sources may exhibit mutual dependence, e.g. in signal power, leading to what has been variously called Multidimensional ICA [3], independent subspace analysis [4], and independent vector analysis [5]; or the dependent subspaces may not be further decomposable into unique components, as is the case with non-Gaussian subspaces with radial symmetry.

In a foundational paper on ICA [6], Comon defined the *contrast functions* to be those (statistical) functions of which are capable of separating or extracting independent sources from a linear mixture. This definition is actually very similar to the idea expressed by P. Huber in his work on *projection pursuit* [7,8]. Basic ICA, i.e. the maximization of a contrast function, is often found to successfully separate sources of the variance dependence type, with the subspace dependence structure ascertainable after the separation. The good performance of basic ICA in the dependent subspace context has led to the conjecture that the minimization of mutual information of the output is able to perform separation of certain dependent sources as well [3].

Theis has considered the ISA problem in a number of articles, proposing the joint block-diagonalization approach [9], considering conditions for separation of non-Gaussian subspaces from Gaussian subspaces in complete mixtures [10], and using the autocorrelation structure of temporally correlated sources or subspaces to perform ISA [11]. Gutch [12] defined the concept of "irreducible subspaces" to be those containing no extractable Gaussian component, and showed that the solution to the ISA problem is unique in this case.

Castella and Comon [13] have also investigated ISA with known dependent subspace structure, and determined specific cases in which cumulant-based contrasts preserve

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 115–122, 2012.

separability and when they fail. Here again the emphasis is on separability of dependent sources (dependent component analysis) rather than the separation of dependent subspaces from one another.

Szabo [14] has shown using the entropy power inequality (EPI), that dependent sources can be separated by minimum mutual information as long as all one-dimensional projections of the dependent sources satisfy the EPI. Szabo's emphasis is on the complete solution to the dependent component analysis problem, the EPI sufficient condition guaranteeing this possibility (in the case of non-radial symmetry). The EPI approach is shown to be successful when the EPI condition is satisfied by the sources, without requiring prior knowledge of the subspace structure or dimensions.

We show here that dependent subspaces can be separated, i.e. the pairwise mutual information can be block-diagonalized, in a more general setting than that considered in [14]. We take the fact that basic ICA can perform ISA (without necessarily further analyzing the subspaces) as significant in and of itself, since it shows that independent subspaces can be separated from one another as a preliminary processing step, with further analysis of the subspaces themselves carried out subsequently.

This result is significant because it immediately provides an answer for the common criticism of ICA-based methods as being naively misspecified, potentially calling in to question the validity of the results. Essentially we are expanding the concept of a source component to be a potentially multidimensional subspace, with the new "ICA model" that is to be presupposed in the often encountered linear model is that *subspaces* of components are independent. Thus we generalize basic ICA in which all subspaces are one-dimensional, and guarantee the ability of ICA approaches to extract independent sources even in the context of interfering dependent subspaces, as well as guaranteeing that estimated dependent subspaces contain all information pertaining to the subspace that is present in the data.

We also define sub- and super-Gaussian subspaces to be those in which all univariate projections are sub- or super-Gaussian in the Benveniste sense, and show using a previous results [16] that ISA of strongly super-Gaussian subspaces is free of local optima.

2 The ISA Problem

Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, be an invertible matrix consisting of m subspaces, and let $\mathbf{s} \in \mathbb{E}_2^n$ be a finite covariance random vector with m corresponding subvectors:

$$\mathbf{A} = [\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_m], \quad \mathbf{s}^T = [\mathbf{s}_1^T \, \mathbf{s}_2^T \, \cdots \, \mathbf{s}_m^T],$$

where $\mathbf{A}_j \in \mathbb{R}^{n \times d_j}$, $\sum_{j=1}^m d_j = n$, and the $\mathbf{s}_j \in \mathbb{E}_2^{d_j}$ are mutually independent, i.e. $p_{\mathbf{s}}(\mathbf{s}) = \prod_{j=1}^m p_{\mathbf{s}_j}(\mathbf{s}_j)$. Let $\mathbf{x}$ be the random vector given by,

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

so that $\mathbf{x} \in \mathbb{E}_2^n$.

The ultimate goal of ISA is to reproduce the source vectors, $\mathbf{s}_j$, $j = 1, \ldots, m$, given a set of observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$. That is, we would like to produce a matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$ such that $\mathbf{WA} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

However, ISA can be divided logically into two problems:

P1: Separate the independent subspaces from one another.
P2: Separate the dependent sources within each subspace.

Most of the work on the ISA problem has been concerned with solving both problem simultaneously. The ISA separation theory of Szabo gives an entropy condition on dependent sources that allows them to be separated using entropy contrasts. We take it, however, that the well known conjecture that basic ICA also performs ISA, is in fact largely concerned with **P1**, separating the subspaces, or in the particular case of extracting an independent source of interest from a mixture that includes interfering dependent source subspaces.

It should also be noted that **P2** of the ISA problem may not have a solution, as in the case of dependent source subspaces with spherically symmetric distributions. These sources are shown in [14] to satisfy the entropy constraint allowing "solution" of the ISA problem. However, as in the case of Gaussian sources in ICA, spherically symmetric subspaces can only be separated up to an arbitrary rotation. Therefore an ISA problem with spherically symmetric dependent subspaces only consists of **P1**.

3 Deflationary Contrast Functions, ICA, and ISA

In the deflationary approach to basic ICA (where $d_j = 1, j = 1, \ldots, m,$) the matrix $\mathbf{W}$ is constructed one row $\mathbf{w}_j^T$ at a time, i.e. the sources are estimated sequentially. This is usually done by sequentially determining maxima of a *contrast function*, $\Phi : \mathbb{E}_2 \to \mathbb{R}$,

$$\hat{\mathbf{w}}_j = \arg \max_{\mathbf{w}^T \mathbf{R}_{\mathbf{xx}} \mathbf{w} = 1} \Phi(\mathbf{w}^T \mathbf{x})$$

The process ensures successive estimates are unique by restricting $\hat{\mathbf{w}}_j^T \mathbf{R}_{\mathbf{xx}} \hat{\mathbf{w}}_{j'} = 0$ for all previously estimated $\hat{\mathbf{w}}_{j'}$.

3.1 Contrast Functions

We define contrast functions as follows [6,8].

Definition 1. *A **contrast function** is a functional,* $\Phi : \mathbb{E}_2 \to \mathbb{R}$*, defined on random variables, satisfying the condition,*

$$\Phi\big(\cos(\theta)X + \sin(\theta)Y\big) \leq \max\big(\Phi(X), \Phi(Y)\big), \; X, Y \, \textit{independent}$$

If the condition is satisfied for all $X, Y \in \mathcal{S} \subset \mathbb{E}_2$*, and strictly satisfied only when* θ *is a multiple of* $\pi/2$*, then the contrast is said to **discriminate** over* $\mathcal{S}$*.*

Examples. Well known contrasts include the following:

1. Inverse entropy power. The entropy power functional, $N(X)$ is defined by

$$N(X) \triangleq (2\pi e)^{-1} \exp(2h(X))$$

where $h(X) \triangleq E\{-\log p_X(X)\}$. The entropy power satisfies, $N(aX) = a^2 N(X)$, and the entropy power inequality (EPI), $N(X+Y) \geq N(X) + N(Y)$. We thus have for the inverse,

$$\begin{aligned} N\big(\cos(\theta)X + \sin(\theta)Y\big)^{-1} &\leq \Big(N\big(\cos(\theta)X\big) + N\big(\sin(\theta)Y\big)\Big)^{-1} \\ &= \Big(\cos^2(\theta)N\big(X\big) + \sin^2(\theta)N\big(Y\big)\Big)^{-1} \\ &\leq \cos^2(\theta)\, N(X)^{-1} + \sin^2(\theta)\, N(Y)^{-1} \\ &\leq \max\big(N(X)^{-1}, N(Y)^{-1}\big) \end{aligned}$$

2. Fisher Information. The Fisher information $J(X)$ is given by,

$$J(X) = E\left\{\frac{d^2}{dx^2} - \log p_X(x)\right\} = E\left\{\left(\frac{d}{dx}\log p_X(x)\right)^2\right\}$$

The Fisher information satisfies, $a^2 J(aX) = J(X)$. It also satisfies an inequality related to the EPI [15]:

$$\begin{aligned} J\big(\cos(\theta)X + \sin(\theta)Y\big) &\leq \cos^2(\theta)J(X) + \sin^2(\theta)J(Y) \\ &\leq \max\big(J(X), J(Y)\big) \end{aligned}$$

3. Cumulant magnitude. The nth cumulant functional $\kappa_n(X)$ is defined by the nth coefficient of the Taylor expansion of the log characteristic function, $\log\varphi(t)$, where $\varphi(t) \triangleq E\{\exp(itX)\}$. Cumulants satisfy the property,

$$\kappa_n\big(\cos(\theta)\, X + \sin(\theta)\, Y\big) \;=\; \cos^n(\theta)\, \kappa_n(X) \;+\; \sin^n(\theta)\, \kappa_n(Y)$$

For n even, and X, Y of the same cumulant sign,

$$\begin{aligned} |\kappa_n\big(\cos(\theta)\, X + \sin(\theta)\, Y\big)| \;&=\; \cos^n(\theta)|\kappa_n(X)| \;+\; \sin^n(\theta)\, |\kappa_n(Y)| \\ &\leq\; \cos^2(\theta)\, |\kappa_n(X)| \;+\; \sin^2(\theta)\, |\kappa_n(Y)| \\ &\leq\; \max\big(|\kappa_n(X)|, |\kappa_n(Y)|\big) \end{aligned}$$

Thus even cumulant magnitude defines a contrast discriminating over sets of random variables with the same cumulant sign.

3.2 Contrasts and ISA

We now show that every contrast function is also a subspace contrast in the sense that deflationary ICA can be used to solve **P1**.

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be invertible, $\mathbf{A} = [\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_m]$, where $\mathbf{A}_j \in \mathbb{R}^{n\times d_j}$, $\sum_{j=1}^m d_j = n$. Let $\mathbf{s} \in \mathbb{E}_2^n$, $\mathbf{s}^T = [\mathbf{s}_1^T\, \mathbf{s}_2^T\, \cdots\, \mathbf{s}_m^T]$, with the $\mathbf{s}_j \in \mathbb{E}_2^{d_j}$ mutually independent. Let $\mathbf{x} = \mathbf{A}\mathbf{s}$.

We prove specifically that for a deflationary contrast Φ, if,

$$\hat{\mathbf{w}}_{j'} = \arg \max_{\mathbf{w}^T \mathbf{R_{xx}} \mathbf{w} = 1} \Phi(\mathbf{w}^T \mathbf{x})$$

then $\mathbf{w}^T \mathbf{A}_j \mathbf{A}_j^T \mathbf{w} = 0$ for all $j \neq j'$.

Theorem 1. *The deflationary contrast method solves* **P1**.

Proof. Let $\mathbf{y} = \mathbf{w}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{x} = \mathbf{c}^T\mathbf{x}$, where $\mathbf{c} \triangleq \mathbf{A}^T\mathbf{w}$. Since $\mathbf{R_{xx}} = E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{A}\mathbf{A}^T$, we have that $\mathbf{w}^T\mathbf{R_{xx}}\mathbf{w} = 1$ implies $\mathbf{c}^T\mathbf{c} = 1$.

We have $\mathbf{y} = \mathbf{c}^T\mathbf{s} = \sum_{j=1}^{m} \mathbf{c}_j^T\mathbf{s}_j$. By the contrast function condition, we have,

$$\Phi(\mathbf{c}^T\mathbf{s}) = \Phi\biggl(\sum_{j=1}^{m}\mathbf{c}_j^T\mathbf{s}_j\biggr) = \Phi\biggl(\sum_{\|\mathbf{c}_j\|\neq 0}\|\mathbf{c}_j\|\frac{\mathbf{c}_j^T\mathbf{s}_j}{\|\mathbf{c}_j\|}\biggr) \leq \max_j \Phi\biggl(\frac{\mathbf{c}_j^T\mathbf{s}_j}{\|\mathbf{c}_j\|}\biggr)$$

with equality only if $\|\mathbf{c}_j\| = 0$ for all but the maximizing j. But this implies that $\mathbf{w}^T\mathbf{A}_j\mathbf{A}_j^T\mathbf{w} = 0$ for all but the maximizing j.

This theorem shows that the solution to a stage of the deflationary ICA process can only be a linear combination of sources from within one and only one of the dependent subspaces. Each subsequent source estimate will either be dependent with a previously estimated source (having positive mutual information) and be a linear combination only of sources in that subspace, or will be independent of previously estimated sources, beginning the estimate of (one direction in) a new independent subspace. At the end of the procedure, the matrix of pairwise mutual information values between the estimated sources will be a block diagonal permutation.

4 Sub- and Super-Gaussian Subspaces

In this section we define a particular classes of dependent subspaces in terms of linear projections, and use a previously derived result on globally optimal ICA [16] to show that the solution to **P1** of the ISA problem is also free of local optima.

We first review the Benveniste definition of (strong) sub- and super-Gaussianity.

Definition 2 (Strongly Sub- and Super-Gaussian Random Variables). *Let X be a random variable with differentiable probability density function, $p_X(x)$. Define $f(x) \triangleq -\log p_X(x)$. Then p_X is a strongly super-Gaussian (sub-Gaussian) if $p_X(x)$ is symmetric about $x = 0$ and $f'(x)/x$ is strictly decreasing (increasing) on $x > 0$.*

We define sub- and super-Gaussian subspaces to be spaces of dependent random variables in which all linear projections are strongly sub- or super-Gaussian respectively.

Definition 3 (Sub- and Super-Gaussian Subspaces). *Let $\mathbf{x} \in \mathbb{R}^d$ be a non-Gaussian dependent random vector. Then $\mathbf{x}$ is a strongly super-Gaussian (sub-Gaussian) random vector if, for all $\mathbf{w} \in \mathbb{R}^d$, we have $y = \mathbf{w}^T\mathbf{x}$ strongly super-Gaussian (sub-Gaussian).*

In previous work [17], we have considered Generalized Gaussian scale mixtures as an example of a non-radially symmetric dependent subspace. As a more general GSM-based formulation dependent subspaces, let s have independent GSM components, i.e. $s_i = \xi^{1/2}z$ for a non-negative finite variance ξ_i and Gaussian z_i. Let,

$$\mathbf{y} = \eta^{1/2}\mathbf{s}$$

where a common nonnegative scalar $\eta^{1/2}$ multiplies each (independent) GSM component of $\mathbf{s} \in \mathbb{E}_2^d$ to form random vector $\mathbf{y}$ with dependent components. Then we have,

$$u \;=\; \mathbf{w}^T\mathbf{y} \;=\; \eta^{1/2}\sum_i w_i\,\xi_i^{1/2} z_i \;\overset{d}{=}\; \eta^{1/2}\big(\xi_1 + \cdots + \xi_d\big)^{1/2} z_1$$

so that u is also a GSM, and thus strongly super-Gaussian. Dependent GSM subspaces are thus strongly super-Gaussian as defined here.

Theorem 2. *The ISA problem* **P1** *with strongly super-Gaussian dependent subspaces has no local optima when solved using a strongly super-Gaussian contrast.*

This follows from the theorem proved in [16].

5 Other Types of Norm Dependence

We finally consider random vectors with somewhat more general dependent densities to inquire as to which types of non-radially symmetric dependent subspaces violate the EPI condition of [14]. That is, what kinds of dependent sources are and are not separated by contrast functions in the solution of **P1**.

Consider a two dimensional dependent subspace with density,

$$p(x_1, x_2) = f\big(g(x_1) + g(x_2)\big)$$

Let $h_y(\theta)$ be the entropy of projections $y = \cos(\theta)x_1 + \sin(\theta)x_2$ as a function of θ.

Theorem 3. *Let f be decreasing, with $-\log f(\sqrt{x})$ concave. Let $g(\sqrt{x})$ be increasing and concave on $x \in (0, \infty)$, then for $\theta \in (0, \pi/4)$, we have,*

$$h_y'(\theta) \geq 0$$

This follows from a derivation similar to that in [16].

Definition 4. *A density, $p(x_1, \ldots, x_n)$, is said to be* sup-sup dependent *(respectively* sub-sub dependent*) if it is of the form,*

$$p(x_1, \ldots, x_n) \;=\; f\big(g_1(x_1) + \cdots + g_n(x_n)\big)$$

with f decreasing on $(0, \infty)$, $-\log f(\sqrt{y})$ concave (respectively convex), $g_i(x_i)$ nonnegative, symmetric, and increasing on $(0, \infty)$, and $g_i(\sqrt{x})$ concave (respectively convex) on $(0, \infty)$, for $i = 1, \ldots, n$. Sup-sub *and* sub-sup dependence *are defined by the concave-convex and convex-concave scenarios respectively.*

Corollary 1. *Sup-sup and sub-sub dependent densities are satisfy the EPI condition of [14] and may thus be separated by contrast functions.*

If the convexity is not "homogeneous" but rather "conflicting" such that one of $-\log f$ and $-\log g$ is concave and one is convex, then we have,

$$h_y'(\theta) \leq 0, \quad \theta \in (0, \pi/4)$$

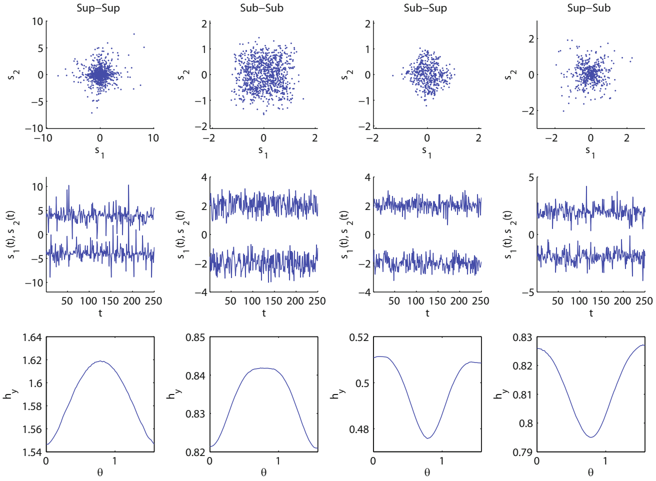


Fig. 1. Examples of dependent densities with various combinations of subgaussian and supergaussian envelope and level curve function

In Figure 1, we present some experiments to verify the theory of this section. We generate four sets of two-dimensional dependent sources, corresponding to the sup-sup, sub-sub, sub-sup, and sup-sub cases respectively. The "sup" density is Laplacian, i.e. $p(x) \propto \exp(-|x|)$, and the "sub" density is Generalized Gaussian with shape parameter 5, $p(x) \propto \exp(-|x|^5)$. The sup-sup data is generated by multiplying i.i.d. Laplacian samples by a common instance dependent scaling, which is Gamma distributed. This creates a supergaussian envelope dependence. The sub-sub data is generated by inducing a slight variance dependence on i.i.d. subgaussian data by multiplying it by a common random Gamma scaling that is tightly concentrated about unity. The sub-sup data is generated by multiplying uniform data over the diamond (Laplacian level curves) by a slight common scaling to induce a subgaussian envelope over Laplacian level curves. The sup-sub data is generated by multiplying i.i.d. uniform data by a strong scaling, to induce a supergaussian envelope on uniform (subgaussian) level curves. The "time series" are shown in the second row, shifted to improve visibility. The bottom row plots the entropy of projections as a function of the rotation angle for $\theta \in (0, \pi/2)$. Symmetry is expected about $\pi/4$, and deviation gives an idea of the noise in the empirical entropy calculation. Entropy is calculated by approximately integrating the histogram.

It can be seen that entropy increases with rotation for the sup-sup and sub-sub dependent sources, while it decreases for the sub-sup and sup-sub dependent sources, as predicted.

References

1. Pham, D.T.: Mutual information approach to blind separation of stationary sources. IEEE Trans. Information Theory 48(7), 1935–1946 (2002)
2. Cardoso, J.-F.: Infomax and maximum likelihood for source separation. IEEE Letters on Signal Processing 4(4), 112–114 (1997)

3. Cardoso, J.-F.: Multidimensional independent component analysis. In: Proceedings of the IEEE International Conference on Acoustics and Signal Processing (ICASSP 1998), Seattle, WA, pp. 1941–1944 (1998)
4. Hyvärinen, A., Hoyer, P.O.: Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Computation 12, 1705–1720 (2000)
5. Kim, T., Eltoft, T., Lee, T.-W.: Independent Vector Analysis: An Extension of ICA to Multivariate Components. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 165–172. Springer, Heidelberg (2006)
6. Comon, P.: Independent component analysis: a new concept? Signal Processing 36(3), 287–314 (1994)
7. Huber, P.J.: Projection pursuit. The Annals of Statistics 13(2), 435–475 (1985)
8. Pham, D.T.: Contrast functions for blind separation and deconvolution of sources. Tech. Rep., Laboratoire de Mod'elisation et Calcul, CNRS, IMAG (2001)
9. Theis, F.J.: Blind signal separation into groups of dependent signals using joint block diagonalization. In: ISCAS (6), pp. 5878–5881 (2005)
10. Theis, F.J., Kawanabe, M.: Uniqueness of Non-Gaussian Subspace Analysis. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 917–925. Springer, Heidelberg (2006)
11. Theis, F.J.: Colored subspace analysis: Dimension reduction based on a signal's autocorrelation structure. IEEE Trans. on Circuits and Systems 57-I(7), 1463–1474 (2010)
12. Gutch, H.W., Theis, F.J.: Independent Subspace Analysis is Unique, Given Irreducibility. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 49–56. Springer, Heidelberg (2007)
13. Castella, M., Comon, P.: Blind Separation of Instantaneous Mixtures of Dependent Sources. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 9–16. Springer, Heidelberg (2007)
14. Szabó, Z., Póczos, B., Lőrincz, A.: Undercomplete Blind Subspace Deconvolution via Linear Prediction. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 740–747. Springer, Heidelberg (2007)
15. Dembo, A., Cover, T.M., Thomas, J.A.: Information theoretic inequalities. IEEE Transactions on Information Theory 37(6), 1501–1518 (1991)
16. Palmer, J.A., Kreutz-Delgado, K., Makeig, S.: Strong Sub- and Super-Gaussianity. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 303–310. Springer, Heidelberg (2010)
17. Palmer, J.A., Kreutz-Delgado, K., Rao, B.D., Makeig, S.: Modeling and Estimation of Dependent Subspaces with Non-Radially Symmetric and Skewed Densities. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 97–104. Springer, Heidelberg (2007)

Distributional Convergence of Subspace Estimates in FastICA: A Bootstrap Study

Jarkko Ylipaavalniemi, Nima Reyhani, and Ricardo Vigário

Department of Information and Computer Science,
Aalto University School of Science,
P.O. Box 15400, FI-00076 Aalto, Finland
first.last@aalto.fi
http://ics.tkk.fi/en/

Abstract. Independent component analysis (ICA) is possibly the most widespread approach to solve the blind source separation (BSS) problem. Many different algorithms have been proposed, together with an extensive body of work on the theoretical foundations and limits of the methods.

One practical concern about the use of ICA with real-world data is the reliability of its estimates. Variations of the estimates may stem from the inherent stochastic nature of the algorithm, or deviations from the theoretical assumptions. To overcome this problem, some approaches use bootstrapped estimates. The bootstrapping also allows identification of subspaces, since multiple separated components can share a common pattern of variation, when they belong to the same subspace. This is a desired ability, since real-world data often violates the strict independence assumption.

Based on empirical process theory, it can be shown that FastICA and bootstrapped FastICA are consistent and asymptotically normal. In the context of subspace analysis, the normal convergence is not satisfied. This paper shows such limitation, and how to circumvent it, when one can estimate the canonical directions within the subspace.

1 Introduction

Blind source separation (BSS) has become a mainstream topic in signal and image processing, with independent component analysis (ICA) as possibly its most widespread solution. Furthermore, it is believed that the basic theoretical foundations of ICA, as well as its various implementations are rather well understood (*c.f.*, [1–3]). In particular, for the FastICA algorithm (c.f. [4, 5]), some theoretical limits have been presented earlier (see, *e.g.*, [6, 4, 7]).

In practice, with real-world data, one persistent concern for the use of ICA is the reliability of the estimated sources. Repeated use of most ICA algorithms results in slight variations in the estimated components. There are many potential factors, including the possibly inherent stochastic nature of the ICA implementation, or some mismatch with the ideal ICA theoretical assumptions [8]. To assess the reliability of the source estimation, several methods have been proposed, often based on a bootstrap analysis of the estimated components [9], or

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 123–130, 2012.

using multiple runs of bootstrapped FastICA, with a subsequent clustering of the estimated sources [8]. Such an approach can lead to very interesting insights for functional magnetic resonance images (fMRI, [10]), *e.g.*, networks of brain activity [11] or independent subspaces [12].

To give a formal theoretical grounds for the practical success of the multiple run approach to FastICA, its limitations and validity were shown in [13], based on a proof of asymptotic normality of FastICA and bootstrapped FastICA, using the method of empirical process theory and Z-estimators [14]. Also, a probabilistic convergence rate was derived. Besides its theoretical importance, the aforementioned results allow for an elegant check of the algorithm's convergence, using a multivariate normality test. Although the HZ-Multivariate Normality Test [15] was used, other, such as the T^2-Hotelling test [16] could also be used. The focus is on FastICA, but several of the considerations could possibly be extended to other ICA methods.

The main focus in this paper is to show that, with real-world data, bootstrapped FastICA is indeed a *consistent estimator* and converges *asymptotically* to a *normal random vector*. The paper shows that the difference between the random estimator and a best guess of the ground truth, converges to a centered Gaussian random variable. Furthermore, since real-world data can violate the strict assumptions of ICA and have independent subspaces, the paper additionally shows that, even in such a subspace situation, it is possible to identify estimates that still fulfill the asymptotic normality, and validates the the testing in this condition as well.

2 Materials and Methods

The following is a short summary of the theoretical results in [13], as they apply in the case of this paper. For a more thorough description and derivations, see the aforementioned reference. In the following, $\mathbb{E}$ is the expectation, Pr denotes a probability, and $\xrightarrow{P}$ means converges in probability. Let $\boldsymbol{z}$ denote the whitened data, $\boldsymbol{w}$ the demixing vectors in the whitened space. More specifically $\boldsymbol{w}_\circ$ is the true solution, $\hat{\boldsymbol{w}}$ the sample estimator and $\hat{\boldsymbol{w}}^*$ the bootstrap estimator.

Theorem 1 (Consistency and Asymptotic Normality of FastICA). *Let us assume* $\mathbb{E}\boldsymbol{z} = 0$ *and* $\boldsymbol{z}$ *has all moments up to the fourth;* $\mathbb{E}\boldsymbol{z}\boldsymbol{z}^\top = I_d$*; and function* $g : \mathbb{R} \to \mathbb{R}$ *and its first and second derivatives, denoted by* g' *and* g''*, are Lipschitz. Further, assume* $g''(\cdot)$ *is bounded;* $\mathbb{E}g'(s_\circ) \neq 0; \mathbb{E}s_\circ^2 g'(s_\circ) \neq 0; \mathbb{E}g^2(s_\circ) \neq 0$*; and* $\mathbb{E}s_\circ^2 g^2(s_\circ) \neq 0$*. Also, let all exist and together with* $\mathbb{E}G(\boldsymbol{w}^\top \boldsymbol{z}), \forall \boldsymbol{w} \in \mathcal{S}^{d-1}$ *be bounded. Then the sequence*

$$\hat{\boldsymbol{w}}_n = \underset{\boldsymbol{w}\in\mathcal{S}^{d-1}}{\operatorname{argmax}}\, \mathbb{E}G(\boldsymbol{w}^\top \boldsymbol{z}),$$

that is produced by the FastICA iteration is consistent and asymptotically normal, i.e.

$$\hat{\boldsymbol{w}}_n \xrightarrow{P} \boldsymbol{w}_\circ$$

$$\sqrt{n}(\hat{\boldsymbol{w}}_n - \boldsymbol{w}_\circ) \rightsquigarrow \mathcal{N}(0, \Sigma_d),$$

where $\Sigma_d = \boldsymbol{A} diag[\frac{\mathbb{E}g^2(s_\circ)}{\mathbb{E}g'(s_\circ)^2}, \ldots, \frac{\mathbb{E}s_\circ^2 g^2(s_\circ)}{(\mathbb{E}s_\circ^2 g'(s_\circ))^2}, \ldots, \frac{\mathbb{E}g^2(s_\circ)}{\mathbb{E}g'(s_\circ)^2}]\boldsymbol{A}^\top$ *and* $\boldsymbol{A}$ *is the true mixing matrix. In addition, the bootstrapped FastICA is also asymptotically normal, i.e.*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| Pr\left\{ \frac{\sqrt{n}}{c}(\hat{\boldsymbol{w}}_n^* - \hat{\boldsymbol{w}}_n) \leq x \right\} - Pr_{\mathcal{X}}\left\{ \mathcal{N}(0, V_{\hat{\boldsymbol{w}}_n}^{-1} U_{\hat{\boldsymbol{w}}_n} (V_{\hat{\boldsymbol{w}}_n}^{-1})^\top) \leq x \right\} \right| \xrightarrow{P} 0,$$

conditioned that $U_{\hat{\boldsymbol{w}}_n} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top g^2(\hat{\boldsymbol{w}}_n^\top \boldsymbol{z}_i)$ *and* $V_{\hat{\boldsymbol{w}}_n} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i^\top g'(\hat{\boldsymbol{w}}_n^\top \boldsymbol{z}_i)$ *exist and are non-singular.*

The theorem justifies the use of FastICA in a multiple run, bootstrap and randomly initialized manner (see [13]). However, after each run, a different set of sources may be estimated and the total number of estimates for each component will vary. Moreover, the whitening step can flip the signs of individual dimensions of the whitened space differently at each run, depending on the bootstrap sampled data. It is, therefore, crucial to identify and group similar components from the various runs. A statistical analysis of each group can then be performed.

To show the theoretical implications in practice, a series of experiments were performed with real-world data. The robust ICA approach used in the experiments is implemented in the Arabica toolbox [17]. Using the toolbox, FastICA can be run multiple times, with varying initial conditions and bootstrap sampling. The algorithm is summarized in Table 1.

Table 1. Arabica Algorithm: Bootstrapped FastICA

- Perform multiple runs of FastICA by repeating the following steps as many times as required.
  1. Draw a bootstrap sample of the given data matrix.
  2. Whiten the bootstrap sample, using PCA, and possibly reduce data dimension.
  3. Randomize the initial conditions for FastICA.
  4. Estimate the desired number of independent components.
- Form groups with the estimates corresponding to the same independent component.
  1. Collect all estimates from the multiple runs.
  2. Calculate the similarity of the estimates taking into account the sign and scale ambiguities in ICA, based on a given similarity measure, *e.g.*, correlation or inner-product.
  3. Cluster the estimates, using given similarity threshold and linkage path length.
  4. Rank the clusters based on the number of estimates and their compactness.

The experiments were done with functional magnetic resonance imaging (fMRI) data from an auditory experiment. A series of whole-head recordings of a single subject were used. In the fMRI study, subjects listened to spoken safety instructions in 30 s intervals, interleaved with 30 s resting periods. All the data were acquired at the Advanced Magnetic Imaging Centre of the Aalto University, using a 3.0 Tesla MRI scanner (Signa EXCITE 3.0T; GE Healthcare, Chalfont St. Giles, UK) with a quadrature birdcage head coil, and using Gradient Echo, Echo Planar Imaging (TR 3 s, TE 32 ms, 96x96 matrix, FOV 20cm, slice thickness 3 mm, 37 axial slices, 80 time points (excl. 4 first ones), flip angle 90°). For further details on the data set, see [8]. Preprocessing of the data included the typical realignment, normalization, smoothing and masking off all areas outside the brain. The resulting data-matrix has a size of 80×263361.

Since the ground truth of the real-world data is unknown, one cannot apply the theory directly. The asymptotic normality of bootstrapped FastICA (see Theorem 1) also means that the difference between the sample estimator and the bootstrap estimator converges to a centered normal vector with nonsingular covariance matrix. Thus, the ground truth was estimated by performing a separate analysis with 100 rounds of ICA, using the whole data without bootstrap, only altering the algorithm's random initial conditions. Each round looked for 15 components in a whitened space with 30 dimensions. The ground truth components were clustered using cosine similarity on the white demixing vectors, with a threshold of 0.99, taking into account only direct links between estimates. This ensures that each group represents a differently estimated demixing vector, even if the components in the original space would be closely related. Additionally, a representative set of initial conditions was found during the whole data analysis. This initialization was kept fixed to those values throughout the bootstrap analysis, so that all the variations in the estimates were due to the bootstrap sampling.

For the bootstrap results, 500 runs of ICA were performed, searching for 15 components from a whitened space with 30 dimensions, and using a bootstrap sampling with 138944 (52.76% out of 263361) samples in each run. The estimates were clustered using correlation between the component time-courses, with a threshold of 0.97 and taking into account only direct links between estimates. The similarity measure is different from the one used in estimating the ground truth, since the aim is to find matching mixing and source vectors, even if they would be produced by a different demixing. Also, the threshold is slightly lower, since the bootstrap sampling will likely add small variations to the estimates.

3 Results

Altogether, 39 independent components were identified from the bootstrap ICA. Figure 1 shows five examples of the found independent components. Each component is shown with the estimate means and variabilities.

The first component has barely any variability, whereas the other four components have a significant one. Only 43 estimates of the fifth component were found during the 500 bootstrap rounds, meaning that the component is very difficult to identify by ICA. The first component represents activation of the primary auditory cortices, whereas the other components split activity along the cingulate gyrus in four different parts. Note that, in some cases the temporal variability is higher around some time points than in other, and also the spatial variance is focused on certain regions. Moreover, the shared locations of the variances, *i.e.* covariance, of the last four components reveals that those components belong to a subspace, as speculated in [12].

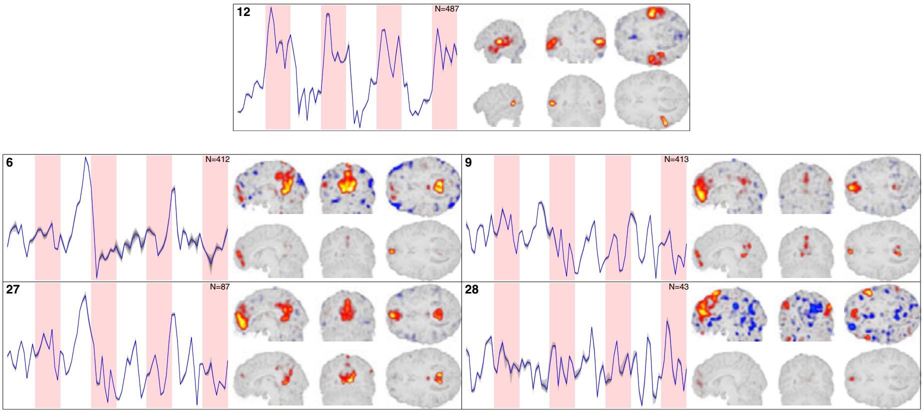
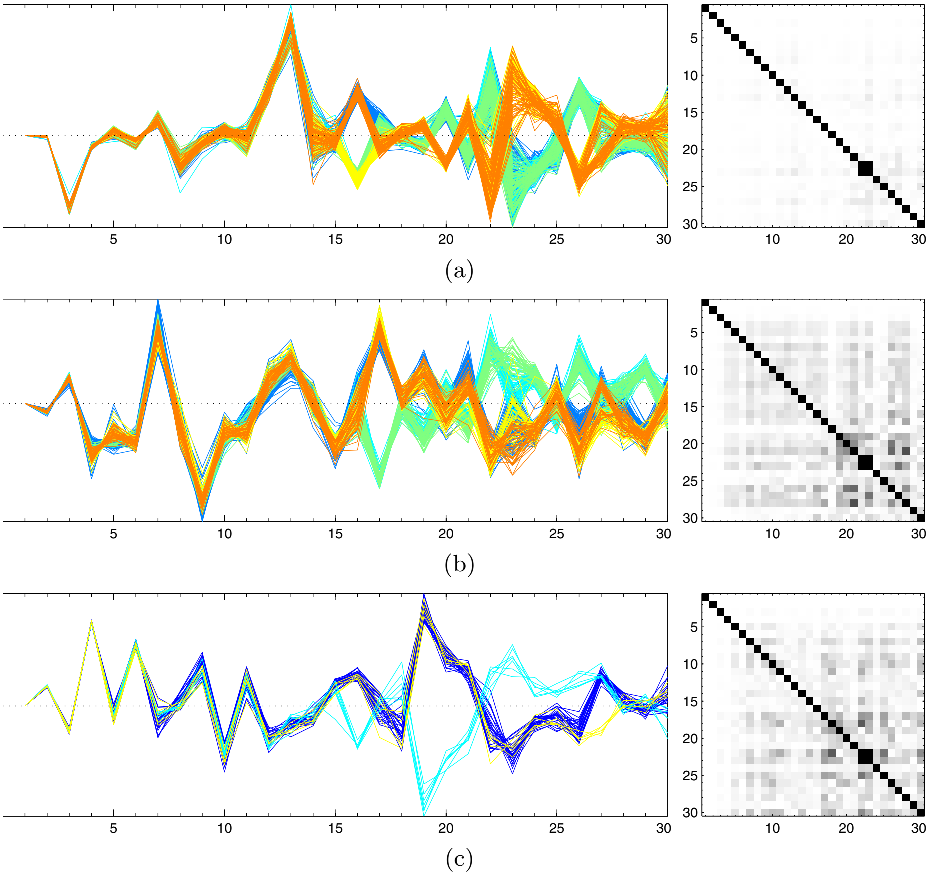


Fig. 1. Examples of independent components estimated with bootstrapped ICA. For each of the 5 components, on the left: the index of the component; the temporal mean, with the temporal quantiles as light shades of gray, all overlaid on the stimulus block reference. On the right: the three slices on top show the spatial mean, overlaid on a structural reference brain; and similarly on the bottom, the spatial variance, overlaid on the same reference. The bootstrap was performed 500 times and N shows how many estimates of each component were found.

To test the convergence of the estimated components, the normality of the difference between the estimated demixing vectors and the ground truth must be assessed. Since the whitening step can flip the signs of individual dimensions among the bootstrap rounds, the results include subgroups of estimated vectors, each of which with their own pattern of signs. To account for this, the estimates were reclustered using cosine similarity, with a suitably high threshold. Figure 2 shows the estimated demixing vectors for some of the example components.

The first example does not belong to a subspace and, apart from the obvious sign flips, the variation is generally quite small. Also, the covariance is nearly equal to the identity matrix. Five subgroups accounting for different configurations of sign flips were identified. Each of the subgroups was then tested against the best matching ground truth component. All except one passed the normality test with P-values of 0.1912, 0.0812, 0.1321 and 0.1615. The group that did not pass the test could include some outliers, even with a high clustering threshold.

The second example is part of a subspace, and the covariance shows clear block structure, *e.g.*, around coordinates 20 and 27. In this case, there are also five subgroups, but they all pass the normality test with P-values of 0.2530, 0.2076, 0.0613, 0.1290 and 0.0512. For the last example, the number of estimates is too small to allow for a normality test, but otherwise the situation seems similar to the previous.



(a)

(b)

(c)

Fig. 2. Estimated demixing vectors and their covariances. The groups of estimated demixing vectors, with different sign configurations are depicted on the left, and the covariance matrix of the estimates on the right. (a) applies to component 12, (b) to component 6, and (c) to component 27. The dashed line is the 0-vector, to make the sign mirroring easier to identify.

For the omitted components 9 and 28, the situation is very similar. All subgroups in component 9 pass the normality test and the covariance shows a weaker structure than in component 6. Component 28 has a similar covariance to component 27, and again too few estimates to allow for normality testing.

The normality tests show that the results are in very good agreement with the theory, even when the components are considered to belong to a subspace. This suggests that, even when some of the variability in the components is due to a covariation within the subspace, bootstrap FastICA is able to estimate reliable directions within the subspace. Although the estimated components passed the normality test, there should be some further evidence of the subspace covariation.

Figure 3 shows coordinate-wise histograms of the largest subgroup of vectors from component 6. Considered as a multivariate vector above, it passed the normality test. As expected, most of the dimensions are Gaussian. However, the estimates of coordinate 27 are clearly bimodal, and suggest that they in fact are from two different local minima. Coordinate 27 is also bimodal in some of the other subgroups and components belonging to the same subspace.

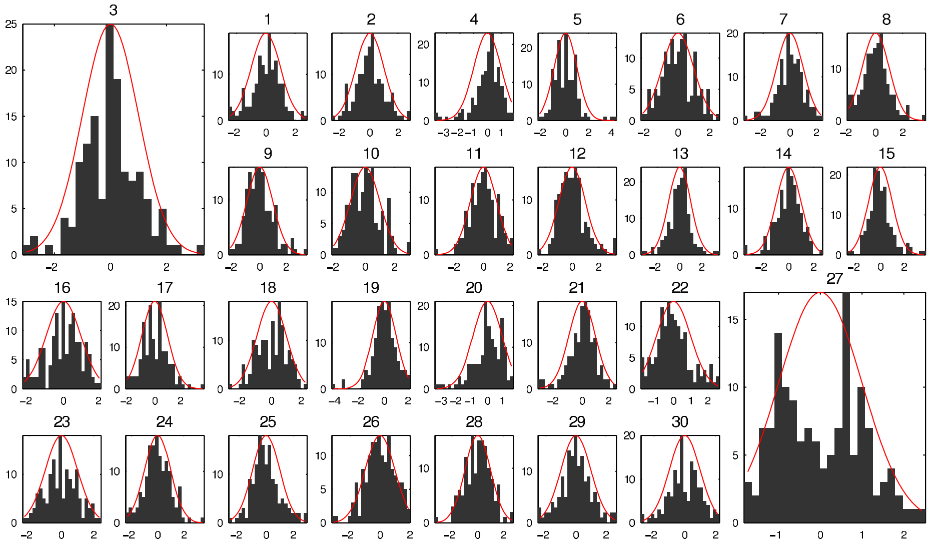


Fig. 3. Coordinate-wise histograms of the estimated demixing vectors. The histograms are calculated from the first subgroup of vectors in component 6. For reference, a fitted Gaussian probability density function is shown with a solid curve overlaid on each histogram. Two of the histograms are enlarged to highlight different shapes.

4 Discussion

In spite of the practical success of the multiple run approach to FastICA, no formal study on its limitations and validity existed before [13]. This paper further shows that the new theory holds with real-world data, and can be very useful at fully understanding the performance of the algorithm and, more importantly, the structure of the data.

However, there are some difficulties with real-world data, as all the assumptions may not be fulfilled. In particular, fMRI data may present non-stationarities and some dependence between components. This may result in a lower rate of convergence and emergence of subspaces. Still, the asymptotic normality of the convergence of both FastICA and bootstrapped FastICA suggests that both random initialization and subsampling decrease the likelihood of finding local optima, which results in more reliable and accurate estimation of the population solution. Also, in the case of subspace estimates, such formulation is still usable, if we have access to canonical directions within the subspace.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis: algorithms and applications. Wiley Interscience (2001)
2. Comon, P.: Independent Component Analysis, a new concept? Signal Processing 36(3), 287–314 (1994); Special issue on Higher-Order Statistics. hal-00417283
3. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. Wiley (2003)
4. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
5. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5), 411–430 (2000)
6. Oja, E., Yuan, Z.: The fastica algorithm revisited: Convergence analysis. IEEE Transactions on Neural Networks 17(6), 1370–1381 (2006)
7. Tichavsky, P., Koldovsky, Z., Oja, E.: Performance analysis of the fastica algorithm and cramer-rao bounds for linear independent component analysis. IEEE Transactions on Signal Processing 54(4), 1189–1203 (2006)
8. Ylipaavalniemi, J., Vigário, R.: Analyzing consistency of independent components: An fMRI illustration. NeuroImage 39(1), 169–180 (2008)
9. Harmeling, S., Meinecke, F., Müller, K.R.: Injecting noise for analysing the stability of ICA components. Signal Processing 84(2), 255–266 (2004)
10. Huettel, S.A., Song, A.W., McCarthy, G.: Functional Magnetic Resonance Imaging, 1st edn. Sinauer Associates, Sunderland (2004)
11. Ylipaavalniemi, J., Savia, E., Malinen, S., Hari, R., Vigário, R., Kaski, S.: Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli. NeuroImage 48(1), 176–185 (2009)
12. Ylipaavalniemi, J., Vigário, R.: Subspaces of Spatially Varying Independent Components in fMRI. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 665–672. Springer, Heidelberg (2007)
13. Reyhani, N., Ylipaavalniemi, J., Vigário, R., Oja, E.: Consistency and asymptotic normality of fastica and bootstrap fastica. Signal Processing (2011) (submitted)
14. van der Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes: With applications to Statistics. Springer, New York (1996)
15. Henze, N., Zirkler, B.: A class of invariant consistent tests for multivariate normality. Commun. Statist.-Theor. Meth. 19(10) (1990)
16. Anderson, T.: An Introduction to Multivariate Statistical Analysis. Wiley Interscience (2003)
17. Ylipaavalniemi, J., Soppela, J.: Arabica: Robust ICA in a Pipeline. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 379–386. Springer, Heidelberg (2009)

New Online EM Algorithms for General Hidden Markov Models. Application to the SLAM Problem

Sylvain Le Corff, Gersende Fort, and Eric Moulines*

LTCI, CNRS and TELECOM ParisTech,
46 rue Barrault 75634 Paris Cedex 13, France

Abstract. In this contribution, new online EM algorithms are proposed to perform inference in general hidden Markov models. These algorithms update the parameter at some deterministic times and use Sequential Monte Carlo methods to compute approximations of filtering distributions. Their convergence properties are addressed in [9] and [10]. In this paper, the performance of these algorithms are highlighted in the challenging framework of Simultaneous Localization and Mapping.

Keywords: Online Expectation-Maximization, Hidden Markov models, Statistical inference, SLAM.

1 Introduction

The Expectation Maximization (EM, [6]) algorithm is a versatile tool for maximum-likelihood based parameter estimation in latent data models. However, when processing large data sets or data stream, EM becomes intractable since it requires the whole data set to be available at each iteration of the algorithm.

In this contribution, we are interested in *online-EM* algorithms designed to deal with data which are available sequentially in time. Online-EM algorithms have been recently proposed. [4,14] address the case of independent and identically distributed (i.i.d.) observations. More complex incomplete data models such as Hidden Markov Models (HMM) are of common use to represent time series in many fields such as statistics, information engineering, signal processing, financial econometrics... [3] provides online-EM algorithms for HMM with finite state space. These algorithms have been extended to general HMM by [3,5] in the case of exponential complete-data likelihood, and by [8] for non exponential and general HMM. Hereafter, we will write "exponential HMM" as a shorthand expression for "HMM with exponential complete-data likelihood".

These online-EM algorithms for HMM are iterative algorithms. Each iteration consists in two steps: *(i)* the E step computes the expectation of the complete log-likelihood under the conditional distribution of the hidden states given

* This work is partially supported by the French National Research Agency, under the programs ANR-07-ROBO-0002 and ANR-08-BLAN-0218.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 131–138, 2012.

the observations (available up to the current time) and the current parameter; *(ii)* the M step updates the parameter as a maximum of this mean complete log-likelihood. Unfortunately, the algorithms mentioned above rely on many approximations. For example, the algorithms by [3,5,8] for general HMM combine stochastic approximation methods, Sequential Monte Carlo (SMC) for the approximation of the filtering distributions and an approximation of the recursive mechanism used to compute particle approximations of the filtering distributions. Therefore, it is really difficult to address the consistency of the estimators and to assert the convergence of these EM-based algorithms.

In this contribution, we propose new online-EM algorithms for general (and non necessarily exponential) HMM. The first algorithm, called *Block Online EM* (BOEM), is designed for exponential HMM such that the filtering distributions can be computed explicitly. Examples of such models are finite HMM and linear Gaussian models. The second algorithm is a SMC approximation of BOEM (so called Particle-BOEM or P-BOEM) designed for HMM with intractable E step. For both algorithms, we also propose *averaged* versions which have better convergence rates. All these algorithms are described in Section 2. The convergence of these algorithms (BOEM, P-BOEM and their averaged versions) is out of the scope of this paper: in [9,10], we provide sufficient conditions for these algorithms to converge to the set of the stationary points of the limiting log-likelihood of the observations. The convergence rates are also derived and it is proved that the averaged versions converge at a faster rate.

We provide in Section 3 an application of the P-BOEM algorithm to non-exponential HMM: P-BOEM is used as a new tool to solve the Simultaneous Localization And Mapping (SLAM) problem. We compare our algorithm to the *OnlineEM SLAM* of [8] and to *MarginalSLAM* of [12]. This numerical section highlights the interest of our algorithm to solve the SLAM problem.

2 New Online EM Algorithms for General HMM

The goal is to fit a HMM model on $\mathbb{Y}$-valued observations $\{\mathbf{Y}_t, t \geq 0\}$ sequentially available. We denote by $\{m_\theta(x,x')\mathrm{d}\lambda(x'), \theta \in \Theta\}$ (resp. $\{g_\theta(x,y)\mathrm{d}\nu(y), \theta \in \Theta\}$) the family of transition kernels onto $\mathbb{X}$ of the hidden states (resp. the conditional distribution of the observation given the hidden state). For simplicity, we assume that $\mathbb{X} \subseteq \mathbb{R}^{n_x}$, $\mathbb{Y} \subseteq \mathbb{R}^{n_y}$ and $\Theta \subseteq \mathbb{R}^{n_\theta}$. The initial distribution χ of the hidden state is assumed to be known. We propose algorithms for the computation of a parameter $\theta_\star$ maximizing the limiting normalized log-likelihood of the observations on this class of model indexed by Θ. We consider the case when m_θ, g_θ describes an exponential HMM i.e. there exist $S : \mathbb{X} \times \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^d$, $\psi : \Theta \to \mathbb{R}$ and $\phi : \Theta \to \mathbb{R}^d$ such that

$$\log\{m_\theta(x,x')g_\theta(x',y)\} = \phi(\theta) + \langle S(x,x',y); \psi(\theta)\rangle \; . \tag{1}$$

For $s \in \mathbb{R}^d$, define

$$\bar{\theta}(s) \overset{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} \; \phi(\theta) + \langle s; \psi(\theta)\rangle \; .$$

Given a set of observations $\mathbf{Y} = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_T\}$, the $(n+1)$-th E step of the *batch EM* algorithm would consist in computing

$$\mathcal{S}_T^{\mathrm{EM}}(\theta_n) \stackrel{\mathrm{def}}{=} \frac{1}{T} \sum_{t=1}^{T} \Phi^0_{\theta_n,t,T}\left(S, \mathbf{Y}\right) \ , \tag{2}$$

where $\Phi^0_{\theta,t,T}\left(S, \mathbf{Y}\right)$ denotes the expectation of the function S under the conditional distribution

$$\Phi^r_{\theta,s,t}(h, \mathbf{y}) \\ \stackrel{\mathrm{def}}{=} \frac{\int \chi(\mathrm{d}x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1})\}\, h(x_{s-1}, x_s, \mathbf{y}_s)\, \mathrm{d}\lambda(x_{r+1:t})}{\int \chi(\mathrm{d}x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1})\}\, \mathrm{d}\lambda(x_{r+1:t})} \ ; \tag{3}$$

and the M-step would update the parameter by $\theta_{n+1} = \bar{\theta}(\mathcal{S}_T^{\mathrm{EM}}(\theta_n))$. A natural extension to deal with sequential data is to update the parameter when a new observation is available. Therefore, the T-th update of this *Online EM* is computed from T observations by the iterative formula $\theta_{T+1} = \bar{\theta}(\mathcal{S}_T^{\mathrm{EM}}(\theta_T))$.

The new ideas of our approach is to update the parameter when a block of observations have been (sequentially) processed: more precisely, every time a new observation is available, the conditional expectation of the complete log-likelihood given the observations from the beginning of the block is updated. Due to the exponential assumption (1), such an update only requires an update of the filtering distribution. Then, at some times, the parameter is updated according to the same rule as in the EM algorithm. Let $\{\tau_n, n \geq 0\}$ be positive integers, and set $T_n \stackrel{\mathrm{def}}{=} T_{n-1} + \tau_n = \sum_{i=1}^{n} \tau_i$, $T_0 \stackrel{\mathrm{def}}{=} 0$. τ_n is the length of block n and the parameter will be updated at times T_n.

Block Online EM (BOEM) is an iterative algorithm: given the parameter θ_n updated at time T_n,

block E step compute the BOEM statistic

$$\mathcal{S}^{\mathrm{BOEM}}_{T_n,\tau_{n+1}}(\theta_n) \stackrel{\mathrm{def}}{=} \frac{1}{\tau} \sum_{t=T_n+1}^{T_n+\tau_{n+1}} \Phi^{T_n}_{\theta_n,t,T_n+\tau_{n+1}}\left(S, \mathbf{Y}\right) \ .$$

M step At time T_{n+1}, update the parameter $\theta_{n+1} = \bar{\theta}\left(\mathcal{S}^{\mathrm{BOEM}}_{T_n,\tau_{n+1}}(\theta_n)\right)$.

Note that the quantity $\mathcal{S}^{\mathrm{BOEM}}_{T_n,\tau_{n+1}}(\theta_n)$ corresponds to the intermediate quantity (2) computed with the observations $(\mathbf{Y}_{T_n+1}, \cdots, \mathbf{Y}_{T_n+\tau_{n+1}})$. This algorithm is fully online if the E-step can be processed online: the observations along block n have to be used once and the algorithm should not ask for a storage of the data. To that goal, the key property is to observe that (see e.g. [3])

$$\mathcal{S}^{\mathrm{BOEM}}_{T,\tau}(\theta) = \phi^\theta_{T,\tau}(R^\theta_{T,\tau}) \tag{4}$$

where $\phi^{\theta}_{T,t}$ is the filtering distribution at time t w.r.t. the parameter θ and the observations $(\mathbf{Y}_{T+1}, \cdots, \mathbf{Y}_{T+t})$, and the functions $R^{\theta}_{T,t} : \mathbb{X} \to \mathbb{R}^d$, $1 \leq t \leq \tau$, satisfy the following equation

$$R^{\theta}_{T,t}(x) = \frac{1}{t}\mathrm{B}^{\theta}_{T,t}\left(x, S(\cdot, x, Y_{T+t})\right) + \frac{t-1}{t}\mathrm{B}^{\theta}_{T,t}\left(x, R^{\theta}_{T,t-1}\right) , \qquad (5)$$

where B^{θ}_{t} denotes the backward smoothing kernel at time t: $\mathrm{B}^{\theta}_{T,t}(x, \mathrm{d}x') \propto m_{\theta}(x', x)\phi^{\theta}_{T,t-1}(\mathrm{d}x')$. By convention, $R^{\theta}_{T,0} = 0$.

When the expectation under the filtering distribution $\phi^{\theta}_{T,t}$ is intractable, it can be replaced by a particle approximation. This yields to the *Particle-BOEM* (P-BOEM) algorithm.

block Particle E step compute the P-BOEM statistic $\mathcal{S}^{\mathrm{P-BOEM}}_{N_{n+1},T_n,\tau_{n+1}}(\theta_n)$, defined as a SMC approximation of $\mathcal{S}^{\mathrm{BOEM}}_{T_n,\tau_{n+1}}(\theta_n)$ computed with N_{n+1} particles.

M step. At time T_{n+1}, update the parameter $\theta_{n+1} = \bar{\theta}\left(\mathcal{S}^{\mathrm{P-BOEM}}_{N_{n+1},T_n,\tau_{n+1}}(\theta_n)\right)$.

Here again, the Particle E step has to be computed online; this can be done by applying the algorithm of [3] (see also [5]), which consists in replacing the filtering distributions in Eqs (4) and (5), by a particle approximation.

Eq. (5) shows that the sufficient statistic along block n follows a stochastic approximation dynamic. It is known that the convergence of such algorithms can be improved by replacing the updated quantity with its *averaged one* (see [9]). In our case, this yields to the *averaged BOEM* algorithm: each block E step and M step of BOEM are followed by

averaged block E step. compute the statistic

$$\widetilde{\mathcal{S}}^{\mathrm{BOEM}}_{n+1} \overset{\mathrm{def}}{=} \frac{T_n}{T_{n+1}}\widetilde{\mathcal{S}}^{\mathrm{BOEM}}_{n} + \frac{\tau_{n+1}}{T_{n+1}}\mathcal{S}^{\mathrm{BOEM}}_{T_n,\tau_{n+1}}(\theta_n) = \frac{1}{T_{n+1}}\sum_{j=1}^{n}\tau_{j+1}\mathcal{S}^{\mathrm{BOEM}}_{T_j,\tau_{j+1}}(\theta_j) .$$

averaged block M step. Update the parameter $\tilde{\theta}_{n+1} = \bar{\theta}\left(\widetilde{\mathcal{S}}^{\mathrm{BOEM}}_{n+1}\right)$.

The same averaged steps can be done for the E and M P-BOEM steps, thus yielding to the *averaged P-BOEM*. The convergence properties of both BOEM and P-BOEM and their averaged versions have been derived in [9] and in [10]. These algorithms are seen as perturbations of a limiting EM recursion and it can be proved that they inherit the asymptotic behavior of this limiting EM. This has to be compared to the online EM of [5] which introduces many approximations and which theoretical analysis remains quite challenging.

3 Experiments

In this section, the performance of the algorithms presented in Section 2 are illustrated through Monte Carlo experiments. The SLAM problem has been addressed in different works [2]. When both the robot motion and the robot perception are perturbed by Gaussian noises, EKF-based algorithms proposed to approximate the joint distribution of the map and the robot pose. It has been successfully applied to numerous SLAM problems. Despite encouraging experimental results, the EKF-based SLAM algorithms do not converge due to the required Taylor expansion and the necessity to approximate a joint distribution between the pose and the map which is a static parameter, see [1,7]. On the other hand, the most famous SLAM solution proposed is the FastSLAM algorithm and its different variants, see [13]. In this case, the model is not linearized and the motion noise is not necessarily Gaussian. In the FastSLAM framework, the joint distribution of the robot trajectories and the map is approximated. The robot path is estimated with sequential Monte Carlo methods and, for each particle representing a trajectory, landmark positions are estimated using EKF steps. A linearization step is required to perform the update of each landmark position. Once again, experimental results and the possibility to keep a map estimate for each possible trajectory made these methods successful. However, the issue of the joint estimation of the static parameter and the robot path still remain: in this case it comes from the well known path degeneracy issue when computing joint distribution with SMC methods. As a map estimate is associated to each particle, after successive resampling steps, all the particles share the same estimation for old landmarks.

To overcome this difficulty, [12] introduced the *MarginalSLAM* algorithm and [8] the *OnlineEM SLAM*. The SLAM problem is seen as an inference task in HMM. The map parameterizes a latent data model and is estimated in the maximum likelihood sense. The localization procedure is answered by SMC methods. In [12], the map is estimated by a stochastic gradient algorithm (see e.g. [11]). In [8], this estimation procedure is replaced by an online EM based algorithm. In this paper, we propose to use the P-BOEM algorithm to sequentially estimate the map and to produce weighted particles to solve the localization problem. As said in Section 1, the convergence properties of P-BOEM have been addressed in [10], justifying the use of this algorithm to give a solution to the SLAM problem.

The robot evolves in a 2-dimensional landmark based map: its pose $x_t \overset{\mathrm{def}}{=} \{x_{t,i}\}_{i=1}^3$ consists in cartesian coordinates $x_{t,1}$ and $x_{t,2}$ and a heading direction $x_{t,3}$. At each time step, the robot motion is controlled by deterministic commands: a velocity v_t and a heading direction ψ_t. The evolution of the robot pose can be written:

$$x_t = f(x_{t-1}, \hat{v}_t, \hat{\psi}_t) \;, \tag{6}$$

where $(\hat{v}_t, \hat{\psi}_t) \sim \mathcal{N}_2(0, Q)$. Q is assumed to be known. From now on, f is the kinematic model of the front wheel of a bicycle (see e.g. [1]):

$$f(x_{t-1}, \hat{v}_t, \hat{\psi}_t) = x_{t-1} + \begin{pmatrix} \hat{v}_t d_t \cos(x_{t-1,3} + \hat{\psi}_t) \\ \hat{v}_t d_t \sin(x_{t-1,3} + \hat{\psi}_t) \\ \hat{v}_t d_t \frac{\sin(\hat{\psi}_t)}{B} \end{pmatrix} \;,$$

where d_t is the time period between two successive poses and B is the robot wheelbase.

Each landmark is represented by a vector θ_j. It is assumed that the total number of landmarks q and the association between observations and landmarks are known. The robot is equipped with range and bearing sensors: it observes the distance and the angular position of all landmarks in its neighborhood denoted by $\mathcal{A}_t$ at time t. The observation $y_{t,i} \in \mathbb{R}^2$ of the landmark i is written $y_{t,i} = h(x_t, \theta_{.,i}) + \delta_{t,i}$, where h is defined by

$$h(x,\tau) = \begin{pmatrix} \sqrt{(\tau_1 - x_1)^2 + (\tau_2 - x_2)^2} \\ \arctan \frac{\tau_2 - x_2}{\tau_1 - x_1} - x_3 \end{pmatrix} .$$

The noise vectors $\{\delta_{t,i}\}_{t,i}$ are i.i.d Gaussian $\mathcal{N}_2(0, R)$, where R is assumed to be known. In this example, the complete-data log-likelihood is not exponential. The marginal log-likelihood is written (up to an additive constant independent from θ),

$$\sum_{i \in \mathcal{A}_t} \ln g_\theta(x_t, y_{t,i}) \propto \sum_{i \in \mathcal{A}_t} [y_{t,i} - h(x_t, \theta_i)]^\star R^{-1} [y_{t,i} - h(x_t, \theta_i)] .$$

P-BOEM cannot be directly applied: therefore, at the beginning of each block, the function $\tau \mapsto h(x,\tau)$ is approximated by its first order Taylor expansion at all the current landmark estimates. This kind of first order approximations is of common use in the SLAM literature (e.g. in EKF-SLAM or in FastSLAM). In our case, this leads to a quadratic approximation of the likelihood of the observation and to an approximate exponential-HMM (see [8]).

Observations are sampled using $R = \text{diag}(\sigma_r^2, \sigma_b^2)$, where $\sigma_r = 0.5$m and $\sigma_b = \frac{\pi}{60}$rad. The robot path is sampled with a given set of controls and using $Q = \text{diag}(\sigma_v^2, \sigma_\phi^2)$ where $\sigma_v = 0.5\text{m.s}^{-1}$ and $\sigma_\psi = \frac{\pi}{60}$rad. In this experiment, the proposed algorithm is compared to the *MarginalSLAM* and to the *OnlineEM SLAM*. The block size sequence is slowly increasing $\{\tau_n \propto n^{1.1}\}_{n \geq 1}$ to allow a sufficiently large number of updates. The number of particles is constant on each block and fixed at 50. For the SMC step, new particles are sampled using the prior model, this method is known in the SMC literature as the Bootstrap filter. The step-size sequence used in the *MarginalSLAM* and in the *OnlineEM SLAM* for the stochastic approximation step are chosen such that $\gamma_n \propto n^{-0.8}$.

For each run the weighted mean of the particles and the estimated map are stored. Figure 1 displays the estimated path given by the *MarginalSLAM* and the P-BOEM SLAM for one of the 50 Monte Carlo runs. The path estimate given by the P-BOEM is clearly better than the one given by the *MarginalSLAM*.

Figure 2 displays boxplots of the landmark estimation error over 50 Monte Carlo runs for the *MarginalSLAM* and the P-BOEM SLAM. Both algorithms give similar results for the estimation of the landmarks observed at the beginning of the experiment. However, when considering the other landmarks, P-BOEM SLAM shows better results. Figure 3 compares the result given by the P-BOEM and the Online EM SLAM. As noted in [10], both algorithms have a similar behaviors.

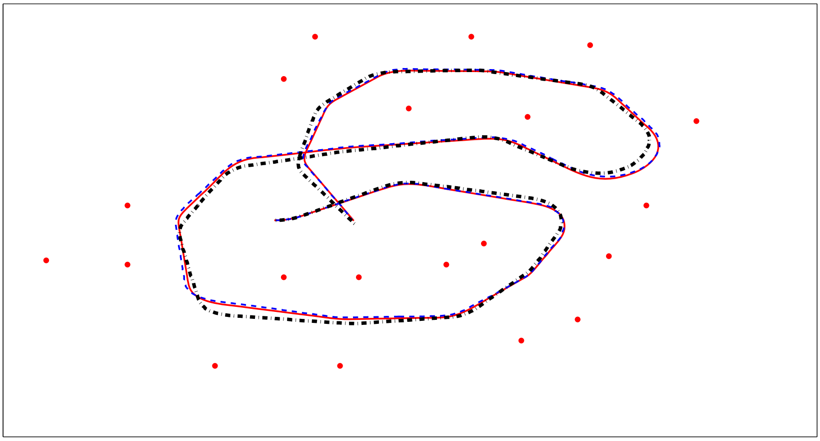


Fig. 1. True trajectory (bold line) and true landmark positions (balls) with the estimated path given by the P-BOEM SLAM (dashed line) and by the *MarginalSLAM* (dashed and dotted line)

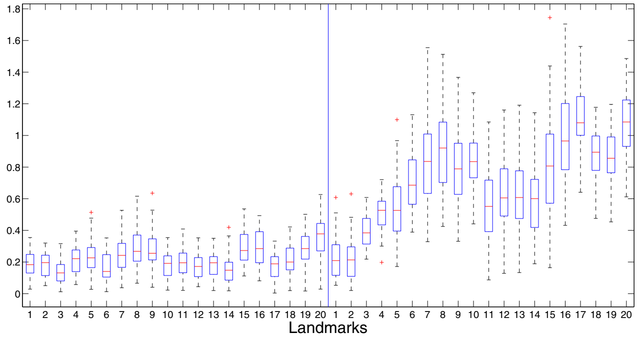


Fig. 2. Distance between the estimate at the end of the loop ($T = 1800$) and the true position using the P-BOEM SLAM (left) and the Marginal SLAM (right)

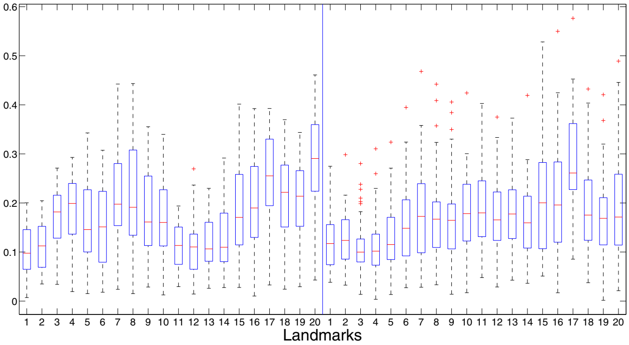


Fig. 3. Distance between the estimate at the end of the loop ($T = 1800$) and the true position using the P-BOEM SLAM (left) and the *OnlineEM SLAM* (right)

4 Conclusion

New algorithms for online Maximum-Likelihood based inference in exponential HMM have been proposed. These new online-EM procedures have been applied to solve the SLAM problem which is a case of non-exponential HMM. The experiments show that the our algorithm provides better result than the *Marginal-SLAM* algorithm when estimating the map online. The results are quite similar to those given by the online EM algorithm of [5]. Nevertheless, the asymptotic behavior of our algorithms has been addressed showing that they answer to the Maximum-Likelihood estimation problem. On the contrary, it remains quite challenging to analyze the convergence properties of the online EM of [5].

References

1. Bailey, T., Nieto, J., Guivant, J., Stevens, M., Nebot, E.: Consistency of the EKF-SLAM algorithm. In: IEEE International Conference on Intelligent Robots and Systems, pp. 3562–3568 (2006)
2. Burgard, W., Fox, D., Thrun, S.: Probabilistic robotics. MIT Press, Cambridge (2005)
3. Cappé, O.: Online EM algorithm for Hidden Markov Models. To Appear in J. Comput. Graph. Statist. (2011)
4. Cappé, O., Moulines, E.: Online Expectation Maximization Algorithm for Latent Data Models. J. Roy. Statist. Soc. B, 593–613 (2009)
5. Del Moral, P., Doucet, A., Singh, S.S.: Forward smoothing using sequential Monte Carlo. arXiv:1012.5390v1 (2011)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B, 1–38 (1977)
7. Julier, S.J., Uhlmann, J.K.: A counter example to the theory of simultaneous localization and map building. In: IEEE International Conference on Robotics and Automation, pp. 4238–4243 (2001)
8. Le Corff, S., Fort, G., Moulines, E.: Online EM algorithm to solve the SLAM problem. In: IEEE Workshop on Statistical Signal Processing (2011)
9. Le Corff, S., Fort, G.: Online Expectation Maximization based algorithms for inference in Hidden Markov Models. arXiv:1108.3968 (2011)
10. Le Corff, S., Fort, G.: Convergence of a particle-based approximation of the Block Online Expectation Maximization algorithm. arXiv:1111.1307 (2011)
11. Le Gland, F., Mevel, L.: Recursive estimation in HMMs. In: IEEE Conference on Decision and Control, pp. 3468–3473 (1997)
12. Martinez-Cantin, R.: Active map learning for robots: insights into statistical consistency. PhD thesis (2008)
13. Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B.: FastSLAM 2.0: An Improved Particle Filtering Algorithm for Simultaneous Localization and Mapping that Provably Converges. In: IJCAI (2003)
14. Titterington, D.M.: Recursive parameter estimation using incomplete data. J. Roy. Statist. Soc. B, 257–267 (1984)

The Role of Whitening for Separation of Synchronous Sources

Miguel Almeida[1,2,⋆], Ricardo Vigário[2], and José Bioucas-Dias[1]

[1] Institute of Telecommunications, Instituto Superior Técnico, Lisbon, Portugal
{miguel.almeida,bioucas}@lx.it.pt

[2] Adaptive Informatics Research Centre, Aalto University School of Science, Finland
ricardo.vigario@aalto.fi

Abstract. The separation of synchronous sources (SSS) is a relevant problem in the analysis of electroencephalogram (EEG) and magnetoencephalogram (MEG) synchrony. Previous experimental results, using pseudo-real MEG data, showed empirically that prewhitening improves the conditioning of the SSS problem. Simulations with synthetic data also suggest that the mixing matrix is much better conditioned after whitening is performed. Unlike in Independent Component Analysis (ICA), synchronous sources can be correlated. Thus, the reasoning used to motivate whitening in ICA is not directly extendable to SSS. In this paper, we analytically derive a tight upper bound for the condition number of the equivalent mixing matrix after whitening. We also present examples with simulated data, showing the correctness of this bound on sources with sub- and super-gaussian amplitudes. These examples further illustrate the large improvements in the condition number of the mixing matrix obtained through prewhitening, thus motivating the use of prewhitening in real applications.

Keywords: whitening, source separation, independent component analysis (ICA), synchrony, phase-locking factor (PLF), condition number.

1 Introduction

Research on the topic of synchrony has gained momentum in recent years. It can be studied under an elegant mathematical framework applicable to many different fields such as laser interferometry, the pull of interstellar objects, and the human brain [9]. Synchrony is believed to play an important role in the interaction of distinct brain regions. For example, a muscle's electromyogram oscillates coherently with several brain regions, when a person is involved in a motor task [8,10]. Memorization, learning, autism, Alzheimer's, Parkinson's, and epilepsy are examples of neuroscience topics associated with synchrony [11].

Inference about the synchrony of the networks present in the brain (or other real-world systems) requires access to the dynamics of the individual oscillators (the "sources"). However, in the brain's electroencephalogram (EEG) and

[⋆] Corresponding author.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 139–146, 2012.

magnetoencephalogram (MEG), the signals from individual oscillators are not directly measurable; one has only access to a superposition of the sources. This is known as the "cross-talk effect" in the field of EEG and MEG research [7]. In this case, spurious synchrony occurs, as has been shown both empirically and analytically [2].

Reversing this superposition is usually called blind source separation (BSS). Usually, it is assumed that the mixing process is linear and instantaneous, a valid approximation in, *e.g.*, brain signals [12]. Let the vector of sources be denoted by $\mathbf{s}(t)$ and the vector of measurements by $\mathbf{y}(t)$. They are related through the model $\mathbf{y}(t) = \mathbf{M}\mathbf{s}(t)$, where $\mathbf{M}$ is a real-valued mixing matrix. The BSS problem has infinitely many solutions. Therefore, assumptions are necessary to adequately pose the problem, such as the independence of the sources, as in Independent Component Analysis (ICA) [6]. However, in the case discussed here, independence is not a valid assumption, because synchronous sources are highly dependent.

We have previously introduced two algorithms to perform Synchronous Source Separation (SSS): Independent Phase Analysis (IPA), a data-driven approach [2], and Phase Locked Matrix Factorization (PLMF), a model-driven approach [3]. Furthermore, we have empirically verified, both with simulated data [2] and with pseudo-real MEG data [1], that prewhitening the data severely improves the quality of the results obtained with these algorithms. However, those empirical findings had no theoretical support. The goal of this paper is to study why prewhitening improves the results of SSS algorithms. We will derive a tight upper bound for the condition number of the problem after prewhitening. We will also present experimental evidence that corroborate this analytical result.

2 Background

2.1 Phase-Locking Factor

Let $\phi_j(t)$ and $\phi_k(t)$, for $t = 1, \ldots, T$, be the time-dependent phases of signals j and k. The real-valued Phase Locking Factor (PLF) between those signals is

$$\varrho_{jk} \equiv \left| \frac{1}{T} \sum_{t=1}^{T} e^{\mathrm{i}[\phi_j(t) - \phi_k(t)]} \right| = \left| \left\langle e^{\mathrm{i}(\phi_j - \phi_k)} \right\rangle \right|, \tag{1}$$

where $\langle \cdot \rangle$ is the time average operator, and $\mathrm{i} = \sqrt{-1}$. Note that $0 \leq \varrho_{jk} \leq 1$. Importantly, the value $\varrho_{jk} = 1$ corresponds to two signals that are fully synchronized: their phase lag, defined as $\phi_j(t) - \phi_k(t)$, is constant. The value $\varrho_{jk} = 0$ is attained if $\phi_j(t) - \phi_k(t)$ is uniformly distributed in $[0,2\pi)$. Values between 0 and 1 represent partial synchrony. Note that a signal's PLF with itself is trivially equal to 1: thus, for all j, $\varrho_{jj} = 1$.

2.2 Whitening

Assume that there is an N by T source matrix $\mathbf{S}$, such that its (j, t) element is the j-th complex-valued source at time t, $s_j(t)$. Each component of $\mathbf{s}(t)$ is assumed to have zero mean, *i.e.*, $\mathrm{E}[\mathbf{s}(t)] = \mathbf{0}$. We also assume that these sources are unknown, but that we can observe a set of measurements $\mathbf{y}(t)$, which are obtained from the sources through $\mathbf{y}(t) = \mathbf{M}\mathbf{s}(t)$ (note that $\mathbf{y}(t)$ also has zero mean), where $\mathbf{M}$ is a square real-valued matrix with full rank.[1] If one also stores successive samples of $\mathbf{y}(t)$ in a matrix $\mathbf{Y}$, then $\mathbf{Y} = \mathbf{MS}$.

Whitening is a process which involves multiplying the data $\mathbf{y}(t)$ by a square matrix $\mathbf{B}$, such that the resulting vector, $\mathbf{z}(t) \equiv \mathbf{B}\mathbf{y}(t)$, has as covariance the identity matrix. There are infinitely many possible matrices $\mathbf{B}$ which can achieve this; one possibility[2] is to have $\mathbf{B} = \mathbf{C}_{\mathbf{Y}}^{-1/2}$, where $\mathbf{C}_{\mathbf{Y}}$ is the covariance matrix of $\mathbf{y}(t)$. The original BSS problem $\mathbf{Y} = \mathbf{MS}$ is thus transformed into an equivalent one $\mathbf{Z} = \mathbf{BMS}$; $\mathbf{BM}$ is called the equivalent mixing matrix.

The BSS community, in particular the users of ICA, have advocated the use of whitening as a preprocessing step [6], because if the sources $\mathbf{s}(t)$ have the identity matrix as their covariance matrix (which is always true if they are independent, up to trivial scalar factors), then the equivalent mixing matrix $\mathbf{BM}$ is necessarily an orthogonal matrix. This means that ICA algorithms can restrict themselves to finding an orthogonal matrix, which makes the ICA problem considerably easier [6]. However, in SSS the sources are highly dependent.

2.3 Condition Number

It is well known that the difficulty of solving linear inverse problems, such as ICA and SSS, can be roughly characterized by the condition number of matrix $\mathbf{M}$ [4]. The condition number of a matrix $\mathbf{M}$ is defined[3] as the quotient $\rho = \frac{\sigma_{max}}{\sigma_{min}}$, where σ_{max} is the largest singular value of $\mathbf{M}$ and σ_{min} is its smallest singular value. The condition number obeys $\rho \geq 1$ for any matrix. Problems with a lower ρ are, in general, easier than problems with a higher ρ, even though this number does not fully characterize the difficulty of these problems [4].

The condition number of a BSS problem depends on the unknown matrix $\mathbf{M}$. In ICA, after prewhitening the inverse problem has $\rho = 1$ [6]. Such is not the case for SSS; however, we will show that an upper bound for this condition number can be derived using prewhitening. We also show experimentally that large improvements on the condition number can be obtained through this process.

3 Upper Bound for Condition Number after Prewhitening

3.1 Notation and Assumptions

Let $\mathbf{S}$ denote a complex-valued N-by-T matrix with the value of the sources, where the (j,t) element of $\mathbf{S}$ contains $s_j(t)$. We decompose $s_j(t) \equiv a_j(t)e^{\mathrm{i}\phi_j(t)}$

[1] This reasoning can be easily extended to $\mathbf{M}$ having more rows than columns [1,6].

[2] In this paper, square roots are taken only of Hermitian positive semidefinite matrices. If $\mathbf{A} = \mathbf{VDV}^{\mathsf{H}}$ is the eigendecomposition of $\mathbf{A}$, we define $\mathbf{A}^{1/2} \equiv \mathbf{VD}^{1/2}\mathbf{V}^{\mathsf{H}}$.

[3] Other definitions of condition number exist. The one presented here is quite common, and will be used throughout the paper.

where $\mathrm{i} = \sqrt{-1}$, $a_j(t)$ is the real-valued, non-negative amplitude, and $\phi_j(t)$ is a real number in the interval $[0, 2\pi)$, called phase. The amplitudes of each of the N sources are considered random variables A_j, *i.e.*, each time point $a_j(t)$ is i.i.d., drawn from a certain probability density distribution. The phases of the sources are also considered random variables, although not independent of each other, as detailed below.

Our goal is to study the simplest case applicable to SSS. We make the following assumptions:

- A_j is independent of A_k for $j \neq k$;
- A_j is independent of ϕ_k for any j and k, including for $j = k$;
- All A_j have the same distribution, which is generally denoted as A;
- ϕ_j and ϕ_k have maximum PLF, *i.e.* they have a constant phase lag;
- The previous point implies that $\phi_j(t) = \phi_j(1) + \Phi(t)$ for all j and t. We assume that $\Phi(t)$ is uniformly distributed in $[0,2\pi)$.

Note that this is still a harder problem than ICA, because $\phi_j(t)$ and $\phi_k(t)$ are strongly dependent. Nevertheless, algorithmic solutions exist that can extract the matrix $\mathbf{S}$ using only information from the observations $\mathbf{Y} = \mathbf{MS}$ [2,3].

3.2 Upper Bound

Multiplying the data by $\mathbf{B} = \mathbf{C}_{\mathbf{Y}}^{-1/2}$ will, in general, result in an equivalent mixing matrix $\mathbf{BM}$ which is complex, even if $\mathbf{M}$ is real. This is a disadvantage, if one aims to use algorithms which search for real separation matrices [2,3]. To take this into account, one can consider the two following formulations, which are equivalent to $\mathbf{Y} = \mathbf{MS}$ with the constraint of real $\mathbf{M}$:

$$\begin{bmatrix} \mathbf{Y}_R \\ \mathbf{Y}_I \end{bmatrix} = \begin{bmatrix} \mathbf{M} \; \mathbf{0} \\ \mathbf{0} \; \mathbf{M} \end{bmatrix} \begin{bmatrix} \mathbf{S}_R \\ \mathbf{S}_I \end{bmatrix} \text{ or } [\mathbf{Y}_R \; \mathbf{Y}_I] = \mathbf{M} [\mathbf{S}_R \; \mathbf{S}_I], \tag{2}$$

where $\mathbf{S}_R \equiv \mathrm{Real}(\mathbf{S})$, $\mathbf{S}_I \equiv \mathrm{Imag}(\mathbf{S})$, and similarly for $\mathbf{Y}_R$ and $\mathbf{Y}_I$. $\mathbf{0}$ is a matrix filled with zeros with the same size as $\mathbf{M}$.

Although both formulations allow the derivations done below, the equivalent mixing matrix is, on average, farther from the bound (and thus, better conditioned) on the second case; therefore, we use that formulation and define $\mathbf{S}_{RI} \equiv [\mathbf{S}_R \; \mathbf{S}_I]$ and similarly for $\mathbf{Y}_{RI}$.[4]

In this formulation, prewhitening involves multiplying the new data $\mathbf{Y}_{RI}$ by a new whitening matrix $\mathbf{B} \equiv \mathbf{C}_{\mathbf{Y}_{RI}}^{-1/2}$ such that $\mathbf{Z}_{RI} \equiv \mathbf{B}\mathbf{Y}_{RI} = \mathbf{BMS}_{RI}$ has as correlation the identity matrix. Since

[4] The second formulation has a trickier interpretation, since $[\mathbf{S}_R \; \mathbf{S}_I]$ is no longer a matrix whose columns are realizations of one random variable. Consequently, the term "correlation matrix" for $\mathbf{C}_{\mathbf{S}_{RI}}$ is somewhat abusive. Formally, we define the "correlation matrix" of $\mathbf{S}_{RI}$ as $\mathbf{C}_{\mathbf{S}_{RI}} \equiv \frac{\mathbf{S}_{RI}\mathbf{S}_{RI}^{\mathsf{T}}}{2T}$; technically, it corresponds to the correlation matrix of a random variable which can take the value $\mathrm{Real}(\mathbf{s})$ and $\mathrm{Imag}(\mathbf{s})$ with equal probability. Similar considerations hold for $\mathbf{C}_{\mathbf{Y}_{RI}}$ and $\mathbf{C}_{\mathbf{Z}_{RI}}$.

$$\mathbf{C}_{\mathbf{Z}_{RI}} = \mathbf{B}\mathbf{M}\mathbf{C}_{\mathbf{S}_{RI}}\mathbf{M}^{\mathsf{T}}\mathbf{B}^{\mathsf{H}} = (\mathbf{B}\mathbf{M}\mathbf{C}_{\mathbf{S}_{RI}}^{1/2})(\mathbf{B}\mathbf{M}\mathbf{C}_{\mathbf{S}_{RI}}^{1/2})^{\mathsf{H}} = \mathbf{I}, \tag{3}$$

one can conclude that $\mathbf{B}\mathbf{M}\mathbf{C}_{\mathbf{S}_{RI}}^{1/2}$ is a unitary matrix, which we denote by $\mathbf{R}$.

We can now study the singular values of the equivalent mixing matrix $\mathbf{BM}$. It holds that $\mathbf{BM} = \mathbf{R}\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}$, and that the singular values of $\mathbf{R}\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}$ are the same as those of $\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}$, since they differ only by a multiplication by a unitary matrix. Therefore, the conditioning of the equivalent source separation problem can be studied by studying the singular values of $\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}$.

Note that $\mathbf{C}_{\mathbf{S}_{RI}} = 1/2(\mathbf{C}_{\mathbf{S}_R} + \mathbf{C}_{\mathbf{S}_I})$. Using the assumptions from Section 3.1 yields, for the diagonal elements of $\mathbf{C}_{\mathbf{S}_{RI}}$,

$$[\mathbf{C}_{\mathbf{S}_{RI}}]_{jj} = \frac{1}{2}\mathrm{E}[A^2], \tag{4}$$

whereas the off-diagonal elements of $\mathbf{C}_{\mathbf{S}_{RI}}$ can be shown to be

$$[\mathbf{C}_{\mathbf{S}_{RI}}]_{jk} = \frac{1}{2}\mathrm{E}[A]^2\cos(\phi_j(1) - \phi_k(1)). \tag{5}$$

Since $\mathrm{E}[A^2] = \mathrm{Var}[A] + \mathrm{E}[A]^2$, for any random variable A, we get

$$\mathbf{C}_{\mathbf{S}_{RI}} = \frac{\mathrm{Var}[A]\mathbf{I} + \mathrm{E}[A]^2\mathbf{F}}{2}, \tag{6}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{F}_{jk} \equiv \cos(\phi_j(1) - \phi_k(1))$.

We now study the eigenvalues of matrix $\mathbf{F}$, which are equal to its singular values, since $\mathbf{F}$ is symmetric and positive semidefinite (as shown below). It is easy to see that $\mathbf{F} = \mathrm{Re}(\mathbf{G})$, with $\mathbf{G} \equiv \mathbf{x}\mathbf{x}^{\mathsf{H}}$, where the vector $\mathbf{x}$ has in its j-th component $x_j \equiv e^{\mathrm{i}\phi_j(1)}$. $\mathbf{G}$ has simple eigenvalue $\lambda_{\mathbf{G}} = N$ (the number of sources), and an eigenvalue $\lambda_{\mathbf{G}} = 0$ with multiplicity $N-1$.

Since the eigenvalues of $\mathbf{G}$ are 0 and N, the eigenvalues of $\mathbf{F}$ necessarily obey $0 \le \lambda_{\mathbf{F}} \le N$. To see this, let $\mathbf{v}$ be any real vector with unit norm. Note that since $\mathbf{v}$ is real, we have $\mathbf{v}^{\mathsf{H}} = \mathbf{v}^{\mathsf{T}}$. Then,

$$\mathbf{v}^{\mathsf{H}}\mathbf{G}\mathbf{v} = \mathbf{v}^{\mathsf{T}}\mathbf{G}\mathbf{v} = \mathbf{v}^{\mathsf{T}}\mathbf{F}\mathbf{v} + \mathbf{v}^{\mathsf{T}}\mathrm{Im}(\mathbf{G})\mathbf{v} = \mathbf{v}^{\mathsf{T}}\mathbf{F}\mathbf{v}, \tag{7}$$

where $\mathbf{v}^{\mathsf{T}}\mathrm{Im}(\mathbf{G})\mathbf{v} = 0$ because $\mathbf{G}$ is Hermitian, thus its imaginary part is skew-symmetric. The leftmost expression is valued between 0 and N, since those are the smallest and largest eigenvalues of $\mathbf{G}$. Thus the rightmost expression must also be within those values. Therefore, the eigenvalues of $\mathbf{F}$ obey $0 \le \lambda_{\mathbf{F}} \le N$.

With these bounds for $\lambda_{\mathbf{F}}$, one can immediately conclude that the eigenvalues of $\mathbf{C}_{\mathbf{S}_{RI}}$ obey

$$\frac{\mathrm{Var}[A]}{2} \le \lambda_{\mathbf{C}_{\mathbf{S}_{RI}}} \le \frac{\mathrm{Var}[A] + N\mathrm{E}[A]^2}{2}. \tag{8}$$

Thus, the condition number of $\mathbf{C}_{\mathbf{S}_{RI}}$ is bounded above by the quotient of these two bounds: $\rho(\mathbf{C}_{\mathbf{S}_{RI}}) \leq 1+N\frac{\mathrm{E}[A]^2}{\mathrm{Var}[A]}$. Also, from simple properties of the condition number, one can conclude that

$$\rho(\mathbf{BM}) = \rho(\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}) = \sqrt{\rho\left(\mathbf{C}_{\mathbf{S}_{RI}}^{-1}\right)} = \sqrt{\rho\left(\mathbf{C}_{\mathbf{S}_{RI}}\right)} \leq \sqrt{1+N\frac{\mathrm{E}[A]^2}{\mathrm{Var}[A]}}. \tag{9}$$

The proof that this upper bound is tight is very simple. It is sufficient to consider the case $\phi_j(1) = \phi_k(1)$ for all j, k, *i.e.*, the situation where all sources have zero phase lag with one another. In that case, $\mathbf{F}$ is a matrix full of ones, and its eigenvalues are exactly 0 and N. It is very simple to see that in that case, $\rho(\mathbf{C}_{\mathbf{S}_{RI}}^{-1/2}) = \sqrt{1+N\frac{\mathrm{E}[A]^2}{\mathrm{Var}[A]}}$ holds.

4 Experiments

The above result is derived for the ideal case, where the assumptions of Section 3.1 are valid. However, in real data these assumptions will never hold, because the number of time points is finite.[5] Therefore, we now study whether this upper bound expression is useful in practice, using small simulated examples.

We generate each set of data in the following way: the initial phase for each source, $\phi_j(1)$, is randomly drawn from a uniform distribution between 0 and 2π. The common phase oscillation, $\Phi(t)$, is given by $\Phi(t) = \omega t$ with $\omega = 0.02\pi$. The amplitudes $a_j(t)$ are independently drawn from the Gamma probability distribution with unit scale parameter and shape parameter equal to $1, 2, 3$. A similar range was used for the Irwin-Hall probability distribution. This corresponds to the sums of one, two, or three Exponential or Uniform distributions, thus representing different values of kurtosis. The mixing matrix $\mathbf{M}$ has each of its elements independently drawn from a Uniform(-1,1) distribution.

Each of these datasets has $T = 10000$ time points and $N = 4$ sources and measurements. We generate 1000 such datasets for each of the six distributions. We then make a scatter plot comparing the condition number of the original mixing matrix $\mathbf{M}$ (in the horizontal axis) with the condition number of the equivalent mixing matrix $\mathbf{BM}$ (in the vertical axis). Each of these plots also shows the theoretical value of the upper bound from eq. (9), drawn as a horizontal line. These plots are shown in Figure 1.

At first glance it might seem unexpected that some points are slightly above the line of the upper bound. This is justified by the difference between the ideal case with $T = \infty$, which was used to derive the bound, and the experimental case with a finite T. In other words, it is a consequence of using a sample covariance matrix instead of the true covariance matrix. Nevertheless, the fraction of points above the line is very small, as is the vertical gap between those points and the line. As the number of time points T approaches infinity, the fraction of points above the line and their gap tends to zero.

[5] This is similar to the ICA case, where although independence of the sources is assumed, it is not verified precisely in real cases.

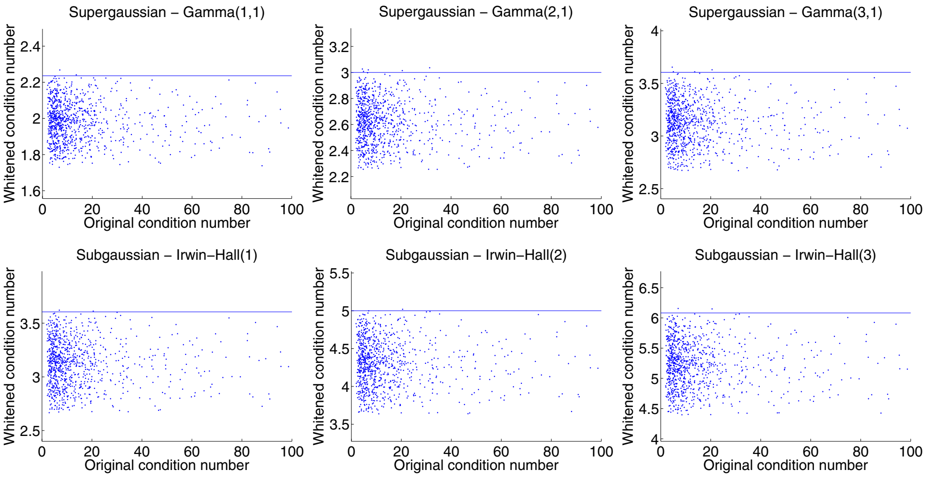


Fig. 1. Experimental confirmation of the upper bound. The condition number of the mixing matrix is displayed on the horizontal axis, and that of the equivalent mixing matrix on the vertical axis, for six different distributions for A. The horizontal line corresponds to the upper bound in eq. (9).

5 Discussion

Several directions can be taken to extend this result. One such direction is to derive an upper bound for cases where the PLF between the sources is smaller than 1. The case of zero PLF includes the ICA case; in that case, the condition number of the equivalent mixing matrix is known to be 1 [6]. In general, we believe that the upper bound for the PLF < 1 case will be smaller than the one derived here; however, it will probably depend on the specific form of the sources' phases: there may be i.i.d. random phase noise, as we studied previously [1], there may be phase slips [9], or other types of imperfect phase-locking.

Another very important direction is to obtain a result on the probability of a case with a finite number of points T, to have a condition number higher than the bound derived here for $T = \infty$. In other words, it would be interesting to know in anticipation how many points will, on average, end up above the horizontal line in Figure 1. This result would necessarily depend on T, on the distribution of the amplitudes A, and on how the mixing matrix $\mathbf{M}$ is generated.

Yet another useful extension is to remove the assumption that all amplitudes are drawn from the same distribution A. As long as the amplitudes are independent of each other and of the phases, the reasoning used throughout this paper stands, although the mathematical expressions involved become less elegant.

Finally, one could also relax the assumption that the amplitudes and phases are independent, since studies have shown that the power of given brain oscillations may be locked to the phases of other oscillations [5]. This requires the assumption of a specific dependency between the amplitudes and phases.

6 Conclusion

We have derived an upper bound for the condition number of an SSS problem, if whitening is performed as pre-processing. Experimental results confirm the validity of this upper bound. The main conclusion is that in virtually any situation, it is advantageous to use whitening as a pre-processing step, even when there is a degree of dependence in the sources.

Acknowledgements. This work was supported by FCT project PEst-OE / EEI / LA0008 / 2011, under internal project DECA-Bio, and by the Academy of Finland through its Centres of Excellence Program 2006-2011.

References

1. Almeida, M., Bioucas-Dias, J., Vigário, R.: Detection and separation of phase-locked subspaces with phase noise. Signal Processing (2011) (submitted)
2. Almeida, M., Schleimer, J.H., Bioucas-Dias, J., Vigário, R.: Source separation and clustering of phase-locked subspaces. IEEE Transactions on Neural Networks 22(9), 1419–1434 (2011)
3. Almeida, M., Vigario, R., Bioucas-Dias, J.: Phase locked matrix factorization. In: Proc. of the EUSIPCO Conference (2011)
4. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. Taylor & Francis (1998)
5. Canolty, R., Edwards, E., Dalal, S., Soltani, M., Nagarajan, S., Kirsch, H., Berger, M., Barbaro, N., Knight, R.: High gamma power is phase-locked to theta oscillations in human neocortex. Science 313, 1626–1628 (2006)
6. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2001)
7. Nunez, P.L., Srinivasan, R., Westdorp, A.F., Wijesinghe, R.S., Tucker, D.M., Silberstein, R.B., Cadusch, P.J.: EEG coherency I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. Electroencephalography and Clinical Neurophysiology 103, 499–515 (1997)
8. Palva, J.M., Palva, S., Kaila, K.: Phase synchrony among neuronal oscillations in the human cortex. Journal of Neuroscience 25(15), 3962–3972 (2005)
9. Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization: A universal concept in nonlinear sciences. Cambridge Nonlinear Science Series. Cambridge University Press (2001)
10. Schoffelen, J.M., Oostenveld, R., Fries, P.: Imaging the human motor system's beta-band synchronization during isometric contraction. NeuroImage 41, 437–447 (2008)
11. Uhlhaas, P.J., Singer, W.: Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology. Neuron 52, 155–168 (2006)
12. Vigário, R., Särelä, J., Jousmäki, V., Hämäläinen, M., Oja, E.: Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans. On Biom. Eng. 47(5), 589–593 (2000)

Simultaneous Diagonalization of Skew-Symmetric Matrices in the Symplectic Group

Frank C. Meinecke

Machine Learning Group, Institute for Computer Science, TU Berlin, Germany
frank.meinecke@tu-berlin.de

Abstract. Many source separation algorithms rely on the approximate simultaneous diagonalization of matrices. While there exist very efficient algorithms for symmetric matrices, the skew-symmetric case turned out to be more difficult. Here we show how the often used whitening/rotation approach for symmetric matrices can be translated to this case. While the former leads to orthogonal transformations in Euclidean space, the latter leads to symplectic transformations in symplectic space. It is demonstrated that the resulting algorithm is more stable than a naïve diagonalization that does not respect the symplectic structure of the problem.

Keywords: Diagonalization, Skew-symmetric matrix, Pairwise Interacting Source Analysis, Symplectic Group.

1 Introduction

Blind source separation (BSS) problems can often be solved by the approximate simultaneous diagonalization of suitably defined square matrices. Given such a set of matrices $\{M(k)|k = 1, \dots, K\}$, the BSS task is then formulated as the search for a transformation matrix B such that for all k

$$BM(k)B^\top \to diag. \tag{1}$$

For instance, consider a multivariate time series $x(t)$, which is a linear mixture of temporally uncorrelated source signals $s(t)$, i.e. $x(t) = As(t)$. Then the algorithms TDSEP [8] and SOBI [1] aim at the simultaneous diagonalization of symmetrized time-lagged covariance matrices $M(\tau) = \Sigma_x(\tau) + \Sigma_x^\top(\tau)$, where

$$\Sigma_x(\tau) = \mathrm{E}[x(t)\, x^\top(t-\tau)]. \tag{2}$$

Temporal uncorrelatedness implies that the time-lagged covariances $\Sigma_s(\tau)$ of the original sources are diagonal. Therefore, diagonalizing $M(\tau)$ corresponds to solving the source separation task and the matrix B is an estimate of the inverse of the mixing matrix A (up to scaling and ordering of its rows). Similarly, in the JADE [2] algorithm, the matrices to be diagonalized are 'slices' of the fourth-order cumulant tensor (or equivalently, eigenmatrices of this tensor). This relies

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 147–154, 2012.

on the assumption that the source signals are statistically independent at zero time lag, which makes these matrices diagonal in the source signal basis.

The matrices mentioned in these examples were all symmetric, i.e. $M^\top = M$. This allows to determine the demixing matrix B in two steps: first a spatial *whitening* transforms one of the matrices (usually the data covariance) to the identity matrix, which constrains the remaining transformation to the set of orthogonal matrices.[1] This reduction to the set of orthogonal matrices stabilizes the optimization in the sense that the BSS algorithm cannot converge to singular or ill-conditioned matrices. Another advantage of this two-step procedure is that there exist very efficient joint diagonalization methods using orthogonal matrices. However, not all source separation problems can be cast in terms of symmetric matrices. In the following, we show how this whitening-rotation scheme can be translated to the case of skew-symmetric matrices.

2 Interacting Sources, Skew-Symmetric Matrices and the Induced Symplectic Geometry

Skew-symmetric matrices play a central role e.g. in Pairwise Interacting Source Analysis (PISA) [6]. In contrast to the BSS algorithms mentioned in the introduction, PISA does not assume independent source signals; instead the sources are assumed to be pairwise synchronized such that for $2n$ source signals, the i–th source ($i < n$) is synchronized with the $(i + n)$–th source, but independent of all other sources. Synchrony between signals can be quantified by their cross-spectrum; however, as shown in [5], only its imaginary part is a *reliable* measure of interactions in the sense that it *definitely* encodes synchronization and *cannot* be explained by linear mixtures of independent sources. Therefore, PISA is based on the imaginary part of the complex cross-spectral matrices $\varGamma_x(\omega) = \Im\,\mathrm{E}[\hat{x}(\omega)\hat{x}(\omega)^\dagger]$ where $\hat{x}(\omega) = \mathcal{F}[x(t)]$ is the Fourier-transformed signal. It is easy to show that the matrices $\varGamma_x(\omega)$ are skew-symmetric. Evaluated on the source signals, the elements of these matrices read

$$\varGamma_s^{ik}(\omega) = \begin{cases} \gamma^i(\omega) & \text{if} \quad k = i + n \\ -\gamma^i(\omega) & \text{if} \quad k = i - n \\ 0 & \text{else} \end{cases} \tag{3}$$

so the $\varGamma_s(\omega)$ matrices in the source basis have the block structure

$$\varGamma_s(\omega) = \begin{bmatrix} 0 & D(\omega) \\ -D(\omega) & 0 \end{bmatrix} \tag{4}$$

with real diagonal sub-matrices $D(\omega)$ for each frequency ω. We will refer to this structure as *skew-double-diagonal* or short *sd-diagonal*. So, given the set of skew-symmetric matrices $\varGamma_x(\omega)$ estimated from the mixed data $x(t)$, the source separation task is to find a real invertible transformation B that recovers this

[1] More precisely, the whitened matrix has to be symmetric and positive definite.

original sd-diagonal structure. Since such sd-diagonal matrices are diagonalized with the constant unitary transformation

$$U := \frac{1}{\sqrt{2}} \begin{bmatrix} I & -iI \\ I & iI \end{bmatrix}, \quad \text{i.e. by} \quad U\mathit{\Gamma}_s(\omega)U^\dagger = \begin{bmatrix} iD(\omega) & 0 \\ 0 & -iD(\omega) \end{bmatrix} \tag{5}$$

the source separation is equivalent to a complex simultaneous diagonalization. It has therefore been proposed to solve the PISA problem in this framework. However, skew-symmetric matrices do not behave as well as symmetric ones. Due to their inherent indeterminacy, a prior whitening is not possible, which means that the optimization cannot profit from the stabilizing effect of a restriction to orthogonal matrices. Another disadvantage is that even if the mixing is purely real-valued, the diagonalization takes a detour over complex matrices, which introduces unnecessary parameters. Also, the diagonal matrices in eq. (5) have distinct internal symmetries and it is not obvious how this can be incorporated into the optimization. Here, we will therefore *directly* tackle the real sd-diagonalizing problem, i.e. the problem of transforming the matrices into the form given by eq. (4).[2] It turns out that this formulation allows us to proceed in a two-step strategy in complete formal analogy to the diagonalization of symmetric matrices.

To understand this, let us revisit the whitening-rotation scheme for symmetric matrices. From a formal perspective, the matrix $\mathit{\Sigma}$ that is used to calculate the whitening defines a positive definite symmetric bilinear form via $(x, y) \mapsto x^\top \mathit{\Sigma} y$. A data space that is equipped with such a bilinear form is called *Euclidean* and there exists a basis in which this bilinear form assumes its normal form, i.e. the identity matrix. The whitening W is just a transformation to such a basis. The remaining transformation R is then such that it leaves this normal form invariant, i.e. $RIR^\top = I$, which means that R is orthogonal.

In the skew-symmetric case, the matrix $\mathit{\Gamma}$ that the whitening-equivalent transformation should be based on is a skew-symmetric bilinear form. The key point is that this defines a different geometric structure in the data space: a space equipped with a skew-symmetric bilinear form is called a *symplectic* space. There also exists a linear transformation to a normal form, which is however not the Identity matrix, but given by

$$\mathit{\Omega} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{6}$$

Like for the whitening in Euclidean space, it is easy to determine a transformation to this normal form e.g. by a singular value decomposition with subsequent scaling of the dimensions. The remaining transformation R should again leave this form invariant, i.e. $R\mathit{\Omega}R^\top = \mathit{\Omega}$, which means that R is a *symplectic transformation*. The following table provides a side-by-side comparison of the symmetric/Euclidean and the skew-symmetric/symplectic case.

[2] Of course this implicitly also solves the simultaneous diagonalization via eq. (5).

SOBI/TDSEP	**PISA**
symmetric matrices $\Sigma, \Sigma(\tau)$	skew-symmetric matrices $\Gamma, \Gamma(\omega)$
Vector space	
Euclidean space	Symplectic space
'Whitening': transformation of bilinear form to ...	
... Euclidean normal form $\Sigma \mapsto I = W\Sigma W^\top$	... symplectic normal form $\Gamma \mapsto \Omega = W\Gamma W^\top$
Normal form preserving transformation	
$RIR^\top = I$ (orthogonal transformation)	$R\Omega R^\top = \Omega$ (symplectic transformation)

It is easy to check that the symplectic matrices form a group, that is they contain the identity, are invertible and the product of two symplectic matrices is also symplectic. All elements of this group have unit determinant, which prevents a symplectic optimization from converging to a trivial solution. The symplectic matrices also define a manifold, which makes it a Lie group, a structure that will play an important role in the following. For any Lie-Group G, the elements $R \in G$ can be written as

$$R = \exp(q)$$

with q being an element of the corresponding Lie-Algebra. Geometrically, the Lie algebra is the tangent space at the identity element. From $R\Omega R^\top = \Omega$, we obtain for the elements q of the symplectic algebra

$$\Omega \exp(q) \Omega^{-1} = \exp(-q^\top) \tag{7}$$

and the power-series definition of the exponential map leads to $\Omega \exp(q)\Omega^{-1} = \exp(\Omega q \Omega^{-1})$. Using $\Omega^{-1} = -\Omega$, we therefore obtain

$$\Omega q \Omega = q^\top . \tag{8}$$

By sub-dividing the matrix q into $n \times n$ submatrices

$$q = \begin{bmatrix} q_1 & q_2 \\ q_3 & q_4 \end{bmatrix} \tag{9}$$

this finally leads to

$$q_4 = -q_1^\top \qquad q_2 = q_2^\top \qquad q_3 = q_3^\top \tag{10}$$

which reflects the symmetries in the symplectic algebra. Given equation (8), we can also project any square matrix M to the symplectic algebra by

$$M \longmapsto \frac{1}{2}\left(M + \Omega M^\top \Omega\right). \tag{11}$$

Based on this structure, we will now derive a gradient-descent based method for simultaneous sd-diagonalization of real skew-symmetric matrices. In the following, we assume that we already performed the symplectic whitening W on a suitably chosen skew-symmetric matrix Γ.

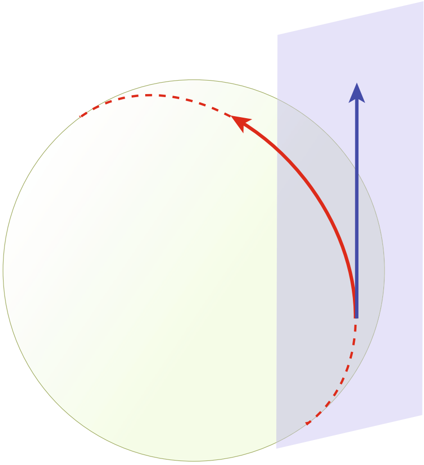


Fig. 1. Cartoon figure, demonstrating geodesic line search in a Lie group. The gradient G (blue arrow) lives in the corresponding Lie algebra (blue plane) and can be mapped to the group manifold (green surface) by the exponential map. The geodesic line $\exp(\lambda G)$ is then given by the dashed red line.

3 Optimizing the Symplectic Transformation

With the remaining symplectic transformation R, we want to simultaneously sd-diagonalize the set of real skew-symmetric matrices Γ_k. This will be achieved by minimizing the sum over all squared off-sd-diagonal entries as measured by the loss function

$$\mathcal{L}(R) = \sum_k \operatorname{trace}\left(\left(R\Gamma_k R^\top \odot \bar{\Psi}\right)^\top \left(R\Gamma_k R^\top \odot \bar{\Psi}\right)\right) \tag{12}$$

where $\odot$ denotes the element-wise Hadamard product and $\Psi := \Omega \odot \Omega$ and $\bar{\Psi} = \mathbb{1} - \Psi$ define pattern matrices which retain (or remove) only the sd-diagonal elements of a square matrix under Hadamard multiplication. This loss function will be minimized with multiplicative updates, i.e. we start with the identity matrix (or any random symplectic matrix) and update in each step the current R_n by the symplectic update Q:

$$R_{n+1} \longleftarrow QR_n. \tag{13}$$

So, at each individual step, we have a loss function in Q given by

$$\mathcal{L}_{R_n}(Q) := \mathcal{L}(QR_n) \tag{14}$$

The gradient of this loss function with respect to Q will always be evaluated at the point $Q = I$, which means that the gradient is always an element of the tangent space at the identity, or in other words, of the symplectic algebra. Every update step, however, should not be in this tangent space, but on the manifold itself. So, given a gradient matrix $G = \nabla_{sp}\,\mathcal{L}_{R_n}$ in the symplectic algebra, we will perform a line search on the one-parameter subgroup

$$Q = \exp(\lambda G) \qquad \lambda \in \mathbb{R} \tag{15}$$

which is the geodesic line in the group $Sp(2n)$ in direction of the gradient G (see also fig. 1).

To calculate the gradient $\nabla_{sp}\,\mathcal{L}_{R_n}$ and define the notion of 'steepest descent', we need a distance measure (i.e. an inner product) in the Lie algebra. In general, the gradient ∇f of a scalar function f is defined as the vector, whose inner product with a unit vector H is equal to the directional derivative of f in the direction of H, i.e.

$$\langle \nabla f, H \rangle = \frac{\partial}{\partial \lambda} f\left(\lambda H\right) \tag{16}$$

For unconstrained matrices with the Euclidean distance (Frobenius norm) and the inner product $\langle M_1, M_2 \rangle = \mathrm{trace}(M_1^\top M_2)$, the gradient is simply the matrix of the partial derivatives $\nabla f = \frac{\partial f}{\partial M}$. However, for constrained sets, this is generally not as simple. Specifically, because of eq. (10), the components of the matrices from the symplectic algebra are not independent, which means that they do not all correspond to different directions in the tangent space. A proper inner product in the symplectic algebra can be defined as (see e.g. [3])

$$\langle G, H \rangle_{sp} := \frac{1}{2}\,\mathrm{trace}\left((G \odot (\mathbb{1} + \Psi))^\top H \right). \tag{17}$$

Note that even though this definition looks asymmetric on a first glance, it is easy to check that indeed $\langle G, H \rangle_{sp} = \langle H, G \rangle_{sp}$.

With this inner product, we can evaluate the left-hand side of eq. (16). Furthermore, if we let the symplectic update Q vary on the line $Q(\lambda) = \exp(\lambda H)$ with a $H \in sp(2n)$ and $\langle H, H \rangle_{sp} = 1$, the right-hand side is given by

$$\begin{aligned}\frac{\partial}{\partial \lambda}\mathcal{L}_{R_n}\left(Q(\lambda)\right) &= \mathrm{trace}\left(\left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right)^\top H Q \right) = \mathrm{trace}\left(Q \left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right)^\top H \right) \\ &= \frac{1}{2}\,\mathrm{trace}\left(\left(Q \left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right)^\top + \Omega \left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right) Q^\top \Omega \right) H \right)\end{aligned} \tag{18}$$

where the last equality is due to eq. (11). Using this, we can now solve eq. (16) for the gradient in the symplectic algebra to obtain

$$\nabla_{sp}\,\mathcal{L}_{R_n} = \left(\left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right) Q^\top + \Omega Q \left(\frac{\partial \mathcal{L}_{R_n}}{\partial Q}\right)^\top \Omega \right) \odot \left(\mathbb{1} - \frac{\Psi}{2} \right). \tag{19}$$

This equation translates the matrix of partial derivatives $\partial \mathcal{L}_{R_n}/\partial Q$ into the gradient for the symplectic algebra. If we write the k-th skew-symmetric matrix at the current update step as $\Gamma'_k := R_n \Gamma_k R_n^\top$, this partial derivative with respect to Q is given by $\partial \mathcal{L}_{R_n}/\partial Q = -4 \sum_k (Q\Gamma'_k Q^\top \odot \bar{\Psi}) Q \Gamma'_k$ so the gradient can be evaluated as

$$\nabla_{sp} \mathcal{L}_{R_n} = -4 \sum_k \left(\left(\Gamma''_k \odot \bar{\Psi} \right) \Gamma''_k + \Omega \Gamma''_k \left(\Gamma''_k \odot \bar{\Psi} \right) \Omega \right) \odot \left(\mathbb{1} - \frac{\Psi}{2} \right) \tag{20}$$

where $\Gamma''_k = Q\, \Gamma'_k\, Q^\top = QR_n\, \Gamma_k\, (QR_n)^\top$. Given this gradient we can now perform a conjugate gradient descent with multiplicative update steps as described in eq. (13). By updating with a symplectic matrix Q at every step, we ensure that we stay on the symplectic manifold, so the scaling of the solution is well controlled. In each step, a search direction is given by a matrix H from the symplectic algebra and the symplectic update Q is determined by a line search in λ for the one-parameter subgroup $Q(\lambda) = \exp(\lambda H)$ (see e.g. [7] for details).

4 Simulations and Conclusion

We will now briefly evaluate the performance of the proposed symplectic optimization algorithm and to compare it with a naïve simultaneous diagonalization in the complex domain, as proposed in [4,6]. The latter approach simply minimizes the sum of the squared off-diagonal elements under the constraint that the transformation B has a unit determinant.

We start with a set of 20 sd-diagonal 6-dimensional matrices and map them with random mixing matrices (i.e. the mixing coefficients are sampled from a standard normal distribution) into a 6, 10, 20, or 40-dimensinal space and add gaussian noise ($\sigma = 0.1$) to the obtained matrices.[3] The goal of the (symplectic) diagonalization is to estimate the mixing matrix A (or, equivalently, its inverse B). We measure the quality of the obtained results in terms of *principal angles* between the three true 2d subspaces, as given by the column-span of $[A_{:,i} A_{:,i+3}]$ and the estimated subspaces determined by the respective algorithm. Principal angles provide information about the relative position of linear subspaces:[4] if all angles are zero, one subspace is a subset of the other. Given the 2 principal angles θ_1, θ_2 between a pair of two-dimensional linear subspaces, we define the *subspace error* as

$$SE := \frac{1}{2} \left(\sin^2 \theta_1 + \sin^2 \theta_2 \right). \tag{21}$$

This error is always between 0 and 1; it is zero only if the two subspaces are identical, it is 1 only if any two vectors from the two subspaces are othogonal. In our simulations, we obtain 3 subspace errors for the three hidden 2d subspaces. Figure 2 shows the mean subspace error (and its standard deviation) for the symplectic and the direct diagonalization in the different settings sketched above over 100 repetitions. To account for possible local minima, each optimization was

[3] To keep the matrices skew-symmetric, the noise was also skew-symmetrized.

[4] See, e.g. http://en.wikipedia.org/wiki/Principal_angles

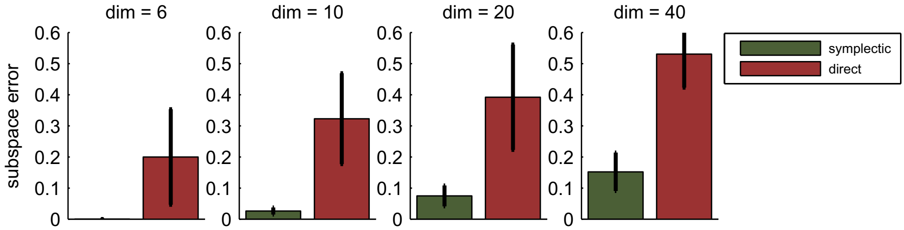


Fig. 2. Comparison between the symplectic and the naive (direct) diagonalization

restarted 5 times with random initializations. From the comparison of the error, it is obvious that respecting the symplectic structure of the underlying problem really pays off. Besides the fact that the symplectic optimization has less free parameters than the naïve approach, it does not run into 'useless' directions. If there exists a subspace that cannot be diagonalized properly, the naive approach would invest much into simply scaling this subspace down with respect to 'better' subspaces. The symplectic optimization does not fall into this trap.

To conclude, we have shown that the whitening/rotation scheme from symmetric diagonalization can be translated to the skew-symmetric case. Compared to a naïve complex diagonalization, the resulting algorithm is more stable and yields better results.

References

1. Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E.: A blind source separation technique using second order statistics. IEEE Trans. on Signal Processing 45(2), 434–444 (1997)
2. Cardoso, J.-F., Souloumiac, A.: Blind beamforming for non gaussian signals. IEE Proceedings-F 140, 362–370 (1993)
3. Meinecke, F.C.: Synchronized? Identifying interactions from superimposed signals. PhD Thesis, TU Berlin (2011)
4. Meinecke, F.C., Ziehe, A., Nolte, G., Müller, K.-R.: Interacting source analysis - identifying interactions in mixed and noisy complex systems. In: Proc. Int. ICA Research Network 2006, Liverpool, pp. 72–79 (2006)
5. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from eeg data using the imaginary part of coherency. Clinical Neurophysiology 115(10), 2292–2307 (2004)
6. Nolte, G., Meinecke, F.C., Ziehe, A., Müller, K.-R.: Identifying interactions in mixed and noisy complex systems. Physical Review E 73 (2006)
7. Plumbley, M.D.: Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras. Neurocomputing 67, 161–197 (2005); Geometrical Methods in Neural Networks and Learning
8. Ziehe, A., Müller, K.-R.: TDSEP – an efficient algorithm for blind separation using time structure. In: Niklasson, L., Bodén, M., Ziemke, T. (eds.) Proceedings of the 8th International Conference on Artificial Neural Networks, ICANN 1998. Perspectives in Neural Computing, pp. 675–680. Springer, Berlin (1998)

Joint Block Diagonalization Algorithms for Optimal Separation of Multidimensional Components

Dana Lahat[1], Jean-François Cardoso[2], and Hagit Messer[1]

[1] School of Electrical Engineering, Tel Aviv University, 69978 Tel Aviv, Israel
[2] LTCI, TELECOM ParisTECH and CNRS, 46 rue Barrault, 75013 Paris, France

Abstract. This paper deals with non-orthogonal joint block diagonalization. Two algorithms which minimize the Kullback-Leibler divergence between a set of real positive-definite matrices and a block-diagonal transformation thereof are suggested. One algorithm is based on the relative gradient, and the other is based on a quasi-Newton method. These algorithms allow for the optimal, in the mean square error sense, blind separation of multidimensional Gaussian components. Simulations demonstrate the convergence properties of the suggested algorithms, as well as the dependence of the criterion on some of the model parameters.

Keywords: Joint block diagonalization, relative gradient, quasi-Newton.

1 Introduction

In this paper, we present two algorithms which (approximately) jointly block-diagonalize a set of weighted real positive-definite matrices. The proposed joint block diagonalization (JBD) is achieved by minimizing the Kullback-Leibler divergence (KLD). The KLD maximizes, under asymptotic conditions, the likelihood of the observations.

The most common criteria which define JBD of a set of matrices are the least squares (LS) criterion and the quadratic criterion. JBD algorithms are usually divided into orthogonal (unitary) and non-orthogonal (non-unitary) ones. In this paper, we focus on non-orthogonal algorithms and only on the case where the de-mixing matrix is invertible. Non-unitary JBD by a LS criterion is discussed, for example, in [1]. The quadratic criterion is minimized using a non-unitary algorithm, for example, by [2]. These criteria are different than our KLD-based criterion, which will be presented shortly. A KLD criterion for JBD, in the context of source separation, has first been suggested by [3], in order to separate one-dimensional sources from their convolutive mixture; however, [3] do not specify the algorithm used to minimize their criterion. The fast algorithm for joint diagonalization via KLD minimization, suggested by Pham [4], is extended for JBD of cyclostationary sources in [5]. However, to the best of our knowledge, an algorithm which guarantees minimal mean square error (MSE) in the sense which will be given in the sequel and for the following data model cannot be found in the literature.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 155–162, 2012.

The data model which motivates our derivation is as follows. Consider a model of T observations of an $m \times 1$ vector $\boldsymbol{x}(t)$, whose latent model is

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) \qquad 1 \leq t \leq T\,, \tag{1}$$

where $\boldsymbol{A}$ is an $m \times m$ invertible matrix and $\boldsymbol{s}(t)$ is a vector of independent *sources*. A natural extension of practical interest is to assume that the m sources can be partitioned into $n \leq m$ groups, with the sources of different groups being statistically independent while the sources in the same group are not independent and cannot be made independent by any linear transform on $\boldsymbol{s}(t)$. We thus denote $\boldsymbol{s}(t) = [\boldsymbol{s}_1(t)^\dagger, \ldots, \boldsymbol{s}_n(t)^\dagger]^\dagger$ with $\boldsymbol{s}_i(t)$ a vector of length m_i, and $\sum_{i=1}^{n} m_i = m$. Let us define a partition $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n]$, where $\boldsymbol{A}_i$, the ith column block of $\boldsymbol{A}$, has dimension $m \times m_i$. Since for any $m_i \times m_i$ invertible matrix $\boldsymbol{Z}_i$, the pair $(\boldsymbol{A}_i,\, \boldsymbol{s}_i(t))$ and the pair $(\boldsymbol{A}_i\boldsymbol{Z}_i^{-1},\, \boldsymbol{Z}_i\boldsymbol{s}_i(t))$ contribute the same quantity $\boldsymbol{x}_i(t) = \boldsymbol{A}_i\boldsymbol{Z}_i^{-1}\boldsymbol{Z}_i\boldsymbol{s}_i(t) = \boldsymbol{A}_i\boldsymbol{s}_i(t)$ to the observations, then source separation can be determined only up to the inherent indeterminacies of a block-diagonal invertible matrix $\boldsymbol{Z}$ with block-pattern $\boldsymbol{m} = [m_1, \ldots, m_n]$, and a block-wise permutation matrix. Note that $\boldsymbol{x}(t) = \sum_{i=1}^{n} \boldsymbol{x}_i(t)$. In the sequel, we denote the scale-invariant vectors $\boldsymbol{x}_i(t)$ *components*, as opposed to the scale-dependent latent *sources* $\boldsymbol{s}_i(t)$.

Let us consider a piecewise stationary model as follows. The observation interval $[1, T]$ is partitioned into Q domains $\mathcal{D}_q$, $q = 1, \ldots, Q$, where $\mathcal{D}_q$ contains n_q samples, so that $\sum_{q=1}^{Q} n_q = T$. We assume that $\boldsymbol{s}(t)$ is independent of $\boldsymbol{s}(t')$ if $t \neq t'$ and that, for any $t \in \mathcal{D}_q$, $\boldsymbol{s}(t) \sim \mathcal{N}(\mathbf{0}_{m\times 1}, \boldsymbol{R}_S^{(q)})$. The linear model (1) implies that $\boldsymbol{R}_X^{(q)} = \boldsymbol{A}\boldsymbol{R}_S^{(q)}\boldsymbol{A}^\dagger$, where $\boldsymbol{R}_X^{(q)} = E\{\boldsymbol{x}(t)\boldsymbol{x}^\dagger(t)\}$ for $t \in \mathcal{D}_q$ and

$$\boldsymbol{R}_S^{(q)} \triangleq \begin{bmatrix} \boldsymbol{R}_{S,11}^{(q)} & \mathbf{0} & \mathbf{0} \\ & & \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{R}_{S,nn}^{(q)} \end{bmatrix} = \operatorname{bdiag}\{\boldsymbol{R}_{S,11}^{(q)}, \ldots, \boldsymbol{R}_{S,nn}^{(q)}\}\,, \tag{2}$$

where $\boldsymbol{R}_{S,ii}^{(q)} = E\{\boldsymbol{s}_i(t)\boldsymbol{s}_i^\dagger(t)\}$ for $t \in \mathcal{D}_q$ and $\operatorname{bdiag}\{\cdot, \ldots, \cdot\}$ denotes a block-diagonal matrix constructed from the matrices in brackets. Analogously, given an $m \times m$ matrix $\boldsymbol{M}$, $\operatorname{bdiag}_{\boldsymbol{m}}\{\boldsymbol{M}\}$ returns the block-diagonal matrix with block pattern $\boldsymbol{m}$ which has the same diagonal blocks as $\boldsymbol{M}$ and has zeros in the off-diagonal blocks. Using the notation

$$D(\boldsymbol{R}_1, \boldsymbol{R}_2) = \frac{1}{2}\big(\operatorname{tr}\{\boldsymbol{R}_1\boldsymbol{R}_2^{-1}\} - \log\det(\boldsymbol{R}_1\boldsymbol{R}_2^{-1}) - m\big) \tag{3}$$

for any two $m \times m$ positive-definite matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, the log-likelihood for the model just described is [6]

$$\begin{aligned} \log p(\{\boldsymbol{x}(t)\}_{t=1}^{T},; \boldsymbol{A}, \{\boldsymbol{R}_S^{(q)}\}_{q=1}^{Q}) &= -\sum_{q=1}^{Q} n_q D(\widetilde{\boldsymbol{R}}_X^{(q)}, \boldsymbol{R}_X(q)) + \kappa \\ &= -T\langle D(\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}_X^{(q)}\boldsymbol{A}^{-\dagger}, \boldsymbol{R}_S^{(q)})\rangle + \kappa \end{aligned} \tag{4}$$

where $\kappa = -\frac{1}{2}(mT + \langle \log\det(2\pi\widetilde{\boldsymbol{R}}_X^{(q)})\rangle)$, $\widetilde{\boldsymbol{R}}_X^{(q)} \triangleq \frac{1}{n_q}\sum_{t\in\mathcal{D}_q}\boldsymbol{x}(t)\boldsymbol{x}^\dagger(t)$, and the last transition in (4) is due to the invariance of (3) to invertible transforms ($\boldsymbol{A}$ in our case), and the notation $\langle \boldsymbol{M}^{(q)}\rangle \triangleq \frac{1}{T}\sum_{q=1}^{Q} n_q \boldsymbol{M}^{(q)}$. By maximizing the likelihood with respect to the nuisance parameters $\{\boldsymbol{R}_S^{(q)}\}_{q=1}^Q$ for fixed $\boldsymbol{A}$,

$$\max_{\boldsymbol{R}_S^{(q)}} \log p(\{\boldsymbol{x}(t)\}_{t=1}^T; \boldsymbol{A}, \{\boldsymbol{R}_S^{(q)}\}_{q=1}^Q) = -T\; C(\boldsymbol{A}, \{\widetilde{\boldsymbol{R}}_X^{(q)}\}_{q=1}^Q) + \kappa\,,$$

we obtain the *contrast function* [7]

$$C(\boldsymbol{A}) = \langle D(\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}_X^{(q)}\boldsymbol{A}^{-\dagger}, \mathrm{bdiag}_{\boldsymbol{m}}\{\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}_X^{(q)}\boldsymbol{A}^{-\dagger}\})\rangle \tag{5}$$

and where, for brevity, the dependence of $C(\boldsymbol{A})$ on the data via $\{\widetilde{\boldsymbol{R}}_X^{(q)}\}_{q=1}^Q$ is not denoted explicitly. The term κ is irrelevant to the maximization of the likelihood with respect to its parameters since it depends only on the data, not on the model. Consequently, maximizing the likelihood is equivalent to minimizing the contrast function (5). Note that (5) is the multidimensional analogue of its one-dimensional counterpart in [8]. The scalar $D(\boldsymbol{R}_1, \boldsymbol{R}_2)$, defined in (3), is the KLD between the distributions $\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_1)$ and $\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_2)$ and thus is a measure of mismatch between two positive matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$. Therefore, in our piecewise stationary model, maximizing (4) is equivalent to minimizing the average mismatch between the sample covariance matrices and their expected counterparts. Since $D(\boldsymbol{R}_1, \mathrm{bdiag}_{\boldsymbol{m}}\{\boldsymbol{R}_1\}) \geq 0$ with equality if and only if $\boldsymbol{R}_1$ is block diagonal with block pattern $\boldsymbol{m}$, then, for any positive matrix $\boldsymbol{R}_1$, the divergence $D(\boldsymbol{R}_1, \mathrm{bdiag}_{\boldsymbol{m}}\{\boldsymbol{R}_1\})$ is a measure of the block-diagonality of $\boldsymbol{R}_1$. Therefore, $C(\boldsymbol{A})$ can be understood as *joint block diagonalization* of the set of covariance matrices $\{\widetilde{\boldsymbol{R}}_X^{(q)}\}_{q=1}^Q$ by matrix $\boldsymbol{A}^{-1}$.

Component separation by minimization of (5) achieves, under asymptotical conditions and when the model holds, the Cramér-Rao lower bound (CRLB) on the mixing model parameters, and is thus optimal in the MSE sense, where MSE(component i)= $\frac{1}{T}\sum_{t=1}^T E\{\|\widehat{\boldsymbol{x}}_i(t) - \boldsymbol{x}_i(t)\|^2\}$, $\|\cdot\|$ denotes the Frobenius norm and $\widehat{\boldsymbol{x}}_i(t)$ denotes an estimate of $\boldsymbol{x}_i(t)$. A closed-form expression for the CRLB and MSE is obtained in [6], where it is shown that this MSE is achievable also for non-Gaussian data.

2 Derivation of the Relative Variations

In this section, we derive the relative gradient (RG) and its first-order variation for the update step of our algorithms. As demonstrated by [9], relative-variation algorithms enjoy equivariant performance and are thus preferred for our problem over their non-relative counterparts.

The first-order variation of $C(\boldsymbol{A})$ when $\boldsymbol{A}$ is replaced by $\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{E})$ (where $\boldsymbol{I}$ denotes the identity matrix and the entries of $\boldsymbol{E}$ are sufficiently small) can always be expressed by the Taylor expansion

$$C(\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{E})) = C(\boldsymbol{A}) + \mathrm{tr}\{(\nabla C(\boldsymbol{A}))^\dagger \boldsymbol{E}\} + \text{higher-order terms in } \boldsymbol{E}\,, \tag{6}$$

for some $m \times m$ matrix $\nabla C(\boldsymbol{A})$, defined as the RG of $C(\boldsymbol{A})$. Similarly to the derivation for the one-dimensional case in [8], one obtains the RG

$$\nabla C(\boldsymbol{A}) = -\langle \mathrm{bdiag}^{-1}_{\boldsymbol{m}}\{\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}\}\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}\rangle + \boldsymbol{I} \tag{7}$$

on which the RG algorithm, explained in Sec. 3, is based.

Another algorithm can be derived based on the Newton method. In order to realize a quasi-Newton (QN) method in the sense of [10], we obtain a first-order approximation of the gradient, using the following steps. First, the ML estimate of $\boldsymbol{A}$ is obtained by setting the RG (7) to zero. This yields the *estimating equations*

$$\langle \mathrm{bdiag}^{-1}_{\boldsymbol{m}}\{\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}\}\ \boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}\rangle = \boldsymbol{I}\,. \tag{8}$$

These estimating equations (8) can be rewritten block-wise as

$$\langle([\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}]_{ii})^{-1}[\boldsymbol{A}^{-1}\widetilde{\boldsymbol{R}}^{(q)}_X\boldsymbol{A}^{-\dagger}]_{ij}\rangle = \boldsymbol{0}_{m_i\times m_j} \quad j \neq i\,. \tag{9}$$

Note that the (i,i)th block of (8), that is, $i=j$ of (9), degenerates into the identity matrix: the diagonal blocks $i=j$ do not yield any constraints, regardless of $\widetilde{\boldsymbol{R}}^{(q)}_X$, reflecting the indeterminacy discussed in Sec. 1. It can be readily verified that the estimating equations (8) are invariant to block-diagonal scale ambiguity.

In the second step, the first-order expansion of the estimating equations (9) under asymptotic conditions ($T\to\infty$ for $\frac{n_q}{T}$ fixed $\forall q$) can be expressed, after some mathematical manipulations analogous to those in [6], as

$$\begin{bmatrix}\mathrm{vec}\{\boldsymbol{\mathcal{E}}_{ij}\}\\ \mathrm{vec}\{\boldsymbol{\mathcal{E}}_{ji}\}\end{bmatrix} = \boldsymbol{\mathcal{H}}_{ij}^{-1}\begin{bmatrix}\mathfrak{g}_{ij}\\ \mathfrak{g}_{ji}\end{bmatrix} + \varOmega(\tfrac{1}{T}) \quad i\neq j\,, \tag{10}$$

where

$$\begin{array}{ll}\boldsymbol{H}^{(q)}_{ij} = \boldsymbol{R}^{(q)}_{S,jj}\otimes\boldsymbol{R}^{-(q)}_{S,ii}\,, & \boldsymbol{\mathcal{H}}_{ij} = \begin{bmatrix}\langle\boldsymbol{H}^{(q)}_{ij}\rangle & \boldsymbol{\mathcal{T}}_{m_j,m_i}\\ \boldsymbol{\mathcal{T}}_{m_i,m_j} & \langle\boldsymbol{H}^{(q)}_{ji}\rangle\end{bmatrix}\\ \boldsymbol{g}^{(q)}_{ij} = -\boldsymbol{R}^{-(q)}_{S,ii}\widetilde{\boldsymbol{R}}^{(q)}_{S,ij}\,, & \mathfrak{g}_{ij} = -\mathrm{vec}\{\langle\boldsymbol{g}^{(q)}_{ij}\rangle\}\end{array} \quad i\neq j\,, \tag{11}$$

$\widetilde{\boldsymbol{R}}^{(q)}_{S,ij} \triangleq \frac{1}{n_q}\sum_{t\in\mathcal{D}_q}\boldsymbol{s}_i(t)\boldsymbol{s}_j^\dagger(t)$, $\boldsymbol{R}^{-(q)}_{S,ii} \triangleq \left(\boldsymbol{R}^{(q)}_{S,ii}\right)^{-1}$ and $\boldsymbol{\mathcal{E}}_{ij}$, the $m_i\times m_j$ blocks of an $m\times m$ matrix $\boldsymbol{\mathcal{E}}$, reflect the relative change in $\boldsymbol{A}$ due to the difference between $\boldsymbol{R}^{(q)}_S$ and $\widetilde{\boldsymbol{R}}^{(q)}_S$. It should be emphasized that since $\mathrm{bdiag}_{\boldsymbol{m}}\{\nabla C(A)\}$ is invariant to changes in $\widetilde{\boldsymbol{R}}^{(q)}_X$ (this is a direct result of (8)), then $\boldsymbol{\mathcal{E}}_{ii}\equiv\boldsymbol{0}_{m_i\times m_i}$. The notation $\varOmega(f)$ in (10) stands for stochastic terms whose standard deviation grows with f, or faster. $\boldsymbol{\mathcal{T}}_{m,n}$ is the transpose operator, where, for an $m\times n$ matrix $\boldsymbol{M}$, $\mathrm{vec}\{\boldsymbol{M}^\dagger\} = \boldsymbol{\mathcal{T}}_{m,n}\mathrm{vec}\{\boldsymbol{M}\}$. We also use the $\mathrm{vec}\{\cdot\}$ operator, which stacks the columns of a $p\times q$ matrix into a $pq\times 1$ vector; the Kronecker product $\otimes$ and the property [11] $\mathrm{vec}\{\boldsymbol{MXN}\} = (\boldsymbol{N}^\dagger\otimes\boldsymbol{M})\mathrm{vec}\{\boldsymbol{X}\}$ for any matrices $\boldsymbol{M},\boldsymbol{N},\boldsymbol{X}$ of compatible dimensions. It is assumed that $\boldsymbol{\mathcal{H}}_{ij}$ is invertible (which is usually the case for randomly-generated data; for further discussion see [12]). The set of matrices $\boldsymbol{\mathcal{E}}_{ij}$, $\forall i\neq j$, obtained from (10), constitute the Newton step. This leads to our QN algorithm, explained in Sec. 3.

3 Algorithms

The pseudocode of the iterative algorithms is given in Algorithm 1, where the part pertaining to each of the RG and QN algorithms is given in Algorithm 2 and 3, respectively. The RG algorithm works as follows: according to (6), if $\boldsymbol{E}$ is a matrix with small enough values to ensure the invertibility of $\boldsymbol{I}+\boldsymbol{E}$, and if $\boldsymbol{A}$ is changed into $\boldsymbol{A}(\boldsymbol{I}+\boldsymbol{E})$, then $C(\boldsymbol{A})$ changes by the amount $\mathrm{tr}\{(\nabla C(\boldsymbol{A}))^\dagger \boldsymbol{E}\}+$ higher-order terms in $\boldsymbol{E}$. Given $\boldsymbol{E} = -\lambda \nabla C(\boldsymbol{A})$ and $\lambda > 0$ a real scalar, the updating rule (line 4 in Algorithm 2) changes $C(\boldsymbol{A})$ into $C(\boldsymbol{A}) - \lambda \|\nabla C(\boldsymbol{A})\|^2+$ higher-order terms in $\nabla C(\boldsymbol{A})$. Hence, the decrease of the contrast function $C(\boldsymbol{A})$ is guaranteed for small enough λ. The updating rule is iterated until $\|\nabla C(\boldsymbol{A})\| \leq$ threshold. In the QN algorithm, the relative change in $\boldsymbol{A}$ is determined directly by $\boldsymbol{\mathcal{E}}$ (10), as explained in Sec. 2. The transformation matrix $\boldsymbol{T}$ in the Algorithms' pseudocodes reflects the relative change in $\boldsymbol{A}$ at each iteration.

The choice of the step-size in a RG algorithm determines its convergence rate; see [2], for example. For the simulations in Sec. 4 we chose to set λ by backtracking line search. Since only $\widetilde{\boldsymbol{R}}_X^{(q)}$ is available to the algorithm, then within the iterations, $\boldsymbol{A}^{-1}\boldsymbol{R}^{(q)}\boldsymbol{A}^\dagger$ is used to approximate both $\widetilde{\boldsymbol{R}}_S^{(q)}$ and $\boldsymbol{R}_S^{(q)}$ of (11). Then, within Algorithm 3, $\boldsymbol{g}_{ij}^{(q)}$ is equal to the evaluated (i, j)th sub-block of $\nabla C(\boldsymbol{A})$.

Algorithm 1. An Iterative JBD Algorithm

```
function JBD($\{\widetilde{\boldsymbol{R}}_X^{(q)}\}_{q=1}^Q$, $\{n_q\}_{q=1}^Q$, $\boldsymbol{m}$, threshold)
    $\boldsymbol{A} \leftarrow \boldsymbol{I}$                                        ▷ Init
    $\boldsymbol{R}^{(q)} \leftarrow \widetilde{\boldsymbol{R}}_X^{(q)}\ \forall q$                               ▷ Init
    while $\|\nabla C(\boldsymbol{A})\| >$ threshold do
        $\nabla C(\boldsymbol{A}) \leftarrow \boldsymbol{I} - \langle \mathrm{bdiag}_{\boldsymbol{m}}^{-1}\{\boldsymbol{R}^{(q)}\}\boldsymbol{R}^{(q)}\rangle$            ▷ (7)
        Evaluate $\boldsymbol{T}$                ▷ Algorithm 2 for RG, Algorithm 3 for QN
        $\boldsymbol{R}^{(q)} \leftarrow \boldsymbol{T}^{-1}\boldsymbol{R}^{(q)}\boldsymbol{T}^{-\dagger}$, $q = 1, \ldots, Q$
        $\boldsymbol{A} \leftarrow \boldsymbol{A}\boldsymbol{T}$                                ▷ For output only
    end while
    return $\boldsymbol{A}$
end function
```

Algorithm 2. Update Step for RG

```
1: $\lambda \leftarrow 1$                    ▷ Choose $\lambda$, e.g. by backtracking line search
2: while $C(\boldsymbol{A}(\boldsymbol{I} - \lambda\nabla C(\boldsymbol{A})) > C(\boldsymbol{A}) - \alpha\lambda\mathrm{tr}\{\|\nabla C(\boldsymbol{A})\|^2\}$ do $\lambda \leftarrow \beta\lambda$
3: end while
4: $\boldsymbol{T} \leftarrow \boldsymbol{I} - \lambda\nabla C(\boldsymbol{A})$
```

Algorithm 3. Update Step for QN

```
for i=1:n, j=1:i-1 do
    $\boldsymbol{g}_{ij} \leftarrow [\nabla C(\boldsymbol{A})]_{ij}$                                  ▷ (11)
    $\boldsymbol{H}_{ij}^{(q)} \leftarrow \boldsymbol{R}_{jj}^{(q)} \otimes \boldsymbol{R}_{ii}^{-(q)}$, $q = 1, \dots, Q$                ▷ (11)
    Evaluate $\boldsymbol{\mathcal{E}}_{ij}, \boldsymbol{\mathcal{E}}_{ji}$                                ▷ (10)
end for
Reconstruct $\boldsymbol{\mathcal{E}}$ from $\{\boldsymbol{\mathcal{E}}_{ij}\}_{i \neq j}$          ▷ $\boldsymbol{\mathcal{E}}_{ii} \equiv \mathbf{0}_{m_i \times m_i}$, see Sec. 2
$\boldsymbol{T} \leftarrow \boldsymbol{I} - \boldsymbol{\mathcal{E}}$
```

It is interesting to compare our algorithm to the Gaussian maximum likelihood independent vector analysis algorithm of [13], which has a QN-type structure similar to Algorithm 3. [13], too, minimize the KLD. However, the purpose of their algorithm is to block-diagonalize a single matrix.

4 Simulations

In this concluding section we compare the convergence rate of the algorithms, as well as the dependence of the criterion (5), on some of the parameters, in numerical experiments. The real positive-definite matrices $\boldsymbol{R}_S^{(q)}$, with block-pattern $\boldsymbol{m} = [4, 3, 2, 1]$, are drawn as $\boldsymbol{R}_{S,ii}^{(q)} = \boldsymbol{U}^\dagger \boldsymbol{U}$, where $\boldsymbol{U}$ is an $m_i \times m_i$ upper triangular matrix whose i.i.d. entries $\sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$, and the condition number of each $\boldsymbol{R}_{S,ii}^{(q)}$ is limited by 500, to assure proper invertibility. Matrices reflecting the latent $\widetilde{\boldsymbol{R}}_S^{(q)}$ are drawn from the Wishart distribution with n_q degrees of freedom, mimicking n_q observations at each $\mathcal{D}_q$. The stopping threshold is set to 10^{-4}. In the RG algorithm we set λ at each iteration using backtracking line search (lines 1–3 in Algorithm 2) with $\alpha = 0.3$, $\beta = 0.2$.

$\boldsymbol{A}$ is realized as $\boldsymbol{A} = \boldsymbol{I} + \boldsymbol{\Upsilon}$, where the entries of $\boldsymbol{\Upsilon}$ are i.i.d. and $\sim \mathcal{U}[-\frac{1}{4}, \frac{1}{4}]$. Since the contrast function (5) is invariant to block-diagonal scale ambiguity, we are concerned only about permutation ambiguity. The said choice of $\boldsymbol{A}$, together with initializing $\boldsymbol{A}$ with $\boldsymbol{I}$ (line 2 in Algorithm 1) allows for sufficient variability in our simulations, while usually assuring convergence to the desired minimum.

The convergence rate of the two algorithms is illustrated in Fig. 1a, on 20 realizations of $\boldsymbol{A}$, with fixed latent $\widetilde{\boldsymbol{R}}_S^{(q)}$. The fast convergence of the QN algorithm is very distinct from that of the RG algorithm. Both algorithms converge, eventually, in all trials, to the same value of the KL divergence, which illustrates the equivariance [9] property of the criterion.

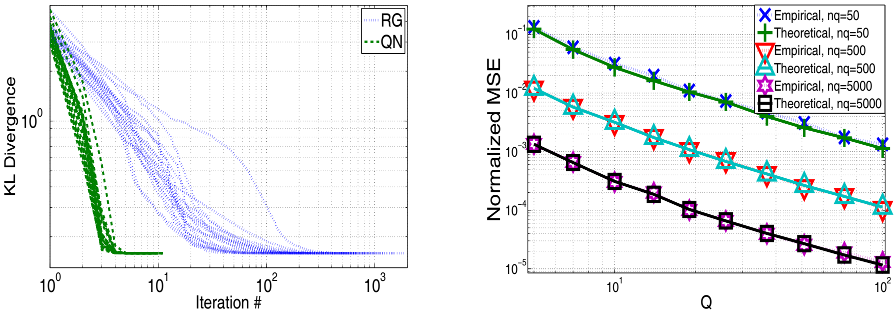
Once we have established that both algorithms converge to the same value, we shall demonstrate the separation quality of the criterion (5) as a function of Q and n_q. Since the QN realization is faster, it is chosen to be used in the following experiment. In the following simulations, for each Q and n_q, 40 trials were run, each with different $\boldsymbol{R}_S^{(q)}$, $\widetilde{\boldsymbol{R}}_S^{(q)}$ and $\boldsymbol{A}$. At each trial, $n_q = 50$, 500 or 500 and is fixed $\forall q$. Since the purpose of our JBD algorithm is component separation, the figure of merit is the MSE. Therefore, for each trial we evaluate the normalized empirical and theoretical MSE. The normalized empirical MSE is defined as (12),

$$\widehat{\mathrm{MSE}} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \frac{1}{T} \sum_{t=1}^{T} \|\widehat{\boldsymbol{x}}_i(t) - \boldsymbol{x}_i(t)\|^2 \quad , \quad \sigma_i^2 \triangleq \frac{1}{T} \sum_{t=1}^{T} E\{\|\boldsymbol{x}(t)\|^2\} \tag{12}$$

$$= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \mathrm{tr}\{\big(\langle \widetilde{\boldsymbol{R}}_X \rangle \otimes \boldsymbol{I}\big) \mathrm{vec}\{\delta \boldsymbol{P}_i\} \mathrm{vec}^{\dagger}\{\delta \boldsymbol{P}_i\}\} \tag{13}$$

and can be shown [6] to be equal to (13), where $\delta \boldsymbol{P}_i \triangleq \widehat{\boldsymbol{P}}_i - \boldsymbol{P}_i$. $\boldsymbol{P}_i$ are the $m \times m$ oblique projection matrices onto Span($\boldsymbol{A}_i$) along Span($\boldsymbol{A}_j$) $\forall j \neq i$. By definition, they satisfy $\boldsymbol{P}_i \boldsymbol{A}_j = \delta_{ij} \boldsymbol{A}_i$ so that $\boldsymbol{x}_i(t) = \boldsymbol{P}_i \boldsymbol{x}(t)$. As opposed to $\boldsymbol{A}$, $\boldsymbol{P}_i$ are invariant to scaling. With this notation, the estimated ith component can be obtained as $\widehat{\boldsymbol{x}}_i(t) = \widehat{\boldsymbol{P}}_i \boldsymbol{x}(t)$, with $\widehat{\boldsymbol{P}}_i = \widehat{\boldsymbol{A}}_i \widehat{\boldsymbol{B}}_i$, $\widehat{\boldsymbol{A}}$ an estimate of $\boldsymbol{A}$ and $\widehat{\boldsymbol{B}}_i$ the ith *horizontal* $m_i \times m$ block of $\widehat{\boldsymbol{B}} \triangleq \widehat{\boldsymbol{A}}^{-1}$. Therefore, the empirical MSE in Fig. 1b is evaluated using (13). As shown in [6], $E\{\widehat{\mathrm{MSE}}\}$ can be expressed explicitly in closed form, up to higher-order terms, as a function only of the model parameters $\boldsymbol{P}_i$ and $\boldsymbol{R}_X^{(q)}$. The "theoretical" data in Fig. 1b was obtained using that expression.

Each data point in Fig. 1b is the *average* of 40 empirical or theoretical MSE values with the same n_q and Q. Fig. 1b illustrates the decreases of the MSE with Q for fixed n_q, as well as with n_q for fixed Q. There is good match between theoretical and empirical values, which validates the convergence of the algorithm to the desired solution. The small values of the normalized MSE demonstrate the separation quality of the signals, in terms of the original multidimensional component separation problem.



(a) Convergence rate

(b) Normalized MSE vs. Q

Fig. 1. (a) Convergence rate of RG and QN algorithms, 20 trials each. Only $\boldsymbol{A}$ varies at each trial. $n_q = 100\ \forall q$. (b) Empirical and theoretical MSE vs. Q; n_q=50, 500 or 5000 $\forall q$. Each data point is averaged over 40 trials. $\boldsymbol{A}$, $\boldsymbol{R}_S^{(q)}$ and $\widetilde{\boldsymbol{R}}_S^{(q)}$ vary at each trial. In both subplots, block-pattern $\boldsymbol{m} = [4, 3, 2, 1]$, threshold=$10^{-4}$.

To conclude, we have presented two non-orthogonal JBD algorithms: QN and RG. These algorithms are capable of optimal, in the MSE sense, separation of multidimensional Gaussian piecewise stationary components, based on minimizing a KLD-based contrast function. Simulations demonstrate the proper convergence of these algorithms, given an appropriate initialization and under asymptotic conditions, to the theoretically-predicted MSE.

References

1. Nion, D.: A tensor framework for nonunitary joint block diagonalization. IEEE Trans. Signal Process. 59(10), 4585–4594 (2011)
2. Ghennioui, H., et al.: Gradient-based joint block diagonalization algorithms: Application to blind separation of FIR convolutive mixtures. Signal Process. 90(6), 1836–1849 (2010)
3. Bousbia-Salah, H., Belouchrani, A., Abed-Meraim, K.: Blind separation of non stationary sources using joint block diagonalization. In: Proc. SSP, pp. 448–451 (August 2001)
4. Pham, D.-T.: Joint approximate diagonalization of positive definite hermitian matrices. SIAM J. Matrix Anal. Appl. 22(4), 1136–1152 (2001)
5. Pham, D.T.: Blind Separation of Cyclostationary Sources Using Joint Block Approximate Diagonalization. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 244–251. Springer, Heidelberg (2007)
6. Lahat, D., Cardoso, J.-F., Messer, H.: Second-order multidimensional ICA: Performance analysis. Submitted to IEEE. Trans. Sig. Proc. (September 2011)
7. Comon, P.: Independent component analysis. In: Proc. Int. Signal Process. Workshop on HOS, Chamrousse, France, pp. 111–120 (July 1991); keynote address. Republished in *HOS*, J.-L. Lacoume ed., Elsevier, 1992, pp. 29–38
8. Pham, D.-T., Cardoso, J.-F.: Blind separation of instantaneous mixtures of non stationary sources. IEEE Trans. Signal Process. 49(9), 1837–1848 (2001)
9. Cardoso, J.-F., Laheld, B.: Equivariant adaptive source separation. IEEE Trans. Signal Process. 44(12), 3017–3030 (1996)
10. Pham, D.-T.: Information approach to blind source separation and deconvolution. In: Emmert-Streib, F., Dehmer, M. (eds.) Information Theory and Statistical Learning, ch.7, pp. 153–182. Springer, Heidelberg (2009)
11. Graham, A.: Kronecker Products and Matrix Calculus with Applications. Mathematics and its Applications. Ellis Horwood Ltd., Chichester (1981)
12. Gutch, H.W., Maehara, T., Theis, F.J.: Second Order Subspace Analysis and Simple Decompositions. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 370–377. Springer, Heidelberg (2010)
13. Vía, J., et al.: A Maximum Likelihood approach for Independent Vector Analysis of Gaussian data sets. In: Proc. MLSP 2011, Beijing, China (September 2011)

On Computation of Approximate Joint Block-Diagonalization Using Ordinary AJD*

Petr Tichavský[1,3], Arie Yeredor[2], and Zbyněk Koldovský[1,3]

[1] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic
tichavsk@utia.cas.cz

[2] Dept. of Electrical Engineering - Systems, School of Electrical Engineering,
Tel-Aviv University, P.O. Box 39040 Tel-Aviv, 69978 Israel

[3] Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic

Abstract. Approximate joint block diagonalization (AJBD) of a set of matrices has applications in blind source separation, e.g., when the signal mixtures contain mutually independent subspaces of dimension higher than one. The main message of this paper is that certain ordinary approximate joint diagonalization (AJD) methods (which were originally derived for "degenerate" subspaces of dimension 1) can also be used successfully for AJBD, but not all are suitable equally well. In particular, we prove that when the set is exactly jointly block-diagonalizable, perfect block-diagonalization is attainable by the recently proposed AJD algorithm "U-WEDGE" (uniformly weighted exhaustive diagonalization with Gaussian iteration) - but this basic consistency property is not shared by some other popular AJD algorithms. In addition, we show using simulation, that in the more general noisy case, the subspace identification accuracy of U-WEDGE compares favorably to competitors.

1 Introduction

Consider a set of square symmetric matrices $\mathbf{M}_i$, $i = 1, \ldots, N$, that are all block diagonal, with K blocks of size $L \times L$ along its main diagonal, $\mathbf{M}_i = \mathrm{Bdiag}(\mathbf{M}_{i1}, \ldots, \mathbf{M}_{iK})$, where $\mathbf{M}_{ik}$ is the $k-$th block of $\mathbf{M}_i$ and the Bdiag$(\cdot)$ operator constructs a block-diagonal matrix from its argument matrices. It follows that the dimension of the matrices is $LK \times LK$. An example of such matrices is illustrated in Figure 2(a) at the end of the paper. Note that the assumption that all blocks are of the same size is only used here to simplify the exposition, and can be relaxed via straightforward generalization.

Next, assume that (possibly perturbed) congruence transformations of these matrices are given as

$$\mathbf{R}_i = \mathbf{A}\mathbf{M}_i\mathbf{A}^T + \mathbf{N}_i, \qquad i = 1, \ldots, N \tag{1}$$

* This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project 102/09/1278.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 163–171, 2012.

where the superscript T denotes a matrix transposition, $\mathbf{A}$ is an unknown square "mixing matrix", and $\mathbf{N}_i$ is a perturbation (or "noise") matrix. We shall refer to the case where all $\mathbf{N}_i = \mathbf{0}$, $i = 1, \ldots, N$ as the "unperturbed" (or "noiseless") case. The choice of symbol $\mathbf{R}$ reflects the fact that the matrices in the set often play a role of (sample-) covariance matrices of a partitioned data, or time-lagged (sample-) covariance matrices.

The goal in Approximate Joint Block Diagonalization (AJBD) is to find a "demixing" matrix $\mathbf{W}$, such that the matrices

$$\widehat{\mathbf{M}}_i = \mathbf{W}\mathbf{R}_i\mathbf{W}^T, \qquad i = 1, \ldots, N \tag{2}$$

are all approximately block diagonal, having the blocks on the main diagonal of the same size as the original matrices $\mathbf{M}_{ik}$. Ideally, one may wish to estimate $\mathbf{W} = \mathbf{A}^{-1}$ and get $\widehat{\mathbf{M}}_i \approx \text{Bdiag}(\widehat{\mathbf{M}}_{i1}, \ldots, \widehat{\mathbf{M}}_{iK})$, where $\widehat{\mathbf{M}}_{ik} \approx \mathbf{M}_{ik}$.

In general, however, it is impossible to recover the original blocks $\mathbf{M}_i$ (even in the "noiseless" case), because of inherent ambiguities of the problem (e.g., [10]), but it is possible to recover "independent subspaces", as explained below.

Let $\mathbf{W}_0 = \mathbf{A}^{-1}$ be partitioned in K blocks $\mathbf{W}_k$ of size $L \times KL$, $\mathbf{W}_0 = [\mathbf{W}_1^T, \ldots, \mathbf{W}_K^T]^T$. Each block $\mathbf{W}_k$ represents a linear space of all linear combinations of its rows. These linear spaces are in general uniquely identifiable [10,4]. Let $\widehat{\mathbf{W}}$ be an estimated demixing matrix. We say that $\widehat{\mathbf{W}}$ is "essentially equivalent" to $\mathbf{W}_0$ (and therefore represents an ideal joint block diagonalization), if there exists a suitable $LK \times LK$ permutation matrix $\boldsymbol{\Pi}$ such that for each $k = 1, \ldots, K$ the subspaces spanned by $\mathbf{W}_k$ and by the respective k-th block of $\boldsymbol{\Pi}\widehat{\mathbf{W}}$ coincide (two subspaces are said to coincide if their mutual angle[1] is zero).

Some existing AJBD algorithms are restricted to the case where $\mathbf{A}$ (and therefore also $\widehat{\mathbf{W}}$) are orthogonal [5], some other algorithms consider a general matrix $\mathbf{A}$ [6,10]. In this paper, we examine the general case.

It is known that reasonable solutions to AJBD can be obtained using a two steps approach, by first applying an ordinary approximate joint diagonalization (AJD) algorithm, and then clustering the separated components (rows of the demixing matrix) [7,12]. In Section 3 we suggest a method for the clustering operation, followed by the main point of this paper: we show that not all AJD algorithms are equally suitable for such a two-steps AJBD approach. More specifically, we prove that unlike several popular AJD approaches, one recently proposed AJD method (U-WEDGE, Uniformly Weighted Exhaustive Diagonalization with Gauss itErations [14]) features a unique ability to attain ideal separation in the unperturbed ("noiseless") case, for general (not necessarily orthogonal) matrices $\mathbf{A}$. Our theoretical results are corroborated with simulation experiments in Section 4, both for the unperturbed and perturbed cases, showing the empirical advantages of U-WEDGE for the latter. We start, however, with a short overview of the AJD methods considered in this work. Their applicability in solving the block AJD problem is studied later in Section 4.

[1] The mutual angle between two subspaces can be obtained in Matlab® using the `subspace` function.

2 Survey of Main AJD Methods

Several well-known AJD methods are based on minimization of one of the three following criteria, possibly subject to one of the two constraints stated below.

$$C_{\mathrm{LS}}(\mathbf{W}) = \sum_{i=1}^{N} \|\mathrm{Off}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\|_F^2 \tag{3}$$

$$C_{\mathrm{LL}}(\mathbf{W}) = \sum_{i=1}^{N} \log \frac{\det \mathrm{Ddiag}(\mathbf{W}\widehat{\mathbf{R}}_i\mathbf{W}^T)}{\det(\mathbf{W}\widehat{\mathbf{R}}_i\mathbf{W}^T)} \tag{4}$$

$$C_{\mathrm{J2}}(\mathbf{W}) = \sum_{i=1}^{N} \|\mathbf{R}_i - \mathbf{W}^{-1}\mathrm{Ddiag}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\mathbf{W}^{-T}\|_F^2 \tag{5}$$

where the operator "Off" nullifies the diagonal elements, whereas "Ddiag" nullifies the off-diagonal elements of a square matrix, $\mathrm{Ddiag}(\mathbf{M}) = \mathbf{M} - \mathrm{Off}(\mathbf{M})$), and "$\|\cdot\|_F$" stands for the Frobenius norm. The possible associated constraints are

1. Each row of the estimated demixing matrix $\widehat{\mathbf{W}}$ has unit Euclidean norm.
2. $\widehat{\mathbf{W}}\mathbf{R}_1\widehat{\mathbf{W}}^T$ has an all-ones main diagonal.

The latter constraint usually corresponds (in the BSS context) to some scaling constraint on the estimated sources.

In the sequel we shall examine five AJD methods: QAJD [15], FAJD [9], LLAJD [11], QRJ2D [2] and WEDGE [14], especially in its unweighted version U-WEDGE. QAJD is based on minimization of the criterion (3) under the constraint 2. FAJD minimizes (3), penalized by a term proportional to $\log|\det \mathbf{W}|$. LLAJD minimizes (4) and QRJ2D minimizes (5), both under the constraint 1 (which is actually immaterial to the minimization in these cases).

WEDGE and its more simple unweighted (or uniformly-weighted) version U-WEDGE, which we consider in here, are different. U-WEDGE seeks a demixing matrix $\mathbf{W}$ which satisfies

$$\mathrm{argmin}_{\mathbf{A}} \sum_{i=1}^{N} \|\mathbf{W}\mathbf{R}_i\mathbf{W}^T - \mathbf{A}\,\mathrm{Ddiag}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\,\mathbf{A}^T\|_{\mathrm{F}}^2 = \mathbf{I} \tag{6}$$

where $\mathbf{I}$ is the $LK \times LK$ identity matrix. Roughly speaking, this implies that the set of matrices $\{\mathbf{W}\mathbf{R}_i\mathbf{W}^T\}$ cannot be jointly-diagonalized any further, since its "residual mixing" matrix, or its "best direct-form diagonalizer" (in the LS sense) is $\bar{\mathbf{A}} = \mathbf{I}$, the identity matrix.

It was shown in [14] that a necessary and sufficient condition for $\mathbf{A} = \mathbf{I}$ to be a stationary point of the criterion in (6) is a simpler set of nonlinear "normal equations",

$$\mathrm{Off}\left[\sum_{i=1}^{N}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\mathrm{Ddiag}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\right] = \mathbf{0}\ . \tag{7}$$

The more general WEDGE algorithm differs from U-WEDGE by incorporating special weight matrices in the quadratic criterion in (6). Although apparently complicated, both versions are computationally very efficient. Particular forms of WEDGE are used successfully in WASOBI for asymptotically optimal blind separation of stationary sources with spectrum diversity, and in BGSEP for separation of nonstationary sources [14]. Although our analytical proof and experiments in the sequel refer to the more simple version U-WEDGE, our experience shows that WEDGE shares the same ability of U-WEDGE to attain exact joint block-diagonalization in the unperturbed case.

3 AJD Methods in the Block Scenario

A natural extension of AJD methods in the block scenario is to replace the criterion (3) by

$$C_{\mathrm{BLS}}(\mathbf{W}) = \sum_{i=1}^{N} \|\mathrm{Boff}(\mathbf{W}\mathbf{R}_i\mathbf{W}^T)\|_F^2 \tag{8}$$

where the operator "Boff" nullifies the elements of a matrix that lie in the diagonal blocks. This is the main idea in [5].

It is obvious that since the criteria (3) and (8) are generally different, their minima differ as well, in general. If the diagonal blocks' sizes L are small, then one may expect the AJD and AJBD solutions to resemble. It is, however, necessary to permute (namely to properly cluster) the rows in the estimated demixing matrix, because the resulting order of rows is arbitrary in plain AJD algorithms.

3.1 Clustering of AJD Components

In this subsection a simple method of clustering the rows of de-mixing matrix is proposed. It allows to reveal (or at least to enhance) the block structure of the result. We suggest the following greedy algorithm: Given the AJD demixing matrix $\mathbf{W}$, compute an auxiliary matrix $\mathbf{B}$ as

$$\mathbf{B} = \sum_{i=1}^{N} |\mathbf{W}\mathbf{R}_i\mathbf{W}^{\mathbf{T}}| \ . \tag{9}$$

where the absolute value is taken elementwise. If the demixing is perfect, $\mathbf{B}$ should have, after arranging columns and rows, the same block structure as the original matrices $\mathbf{M}_i$. Take the first column of $\mathbf{B}$ and sort its elements decreasingly. Let $i_1, \ldots, i_L$ be the indices of the column elements with the L largest values. Then $\mathbf{W}_1$ is built of the rows of $\mathbf{W}$ with these indices. The rows and columns of $\mathbf{B}$ at the positions $i_1, \ldots, i_L$ are set to zero, and the procedure iterates further sorting of the column of $\mathbf{B}$ with the next nonzero elements, until all subspaces (blocks) $\mathbf{W}_k$, $k = 1, \ldots, K$, have been determined.

3.2 U-WEDGE Provides Perfect Separation of the Blocks

In this subsection we prove that in the unperturbed (“noiseless") case, U-WEDGE provides, upon convergence, perfect separation of the blocks.[2] Let $\mathbf{V}_k$ be the result of the hypothetical operation of applying U-WEDGE to each of the blocks-sets $\mathbf{M}_{1k}, \ldots, \mathbf{M}_{Nk}$ for $k = 1, \ldots, K$, where $\mathbf{M}_{ik}$ is the kth diagonal block of $\mathbf{M}_i$, $i = 1, \ldots, N$. It follows from (7) that each $\mathbf{V}_k$ obeys

$$\mathrm{Off}\left[\sum_{i=1}^{N}(\mathbf{V}_k\mathbf{M}_{ik}\mathbf{V}_k^T)\,\mathrm{Ddiag}(\mathbf{V}_k\mathbf{M}_{ik}\mathbf{V}_k^T)\right] = \mathbf{0}\ . \tag{10}$$

Now, define $\mathbf{W}_U$ as

$$\mathbf{W}_U = \mathrm{Bdiag}(\mathbf{V}_1, \ldots, \mathbf{V}_K)\,\mathbf{A}^{-1}\ . \tag{11}$$

It is straightforward to see that $\mathbf{W}_U$ is a U-WEDGE block diagonalizer of the original matrix set $\mathbf{R}_i = \mathbf{A}\mathbf{M}_i\mathbf{A}^T$, because it obeys the corresponding normal equation

$$\mathrm{Off}\left[\sum_{i=1}^{N}(\mathbf{W}_U\mathbf{R}_i\mathbf{W}_U^T)\,\mathrm{Ddiag}(\mathbf{W}_U\mathbf{R}_i\mathbf{W}_U^T)\right] = \mathbf{0}\ , \tag{12}$$

and on the other hand, that $\mathbf{W}_U\mathbf{R}_i\mathbf{W}_U^T$ has the perfect block-diagonal structure,

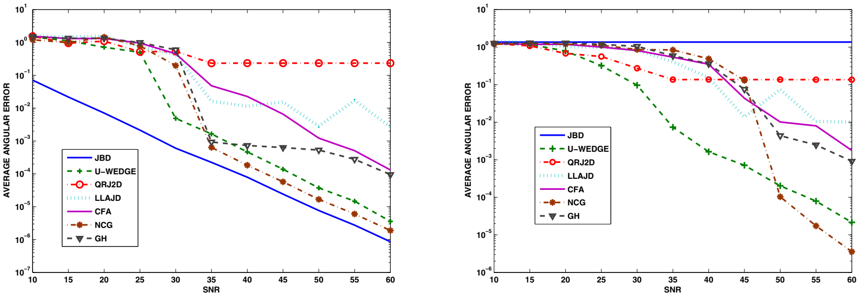
$$\mathbf{W}_U\mathbf{R}_i\mathbf{W}_U^T = \mathrm{Bdiag}(\mathbf{V}_1\mathbf{M}_{i1}\mathbf{V}_1^T, \ldots, \mathbf{V}_K\mathbf{M}_{iK}\mathbf{V}_K^T), \qquad i = 1, \ldots, N\ . \tag{13}$$

We note in passing, that since, as mentioned in [14], (7) is also a necessary condition for a solution of the FFDiag AJD algorithm [16], this property is shared by the latter as well.

4 Simulation Experiments

We first consider an experiment reflecting the unperturbed case, as shown in Figure 2 at the end of the paper. We generated $N = 3$ block-diagonal matrices $\mathbf{M}_i$, $i = 1, 2, 3$, of dimension 20×20, each containing four symmetric 5×5 blocks $\mathbf{M}_{ik}$ generated as $\mathbf{M}_{ik} = \mathbf{H}_{ik}\mathbf{H}_{ik}^T$, $\mathbf{H}_{ik}$ being random 5×5 matrices with independent standard Gaussian elements. The matrices $\mathbf{M}_i$ are shown in diagram (a). The 20×20 mixing matrix $\mathbf{A}$ was generated as random orthogonal, via QR decomposition of a random matrix. Diagram (b) shows raw results of applying U-WEDGE to the unperturbed set $\mathbf{R}_i = \mathbf{A}\mathbf{M}_i\mathbf{A}^T$, $i = 1, 2, 3$. Obviously, the block-diagonal structure of the results is obscured by residual random permutations in these

[2] Theoretically U-WEDGE can be stacked in a false solution [14], but in practice it is very rare, and the solution is unique up to well known permutation ambiguity.



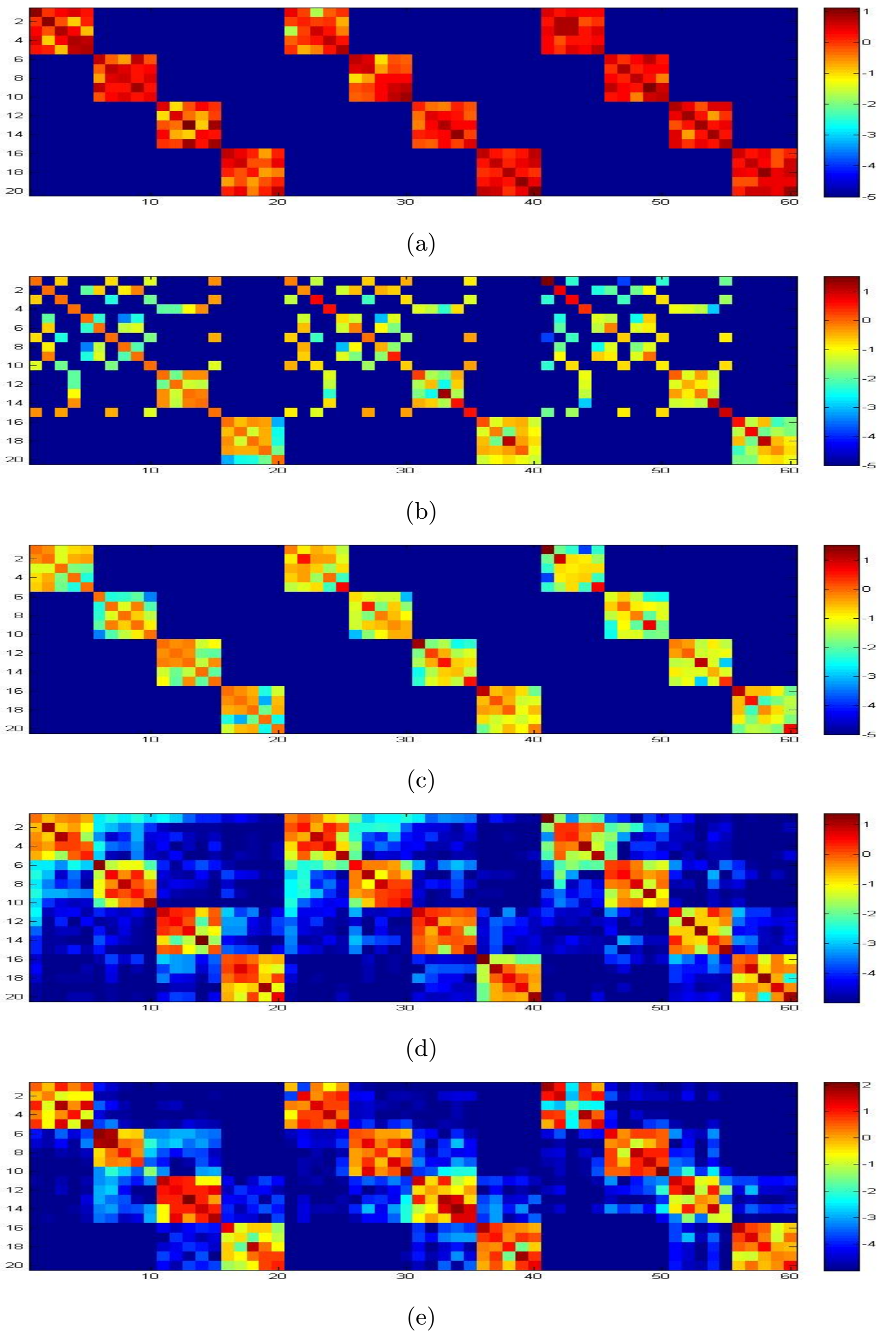
(a)

(b)

Fig. 1. Average subspace angular error for different AJD techniques versus SNR. (a) orthogonal mixing matrix, (b) random mixing matrix.

matrices. Diagram (c) shows the same matrices after applying the re-ordering procedure described in section 3.1. The angular error between the estimated and original subspaces (blocks of $\mathbf{W}$ and $\mathbf{A}^{-1}$) are zeros. Diagrams (d) and (e) show results obtained using the same procedure with the AJD algorithms QRJ2D and LLAJD. The average angular errors of the estimated subspaces were 4.6x10^{-3} (rad) and 1.3x10^{-4}(rad), respectively. The algorithms QAJD and FAJD were excluded, as they did not converge properly in this experiment.

In Figure 1 we proceed to compare the performance in the perturbed ("noisy") case. We plot the average angular subspace errors vs. the Signal-to-Noise Ratio (SNR) for the two-steps method using U-WEDGE, QRJ2D, LLAJD. For reference, we also compare to a unitary JBD algorithm [5], and three non-unitary algorithms: the closed form algorithm, utilizing only the first two matrices, labelled as CFA [10], the algorithm of Ghennioui et al, labelled as GH, [6], and the nonlinear conjugent gradient (NCG) of Nion [10]. The random noise matrices $\mathbf{N}_i$ were taken as symmetric with zero-mean entries, Gaussian-distributed with variance $10^{-\mathrm{SNR}/10}$. The average of the angular error is taken with respect to the four block and over 10 independent trials (with newly generated blocks and the noise, and the same mixing matrix $\mathbf{A}$). We consider both the case of orthogonal (Fig.1(a)) and non-orthogonal (Fig.1(b)) $\mathbf{A}$. We note that JBD (which assumes orthogonality) performs best in the former but fails in the latter. Among the AJD-based methods, U-WEDGE based AJBD usually attains the best results for moderate SNR's. It is outperformed by NCG, when the SNR is high. The worse performance of NCG at low SNR is probably due to getting the algorithm stacked in side local minima. Note a huge difference in computation speed. While one run of NCG takes cca 90 s, one run of U-WEDGE takes about 0.01 s of matlab running time on an ordinary PC with a 3GHz processor.



(a)

(b)

(c)

(d)

(e)

Fig. 2. Original and demixed matrices, displayed as $\log_{10}(|\mathbf{M}_i| + 10^{-5})$: (a) Original block-diagonal matrices (b) the matrices after mixing and de-mixing by U-WEDGE (c) the matrices after sorting row and columns (d) result for QRJ2D (e) result for LLAJD

5 Conclusions

We have shown theoretically and demonstrated in simulations that in the context of AJD-based AJBD, U-WEDGE attains an exact solution in the unperturbed case (with general mixing matrices), and usually performs better than other AJD algorithms in the perturbed case. The paper gives an explanation why the BG-WEDGE algorithm (which is similar) works so well in the time domain blind audio source separation [8].

Acknowledgements. The authors wish to give thanks to Dr. Dimitri Nion for sending them a matlab code of his algorithm JBD-NCG.

References

1. Afsari, B.: Sensitivity analysis for the problem of matrix joint diagonalization. SIMAX 30(3), 1148–1171 (2008)
2. Afsari, B.: Simple LU and QR Based Non-orthogonal Matrix Joint Diagonalization. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 1–7. Springer, Heidelberg (2006)
3. Abed-Meraim, K., Belouchrani, A.: Algorithms for Joint Block Diagonalization. In: Proc. of EUSIPCO 2004, Vienna, Austria, pp. 209–212 (2004)
4. de Lathauwer, L.: Decomposition of higher-order tensor in block terms - Part II: definitions and uniqueness. SIAM J. Matrix Anal. and Appl. 30(3), 1033–1066 (2008)
5. Févotte, C., Theis, F.J.: Pivot Selection Strategies in Jacobi Joint Block-Diagonalization. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 177–184. Springer, Heidelberg (2007)
6. Ghennioui, H., et al.: A Nonunitary Joint Block Diagonalization Algorithm for Blind Separation of Convolutive Mixtures of Sources. IEEE Signal Processing Letters 14(11), 860–863 (2007)
7. Koldovský, Z., Tichavský, P.: A Comparison of Independent Component and Independent Subspace Analysis Algorithms. In: EUSIPCO 2009, Glasgow, Scotland, April 24-28, pp. 1447–1451 (2009)
8. Koldovský, Z., Tichavský, P.: Time-domain blind separation of audio sources based on a complete ICA decomposition of an observation space. IEEE Tr. Audio, Speech, and Language Processing 19(2), 406–416 (2011)
9. Li, X.-L., Zhang, X.D.: Nonorthogonal joint diagonalization free of degenerate solutions. IEEE Tr. Signal Processing 55(5), 1803–1814 (2007)
10. Nion, D.: A Tensor Framework for Nonunitary Joint Block Diagonalization. IEEE Tr. Signal Processing 59(10), 4585–4594 (2011)
11. Pham, D.-T.: Joint approximate diagonalization of positive definite Hermitian matrices. SIAM J. Matrix Anal. and Appl. 22(4), 1136–1152 (2001)
12. Szabó, Z., Póczos, B., Lörincz, A.: Separation theorem for independent subspace analysis and its consequences. Pattern Recognition 45(4), 1782–1791 (2012)

13. Tichavský, P.: Matlab code for U-WEDGE, WEDGE, BG-WEDGE and WASOBI, http://si.utia.cas.cz/Tichavsky.html
14. Tichavský, P., Yeredor, A.: Fast Approximate Joint Diagonalization Incorporating Weight Matrices. IEEE Tr. Signal Processing 57(3), 878–891 (2009)
15. Vollgraf, R., Obermayer, K.: Quadratic optimization for simultaneous matrix diagonalization. IEEE Tr. Signal Processing 54(9), 3270–3278 (2006)
16. Ziehe, A., Laskov, P., Nolte, G., Müller, K.-R.: A Fast Algorithm for Joint Diagonalization with Non-orthogonal Transformations and its application to Blind Source Separation. Journal of Machine Learning Research 5, 777–800 (2004)

Joint Diagonalization of Several Scatter Matrices for ICA

Klaus Nordhausen[1], Harold W. Gutch[2,3], Hannu Oja[1], and Fabian J. Theis[3,4]

[1] University of Tampere, Finland
[2] Max Planck Institute for Dynamics and Self-Organization, Germany
[3] Technical University Munich, Germany
[4] Helmholtz-Institute Neuherberg, Germany

Abstract. Procedures such as FOBI that jointly diagonalize two matrices with the independence property have a long tradition in ICA. These procedures have well-known statistical properties, for example they are prone to failure if the sources have multiple identical values on the diagonal. In this paper we suggest to diagonalize jointly $k \geq 2$ scatter matrices having the independence property. For the joint diagonalization we suggest a novel algorithm which finds the correct direction in an deflation based manner, one after another. The method is demonstrated in a small simulation study.

Keywords: ICA, scatter matrix, independence property, joint diagonalization.

1 Introduction

The independent component (IC) model is a well-established semiparametric model: Let $\mathbf{x}$ be a p-variate random vector. Then the most basic IC model is

$$\mathbf{x} = \boldsymbol{\Omega}\mathbf{z}, \tag{1}$$

where $\mathbf{z}$ is an unobservable p-variate source vector having independent components and $\boldsymbol{\Omega}$ is an unknown $p \times p$ full-rank mixing matrix. Any $p \times p$ matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}\mathbf{x}$ has independent components is called an unmixing matrix. Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be a random sample from a distribution obeying the model (1). The goal of independent component analysis (ICA) is then to find an estimate $\hat{\boldsymbol{\Gamma}}$ (based on $\mathbf{X}$) for some unmixing matrix $\boldsymbol{\Gamma}$. For an overview for different estimation procedures, see for example [1,2]. A family of estimates based on the joint diagonalization of two scatter matrices with the so called independence property has been recently proposed [3,4,5]. In this paper we extend this family by jointly diagonalizing k scatter matrices, $k \geq 2$. We motivate also why this is an improvement over diagonalizing only two matrices. For the joint diagonalization we use a new algorithm which finds the directions of the unmixing matrix in a deflation based manner. This allows us to develop the statistical theory (convergence, asymptotic normality) for the new estimates.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 172–179, 2012.

The structure of the paper is as follows. In Sections 2 and 3, we first define the concepts of scatter functionals and independent component (IC) functionals, and recall how two scatter matrices with the independence property can be used to find an IC functional. Then, in Section 4, we introduce our new family of IC functionals and estimates based on $k \geq 2$ scatter matrices, and describe the statistical properties of the estimates. The new algorithm for this procedure is introduced in Section 5, and the paper is concluded with a small simulation study.

Due to space restrictions proofs and further results will be published in an extended version of this paper.

2 Scatter Functionals

Location and scatter functionals are generally used to describe the properties of multivariate distributions in wide nonparametric and semiparametric models. Let $\mathbf{x}$ be a p-variate random vector with cumulative distribution function (cdf) $F_{\mathbf{x}}$. A *location functional* is then a p-vector valued functional $\mathbf{T}(F_{\mathbf{x}})$ that is affine invariant in the sense that

$$\mathbf{T}(F_{\mathbf{Ax+b}}) = \mathbf{AT}(F_{\mathbf{x}}) + \mathbf{b},$$

for all full-rank $p \times p$ matrices $\mathbf{A}$ and all p-vectors $\mathbf{b}$. A *scatter functional* is a $p \times p$ matrix valued functional $\mathbf{S}(F_{\mathbf{x}})$ which is symmetric, psd and affine invariant in the sense that

$$\mathbf{S}(F_{\mathbf{Ax+b}}) = \mathbf{AS}(F_{\mathbf{x}})\mathbf{A}^T,$$

again for all full rank $p \times p$ matrices $\mathbf{A}$ and p-vectors $\mathbf{b}$. If only $\mathbf{S}(F_{\mathbf{Ax+b}}) \propto \mathbf{AS}(F_{\mathbf{x}})\mathbf{A}^T$ is true then $\mathbf{S}$ is called a *shape functional*. For any scatter functional $\mathbf{S}$, functionals $(p/\text{trace}(\mathbf{S}))\mathbf{S}$ and $(\det(\mathbf{S}))^{-1/p}\mathbf{S}$ are shape functionals, for example. If $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a random sample from $F_{\mathbf{x}}$ and F_n is the empirical cdf based on $\mathbf{X}$, then the sample statistics $\mathbf{T}(F_n) = \mathbf{T}(\mathbf{X})$ and $\mathbf{S}(F_n) = \mathbf{S}(\mathbf{X})$ are natural estimates of $\mathbf{T}(F_{\mathbf{x}})$ and $\mathbf{S}(F_{\mathbf{x}})$, respectively.

There are several families of location and scatter functionals proposed in the literature. The simultaneous M-functionals for location and scatter are often defined by the implicit equations

$$\mathbf{T}(\mathbf{x}) = E(w_1(r))^{-1}\mathbf{E}(w_1(r)\mathbf{x}) \text{ and } \mathbf{S}(\mathbf{x}) = \mathbf{E}(w_2(r)(\mathbf{x} - \mathbf{T}(\mathbf{x}))(\mathbf{x} - \mathbf{T}(\mathbf{x}))^T),$$

where $r = [(\mathbf{x} - \mathbf{T}(\mathbf{x}))^T\mathbf{S}(\mathbf{x})^{-1}(\mathbf{x} - \mathbf{T}(\mathbf{x}))]^{1/2}$, and $w_1(r)$ and $w_2(r)$ are two nonnegative continuous weight functions. Some interesting special cases are

1. the mean vector and the covariance matrix ($w_1(r) = w_2(r) = 1$),
2. Hettmansperger-Randles (HR) functional, see [6], with $w_1(r) = 1/r$ and $w_2(r) = p/r^2$,
3. Huber's functional, $w_1(r) = \begin{cases} 1 & r \leq c \\ c/r & r > c \end{cases}$ and $w_2(r) = \begin{cases} 1/\sigma^2 & r \leq c \\ c/(r^2\sigma^2) & r > c \end{cases}$, where c is a tuning constant and σ^2 a scaling parameter (see [7]), and

4. ML-estimate for the Cauchy distribution, see [8], with $w_1(r) = w_2(r) = (p+1)/(r^2+1)$.

For other families of location and scatter functionals such as MM-, CM- and S-functionals, see [9].

In practice, M-estimates, that is, the values of the M-functionals at sample cdf, are computed using a fixed point algorithm based on the estimating equations above. One-step M-functionals $(\mathbf{T}_1, \mathbf{S}_1)$ are then obtained using initial functionals $(\mathbf{T}_0, \mathbf{S}_0)$ and one step

$$\mathbf{T}_1(\mathbf{x}) = E(w_1(r))^{-1}\mathbf{E}(w_1(r)\mathbf{x}) \text{ and } \mathbf{S}_1(\mathbf{x}) = \mathbf{E}(w_2(r)(\mathbf{x}-\mathbf{T}_0(\mathbf{x}))(\mathbf{x}-\mathbf{T}_0(\mathbf{x}))^T),$$

where now $r = [(\mathbf{x}-\mathbf{T}_0(\mathbf{x}))^T\mathbf{S}_0(\mathbf{x})^{-1}(\mathbf{x}-\mathbf{T}_0(\mathbf{x}))]^{1/2}$. Important special cases are

1. a scatter matrix based on fourth moments $\mathbf{COV}_4$ starting with $(\mathbf{E}, \mathbf{COV})$ and using $w_2(r) = r^2/(p+2)$, and
2. one-step Hallin-Paindaveine (HP) functional starting with the HR estimate and using $w_2(r) = \psi_p^{-1}(F_r(r))/r$ where ψ_p is the cdf of a χ^2_p-distribution, see [10].

If $\mathbf{x}$ has an elliptically symmetric distribution, then $\mathbf{S}(F_\mathbf{x}) \propto \mathbf{COV}(\mathbf{x})$, but this is not true in the IC model. For ICA, it is important that the scatter functional possesses the so called *independence property* meaning that, if $\mathbf{x}$ has independent components then $\mathbf{S}(\mathbf{x})$ is a diagonal matrix. For most families of scatter functionals mentioned above, this is not generally true. It is easy to see that $\mathbf{COV}$ and $\mathbf{COV}_4$ have the independence property. However, for any scatter functional $\mathbf{S}$ not having this property, its symmetrized version $\mathbf{S}_{sym}$ has the property. ($\mathbf{S}_{sym}(\mathbf{x}) = \mathbf{S}(\mathbf{x}_1 - \mathbf{x}_2)$ where $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent copies of $\mathbf{x}$, see [11,12].) Unfortunately, symmetrized scatter estimates are computational intensive. Note however that, if $\mathbf{x}$ has independent components and at most one component is skew then all scatter functionals will be diagonal [13].

3 Independent Component (IC) Functionals

Let $\mathcal{C}$ be the set of $p \times p$ matrices that have exactly one non-zero element in each row and each column. Assume that $\mathbf{x} = \boldsymbol{\Omega}\mathbf{z}$ obeys the model (1). Then an *independent component (IC) functional* $\boldsymbol{\Gamma}(F)$ is a $p \times p$ matrix valued functional that satisfies

$$(i)\ \boldsymbol{\Gamma}(F_\mathbf{x})\boldsymbol{\Omega} \in \mathcal{C} \quad \text{and} \quad (ii)\ \boldsymbol{\Gamma}(F_{\mathbf{Ax+b}}) = \boldsymbol{\Gamma}(F_\mathbf{x})\mathbf{A}^{-1},$$

for all full-rank $\mathbf{A}$ and all $\mathbf{b}$. IC functionals can be constructed, however, only in submodels of (1); it is well known for example that at most one independent component can be gaussian.

The IC functional $\boldsymbol{\Gamma}(F)$ based on the two scatter matrix functionals $\mathbf{S}_1(F)$ and $\mathbf{S}_2(F)$ with the independence property is defined by the estimating equations

$$\boldsymbol{\Gamma}\mathbf{S}_1\boldsymbol{\Gamma}^T = \mathbf{I}_p \text{ and } \boldsymbol{\Gamma}\mathbf{S}_2\boldsymbol{\Gamma}^T = \boldsymbol{\Lambda},$$

where $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(F)$ is a diagonal matrix (functional) with diagonal elements $\lambda_1 \geq \ldots \geq \lambda_p > 0$. The matrix $\boldsymbol{\Gamma}$ then jointly diagonalizes both $\mathbf{S}_1$ and $\mathbf{S}_2$. Unfortunately, $\boldsymbol{\Gamma}$ is an IC functional (up to sign changes of its rows) only in the submodel (1) with distinct eigenvalues $\lambda_1 > \ldots > \lambda_p > 0$ of $\mathbf{S}_1^{-1}\mathbf{S}_2$. (Note that these eigenvalues, do not depend on $\boldsymbol{\Omega}$.) This may be seen a serious restriction as then only the components with distinct eigenvalues can be recovered. To avoid this problem, one can then try to diagonalize simultaneously $k > 2$ scatter matrices.

The use of two scatter matrices in ICA has been studied in [3,4] (real data), and in [5] (complex data). The limiting statistical properties of the estimates have been studied in [14]. One of the first solutions for the ICA problem, the FOBI functional [15], is obtained with scatter functionals $\mathbf{COV}$ and $\mathbf{COV}_4$, respectively. Eigenvalues in $\boldsymbol{\Lambda}$ are simple functions of the classical moment based kurtosis measures. FOBI is highly non-robust as it is based on fourth moments - robust ICA estimates are obtained with robust choices of $\mathbf{S}_1$ and $\mathbf{S}_2$. JADE [16] may be seen as an extension of FOBI as it jointly diagonalizes several cumulant matrices in order to avoid the problem of identical eigenvalues. JADE has however two drawbacks: (i) It is not affine invariant (and therefore not an IC functional) and (ii) it is highly non-robust.

4 Joint Diagonalization of Several Scatter Functionals

Let now $\mathbf{S}_1, ..., \mathbf{S}_k$ be k scatter functionals with the independence property, $k \geq 2$. The general idea is to find a $p \times p$ matrix $\boldsymbol{\Gamma}$ that minimizes $\sum_{i=2}^{k} ||\text{off}(\boldsymbol{\Gamma}\mathbf{S}_i\boldsymbol{\Gamma}^T)||^2$ under the constraints $\boldsymbol{\Gamma}\mathbf{S}_1\boldsymbol{\Gamma}^T = \mathbf{I}_p$. As $\mathbf{S}_1, ..., \mathbf{S}_k$ are scatter matrices, it is equivalent to maximize $\sum_{i=2}^{k} ||\text{diag}(\boldsymbol{\Gamma}\mathbf{S}_i\boldsymbol{\Gamma}^T)||^2$ under the same constraints. To fix the order of the independent components (rows of $\boldsymbol{\Gamma}$) we require that the diagonal elements of $\sum_{i=2}^{k}(\text{diag}(\boldsymbol{\Gamma}\mathbf{S}_i\boldsymbol{\Gamma}^T))^2$ are in decreasing order. The functional $\boldsymbol{\Gamma}$ is then an independent component (IC) functional in a wider submodel (1) than any of the IC functionals based on pairs of scatter matrices $(\mathbf{S}_1, \mathbf{S}_i)$, $i = 2, ..., k$, only. It is sufficient that only $\mathbf{S}_1$ is a scatter matrix and $\mathbf{S}_2, ..., \mathbf{S}_k$ are shape matrices with the same trace or determinant. This guarantees that none of the matrices dominates too much the others.

Write $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)^T$. The columns $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p$ can then be solved one by one so that $\boldsymbol{\gamma}_j$, $j = 1, ..., p-1$ maximizes

$$G_j(\boldsymbol{\gamma}_j) = \sum_{i=2}^{k} (\boldsymbol{\gamma}_j^T \mathbf{S}_i \boldsymbol{\gamma}_j)^2$$

under the constraints $\boldsymbol{\gamma}_r^T \mathbf{S}_i \boldsymbol{\gamma}_j = \delta_{rj}$, $r = 1, ..., j$. Using Lagrangian multiplier technique, one then obtains estimating equations equations

$$\mathbf{T}(\boldsymbol{\gamma}_j) = \mathbf{S}_1(\sum_{r=1}^{j} \boldsymbol{\gamma}_r \boldsymbol{\gamma}_r^T)\mathbf{T}(\boldsymbol{\gamma}_j), \quad j = 1, ..., p-1,$$

where $\mathbf{T}(\gamma) = \sum_{i=2}^{k}[(\gamma^T \mathbf{S}_i \gamma)\mathbf{S}_i]\gamma$. One can then show that, under general assumptions, if $\sqrt{n}(vec(\hat{\mathbf{S}}_1, ..., \hat{\mathbf{S}}_k) - vec(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k))$ has a limiting multivariate normal distribution then so does $\sqrt{n}(\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma})$.

5 Practical Implementation

The implementation of the method described in the previous section is straightforward and follows the ideas of the deflation-based fastICA algorithm (see for example [1]). The basic idea is use $\mathbf{S}_1$ to whiten the data and then find the rows of an orthogonal transformation matrix one by one to jointly diagonalize the reminding $k-1$ scatter (or shape) matrices computed for the whitened data. To fix the extraction order in practice, the initial value of the orthogonal matrix is the matrix of the eigenvectors of $\mathbf{S}_1^{-1/2}\mathbf{S}_2\mathbf{S}_1^{-1/2}$. The algorithm is then as follows:

```
program k-scatter for ICA
    S1 = S1(X) # S1 for the original data
    Z = S1^{-1/2} X # whiten the data using S1
    SK = array(S_i(Z)) # i=2,...,k.
    SK[,,1] = UDU’ # eigendecomposition of S2
    W = 0 # pxp matrix with 0’s for storing the results
    for (i in 1:p){
        wn <- U’[,i]
        for (it in 1:maxiter){
            w <- wn
            wn <- matrix(0,p,1)
            for (mi in 1:k-1){
                # calculate the gradient
                wn <- wn +  SK[,,mi] * w * w’ * SK[,,mi] * w
            }
            wn <- wn - W’ * W * wn
            wn <- wn / norm(wn)
            if (norm(w-wn) < eps || norm(w+wn) < eps) break
            if (it == maxiter) stop("no convergence reached")
        }
        W[i,] <- wn’
    }
    W <- W*S1^{-1/2}
   return W
```

6 Simulation

A small simulation study was used to evaluate the finite sample performance and robustness of the new procedure. Comparisons were made to the procedures based on two scatter matrices only as well as to classical fastICA and JADE algorithms. The performance criterion used in the comparisons was the minimum distance criterion,

$$MD(\hat{\boldsymbol{\Gamma}}, \boldsymbol{\Omega}) = \frac{1}{\sqrt{p-1}} \inf_{\mathbf{C} \in \mathcal{C}} ||\mathbf{C}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Omega} - \mathbf{I}_p||.$$

The range of the MD index is $[0,1]$, and $MD = 0$ means an optimal separation. For details about the index, see [17].

We compare the performance of the IC estimates in four different 4-variate settings,

1. z_1 has a χ^2_4 distribution, z_2 a t_7 distribution, z_3 has a logistic distribution, and z_4 has a $N(0,1)$ distribution. The variables are standardized so that $E(z_i) = 0$ and $Var(z_i) = 1$, $i = 1, ..., 4$. Marginal kurtosis values are $\kappa_1 = 3$, $\kappa_2 = 2$, $\kappa_3 = 1.2$, and $\kappa_4 = 0$.
2. As setting 1 but 5 % of the $\mathbf{z} = (z_1, ..., z_4)^T$ are replaced by outliers having $N_4(\mathbf{0}, \boldsymbol{\Sigma})$ distribution with $\boldsymbol{\Sigma} = \mathbf{UDU}'$ where $\mathbf{U}$ is a random orthogonal matrix and $\mathbf{D} = diag(100, 50, 20, 0.001)$.
3. z_1 has a χ^2_{10} distribution, z_2 has a t_9 distribution, z_3 has a logistic distribution, and z_4 has a power exponential distribution with shape parameter 1.3401. Again, after the standardization, $E(z_i) = 0$ and $Var(z_i) = 1$, $i = 1, ..., 4$. Now $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 1.2$ meaning, for example, that FOBI should not work.
4. As setting 3 but with outliers as in setting 2.

Note that there is only one skew component in each setting. Therefore all the scatter matrices have the independence property here. The estimation methods to be compared are (i) non-robust `2S_FOBI` using **COV** and $\mathbf{COV}_4$, (ii) robust `2S_ROB` using HR and Huber scatter matrices, (iii) non-robust `kS_FOBI` using **COV**, $\mathbf{COV}_4$, HR, and Huber scatter matrices, and (iv) robust `kS_ROB` using HP, spatial rank (see [3]), HR, Huber and Cauchy scatter matrices. The scale differences were eliminated so that, in all cases, $\mathbf{S}_2, ..., \mathbf{S}_5$ were standardized to have determinant 1. The comparisons were made also to (v) the deflation-based fastICA algorithm with nonlinearity function tanh and a random initial matrix, to (vi) the JADE estimate, and to (vii) a (reference) random guess matrix where the elements $\hat{\boldsymbol{\Gamma}}$ are iid from a $N(0,1)$ distribution.

In all settings we choose $\mathbf{A} = \mathbf{I}_p$. Note that the methods based on scatter functionals do not depend on the choice of $\mathbf{A}$. This is not true for JADE which complicates the comparison. Note also that the performance of deflation-based fastICA depends on how the initial value of the estimate is chosen in the algorithm. With a random initial value, the estimate is not affine equivariant [18].

Figure 1 shows the average performance of the 7 estimates in four settings with samples sizes $n = 200, 500, 1000, 2000$ and with 1000 repetitions in each case. The results are remarkable. The estimates based on five scatter matrices clearly outperform those based on two matrices only. The new estimates seem to work well already for relative small sample sizes. Only in setting 1 and with large sample sizes, fastICA and JADE are better than the new procedures. JADE,

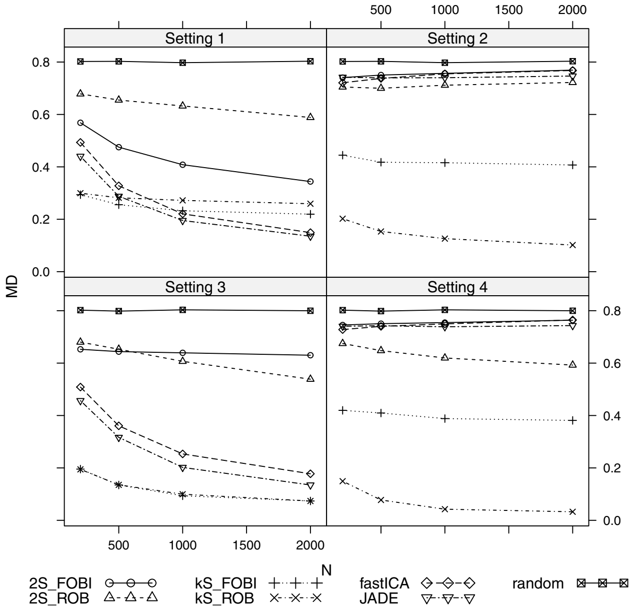


Fig. 1. Average MD values in the simulations over 1000 repetitions

fastICA, `2S_FOBI` and `kS_FOBI` suffer severely when outliers are present. The average MD values of both `2S_FOBI` and `2S_ROB` converge extremely slowly to zero. In fact, there is no convergence for `2S_FOBI` in setting 3 (and 4) as all the four kurtosis values are the same. In general the advantage of using more than 2 (robust) scatter matrices is overwhelming in this small simulation study.

7 Conclusions

We suggested in this paper a novel method for ICA which is based on the joint diagonalization of $k > 2$ scatter matrices. This new method can be seen as an extension of the ICA method based on two scatter functionals as well as an extension of JADE. The new approach is valid for a larger family of distributions than the two scatter functionals approach. Compared to JADE, the new estimates have the advantage of being affine equivariant. Our novel deflation-based algorithm for joint diagonalization makes the method also analytically tractable. Due to space restrictions the theoretic properties like limiting normality are just outlined and detailed results with limiting covariance matrices of the estimates, for example, will be presented in an extended version of the paper. A small simulation study showed that the new approach has a high efficiency and it is seems highly robust if only robust scatter functionals are employed. For an extended version of the paper, larger simulations with other model selections and several other choices of scatter matrices are necessary for a better understanding of the estimation procedure.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley & Sons, New York (2001)
2. Theis, F.J., Inouye, Y.: On the Use of Joint Diagonalization in Blind Signal Processing. In: Proceedings of IEEE International Symposium on Circuits and Systems, ISCAS 2006, pp. 3586–3589 (2006)
3. Oja, H., Sirkiä, S., Eriksson, J.: Scatter Matrices and Independent Component Analysis. Austrian Journal of Statistics 35, 175–189 (2006)
4. Nordhausen, K., Oja, H., Ollila, E.: Robust Independent Component Analysis Based on Two Scatter Matrices. Austrian Journal of Statistics 37, 91–100 (2008)
5. Ollila, E., Oja, H., Koivunen, V.: Complex-valued ICA Based on a Pair of Generalized Covariance Matrices. Computational Statistics & Data Analysis 52, 3789–3805 (2008)
6. Hettmansperger, T.P., Randles, R.H.: A Practical Affine Equivariant Multivariate Median. Biometrika 89, 851–860 (2002)
7. Huber, P.J.: Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics 35, 73–101 (1964)
8. Kent, J.T., Tyler, D.E., Vardi, Y.: A Curious Likelihood Identity for the Multivariate t-distribution. Communications in Statistics, Theory and Methods 23, 441–453 (1994)
9. Maronna, R.A., Martin, D.M., Yohai, V.J.: Robust Statistics. Theory and Methods. Wiley & Sons, Chichester (2006)
10. Hallin, M., Paindaveine, D.: Semiparametrically Efficient Rank-based Inference for Shape. I. Optimal Rank-based Tests for Sphericity. Annals of Statistics 34, 2707–2756 (2006)
11. Sirkiä, S., Taskinen, S., Oja, H.: Symmetrised M-estimators of Multivariate Scatter. Journal of Multivariate Analysis 98, 1611–1629 (2007)
12. Roelant, E., Van Aelst, S., Croux, C.: Multivariate Generalized S-estimators. Journal of Multivariate Analysis 100, 876–887 (2009)
13. Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H.: Invariant co-ordinate selection. Journal of the Royal Statistical Society 71, 549–592 (2009)
14. Ilmonen, P., Nevalainen, J., Oja, H.: Characteristics of Multivariate Distributions and the Invariant Coordinate System. Statistics & Probability Letters 80, 1844–1853 (2010)
15. Cardoso, J.F.: Source Separation Using Higher Order Moments. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, pp. 2109–2112 (1989)
16. Cardoso, J.-F., Souloumiac, A.: Blind Beamforming for non Gaussian Signals. IEE Proceedings-F 140, 362–370 (1993)
17. Ilmonen, P., Nordhausen, K., Oja, H., Ollila, E.: A New Performance Index for ICA: Properties, Computation and Asymptotic Analysis. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 229–236. Springer, Heidelberg (2010)
18. Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., Ollila, E.: Deflation-based FastICA Reloaded. In: Proceedings of 19th European Signal Processing Conference 2011 (EUSIPCO 2011), pp. 1854–1858 (2011)

To Infinity and Beyond: On ICA over Hilbert Spaces

Harold W. Gutch[1,2] and Fabian J. Theis[2,3]

[1] Max Planck Institute for Dynamics and Self-Organization,
Department of Nonlinear Dynamics, Göttingen, Germany
[2] Technical University Munich, Germany
[3] Helmholtz-Institute Neuherberg, Germany

Abstract. The original Independent Component Analysis (ICA) problem of blindly separating a mixture of a finite number of real-valued statistically independent one-dimensional sources has been extended in a number of ways in recent years. These include dropping the assumption that all sources are one-dimensional and some extensions to the case where the sources are not real-valued. We introduce an extension in a further direction, no longer assuming only a finite number of sources, but instead allowing infinitely many. We define a notion of independent sources for this case and show separability of ICA in this framework.

1 Introduction

Independent Component Analysis (ICA) has become a standard approach to the Blind Source Separation (BSS) problem: Assume N real valued signal sources $\mathbf{S} = (S_1, \ldots, S_N)$ that are not directly given, but instead only an M-dimensional mixture $\mathbf{X} = f(\mathbf{S})$ for some invertible function f of the sources can be observed. Given now only $\mathbf{X}$, the task is reconstruction of $\mathbf{S}$. In the most simple setting $N = M$ and f is linear (so it can be written as a square matrix $\mathbf{A}$). On a formal level, $\mathbf{S}$ is here modeled as an N-dimensional real valued random vector, and ICA assumes that the N components of $\mathbf{S}$ are stochastically independent. In this case, it is known [1] that if at most one of the sources has a Gaussian distribution one can indeed uniquely reconstruct $\mathbf{S}$ from $\mathbf{X}$ up to possible scaling and permutation indeterminacies. We show that if one instead models $\mathbf{S}$ as a random vector taking values in a real Hilbert space (intuitively, this can be visualized as the limit case of $N \to \infty$), under some mild conditions (which correspond to no additional assumptions in the finite dimensional case) $\mathbf{S}$ can be again recovered given only $\mathbf{X}$, up to possible scaling and permutation indeterminacies. While more application-oriented, functionalPCA is a related approach.

This paper is organized as follows. In Section 2 we repeat the definition of Hilbert spaces and their most elementary properties. Section 3 contains an introduction into random variables with values in Hilbert spaces, and how many key concepts there can be reduced to real valued random variables. We then have the tools to prove separability of the ICA model in this setting in Section 4 before concluding with a short discussion, open questions and further directions of the model in Section 5.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 180–187, 2012.

2 Vector Spaces and Hilbert Spaces

For any integer $d > 0$, the d-dimensional real vector space $\mathbf{V} = \mathbb{R}^d$ can be visualized as the set of ordered d-fold tuples of reals, $\mathbf{v} = (v_1, \ldots, v_d)$ where addition of two of these is defined component-wise, $(\mathbf{v}+\mathbf{w})_k := v_k + w_k$, and multiplication of a $\mathbf{v} \in \mathbf{V}$ with a $\lambda \in \mathbb{R}$ is defined by $(\lambda\mathbf{v})_k := \lambda v_k$. A set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_l\}$ is said to be linearly independent if $\sum_{k=1}^{l} \lambda_k \mathbf{v}_k = 0$ implies $\lambda_k = 0$ for all k. Such a set can contain at most d vectors, and in that case it is said to be a basis. Fixing a basis $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_d\}$, any $\mathbf{v} \in \mathbf{V}$ can be written as a weighted sum of the basis vectors, $\mathbf{v} = \sum_{k=1}^{d} \lambda_k \mathbf{e}_k$, with unique coefficients λ_k. A linear function $f : \mathbf{V} \to \mathbf{V}$ (i.e., $f(\mathbf{v}+\mathbf{w}) = f(\mathbf{v}) + f(\mathbf{w})$ and $f(\lambda\mathbf{v}) = \lambda f(\mathbf{v})$ for all $\mathbf{v}, \mathbf{w} \in \mathbf{V}$, $\lambda \in \mathbb{R}$) is uniquely determined by the values it takes on a basis. We can furthermore employ $\mathbf{V}$ with an inner product by letting $\langle \sum_{k=1}^{d} \lambda_k \mathbf{e}_k, \sum_{k=1}^{d} \mu_k \mathbf{e}_k \rangle := \sum_{k=1}^{d} \lambda_k \mu_k$. This operation is bilinear, symmetric and positive definite so it induces a norm by $||\mathbf{v}|| := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ and gives us a notion of orthogonality by $\mathbf{v} \perp \mathbf{w} :\Leftrightarrow \langle \mathbf{v}, \mathbf{w} \rangle = 0$. The basis is said to be *orthonormal*, as for all $i \neq j$ then $\mathbf{e}_i \perp \mathbf{e}_j$ and $||\mathbf{e}_i|| = 1$. Note that the inner product depends on the basis.

2.1 Hilbert Spaces

Generally, a real vector space $\mathbf{V}$ is defined as a set that, together with the inner sum $+$, forms a commutative group (i.e., $\mathbf{v}+\mathbf{w} = \mathbf{w}+\mathbf{v}$ and $\mathbf{u}+(\mathbf{v}+\mathbf{w}) = (\mathbf{u}+\mathbf{v})+\mathbf{w}$ for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{V}$, existence of some $0 \in \mathbf{V}$ such that $0+\mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in \mathbf{V}$ and for every $\mathbf{v} \in \mathbf{V}$ existence of some $\mathbf{w} \in \mathbf{V}$ such that $\mathbf{v}+\mathbf{w} = 0$) and that is employed with a scalar multiplication such that for all $\lambda, \mu \in \mathbb{R}$, $\mathbf{v}, \mathbf{w} \in \mathbf{V}$: a) $(\lambda\mu)\mathbf{v} = \lambda(\mu\mathbf{v})$ (associativity), b) $(\lambda+\mu)\mathbf{v} = \lambda\mathbf{v} + \mu\mathbf{v}$ and $\lambda(\mathbf{v}+\mathbf{w}) = \lambda\mathbf{v} + \lambda\mathbf{w}$ (distributivity), and c) $1\mathbf{v} = \mathbf{v}$. This definition allows to define linear independence as before, and if there then is an integer d such that there is a set of d linearly independent vectors, but every set of $d+1$ vectors is dependent, we say that $\mathbf{V}$ has dimension d. In this case this formal definition fully corresponds to the intuitive visualization from above. However the vector space axioms do not demand existence of such an integer. Assume for example the set of all polynomials in one variable t with real coefficients, which easily can be verified to be a real vector space. Obviously for any d, the set $\{0, t, \ldots, t^d\}$ contains $d+1$ linearly independent elements.

It turns out that in this setting it makes more sense not to first fix an orthonormal basis and then to let this basis induce a scalar product, but rather to directly fix a scalar product, and then possibly choose a basis that is orthonormal with respect to the scalar product.

Definition 1. *Let $\mathcal{H}$ be a real vector space and $\langle .,. \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ a scalar product on $\mathcal{H}$, i.e., $\langle .,. \rangle$ is symmetric, bilinear and positive definite. If then for any sequence $(\mathbf{x}_n)_{n\in\mathbb{N}}$ of elements of $\mathcal{H}$ where* $\lim_{m,n\to\infty} ||\mathbf{x}_m - \mathbf{x}_n|| = 0$ *(this limit is understood to hold for any sequence of indices $m, n \to \infty$) there exists* *some* $\mathbf{x} \in \mathcal{H}$ *such that* $\lim_{n\to\infty} ||\mathbf{x} - \mathbf{x}_n|| = 0$ *(where* $||.||$ *is defined as* $||\mathbf{x}|| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ *for arbitrary* $\mathbf{x} \in \mathcal{H}$*), then* $\mathcal{H}$ *is said to be a Hilbert space*[1]*.*

Before we look at random variables on Hilbert spaces, we repeat some well known facts and conventions. A set $\mathcal{I}$ is said to be countable if there is a one-to-one correspondence between $\mathcal{I}$ and the natural numbers $\mathbb{N}$[2]. A set $\mathbf{A} \subset \mathcal{H}$ is said to be an orthonormal set if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ for all $\mathbf{x} \neq \mathbf{y}$ in $\mathbf{A}$ and if $||\mathbf{x}|| = 1$ for all $\mathbf{x} \in \mathbf{A}$. If $\mathbf{A}$ is an orthonormal set, then for every $\mathbf{x} \in \mathcal{H}$ there is at most a countable number of $\mathbf{y} \in \mathbf{A}$ such that $\langle \mathbf{x}, \mathbf{y} \rangle \neq 0$. In this case the expression $\sum_{\mathbf{y}\in\mathbf{A}} \langle \mathbf{x}, \mathbf{y} \rangle \mathbf{y} =: \mathbf{P}_\mathbf{A}(\mathbf{x})$ converges and is independent of the order of terms in the sum and we call $\mathbf{P}_\mathbf{A}$ the orthogonal projection onto the subspace generated by $\mathbf{A}$. If $\mathbf{P}_\mathbf{A}(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in \mathcal{H}$, then $\mathbf{A}$ is said to be an orthonormal basis. In this case the coefficients $\langle \mathbf{x}, \mathbf{y} \rangle$ are called the *Fourier coefficients* of $\mathbf{x}$ with respect to $\mathbf{A}$. Generally, for an orthonormal sequence $(\mathbf{y}_i)_{i\in\mathbb{N}}$ and reals $(\lambda_i)_{i\in\mathbb{N}}$, the expression $\sum_{i\in\mathbb{N}} \lambda_i \mathbf{y}_i$ converges iff $\sum_{i\in\mathbb{N}} \lambda_i^2 < \infty$, so the sum over the squares of the Fourier coefficients of any $\mathbf{x} \in \mathcal{H}$ is finite. The Hilbert space $\mathcal{H}$ is said to be separable if there is a countable sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ such that for any $\mathbf{x} \in \mathcal{H}$ and any $\varepsilon \in \mathbb{R}$ there is some $\mathbf{x}_k$ such that $||\mathbf{x} - \mathbf{x}_k|| < \varepsilon$. This is the case if and only if $\mathcal{H}$ has a countable basis. In applications, Hilbert spaces are often encountered as *function spaces*, i.e., the elements themselves are functions, e.g. on $\mathbb{R}$ or $\mathbb{C}$. In order to distinguish between these elements of $\mathcal{H}$ and functions $f : \mathcal{H}_1 \to \mathcal{H}_2$ *between* Hilibert spaces, the latter are denoted *operators*. As usual, an operator f is said to be continuous if for every ε there is some δ such that $||f(\mathbf{x}) - f(\mathbf{y})|| < \varepsilon$ whenever $||\mathbf{x} - \mathbf{y}|| < \delta$ (i.e., if small changes in the inputs cause only small changes of the image). In the finite dimensional case every linear operator is continuous, but in the infinite dimensional case a linear operator $\mathbf{A}$ is continuous if and only if it is bounded, i.e., if $||\mathbf{A}|| := \sup\{||\mathbf{A}(\mathbf{x})|| : ||\mathbf{x}|| \leq 1\} < \infty$. This is not always the case, take for example the "differential operator" $D : \mathbf{e}_k \mapsto k\mathbf{e}_{k-1}$. For every linear and bounded operator $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ there is a linear and bounded operator $\mathbf{B} : \mathcal{H} \to \mathcal{H}$ such that $\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{B}\mathbf{y} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, and $\mathbf{B}$ is called the *adjoint of* $\mathbf{A}$, in symbols: $\mathbf{B} = \mathbf{A}^\dagger$. If $\mathbf{A} : \mathcal{H} \to \mathbb{R}$ is linear and bounded, there is some $\mathbf{y} \in \mathcal{H}$ such that $\mathbf{A}(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$ for every $\mathbf{x} \in \mathcal{H}$ (Riesz' representation theorem). An operator $f : \mathcal{H} \to \mathbb{R}$ is said to be (weakly) differentiable at $\mathbf{x}$ if for every direction $\mathbf{h} \in \mathcal{H}$ $\lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{h}) - f(\mathbf{x})}{t}$ exists, and it is called the derivative of f at $\mathbf{x}$ with respect to the direction $\mathbf{h}$. If $f : \mathcal{H} \to \mathbb{R}$ is linear and bounded, i.e., $f(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$ for some $\mathbf{x} \in \mathcal{H}$, then $f' \equiv \mathbf{y}$.

3 Statistics on Hilbert Spaces

We now give an overview of statistics on Hilbert spaces. We restrict this introduction to only the essentials and refer to [2], Ch. 8 for a good explanation on

[1] Sometimes Hilbert spaces are assumed to be infinite dimensional in order to distinguish them from "usual" vector spaces, but we do not make this distinction – for us, a finite dimensional vector space just is a special case of a Hilbert space.

[2] Some authors assume only an injection from $\mathcal{I}$ to $\mathbb{N}$, also calling finite sets "countable" and then also using the term "countably infinite" for sets with a bijection to $\mathbb{N}$.

the basics of statistics on Hilbert spaces, and to [3] for a more technical and axiomatic approach.

3.1 Random Variables on Hilbert Spaces

Let $\mathcal{H}$ be a real, separable Hilbert space with scalar product $\langle .,. \rangle$ (and induced norm $||.||$). We fix a countable orthonormal basis of $\mathcal{H}$ and denote it as $\{\mathbf{e}_k | k \in \mathbb{N}\} = \{\mathbf{e}_1, \mathbf{e}_2, \dots\}$. Just like in the finite dimensional case, open spheres around some $\mathbf{x} \in \mathcal{H}$ with radius $r \in \mathbb{R}$ are defined as $B_r(\mathbf{x}) := \{\mathbf{y} \in \mathcal{H} : ||\mathbf{y} - \mathbf{x}|| < r\}$. We then also define $\mathcal{B}(\mathcal{H})$, the *Borel σ-algebra* (or σ-field) generated by the open spheres in $\mathcal{H}$, i.e., the smallest family of sets that contains all open spheres in $\mathcal{H}$ and is closed under taking of complements and countable unions.

An $\mathcal{H}$-valued random variable is a measurable function $\mathbf{X}$ from a probability space $(\Omega, \mathcal{F}, P)$ to $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ and $\mathcal{B}(\mathcal{H})$ is the set of events of $\mathbf{X}$. The probability of an event $A \in \mathcal{B}(\mathcal{H})$ is given via the induced measure on $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$, defined as $P_{\mathbf{X}}(A) := P(\mathbf{X}^{-1}(A)) = P(\{\omega \in \Omega : \mathbf{X}(\omega) \in A\})$. The induced measure of a random variable $\mathbf{X}$ is also called its *distribution* or *law*.

3.2 Independence of Infinitely Many Components

While independence of a finite number of random variables is a well-known concept, less is known about this idea in the infinite setting. In this case independence of an infinite number of objects reduces to the definition in the finite case as follows.

Let $\mathcal{I}$ be an index set (possibly infinite) for a family of events $\mathcal{A} := \{A_i \in \mathcal{B}(\mathcal{H}) : i \in \mathcal{I}\}$. Then $\mathcal{A}$ is said to be independent if for every finite index subset $\{i_1, \dots, i_N\} \subseteq \mathcal{I}$, the probability of the joint event factorizes into the probabilities of the single events:

$$P_{\mathbf{X}}(A_{i_1} \cap \ldots \cap A_{i_N}) = P_{\mathbf{X}}(A_{i_1}) \cdots P_{\mathbf{X}}(A_{i_N}) \ .$$

Of particular interest in ICA are the components of a random vector, i.e., the projections of $\mathbf{X}$ to the basis vectors $\mathbf{e}_k$. These are given by the (real) random variables $X_k := \langle \mathbf{X}, \mathbf{e}_k \rangle$, and if $\{i_1, \dots, i_N\}$ is a set of N indices, we say that the components $X_{i_1}, \dots, X_{i_N}$ are independent if for every choice of events $(A_1, \dots, A_N)$ (where A_k is an event of X_{i_k}) the events $A_1, \dots, A_N$ are independent. Independence of an *infinite* number of components $\mathcal{I} \subset \mathbb{N}$ again is defined by demanding that every *finite* subset of components $\{i_1, \dots, i_N\} \subset \mathcal{I}$ be independent. This is the case if and only if the probability for *any finite* tuple of joint events of these N random variables factorizes into the product of the probabilities of the single events, or, in other words, if the law of $(X_{i_1}, \dots, X_{i_N})$ factorizes into the single laws:

$$P_{(X_{i_1}, \dots, X_{i_N})} = P_{X_{i_1}} \cdots P_{X_{i_N}} \ .$$

We propose the following definition of *independence of a random vector* on $\mathcal{H}$:

Definition 2. *Let $\mathcal{H}$ be an infinite dimensional, separable Hilbert space with basis $\{\mathbf{e}_i : i \in \mathbb{N}\}$ and $\mathbf{X} : (\Omega, \mathcal{F}, P) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$ a random variable with values in $\mathcal{H}$. We say that $\mathbf{X}$ is independent if for any finite index set $\mathcal{I} = \{i_1, \ldots, i_N\}$*

1. *The components $X_{i_1}, \ldots, X_{i_N}$ are independent*
2. *$\mathbf{P}_{\mathbf{e}_\mathcal{I}}\mathbf{X}$ and $(\mathbf{Id} - \mathbf{P}_{\mathbf{e}_\mathcal{I}})\mathbf{X}$ are independent, (where $\mathbf{P}_{\mathbf{e}_\mathcal{I}}(\mathbf{x}) = \sum_{i \in \mathcal{I}} \langle \mathbf{x}, \mathbf{e}_i \rangle \mathbf{e}_i$).*

Note that in the finite dimensional case only the first assumption is required. The reason to include the second assumption is as follows. Simply demanding the components $\{X_1, X_2, \ldots\}$ to be independent is per definition equivalent to independence of any finite set of components. But every such set obviously is a true subset of the components of $\mathbf{X}$. In the finite dimensional case, an N-dimensional random vector $\mathbf{X}$ is independent if and only if every finite subset (including the set $\{1, \ldots, N\}$ itself) is independent, but it is *not* sufficient if every *true* subset of components is independent. Consider for example a random vector where the first $N-1$ components are i.i.d. samples from $\mathcal{N}(0, 1)$. The last component then is constructed by first (independently) sampling again from $\mathcal{N}(0, 1)$. It then is either multiplied by (-1) or not, such that after this the number of components of this sample with non-negative values is even. Projection of this random vector onto the subspace given by any set of $N-1$ axes results in an independent $N-1$ dimensional Gaussian, but the whole random vector clearly is not independent. As we cannot rule out existence of infinite dimensional random variables similar to the ones we just described, where every finite dimensional restriction is independent, but which are not independent in the sense of our definition, we also demand independence of the restriction to the rest, as in the second assumption of the definition.

3.3 Moments

As in the finite dimensional case, statistics on Hilbert spaces are intrinsically connected to integration with respect to a distribution. If $\mathbf{X}$ is a finite random variable on $\mathcal{H}$ (i.e., one taking only finitely many values), say $\mathbf{X}(A_k) = h_k \in \mathcal{H}$ for disjoint sets $A_1, \ldots, A_N$ where $\cup_{k=1}^N A_k = \Omega$, then one defines for all $M \in \mathcal{F}$

$$\int_M \mathbf{X} dP := \int_M \mathbf{X}(\omega) P(d\omega) := \sum_{k=1}^N h_k P(M \cap A_k) \ .$$

This integral has the usual properties of additivity and linearity, and as expected, $|| \int_M \mathbf{X} dP || \leq \int_M ||\mathbf{X}|| dP$ (note that $||\mathbf{X}||$ is a real-valued random variable, so the latter expression is the usual Lebesgue integral). We are mostly interested in integrals over the whole probability space Ω, and the random variable $\mathbf{X}$ is said to be *Bochner integrable* or simple *integrable* if $\int_\Omega ||\mathbf{X}|| dP < \infty$. Assume in this case a sequence of finite random variables $(\mathbf{X}_k)_{k \in \mathbb{N}}$ on $\mathcal{H}$ that converges pointwise to $\mathbf{X}$ (i.e., $\lim_{k \to \infty} \mathbf{X}_k(\omega) = \mathbf{X}(\omega)$ for every $\omega \in \Omega$). Then the expectation (or mean) of $\mathbf{X}$ is

$$E[\mathbf{X}] := \int_\Omega \mathbf{X} dP := \lim_{k \to \infty} \int_\Omega \mathbf{X}_k dP < \infty$$

(one can show that if $\mathbf{X}$ is integrable, this limit is independent of the choice of the sequence $(\mathbf{X}_k)_{k\in\mathbb{N}}$, so this expression is well-defined). Equivalently, $E[\mathbf{X}]$ is the unique value $\mathbf{m}_1 \in \mathcal{H}$ for which $\langle \mathbf{x}, \mathbf{m}_1 \rangle = \int_\Omega \langle \mathbf{X}, \mathbf{m}_1 \rangle dP$ reducing the mean to a Lebesgue integral over the reals (note that the expression in the integral on the right hand side is real valued). If $\mathbf{A}$ is a bounded, linear operator and $\mathbf{A}(\mathbf{X})$ is integrable, then $\mathbf{A}E[\mathbf{X}] = E[\mathbf{A}(\mathbf{X})]$.

More generally, existence of the p-th moment (for an integer p) is defined as $\int_\Omega ||\mathbf{X}||^p dP < \infty$ and if the p-th moment exists, then the q-th moment exists for every $q < p$. If the second moment exists, the covariance of $\mathbf{X}$ is a symmetric, positive definite, bilinear form and its value at $(\mathbf{x}, \mathbf{y})$ is defined as

$$\mathrm{Cov}(\mathbf{X})(\mathbf{x}, \mathbf{y}) := \int_\Omega \langle \mathbf{X} - E[\mathbf{X}], \mathbf{x} \rangle \langle \mathbf{X} - E[\mathbf{X}], \mathbf{y} \rangle dP \ .$$

This expression corresponds to $\mathbf{x}^\top \mathrm{Cov}(\mathbf{X})\mathbf{y}$ in the finite dimensional case, where the main interest lies in the entries of $\mathrm{Cov}(\mathbf{X})$, the finite dimensional counterparts to expressions of the kind $\mathrm{Cov}(\mathbf{X})(\mathbf{e}_i, \mathbf{e}_j)$. One can show that the covariance has finite trace $\sum_{k=1}^\infty \mathrm{Cov}(\mathbf{X})(\mathbf{e}_k, \mathbf{e}_k) < \infty$ if it exists. This prohibits the standard ICA preprocessing step of assuming unit covariance of the sources and rescaling of the observations to unit covariance.

3.4 Characteristic Function

The characteristic function of a random variable $\mathbf{X}$ with values on a real Hilbert space $\mathcal{H}$ is the complex valued function on $\mathcal{H}$ defined as

$$\widehat{\mathbf{X}}(\mathbf{x}) := E[\exp(i\langle \mathbf{X}, \mathbf{x} \rangle)] \ .$$

For every $\mathbf{x} \in \mathcal{H}$, the characteristic function of $\mathbf{X}$ is reduced to the value of the characteristic function of the (real) random variable $\langle \mathbf{X}, \mathbf{x} \rangle$ at 1. Therefore the characteristic function of $\mathbf{X}$ always exists and just as in the real case $\widehat{\mathbf{X}}(0) = 1$. If $\mathbf{X}$ is independent, then $\widehat{\mathbf{X}}(\mathbf{x}) = \sum_{k=1}^\infty \widehat{X_k}(x_k)$ where $X_k := \langle \mathbf{X}, \mathbf{e}_k \rangle$ is the k-th component of $\mathbf{X}$ and $x_k := \langle \mathbf{x}, \mathbf{e}_k \rangle$ is the k-th component of the vector $\mathbf{x} \in \mathcal{H}$.

4 Separability of ICA on Hilbert Spaces

Let now $\mathbf{S} : \Omega \to \mathcal{H}$ be a random variable with values in $\mathcal{H}$. Assume $\mathbf{S}$ to have independent components, none of these to be normally distributed, and let $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ be an arbitrary invertible bounded linear operator. We will show that if the components of $\mathbf{AS}$ again are independent, then $\mathbf{A}$ can only map single components of $\mathbf{S}$ to single components of $\mathbf{AS}$, but it can not perform any mixing of components. In formal terms: For every index i there is exactly one j such that $\langle \mathbf{A}\mathbf{e}_i, \mathbf{e}_j \rangle \neq 0$. This guarantees that reconstruction of $\mathbf{S}$ is possible (apart from permutation and scaling) given only $\mathbf{AS}$, as it suffices to find an invertible, linear and bounded $\mathbf{W}$ such that $\mathbf{WAS}$ again is independent. Indeed, in this case

(**WA**) will then not perform any mixing, so it can be represented as at most a permutation followed by some scaling.

We will make use twice of the following lemma in the main theorem. The proof of the lemma is straight-forward.

Lemma 1. *Assume twice differentiable operators $f_k : \mathcal{H} \to \mathbb{C}$, where $k \in \mathbb{N}$ such that the infinite product $\prod_{k=1}^{\infty} f(\mathbf{x})$ converges to a twice differentiable function $f(\mathbf{x})$. Then*

$$f \partial_i \partial_j f - (\partial_i f)(\partial_j f) \equiv \sum_{k=1}^{\infty} \Big(\prod_{l \neq k} f_l \Big)^2 \Big[f_k (\partial_i \partial_j f_k) - (\partial_i f_k)(\partial_j f_k) \Big] .$$

Using this, we can now proceed to the main theorem.

Theorem 1. *Let $\mathcal{H}$ be a real, separable Hilbert space and $\mathbf{S}$ an independent random variable with values in $\mathcal{H}$ whose characteristic function is twice differentiable. Let $\mathbf{A}$ be an invertible bounded linear operator on $\mathcal{H}$ and assume that $\mathbf{X} := \mathbf{AS}$ again is independent. Whenever $k \in \mathbb{N}$ such that there are two indices $i \neq j \in \mathbb{N}$ fulfilling $\langle \mathbf{e}_i, \mathbf{Ae}_k \rangle \neq 0 \neq \langle \mathbf{e}_j, \mathbf{Ae}_k \rangle$, then the k-th component of $\mathbf{S}$ has to be normally distributed.*

Proof. As $\mathbf{X}$ has independent components, $\widehat{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{\infty} \widehat{X_k}(x_k)$, therefore evaluating the right hand side of Lemma 1 with $f_k(\mathbf{x}) = \widehat{X_k}(\langle \mathbf{e}_k, \mathbf{x} \rangle)$ tells us that $\widehat{\mathbf{X}} \partial_i \partial_j \widehat{\mathbf{X}} - (\partial_i \widehat{\mathbf{X}})(\partial_j \widehat{\mathbf{X}}) \equiv 0$ whenever $i \neq j$. Furthermore

$$\widehat{\mathbf{X}}(\mathbf{x}) = E[\exp(i \langle \mathbf{AS}, \mathbf{x} \rangle)] = \widehat{\mathbf{S}}(\mathbf{A}^{\dagger} \mathbf{x}) = \prod_{k=1}^{\infty} \widehat{S_k}(\langle \mathbf{A}^{\dagger} \mathbf{x}, \mathbf{e}_k \rangle) = \prod_{k=1}^{\infty} \widehat{S_k}(\langle \mathbf{x}, \mathbf{Ae}_k \rangle)$$

so defining $g_k(\mathbf{x}) := \widehat{S_k}(\langle \mathbf{x}, \mathbf{Ae}_k \rangle)$ and again applying Lemma 1 to the equality $\widehat{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{\infty} g_k(\mathbf{x})$ yields

$$0 \equiv \sum_{k=1}^{\infty} \Big(\prod_{l \neq k} g_l \Big)^2 \Big[g_k (\partial_i \partial_j g_k) - (\partial_i g_k)(\partial_j g_k) \Big] . \tag{1}$$

In order to further simplify this expression we calculate the partial derivatives of g_k. Note that $\partial_i \langle \mathbf{x}, \mathbf{Ae}_k \rangle = \langle \mathbf{e}_i, \mathbf{Ae}_k \rangle$, so the chain rule tells us $(\partial_i g_k)(\mathbf{x}) = \widehat{S_k}'(\langle \mathbf{x}, \mathbf{Ae}_k \rangle) \langle \mathbf{e}_i, \mathbf{Ae}_k \rangle$ and $(\partial_i \partial_j g_k)(\mathbf{x}) = \widehat{S_k}''(\langle \mathbf{x}, \mathbf{Ae}_k \rangle) \langle \mathbf{e}_i, \mathbf{Ae}_k \rangle \langle \mathbf{e}_j, \mathbf{Ae}_k \rangle$. Setting $s_k := \langle \mathbf{x}, \mathbf{Ae}_k \rangle$ and plugging these expressions into Eqn. (1) then yields

$$0 = \sum_{k=1}^{\infty} \Big(\prod_{l \neq k} \widehat{S}_l(s_l) \Big)^2 \langle \mathbf{e}_i, \mathbf{Ae}_k \rangle \langle \mathbf{e}_j, \mathbf{Ae}_k \rangle \Big[\widehat{S_k}(s_k) \widehat{S_k}''(s_k) - \widehat{S_k}'(s_k) \widehat{S_k}'(s_k) \Big] \tag{2}$$

for all $\mathbf{x} \in \mathcal{H}$. Let $\mathbf{s} := \sum_{k=1}^{\infty} s_k \mathbf{e}_k$, then one readily verifies that $\mathbf{A}^{\dagger} \mathbf{x} = \mathbf{s}$. As $\mathbf{A}$ (and therefore $\mathbf{A}^{\dagger}$) is invertible, this shows that Eqn. (2) also holds for all possible $\mathbf{s} \in \mathcal{H}$. Assume now some $\mathbf{s}_0$ such that $\widehat{\mathbf{S}}(\mathbf{s}_0) \neq 0$ (such a point exists as

$\widehat{\mathbf{S}}(0) = 1$), which then also holds in a neighborhood U of $\mathbf{s}_0$, as this is an open condition. Dividing Eqn. (2) by $\widehat{\mathbf{S}}(\mathbf{s})^2$ shows

$$0 = \sum_{k=1}^{\infty} \widehat{S_k}(s_k)^{-2} \langle \mathbf{e}_i, \mathbf{A}\mathbf{e}_k \rangle \langle \mathbf{e}_j, \mathbf{A}\mathbf{e}_k \rangle \left[\widehat{S_k}(s_k) \widehat{S_k}''(s_k) - \left(\widehat{S_k}'(s_k) \right)^2 \right] .$$

This holds for all possible values of $\mathbf{s} \in U$. As we are free to independently modify every single coordinate in this equality, every single summand already has to be 0:

$$0 = \widehat{S_k}(s_k)^{-2} \langle \mathbf{e}_i, \mathbf{A}\mathbf{e}_k \rangle \langle \mathbf{e}_j, \mathbf{A}\mathbf{e}_k \rangle \left[\widehat{S_k}(s_k) \widehat{S_k}''(s_k) - \left(\widehat{S_k}'(s_k) \right)^2 \right] .$$

Here $\widehat{S_k}(s_k)^{-2} \neq 0$, as $\widehat{\mathbf{S}}(\mathbf{s}) \neq 0$. But if now also $\langle \mathbf{e}_i, \mathbf{A}\mathbf{e}_k \rangle \langle \mathbf{e}_j, \mathbf{A}\mathbf{e}_k \rangle \neq 0$, then $\widehat{S_k}(s_k) \widehat{S_k}''(s_k) - \left(\widehat{S_k}'(s_k) \right)^2 = 0$, so $\widehat{S_k}(s_k) = \exp(as_k^2 + bs_k + c)$ for some parameters $a, b, c \in \mathbb{C}$. Due to continuity this then holds not only in U, but also in its closure $\overline{U}$, therefore in all $\mathcal{H}$, proving that S_k has a Gaussian distribution.

We have shown that if for an independent random vector $\mathbf{S}$ also $\mathbf{X} := \mathbf{A}\mathbf{S}$ is independent for an invertible, bounded operator $\mathbf{A}$, then any source S_k of $\mathbf{S}$ that gets mixed into more than one of observations (the components of $\mathbf{X}$), has to be normally distributed (note that this includes deterministic components as a special case). Conversely, if none of the sources has a Gaussian distribution, $\mathbf{A}$ can only map single sources to single observations.

5 Discussion

We have introduced a definition of independence of a random vector on a Hilbert space that is suitable for the ICA application. Using this, we have shown separability of the ICA model in the context of infinite dimensional random variables. Apart from the obvious generalization of our model to complex Hilbert spaces future work can be done towards generalizations of the model to also include subspace structures. Here there are three possible settings: an infinite number of subspaces all of which have finite size; a finite number of subspaces, where at least some have infinite size; and finally an infinite number of subspaces where at least some have infinite size. Finally, the two questions of actual implementation and application to a (possibly highly idealized) real world setting are still completely open.

References

1. Comon, P.: Independent component analysis - a new concept? Signal Processing 36, 287–314 (1994)
2. Bonaccorsi, S., Priola, E.: From Brownian Motion to Stochastic Differential Equations. In: 10th Internet Seminar
3. da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Cambridge University Press (1992)

Regularized Sparse Representation for Spectrometric Pulse Separation and Counting Rate Estimation

Tom Trigano and Yann Sepulcre

Shamoon College of Engineering,
Department of Electrical Engineering
Jabotinski 84, 77245 Ashdod, Israel
{thomast,yanns}@sce.ac.il
http://www.sce.ac.il

Abstract. One of the objectives of nuclear spectroscopy is to estimate the varying counting rate activity of unknown radioactive sources. When this activity is high, however, nonparalyzable detectors suffer from a type of distortion called pile-up effect, when pulses created from different sources tend to overlap. This distortion leads to an underestimation of the activity, which explains the interest of methods for individual pulse separation. We suggest in this paper a two-step method for a better counting rate estimation: the signal is first approximated using a block-sparse regression method, allowing to separate individual pulses quite well. We then estimate their arrival times and plug them into a known activity estimator. Results on simulations and real data illustrate the efficiency of the proposed approach.

Keywords: Gamma spectrometry, Group LASSO, sparse representation, pileup separation.

1 Introduction

Gamma spectrometry experiments aim to identify radioactive sources as well as their activity. In a given experiment, photons interact with a detector at random times, creating electrical pulses which are afterwards analyzed [4]. However, when the activity of the radioactive source is high, generated pulses may start at very close times and overlap, thus leading to an underestimation of the activity. An example of the pileup phenomenon is presented in Figure 1: if a simple threshold is used on the red part of the signal, we detect only two arrivals and underestimate the activity. The pileup phenomenon motivates the search for algorithms which allow to separate clusters of electrical pulses for a better identification of radioactive sources and activity estimation[7], and is known in this framework as dead-time correction. However, most methods are not fitted to high counting rates, since they do not rely on any shape information of the time signal. Since we focus in this paper on counting rate estimation (that is, retrieving a vector of arrival times from a vector representing the time signal),

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 188–195, 2012.

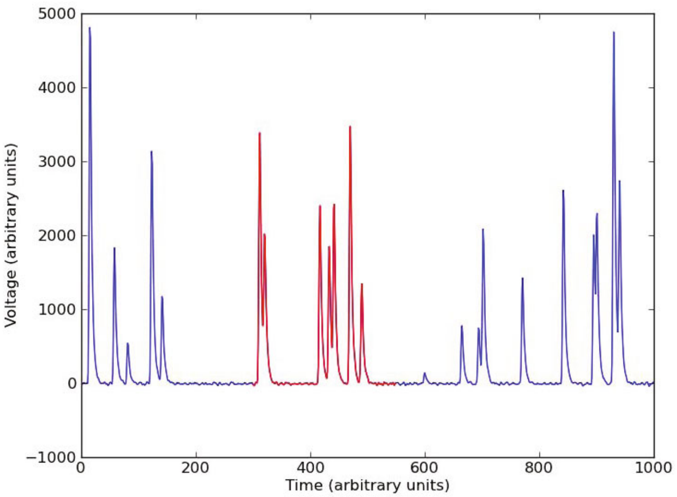


Fig. 1. Example of a spectrometric signal. The red part is made of pileups.

the problem of pileup correction can be formally viewed as a regression problem, which is moreover sparse since the arrival times are usually modelled by a simple Poisson process. Since the seminal papers [2], representation of sparse signals has received a considerable attention, and significant advances have been made both from the theoretical and applied point of view [3]. The paper is organized as follows: we present in section 2 the model and describe a post-processed version of the Group-LASSO in order to estimate the arrival times of individual pulses and counting rate activity. Results are presented in Section 3, showing that the proposed approach outperforms the standard counting rates estimation techniques in the field.

2 Methodology

2.1 Model Description

We consider the following sampled version of the shot-noise model, which is often used in nuclear science to model the recorded spectrometric signal:

$$y_i = \sum_{n \geq 1} E_n \Phi_n(t_i - T_n) + \varepsilon_i, \; i = 1 \dots N, \tag{1}$$

where $\{T_n, \; n \geq 1\}$ are the photon arrival times on $[0, T]$, and form a sample path of a nonhomogeneous Poisson process (NHPP) with intensity $\lambda(t)$; $\{(E_n, \Phi_n), \; n \geq 1\}$ are respectively the energy and the shape pulse created by the $n-$th photon impinging the detector; $\{\varepsilon_i, \; i \geq 1\}$ are independent and identically distributed normal variables with known variance σ^2, which model the additive noise. We wish to estimate $\lambda(t)$ given a sample of signal points $\{y_i,\ i = 1 \dots N\}$.

On most detectors, an electrical pulse created by a single photon has a characteristic shape created by the charge collection and migration in the detector, making relevant to assume that $\{\Phi_n,\ n \geq 1\}$ belong to some parametric family of Gamma functions

$$\Gamma_{\boldsymbol{\theta}}(t) = t^{\theta_1} \cdot e^{-\theta_2\, t}\, 1(t \geq 0)\ ; \tag{2}$$

where the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ belong to a discrete domain $\mathcal{S}_\Gamma$ of cardinal p. Denote by $\mathcal{T} \stackrel{\Delta}{=} \{0 = t_1, t_2, \dots, t_N = T\}$ the subdivision created by the sampling, and by $\mathbf{A}_j$ the following dictionary of shapes translated by $t_j \in \mathcal{T}$: $\mathbf{A}_j \stackrel{\Delta}{=} [\Gamma_{\boldsymbol{\theta}_k}(t_i - t_j)]_{1 \leq i \leq N, 1 \leq k \leq p}$ .Each column of $\mathbf{A}_j$ is a translated and sampled basis signal. Our global dictionary $\mathbf{A}$ is obtained by concatenating the blocks $\mathbf{A}_j$ ordered by increasing times t_j, that is $\mathbf{A} \stackrel{\Delta}{=} [\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{N-1}]$. Observe that $\mathbf{A}_N = 0$, therefore it is not included in the global dictionary. Hence (1) can be rewritten block-wise as

$$\mathbf{y} = \sum_{j=1}^{N-1} \mathbf{A}_j\, \boldsymbol{\beta}_j\, +\, \boldsymbol{\delta}\, +\, \boldsymbol{\varepsilon}\,, \tag{3}$$

where we define the signal $\mathbf{y} \stackrel{\Delta}{=} [\mathbf{y}_1\ \mathbf{y}_2\ \dots\ \mathbf{y}_N]^T$, the regressor to be estimated $\boldsymbol{\beta} \stackrel{\Delta}{=} [\boldsymbol{\beta}_1\ \boldsymbol{\beta}_2\ \dots\ \boldsymbol{\beta}_{N-1}]$, the rounding errors due to discretization $\boldsymbol{\delta} \stackrel{\Delta}{=} [\boldsymbol{\delta}_1\ \boldsymbol{\delta}_2\ \dots\ \boldsymbol{\delta}_N]^T$, and the random noises $\boldsymbol{\varepsilon} \stackrel{\Delta}{=} [\boldsymbol{\varepsilon}_1\ \boldsymbol{\varepsilon}_2\ \dots\ \boldsymbol{\varepsilon}_N]^T$. Note that in the latter definitions all the $\boldsymbol{\beta}_j$'s, $\boldsymbol{\delta}_j$'s and $\boldsymbol{\varepsilon}_j$'s are vectors of size p. Define $J(\boldsymbol{\beta}) \stackrel{\Delta}{=} \{m,\ \boldsymbol{\beta}_m \neq 0\}$. The estimation of the arrival times and the separation of individual pulses can both be related to the estimation of $J(\boldsymbol{\beta})$. Provided the sampling frequency is high this set can be considered as sparse, a condition referred to as "block sparsity" in this paper. However, the actual T_k's in (1) do not belong to $\mathcal{T}$ almost surely, and actual pulses are just approximated by (2). Both facts yield the use of a "block sparse" regression method which reconstructs the signal by a parsimonious use of the blocks $\mathbf{A}_j$.

2.2 Overview on the group LASSO

The linear regression problem raised by (3), which is "block sparse" in the meaning exposed above, is solved by a specific version of the group LASSO [10]:

$$\boldsymbol{\beta}_{GL}(r) = \mathop{\arg\min}_{\boldsymbol{\beta} \in \mathbb{R}^{(N-1)p}}\ \frac{1}{2N} \left\|\mathbf{y} - A\boldsymbol{\beta}\right\|_2^2 + r \sum_{j=1}^{N-1} \left\|\boldsymbol{\beta}_j\right\|_2\ \ ; \tag{4}$$

The mixed ℓ_2/ℓ_1 penalization right term encourages block sparsity, and not sparsity inside blocks. The group LASSO is the good candidate to find block sparse solutions to (3), and it is efficiently computed [1].Furthermore, a numerical condition known as *irrepresentability*, introduced originally for the LASSO [9], guarantees selection consistency; this condition states roughly that in order to recover $J(\beta)$, correlations between active and inactive blocks should be rather small when compared to the correlations between active blocks only. Unfortunately, this condition is not satisfied in our framework, so for our purpose we could not accept all the time blocks selected. This justifies the use of an additional post-processing step in the proposed method for counting rate estimation, which is now described.

2.3 Proposed Algorithm for Pileup Separation and Activity Estimation

In this section we present our estimate $\widehat{\lambda}(t)$ of $\lambda(t)$, based on the group LASSO estimate $\boldsymbol{\beta}_{GL}(r)$ obtained from (4). Since irrepresentability condition does not hold, and since in real world data (2) is only an approximation, it is likely that $J(\boldsymbol{\beta}) \subset J(\boldsymbol{\beta}_{GL}(r))$. However, selected blocks which are not connected to actual T_n's are typically consecutive to the correct time blocks, because time proximity is equivalent to highly correlated signals. Therefore two post-processing steps are performed on the group LASSO regressor $\widehat{\boldsymbol{\beta}}$: first we apply a threshold on each block of $\widehat{\boldsymbol{\beta}}$ to discard some unwanted selected blocks; then we consider consecutive active blocks as a single event. Mathematically, defining $\boldsymbol{\Psi}(\mathbf{x}) \overset{\Delta}{=} \min_{1\leq k\leq p} x_p$ for some threshold η, the estimated arrival times are computed recursively:

$$\widehat{T}_n = \min\{t_j > \widehat{T}_{n-1} \ ; \ \boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_j) > \eta, \ \boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_{j-1}) < \eta\}; \tag{5}$$

The present choice of $\boldsymbol{\Psi}$ is more suitable than other choices (e.g. the average) which were more adapted to the LASSO estimator [8,6], but one has to choose η accordingly. Finally, denote by $\widehat{M}$ the number of estimated times $\widehat{T}_1, \widehat{T}_2, \cdots$, and let W be a nonnegative kernel which integrates to 1 and $h(\widehat{M})$ a standard bandwith parameter. The intensity is estimated by plugging the latter values in a known nonparametric kernel estimate of $\lambda(t)$, e.g. [5]:

$$\widehat{\lambda}(t) = \frac{1}{h(\widehat{M})} \sum_{n=1}^{\widehat{M}} W\left(\frac{t - \widehat{T}_n}{h(\widehat{M})}\right). \tag{6}$$

3 Results and Discussion

We present in this section results on simulations and real data which illustrate the efficiency of the proposed approach.

3.1 Simulations Protocol

The simulations are carried out similarly to [8]: the arrival times $\{T_n, n \geq 0\}$ are governed by a NHPP with intensity $\lambda(t) \overset{\Delta}{=} 0.2e^{-(t-30)^2/75} + \alpha e^{-(t-60)^2/50}$

on $[0, 100]$, where $0.1 < \alpha < 1$ is a variable parameter in order to investigate performances for high to very high counting rates. The sampling rate is set to 1, the energies $\{E_n, n \geq 0\}$ are drawn according to a truncated Gaussian density of mean 20 and variance 9; the noise level is $\sigma = 0.1$. The kernel W is Gaussian, the bandwith $b = 2$. Furthermore, we choose the gamma parameters of $\mathbf{A}$ such that $\theta_1 \in \{0.4, 0.5, \ldots, 1.3\}$ and $\theta_2 = \theta_1\}$. Simulations are performed for two different cases:

(I) the pulse shapes are drawn randomly from the dictionary $\mathbf{A}$;
(II) the pulse shapes are Gamma functions whose parameters are randomly drawn from a uniform distribution on $[0, 2] \times [0, 2]$, which includes $\mathcal{S}_\Gamma$.

Note that case I is the standard regression framework, whereas case II is more realistic, since the artificial data generated do not belong exactly to the Gamma model assumed in (2). The threshold η is chosen in order to exhibit performances similar to [8] in case I, so that we can compare accordingly the performances in case II. In our simulations, The Mean Integrated Squared Error (MISE) $\mathbb{E}(\|\lambda - \hat{\lambda}\|_2^2)$ is approximated by a Monte-Carlo method (10^4 draws) for four different approaches:

1. the post-processed group LASSO suggested in this paper (denoted by *PPGL* in the results);
2. the post-processed LASSO described in [8], denoted by *PPL*;
3. the kernel estimator obtained by plugging $\{[T_n] + 1,\ n = 1 \ldots M\}$ in (6), referred to as *oracle*. The oracle is based on the knowledge of the T_n's and is in practice impossible to compute; nevertheless, since it is the best value attainable from the knowledge of the arrival times and the sampled signal, it will be useful to compare it with PPL and PPGL;
4. the kernel estimator obtained by simple thresholding (denoted by *uncorrected*).

3.2 Results and Discussion

In figure 2 the empirical MISE of $\widehat{\lambda}(t)$ is reported for different values of α, for every four methods in case I; whereas figure 3 displays analog results in case II. Figure 4 represents the MISE for case II.

The previous graphs show clearly in both cases that sparse methods (PPGL and PPL) outperform by far the "uncorrected" method: indeed, their performances are close to the "oracle", which is the best attainable estimator given the sampled signal $\mathbf{y}$. Unsurprisingly figure 3 shows lower performance in the non standard regression problem (case II), but still close to the oracle and better than the uncorrected estimator. Furthermore figures 3, 4 show that PPGL provides a slightly better $\widehat{\lambda}(t)$ than PPL, validating the proposed approach.

Interestingly Figure 3 shows that both PPL and PPGL overestimate the activity in case II for lower intensities, while PPGL proves to be better than PPL: indeed the incompleteness of $\mathbf{A}$ is counterbalanced by the "grouping effect" of PPGL, resulting in less false detections. In high intensity consecutive pulses are

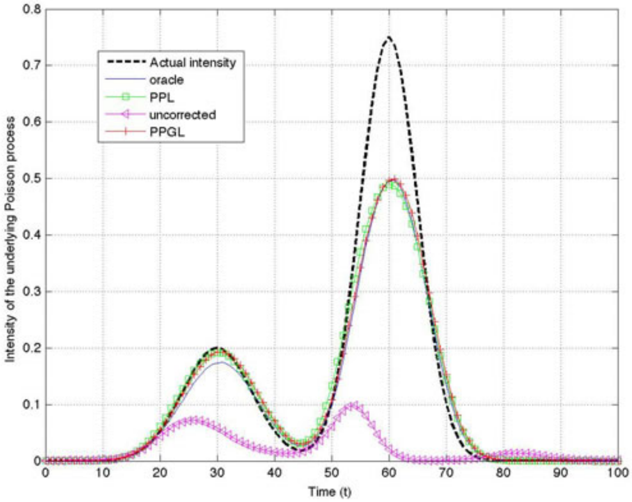


Fig. 2. Nonparametric estimation of $\lambda(t)$ after pulse separation – Case I

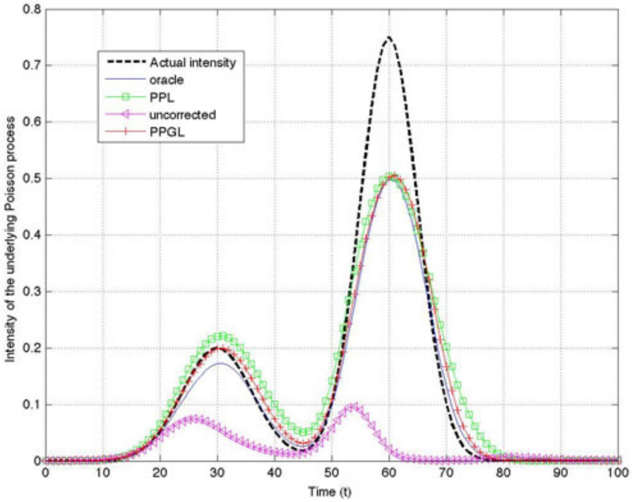


Fig. 3. Nonparametric estimation of $\lambda(t)$ after pulse separation – Case II

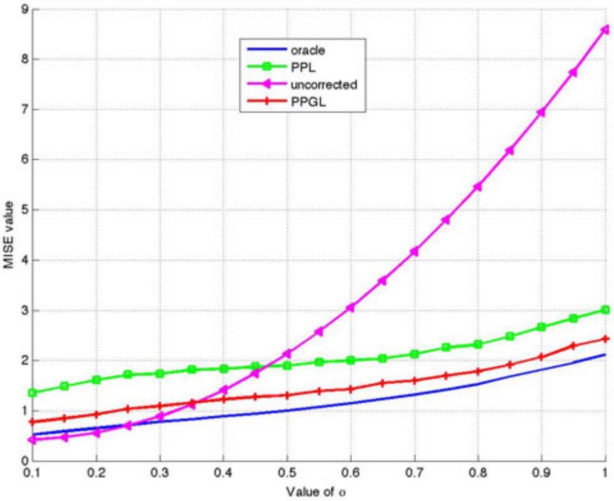


Fig. 4. MISE approximation of the different methods – Case II

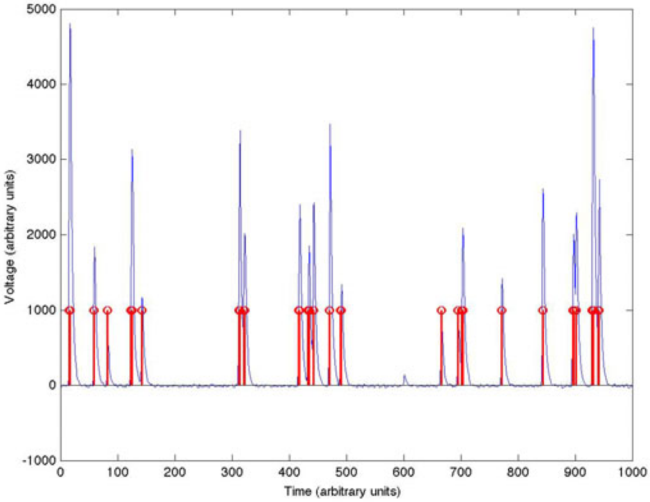


Fig. 5. Decomposition of the signal using PPGL and comparison between the real signal (blue) and active blocks after PPGL (red). The electrical pulses are well separated.

often confused with single ones, and the advantage of grouping is less striking. It should be noted however that PPGL behaves much better than PPL, and separates individual pulses as well as provides a reliable estimate of the activity.

An application on real Gamma spectrometric data is shown in figure 5: when the group LASSO and the post-processing steps are applied on real data, we observe that pulses are well separated in most experiments. However, in our example, an electrical pulse with small amplitude (around 600) goes undetected. Since this is connected to both choices of r and η, this stresses that the sparsity parameter r should be carefully chosen. However, it is noted from figure 5 that the proposed method allows to separate electrical pulses inside pileups, which makes it a valuable approach for high counting rates estimation.

4 Conclusion

We illustrated in this paper the advantages of recent sparse representation developement for the pulse separation and counting rate estimation in the field of nuclear spectrometry. The PPGL method proposed outperforms current state-of-the-art methods in our simulations and real datasets. The theoretical results on this approach is currently under investigation, and should appear in future contributions.

References

1. Beck, A., Teboulle, M.: A fast iterative Shrinkage-Thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2, 183 (2009)
2. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20, 33–61 (1998)
3. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32(2), 407–499 (2004)
4. Knoll, G.F.: Radiation Detection and Measurement, 2nd edn. Wiley (1989)
5. Lewis, P.A.W., Shedler, G.S.: Statistical analysis of non-stationary series of events in a data base system. IBM J. Res. Dev. 20(5), 465–482 (1976)
6. Meinshausen, N., Yu, B.: Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics 37(1), 246–270 (2009)
7. Michotte, C., Nonis, M.: Experimental comparison of different dead-time correction techniques in single-channel counting experiments. Nuclear Instruments and Methods in Physics Research Section A 608(1), 163–168 (2009)
8. Trigano, T., Sepulcre, Y., Roitman, M., Aferiat, U.: On nonhomogeneous activity estimation in gamma spectrometry using sparse signal representation. In: 2011 IEEE Statistical Signal Processing Workshop (SSP), pp. 649–652. IEEE (June 2011)
9. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). IEEE Trans. Inf. Theor. 55(5), 2183–2202 (2009)
10. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, B 68(1), 49–67 (2006)

Some Uniqueness Results in Sparse Convolutive Source Separation

Alexis Benichoux[1], Prasad Sudhakar[2], Fréderic Bimbot[1], and Rémi Gribonval[1]

[1] METISS Team, INRIA Rennes-Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes CEDEX, France
[2] ICTEAM/ELEN, Université catholique de Louvain
B-1348, Louvain-la-Neuve, Belgium
firstname.secondname@{inria.fr,uclouvain.be}

Abstract. The fundamental problems in the traditional frequency domain approaches to convolutive blind source separation are 1) arbitrary permutations and 2) arbitrary scaling in each frequency bin of the estimated filters or sources. These ambiguities are corrected by taking into account some specific properties of the filters or sources, or both. This paper focusses on the filter permutation problem, assuming the absence of the scaling ambiguity, investigating the use of temporal sparsity of the filters as a property to aid permutation correction. Theoretical and experimental results bring out the potential as well as the extent to which sparsity can be used as a hypothesis to formulate a well posed permutation problem.

Keywords: sparse filters, convolutive blind source separation, permutation ambiguity, ℓ^p minimization, Hall's Marriage Theorem, bi-stochastic matrices.

1 Introduction

Let $x_i[t]$, $1 \leq i \leq M$ be M mixtures of N source signals $s_j[t]$, resulting from the convolution with filters $a_{ij}[t]$, each of length L such that:

$$x_i[t] = \sum_{j=1}^{N} (a_{ij} \star s_j)[t], \quad 1 \leq i \leq M. \tag{1}$$

where $\star$ denotes convolution. The filter $a_{ij}[t]$ typically models the impulse response between the j^{th} source and the i^{th} sensor. By abuse of notation, $\mathbf{F}a_{ij} = \{a_{ij}[\omega]\}_{0\leq\omega<L}$ denotes the discrete Fourier transform of the filter seen as a vector $a_{ij} = \{a_{ij}[t]\}_{0\leq t<L} \in \mathbb{C}^L$. Also, the mixing equation (1) can be rewritten as $\mathbf{X} = \mathbf{A} \star \mathbf{S}$, with $\mathbf{A}$ the matrix of filters $\mathbf{A} := (\{a_{ij}[t]\}_{0\leq t<L})_{1\leq i\leq M,\ 1\leq j\leq N}$, $\mathbf{X}$ the observation matrix and $\mathbf{S}$ the source matrix.

In this context, blind filter estimation refers to the problem of obtaining estimates of the filters $\mathbf{A}$ from the mixtures $\mathbf{X}$, without any explicit knowledge about the sources $\mathbf{S}$. Filter estimation is relevant for tasks such as deconvolution, source localisation, etc. It also has a relationship with the problem of Multiple-Input-Multiple-Output system identification in communications engineering.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 196–203, 2012.

2 Permutation and Scaling Ambiguities in Frequency Domain Filter Estimation

A widely used approach for filter estimation relies on the transformation of the mixing model in Eq. (1) into the time-frequency domain, converting a single convolutive filter estimation problem into several complex instantaneous filter estimation problems. Using standard techniques for instantaneous mixing parameter estimation [1], complex mixing filter coefficients $\widetilde{\mathbf{A}}[\omega] = \{\tilde{a}_{ij}[\omega]\}_{1 \leq i \leq M,\ 1 \leq j \leq N}$ are estimated for each frequency bin $0 \leq \omega < L$.

However, without any further assumption on either the filters $a_{ij}[t]$ or the sources $s_j[t]$, one can find filter estimates $\widetilde{\mathbf{A}} = (\widetilde{a}_{ij})$, only up to a global permutation and scaling. That is, for every frequency ω we have

$$\widetilde{a}_{ij}[\omega] = \lambda_j[\omega] a_{i\sigma_\omega(j)}[\omega], \tag{2}$$

where $\lambda_j[\omega]$ and $\sigma_\omega \in \mathfrak{S}_N$ are the unknown scaling and permutation, with $\mathfrak{S}_N$ being the set of permutations of the integers between 1 and N. Several methods [2] attempt to solve these ambiguities by exploiting properties of either $\mathbf{S}$ or $\mathbf{A}$.

2.1 Exploiting Sparsity to Solve the Permutation Ambiguity

In this paper, we hypothesize that the filters $\mathbf{A}$ are *sparse* in the time domain and use this property to solve the permutation ambiguity, in the absence of scaling ($\lambda_j[\omega] = 1$). The assumption that $\mathbf{A}$ is sparse means that each filter a_{ij} has few nonzero coefficients, typically measured by the ℓ^0 pseudo-norm

$$\|a_{ij}\|_0 := \sharp\{0 \leq t < L,\ a_{ij}[t] \neq 0\} = \sum_t |a_{ij}[t]|^0.$$

Besides the ℓ^0 pseudo-norm, the following ℓ^p quasi-norms will be used to quantify the sparsity of $\mathbf{A}$:

$$\|\mathbf{A}\|_p^p := \sum_{ij} \|a_{ij}\|_p^p = \sum_{ijt} |a_{ij}[t]|^p, \quad 0 < p \leq 1.$$

The underlying approach in this work is to seek permutations $\widehat{\sigma}_0, \dots \widehat{\sigma}_{L-1}$ which yield the sparsest estimated time-domain matrix of filters $\widehat{\mathbf{A}} = (\widehat{a}_{ij})$, where $\widehat{a}_{ij}[\omega] := \widetilde{a}_{i\widehat{\sigma}_\omega(j)}[\omega]$.

2.2 Main Result and Structure of the Paper

As the main result, we show (Theorem 2) that when the filter length L is prime, and if $\frac{k}{L} \leq \alpha(N)$, with N the number of sources, then k-sparse filters (i.e., $\|a_{ij}\|_0 \leq k$) uniquely minimize the ℓ^0 norm of $\mathbf{A}$ (up to a global permutation).

In Sec. 3, we investigate the interplay of sparsity and frequency permutations of filters and present our main result. We omit the proof of our main result due to the space constraints, but we describe the main ingredients of the same[1]. In Sec. 4 we propose a combinatorial ℓ^1 minimization algorithm to resolve filter permutations. The effectiveness and limitations of the algorithm is empirically shown, and the observations are related with the theoretical results.

3 Theoretical Guarantees

Given an $M \times N$ filter matrix $\mathbf{A}$, made of filters of length L, and an L-tuple $(\sigma_0, \ldots, \sigma_{L-1}) \in \mathfrak{S}_N$ of permutations, we let $\widetilde{\mathbf{A}}$ be the matrix obtained from $\mathbf{A}$ by applying the permutations in the frequency domain, as in (2), without scaling $(\lambda_j[\omega] = 1)$.

The effect of the permutations is said to coincide with that of a global permutation $\pi \in \mathfrak{S}_N$ of the columns of $\mathbf{A}$ if $\widetilde{a}_{ij} = a_{i\pi(j)}$, $\forall i, j$, or equivalently in the frequency domain:

$$\widetilde{a}_{ij}[\omega] := a_{i\sigma_\omega(j)}[\omega] = a_{i\pi(j)}[\omega],\ 0 \leq \omega < L,\ \forall i, j.$$

This is denoted $\mathbf{A} \equiv \widetilde{\mathbf{A}}$. First, we show that for filters with disjoint time-domain supports, permutations cannot decrease the ℓ^p norm, $0 \leq p \leq 1$:

Theorem 1. *Let $\Gamma_{ij} \subset \{0, \ldots, L-1\}$ be the time-domain support of a_{ij}. Suppose that for all i and $j_1 \neq j_2$ we have*

$$\Gamma_{i,j_1} \cap \Gamma_{i,j_2} = \emptyset. \tag{3}$$

Then, for $0 \leq p \leq 1$, we have $\|\widetilde{\mathbf{A}}\|_p \geq \|\mathbf{A}\|_p$.

Note that filters with disjoint supports need not be very sparse: M filters of length L can have disjoint supports provided that $\max_j \|a_{ij}\|_0 \leq L/M$. Yet, disjointness of supports is a strong assumption, and Theorem 1 only indicates that frequency permutations cannot decrease the ℓ^p norm. Thus, the minimum value of the ℓ^p norm might not be uniquely achieved (up to a global permutation). In our main result, we consider k-sparse filters of prime length, and $p = 0$:

Theorem 2. *Let $\mathbf{A}$ be a $M \times N$ matrix of filters of prime length L. Assume that*

$$\max_{ij} \|a_{ij}\|_0 \leq k. \tag{4}$$

where

$$\frac{k}{L} \leq \alpha(N) := \begin{cases} \frac{2}{N(N+2)} & \text{if } N \text{ is even,} \\ \frac{2}{(N+1)^2} & \text{if } N \text{ is odd.} \end{cases} \tag{5}$$

Then, up to a global permutation, $\mathbf{A}$ uniquely minimises the ℓ^0 pseudo-norm among all possible frequency permutations.

Noticeably, the uniqueness condition does not depend on the number M of mixtures. In order to prove Theorem 2, it is important to quantify the amount of permutations incurred. We use the following definition of the "size" of permutations in the rest of the paper.

[1] An extended version of this paper, containing all proofs, has been submitted for possible publication and is available as INRIA Technical Report No 7782.

3.1 Quantification of Permutations

Given a reference global permutation π, we define the maximum number of frequencies where each estimated filter actually differs from the (globally permuted) original filters, by:

$$\Delta(\widetilde{\mathbf{A}}, \mathbf{A}|\pi) := \max_{i,j} \|\mathbf{F}(\widetilde{a}_{ij} - a_{i\pi(j)})\|_0 \tag{6}$$

The "size" of permutations is then defined as:

$$\Delta(\widetilde{\mathbf{A}}, \mathbf{A}) := \min_{\pi \in \mathfrak{S}_N} \Delta(\widetilde{\mathbf{A}}, \mathbf{A}|\pi). \tag{7}$$

Note that $\Delta(\widetilde{\mathbf{A}}, \mathbf{A}) = 0$ iff $\widetilde{\mathbf{A}} \equiv \mathbf{A}$. We also use the symbol Δ to denote $\Delta(\widetilde{\mathbf{A}}, \mathbf{A})$.

3.2 Exploitation of an Uncertainty Principle

Along with the quantification of permutations, the following lemma, which exploits an uncertainty principle, is an intermediate result to prove Theorem 2.

Lemma 1. *Assume that $\widetilde{\mathbf{A}} \not\equiv \mathbf{A}$, that L is a prime integer, and that* (4) *holds with*

$$2k + \Delta \leq L. \tag{8}$$

Then $\|\widetilde{\mathbf{A}}\|_0 > \|\mathbf{A}\|_0$ and $\|\widetilde{a}_{ij}\|_0 \geq \|a_{ij}\|_0, \forall i, j$. The latter inequality is strict when $\widetilde{a}_{ij} \neq a_{ij}$. For a general L (not necessarily prime), the same conclusions hold when the assumption (8) *is replaced with*

$$2k \cdot \Delta < L. \tag{9}$$

This lemma states that when the original filters $\mathbf{A}$ are sufficiently sparse and if the size Δ of permutations are controlled, in relation to the filter length L, then the resulting permuted filters $\widehat{\mathbf{A}}$ have a larger ℓ^0 norm than $\mathbf{A}$. Moreover, it also states that each individual filter $\widetilde{a}_{ij}$ will have an ℓ^0 norm that is at least as large as that of the corresponding a_{ij}. The skilled reader will rightly sense the role of uncertainty principles [3,4], [5, Theorem 1] in the above lemma.

As opposed to Lemma 1, Theorem 2 does not use an explicit quantification of the permutations, Δ. In fact, this quantity is buried inside the constant $\alpha(N)$. It is actually necessary to make some combinatorial arguments concerning the permutations to arrive at the constant $\alpha(N)$, starting from Δ. The objective of these arguments is to bound Δ from above.

3.3 Combinatorial Arguments

Using Lemma 1 with prime L, a simple combinatorial argument can be used to obtain a weakened version of Theorem 2, with the more conservative constant

$\alpha'(N) := 1/2N!$: by the pigeonhole principle, for any L-tuple of frequency permutations among N sources, at least $L/N!$ permutations are identical; as a result, $\Delta(\widetilde{\mathbf{A}}, \mathbf{A})$ is universally bounded from above by $L - L/N!$; hence if $k \leq L/2N!$ we obtain $2k + \Delta \leq L$ and we can conclude thanks to Lemma 1.

The proof of Theorem 2 with the constant $\alpha(N)$ exploits a stronger universal upper bound $\Delta(\widetilde{\mathbf{A}}, \mathbf{A}) \leq L(1 - 2\alpha(N))$, obtained through an apparently new quantitative application of Hall's Marriage Theorem [6] to bi-stochastic matrices. Bi-stochastic matrices are defined as:

Definition 1 (Bi-stochastic matrix). *An $N \times N$ matrix $\mathbf{B}$ is called* bi-stochastic *if all its entries are non-negative, and the sum of the entries over each row as well as the sum of the entries over each column is one.*

The following lemma connects permutation matrices, which define permutations, and bi-stochastic matrices through Hall's marriage theorem. The subsequent corollary (Corollary 1) provides the bound $\Delta(\widetilde{\mathbf{A}}, \mathbf{A}) \leq L(1 - 2\alpha(N))$, which is crucial to Theorem 2.

Lemma 2. *Let $\mathbf{B}$ be an $N \times N$ bi-stochastic matrix: there exists a permutation matrix $\mathbf{P}$ such that all the entries of $\mathbf{B}$ on the support of $\mathbf{P}$ exceed the threshold*

$$2\alpha(N) = \begin{cases} \frac{4}{N(N+2)} & \text{if } N \text{ is even,} \\ \frac{4}{(N+1)^2} & \text{if } N \text{ is odd.} \end{cases} \tag{10}$$

Corollary 1. *Let $\sigma_0, \ldots, \sigma_{L-1} \in \mathfrak{S}_N$ be L permutations. There exists a global permutation π such that*

$$C_{j\pi(j)} = \sharp\{\ell : \sigma_\ell(j) = \pi(j)\} \geq 2L\alpha(N), \quad \forall 1 \leq j \leq N.$$

The reader may have noticed that Theorem 2, while dropping the *disjoint support* assumption from Theorem 1, introduces new restrictions: the assumption that L is prime, and the restriction to $p = 0$ compared to $0 \leq p \leq 1$ in Theorem 1. How critical are these restrictions? Could they be extended to filters of arbitrary length L and $0 \leq p \leq 1$? This is discussed in the following section.

3.4 Extending Theorem 2 to Non-prime Filter Length *L*?

As indicated by Lemma 3 below, even for $L \geq 4$, there exists sparse matrices of filters that are the sparsest *but not unique (even up to a global permutation)* solution of the considered problem: certain frequency permutations provide an *equally sparse, but not equivalent, solution.*

Lemma 3. *For any integer k such that $2k$ divides L, there exists a matrix of k-sparse filters $\mathbf{A}$ and a set of $L/2k$ frequency permutations resulting in $\widetilde{\mathbf{A}} \not\equiv \mathbf{A}$ such that for all $0 \leq p \leq \infty$: $\|\widetilde{\mathbf{A}}\|_p = \|\mathbf{A}\|_p$, and*

$$\|\widetilde{a}_{ij}\|_p = \|a_{ij}\|_p, \quad \forall i, j. \tag{11}$$

We have $2k \cdot \Delta(\widetilde{\mathbf{A}}, \mathbf{A}) = L$.

The fact that the filter matrices $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ satisfy $2k \cdot \Delta(\widetilde{\mathbf{A}}, \mathbf{A}) = L$ shows the sharpness of Lemma 1 for the case when L is even: the strict inequality in (9) cannot be improved.

Specializing Lemma 3 to $k = 1$ for even $L \geq 4$ yields ideally 1-sparse filters a_{ij} and a set of $L/2$ frequency permutations such that: $\widetilde{a}_{ij}$ are 1-sparse; $\widetilde{\mathbf{A}}$ is not equivalent to $\mathbf{A}$ and cannot be discriminated from it by any ℓ^p norm.

Lemma 3 actually gives a worst case well-posedness bound for filters with arbitrary lengths, and is pessimistic. But, such a bound is achieved in cases when the filters are associated to Dirac combs, which are highly structured. However, existing probabilistic versions of uncertainty principles (see, e.g., the nice survey [7]) lead us to conjecture that if the sparse filters in $\mathbf{A}$ are drawn at random (e.g. from Bernoulli-Gaussian distribution), the uniqueness guarantee of Theorem 2 will hold except with small probability $O(L^{-\beta})$, provided that $k < c(\beta)L/\log L$, for large L.

4 Numerical Experiments

The results achieved so far are theoretical well-posedness guarantee, but do not quite provide algorithms to compute the potentially unique (up to global permutation) solution of the frequency permutation problem. We conclude this paper with the description of a relatively naive optimization algorithm, an empirical assessment of its performance with Monte-Carlo simulations, and a discussion of how this compares with the theoretical uniqueness guarantees achieved above.

4.1 Proposed Combinatorial Algorithm

Given a "permuted" matrix $\widetilde{\mathbf{A}}$, one wishes to find a set of frequency permutations yielding a new matrix $\widehat{\mathbf{A}}$ with minimum ℓ^p norm. The proposed algorithm starts from $\widehat{\mathbf{A}}_0 = \widetilde{\mathbf{A}}$. Given $\widehat{\mathbf{A}}_n$, a candidate matrix $\widehat{\mathbf{A}}_{n+1,\pi}$ can be obtained by applying a permutation π at frequency $\omega_n \equiv n\ [\mathtt{mod}\ L]$. Testing each possible permutation π and retaining the one π_n which minimizes $\|\widehat{\mathbf{A}}_{n+1,\pi}\|_p$ yields the next iterate $\widehat{\mathbf{A}}_{n+1} := \widehat{\mathbf{A}}_{n+1,\pi_n}$. The procedure is repeated until the ℓ^p norm $\widehat{\mathbf{A}}_n$ ceases to change. Since there is a finite number of permutations to try, the stopping criterion is met after sufficiently many iterations.

In theory, it could happen that the stopping criterion is only met after a combinatorially large number of iterations. However, the algorithm stops much sooner in practice. In fact, if we were to use the ℓ^0 norm, the algorithm would typically stop after just one iteration, because the ℓ^0 norm attains its maximum value $M \times N \times L$ for most frequency permutations except a few very special ones. For this reason, we chose to test the algorithm using ℓ^p norms $p > 0$, which are not as "locally constant" as the ℓ^0 norm. To our surprise, the experiments below will show that the best performance is not achieved for small p, but rather for $p = 2 - \epsilon$ with small $\epsilon > 0$. For $p = 0$ and for $p \geq 2$, the algorithm indeed completely fails.

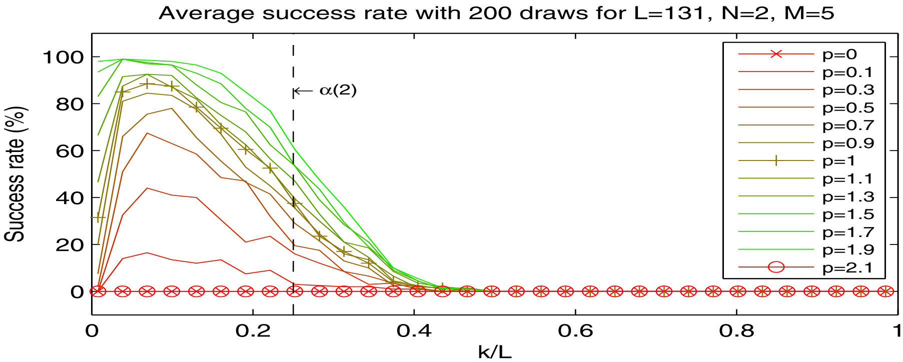


Fig. 1. Filter recovery success as a function of p, $0 \leq p \leq 1.9$

4.2 Monte-Carlo Simulations

For various sparsity levels k and dimensions M, N, random sparse filter matrices $\mathbf{A}$ made of independent random k-sparse filters of length $L = 31$ were generated. Each filter was drawn by choosing: a) a support of size k uniformly at random; b) i.i.d. Gaussian coefficients on this support. For each configuration (k, M, N), 200 random sparse filter matrices $\mathbf{A}$ were drawn. For each $\mathbf{A}$, independent random frequency permutations were applied to obtain $\widetilde{\mathbf{A}}$. The algorithm was then applied to obtain $\widehat{\mathbf{A}}$. The rate of recovery was then computed for each configuration (k, M, N), with an SNR threshold of 100 dB to consider the estimation as a success. We observed that in case of success the SNR was actually more than 300 dB, while in case of failure it was essentially 0 dB.

Figure 1 displays the success rate as a function of the relative sparsity k/L, for various choices of the ℓ^p criterion, with filters of prime length $L = 131$, $N = 2$ sources and $M = 5$ channels. The vertical dashed line indicates the threshold $k/L \leq \alpha(2)$ associated with the well-posedness guarantee (using an ℓ^0 criterion) of Theorem 2. Surprisingly, one can observe that the success rate increases when $0 < p < 2$ is increased. The maximum success rate is achieved when $p = 2 - \epsilon$ with small $\epsilon > 0$.

Beyond the well-posedness regime suggested by the theory (i.e., to the right of the vertical dashed line), the algorithm can succeed, but at a rate that rapidly decreases when the relative sparsity k/L increases. In the regime of well-posed problems, the proposed algorithm is often successful but can still fail to perfectly recover the filters, especially –and surprisingly– for small values of k. This phenomenon is strongly marked for $p < 1$ and essentially disappears for $p > 1$. It remains an open question to determine the respective roles of the ℓ^p criterion and of the naive greedy optimization algorithm in this limited performance for $k/L \ll 1$ when the problem is well-posed with respect to the ℓ^0 norm.

5 Conclusions

It is well known that a sufficient sparsity assumption can be used to make underdetermined linear inverse problems well-posed: without the sparsity assumption, the problem admits an affine set of solutions, which intersects at only one point with the set of sparse vectors. Besides this well-posedness property, a key factor that has lead to the large deployment of sparse models and methods in various fields of science is the fact that a convex relaxation of the NP-hard ℓ^0 minimization problem can be guaranteed to find this unique solution under certain sparsity assumptions. The availability of efficient convex solvers then really makes the problem tractable.

The problem considered in this paper is not a linear inverse problem. Even though it is a simplification of the original permutation *and scaling* problem arising from signal processing, it remains a priori a much harder problem than linear inverse problems in terms of the structure of the solution set: each solution comes with a herd of solutions that are equivalent up to a global permutation.

It is encouraging that we have obtained well-posedness results in this context, but this is at best the beginning of the story: even if the solution is unique, how do we efficiently compute it? Can these results be extended to the original permutation and scaling problem? Why does the proposed naive algorithm perform better for $p > 1$? Answers to these questions can have an impact in fields like blind source separation with sparse multipath channels.

Acknowledgment. The authors acknowledge the support by the European Community's FP7-FET program, SMALL project, (agreement no. 225913), and the support by Agence Nationale de la Recherche (ANR), project ECHANGE (ANR-08-EMER-006).

References

1. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press (2010)
2. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. In: Multichannel Speech Processing Handbook, Citeseer
3. Donoho, D.L., Stark, P.B.: Uncertainty Principles and Signal Recovery. SIAM Journal on Applied Mathematics 49(3), 906–931 (1989)
4. Elad, M., Bruckstein, A.M.: A generalized uncertainty principle and sparse representation in pairs of bases. IEEE Trans. on Information Theory 48(9), 2558–2567 (2002)
5. Tao, T.: An Uncertainty Principle for Cyclic Groups of Prime Order. Mathematical Research Letters 12, 121–127 (2005)
6. Hall, P.: On Representatives of Subsets. J. London Math. Soc. 10(1), 26–30 (1935)
7. Tropp, J.: On the Linear Independence of Spikes and Sines. Journal of Fourier Analysis and Applications 14(5), 838–858 (2008)

Ternary Sparse Coding

Georgios Exarchakis[1], Marc Henniges[1], Julian Eggert[2], and Jörg Lücke[1]

[1] FIAS, Goethe-Universität Frankfurt am Main, Germany
[2] Honda Research Institute Europe, Offenbach am Main, Germany

Abstract. We study a novel sparse coding model with discrete and symmetric prior distribution. Instead of using continuous latent variables distributed according to heavy tail distributions, the latent variables of our approach are discrete. In contrast to approaches using binary latents, we use latents with three states (-1, 0, and 1) following a symmetric and zero-mean distribution. While using discrete latents, the model thus maintains important properties of standard sparse coding models and of its recent variants. To efficiently train the parameters of our probabilistic generative model, we apply a truncated variational EM approach (Expectation Truncation). The resulting learning algorithm infers all model parameters including the variance of data noise and data sparsity. In numerical experiments on artificial data, we show that the algorithm efficiently recovers the generating parameters, and we find that the applied variational approach helps in avoiding local optima. Using experiments on natural image patches, we demonstrate large-scale applicability of the approach and study the obtained Gabor-like basis functions.

1 Introduction

Since it has first been introduced, Sparse Coding (SC) [1] has become a standard model to explain the behavior of simple cells in the primary visual cortex. However, the derivation of sparse coding algorithms typically involves analytical intractabilities, e.g., because involved posterior distributions and their expectation values have no closed-form solutions. The standard approach to overcome this difficulty is the use of MAP estimates for inference and learning (e.g.,[1,2] but also compare [3] or reviews on variational methods). Using binary hidden variables, this analytical intractability can be avoided altogether because learning rules using Expectation Maximization (EM) are closed-form. However, in contrast, e.g., to a Laplace prior, the Bernoulli distribution used for binary variables is not symmetric and not zero-mean (see Fig. 1). These differences to conventional priors can have implications for the used preprocessing and the inferred basis functions, which may add to effects caused by the introduction of discrete hidden variables. To study the implications of discrete hidden variables independently of differences in prior symmetries, we investigate in this work a generative model with a symmetric and discrete prior distribution. Such a prior can be well suited for different types of data, and it directly relates to recent Sparse Coding versions with hard-sparseness constraints (compare, e.g., [4,5]).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 204–212, 2012.

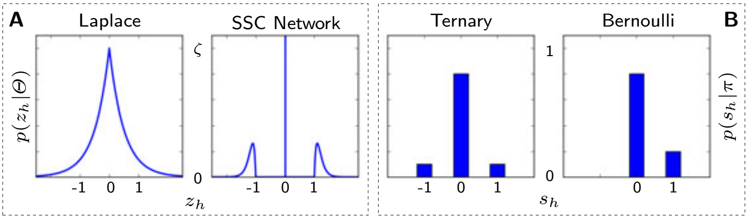


Fig. 1. Comparison of different prior distributions. **A** shows distributions for a continuous variable z_h (Laplace as used in [1] and the SSC network prior [4]) with model parameters Θ. In **B** the symmetrical Ternary Sparse Coding prior of this paper and the unsymmetrical Bernoulli prior (e.g., [13,6]) are displayed for the discrete variable s_h with model parameter π.

2 Sparse Coding with Ternary Hidden Variables

Let $\{\boldsymbol{y}^{(n)}\}_{n=1,\ldots,N}$ be a set of N independent data points, $\boldsymbol{y}^{(n)} \in \mathbb{R}^D$, i.e., the number of observed variables is D. We want to find the parameters $\Theta = (W, \sigma, \pi)$ that maximize the data likelihood $\mathcal{L} = \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)} \,|\, \Theta)$, under the generative model:

$$p(\boldsymbol{s}|\pi) \;=\; \textstyle\prod_{h=1}^{H} \left(\frac{\pi}{2}\right)^{|s_h|} \left(1-\pi\right)^{1-|s_h|}, \qquad p(\boldsymbol{y}\,|\,\boldsymbol{s}\,,\,W,\sigma) \;=\; \mathcal{N}(\boldsymbol{y};\, W\boldsymbol{s}\,,\sigma^2\mathbb{1})\,, \quad (1)$$

where $W \in \mathbb{R}^{D\times H}$ denotes the H basis vectors for the data points, and $s_h \in \{-1, 0, 1\}$, with $h = 1, \ldots, H$, is the latent variable that specifies how W_h contributes to a data point. That is, we can assume that each observed variable can take the form $y_d = \sum_{h=1}^{H} W_{dh} s_h$ plus a Gaussian noise with variance σ^2. The π parameter in the prior is the probability of a cause being active. Since our prior is symmetrical, the probability of a cause being added to generate a data point is equal to the probability of a cause being subtracted: $p(s_h = -1) = p(s_h = 1) = \frac{\pi}{2}$. The difference between previous SC generative models and this model is the choice of a symmetrical and discrete prior for latents taking values −1, 0, and 1. As a consequence, Ternary Sparse Coding (TSC) can explain data generated by Binary Sparse Coding (BSC,[13,6]) while the opposite does not apply (see Fig. 2). To optimize the parameters Θ, we apply with Expectation Truncation (ET) a truncated variational EM approach [7]. Instead of optimizing the log-likelihood directly, variational EM optimizes the free-energy, a lower-bound which depends on the parameters and an approximate posterior distribution q:

$$\mathcal{F}(q,\Theta)=\sum_{n=1}^{N}\left[\sum_{\boldsymbol{s}} q^{(n)}(\boldsymbol{s}\;;\Theta^{\mathrm{old}})\;\Big[\log\big(p(\boldsymbol{y}^{(n)}\,|\,\boldsymbol{s}\,,W,\sigma)\big)+\log\big(p(\boldsymbol{s}\;|\,\pi)\big)\Big]\right]+H(q)\,. \quad (2)$$

Instead of conventional variational EM, ET does not assume a factored form of $q^{(n)}$ but uses truncated sums to set $q^{(n)}$ proportional to the exact posterior on

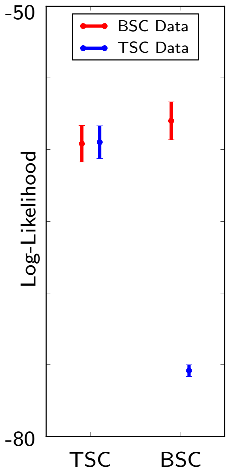


Fig. 2. Comparison of model performance on different data types. We created 100 sets of $N = 1000$ data points with the BSC model and the TSC model, then ran both algorithms on all data sets and compared the likelihood values. The figure shows the mean Log-likelihood values with standard deviation for the TSC model and the BSC model both on TSC data (blue, right) and BSC data (red, left). Evidently, TSC explains both data types equally well, whereas BSC describes data points created with TSC worse than its own data points. While the latter is an expected result for a model mismatch, the equal performance of TSC on both types of data points out the more general character of TSC. More particularly, BSC can be understood as a special case of TSC.

subsets $\mathcal{K}_n$ (see [7] for details). Consequently, the expectation values w.r.t. $q^{(n)}$ amount to:

$$\langle g(\boldsymbol{s})\rangle_{q^{(n)}} = \frac{\sum_{\boldsymbol{s}} p(\boldsymbol{s}, \boldsymbol{y}^{(n)} \,|\, \Theta^{\mathrm{old}})\; g(\boldsymbol{s})}{\sum_{\tilde{\boldsymbol{s}}} p(\tilde{\boldsymbol{s}}, \boldsymbol{y}^{(n)} \,|\, \Theta^{\mathrm{old}})} \approx \frac{\sum_{\boldsymbol{s}\in\mathcal{K}_n} p(\boldsymbol{s}, \boldsymbol{y}^{(n)} \,|\, \Theta^{\mathrm{old}})\; g(\boldsymbol{s})}{\sum_{\tilde{\boldsymbol{s}}\in\mathcal{K}_n} p(\tilde{\boldsymbol{s}}, \boldsymbol{y}^{(n)} \,|\, \Theta^{\mathrm{old}})}, \quad (3)$$

where $g(\boldsymbol{s})$ is a function of $\boldsymbol{s}$ (and potentially the parameters). Eqn. 3 represents a good approximation if the set $\mathcal{K}_n$ contains most of the posterior probability mass w.r.t. a given data point $\boldsymbol{y}^{(n)}$. For our model, $\mathcal{K}_n$ in (3) is chosen to contain hidden states $\boldsymbol{s}$ with at most γ active causes, i.e., $\|\boldsymbol{s}\|_1 \leq \gamma$. Furthermore, we only consider the combinatorics of $H' \geq \gamma$ hidden variables that are likely to have contributed to generating a given data point $\boldsymbol{y}^{(n)}$. More formally we define:

$$\mathcal{K}_n = \{\boldsymbol{s} \mid \|\boldsymbol{s}\|_1 \leq \gamma \text{ and } \forall i \notin I : s_i = 0\}, \quad (4)$$

where the index set I contains those latent indices h with the H' largest values of a *selection function* $\mathcal{S}_h(\boldsymbol{y}^{(n)})$. In our case this function is given by:

$$\mathcal{S}_h(\boldsymbol{y}) = \max\left\{p(s_h = -1, \boldsymbol{y} \,|\, \Theta^{\mathrm{old}}), p(s_h = 1, \boldsymbol{y} \,|\, \Theta^{\mathrm{old}})\right\}\ \forall \boldsymbol{s} : \|\boldsymbol{s}\|_1 = 1. \quad (5)$$

A large value of $\mathcal{S}_h(\boldsymbol{y})$ signals a high likelihood that $\boldsymbol{y}$ contains the basis function $\boldsymbol{W}_h$ as a component. In numerical experiments on ground-truth data we can verify that for most data points the approach (3) with (4) and (5) indeed approximates the true expectation values with high accuracy. By applying the ET approximation, exact EM (which scales exponentially with H) is altered to an algorithm which scales polynomially with H' (approximately $\mathcal{O}\left(H'^{\gamma}\right)$) and linearly with H. Note, however, that in general larger values of H also require larger amounts of data points.

With the tractable approximations for the expectation values $\langle g(\boldsymbol{s})\rangle_{q^{(n)}}$ computed with (3) to (5) the update equations for W and σ are given by:

$$W^{\mathrm{new}} = \left(\sum\nolimits_{n\in\mathcal{M}} \boldsymbol{y}^{(n)} \langle \boldsymbol{s} \rangle_{q_n}^{T}\right) \left(\sum\nolimits_{n'\in\mathcal{M}} \left\langle \boldsymbol{s}\,\boldsymbol{s}^{T}\right\rangle_{q_{n'}}\right)^{-1} \tag{6}$$

$$\sigma^{\mathrm{new}} = \sqrt{\tfrac{1}{|\mathcal{M}|\,D} \sum\nolimits_{n\in\mathcal{M}} \left\langle \left|\left|\boldsymbol{y}^{(n)} - W\,\boldsymbol{s}\right|\right|^2 \right\rangle_{q_n}} \tag{7}$$

Note that we do not sum over all data points $\boldsymbol{y}^{(n)}$ but only over those in a subset $\mathcal{M}$ (note that $|\mathcal{M}|$ is the number of elements in $\mathcal{M}$). The subset contains those data points for which (3) finally represents a good approximation. It is defined to contain the N^{cut} data points with the highest values for $\sum_{\tilde{\boldsymbol{s}}\in\mathcal{K}_n} p(\tilde{\boldsymbol{s}}, \boldsymbol{y}^{(n)}\,|\,\Theta^{\mathrm{old}})$, i.e., with the highest values for the denominator in (3). N^{cut} is hereby the expected number of data points that have been generated by states with less or equal γ non-zero entries: $N^{\mathrm{cut}} = N \sum_{\boldsymbol{s},\,\|\boldsymbol{s}\|_1\leq\gamma} p(\boldsymbol{s}\,|\,\pi) = N \sum_{\gamma_1=0}^{\gamma} \sum_{\gamma_2=0}^{\gamma-\gamma_1} \binom{H}{\gamma_1\gamma_2(H-\gamma_1-\gamma_2)} \left(\frac{\pi}{2}\right)^{\gamma_1+\gamma_2} (1-\pi)^{H-\gamma_1-\gamma_2}$.

Update equations (6) and (7) were obtained by setting the derivatives of Eqn. 2 (w.r.t. W and σ) to zero. Similarly, we can derive the update equation for π. However, as the approximation only considers states $\boldsymbol{s}$ with a maximum of γ non-zero entries, the update has to correct for an underestimation of π. If such a correction is taken into account (compare [6,14]), we obtain the update rule:

$$\pi^{\mathrm{new}} = \frac{A(\pi)\,\pi}{B(\pi)}\,\frac{1}{|\mathcal{M}|} \sum_{n\in\mathcal{M}} \langle \|\boldsymbol{s}\|_1 \rangle_{q_n} \quad \text{with} \tag{8}$$

$$A(\pi) = \sum_{\gamma_1=0}^{\gamma} \sum_{\gamma_2=0}^{\gamma-\gamma_1} \binom{H}{\gamma_1\gamma_2\,(H-\gamma_1-\gamma_2)} \left(\frac{\pi}{2}\right)^{\gamma_1+\gamma_2} (1-\pi)^{H-\gamma_1-\gamma_2} \quad \text{and}$$

$$B(\pi) = \sum_{\gamma_1=0}^{\gamma} \sum_{\gamma_2=0}^{\gamma-\gamma_1} (\gamma_1+\gamma_2) \binom{H}{\gamma_1\gamma_2\,(H-\gamma_1-\gamma_2)} \left(\frac{\pi}{2}\right)^{\gamma_1+\gamma_2} (1-\pi)^{H-\gamma_1-\gamma_2}$$

Note that if we allow all possible states (i.e., $\gamma = H$), the correction factor $\frac{A(\pi)\,\pi}{B(\pi)}$ in (8) is equal to $\frac{1}{H}$ and the set $\mathcal{M}$ becomes equal to the set of all data points (because $N^{\mathrm{cut}} = N$). Eqn. 8 then falls back to the exact EM update rule for π. The same applies for Eqns. 6 and 7 for $\gamma = H$. By choosing a value for γ between one and H we can thus choose the accuracy of the used approximation. The higher the value of γ the more accurate is the approximation but the larger are also the computational costs. For intermediate values of γ we can obtain very good approximations with small computational costs.

3 Numerical Experiments

Linear Bars Test. In order to evaluate its performance, we applied the algorithm to artificial data points as shown in Fig. 3A. These data were generated

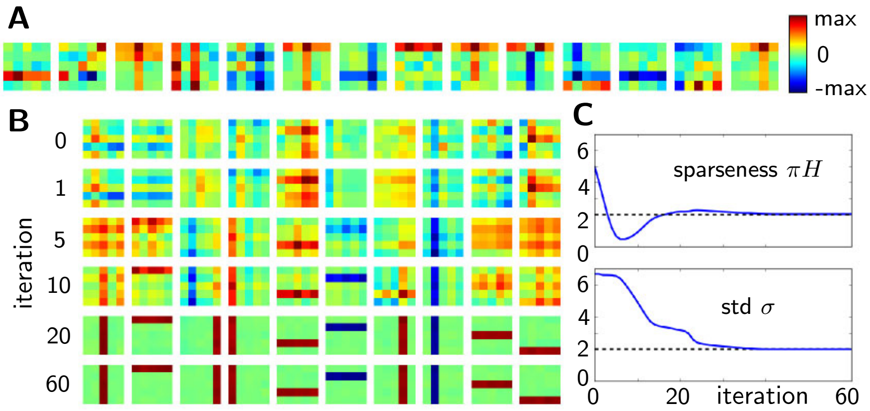


Fig. 3. Linear bars test with $H = 10$, $D = 5 \times 5$, and $N = 1000$. **A** 14 example data points. **B** Basis functions for iterations given on the left. **C** Sparseness and standard deviation plotted over the iterations. Ground-truth indicated by dashed horizontal line.

by $H = 10$ basis functions $\boldsymbol{W}_h$ as described by the generative model in (1). The prior parameter was set to $\pi = 0.2$ which results in $\pi H = 2$ basis functions contributing to one data point on average. Each basis function represents one $D = 25$ pixel image forming a vertical or a horizontal bar on a 5×5 grid. In particular, $W_h^d \in \{0, 10\}$ where the value 10 denotes a bar pixel and 0 represents a background pixel. To these data we added iid Gaussian noise (mean = 0, std = 2). ET approximation parameters described in (4) were set to $H' = 5$ and $\gamma = 3$. An annealing temperature (see [8,6]) was kept constant at $T = 2$ for the first 10 iterations, and then decreased linearly to $T = 1$ at iteration 40, where it was kept until termination of the algorithm at iteration 60. For the first 20 iterations we set the amount of used data points to the number of all data points, i.e. $|\mathcal{M}| = N$, then linearly decreased it to N^{cut} until iteration 40, where we kept it until the end (compare [7]). We set the initial values for each basis function to be the average over the data points plus a Gaussian white noise with standard deviation 0.05. Sparseness was initialized at $\pi H = 5$, thus assuming that five of the causes contributed to an image on average. The standard deviation was initialized as the average variance of all pixels which led to a value of $\sigma \approx 6$. After each iteration we added Gaussian noise to the learned basis functions (mean = 0, std = 0.05), which we decreased linearly to zero from iteration 20 to 40. We ran the algorithm with the above parameters 1000 times, each time using a newly generated set of N = 1000 data points. In 913 of these trials we recovered all bars (≈ 91.3% reliability) and obtained a mean value of $\pi H = 2$ (0.05 std) for the sparseness and $\sigma = 2$ (0.1 std) for the data noise (i.e., the algorithm successfully recovered the generating parameters). Figs. 3B and 3C show the typical development of the model parameters over the 60 iterations.

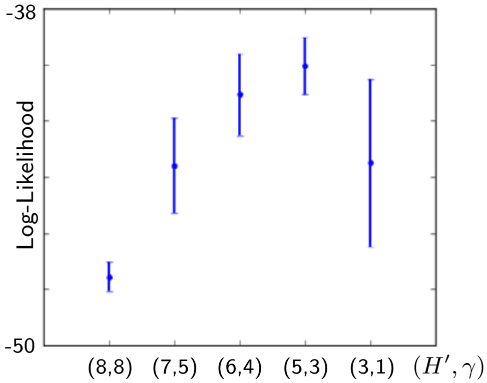


Fig. 4. Final log-likelihood values with standard deviation for TSC runs with different sets of ET parameters. The algorithms were applied to linear bars tests with $H = 8$, $\pi = \frac{2}{H}$, $D = 4 \times 4$. For each approximation parameter pair (H', γ), we ran the algorithm 100 times on $N = 1000$ newly generated data point for each run. Model parameters were initialized as described in Sec. 2.

Effect of ET on Local Optima. To examine the effect of different approximation parameters on the convergence of the algorithm, we applied the TSC algorithms with different sets of values for H' and γ to a linear bars test (Fig. 4). As can be observed in Fig. 4, the highest likelihood values were neither obtained for very high values of H' and γ nor for very low ones. For low values the approximation gets too coarse and the number of considered data points, N^{cut}, gets too small. For high values the approximation accuracy is high but the algorithm has a strong tendency to converge to local optima. The local optima usually correspond to solutions with, on average, more basis functions per data point than in the generation process. The highest likelihood values are obtained for intermediate values of H' and γ. For such values the approximation quality is high and, at the same time, local optima are avoided much more frequently. The latter effect can be explained by ET keeping a subset of all possible fixed-points of learning stable. In our case, fixed-points of too dense solutions are avoided.

Natural Image Patches. As an example for large-scale applicability, we applied the TSC algorithm to $N = 200\,000$ preprocessed[1] patches of natural images, of size $D = 26 \times 26$, taken from the van Hateren image database [9]. The algorithm assumed $H = 700$ basis functions and an initial value for sparseness $\pi H = 1$. The parameters W and σ were initialized as in the linear bars test. The approximation parameters were set to $\gamma = 8$ and $H' = 10$. For the first 20 iterations, the algorithm used all data points for learning, $|\mathcal{M}| = N$, then linearly decreased the number of data points to $|\mathcal{M}| = N^{\mathrm{cut}}$ for the next 20 iterations, and kept it at this value until termination of the algorithm at iteration 200. After each iteration, the basis functions were slightly perturbed using additive Gaussian noise which was linearly decreased from iteration 30 to 48.

Fig. 5A shows a random selection of 200 of the 700 obtained basis functions. In Fig. 5B the most globular of the 700 basis functions are displayed. The monitored time-course of the data sparseness $(\pi\, H)$ and the time-course of the data noise

[1] We used a Difference of Gaussians filter where filter parameters were chosen as in [10]; before the brightest 2% of the pixels were clamped to the maximal value of the remaining 98% (influence of light-reflections were reduced in this way).

(σ) are displayed in Fig. 5C. As can be observed, we obtain Gabor-like basis functions with different orientations, spatial frequencies, and phase as well as globular basis functions with no or very little orientation preferences (compare [11]). Along with the basis functions we obtain an estimate for the noise ($\sigma = 1.8$) and, more importantly, for the data sparseness of $\pi H = 6.9$ active causes per 26×26 patch.

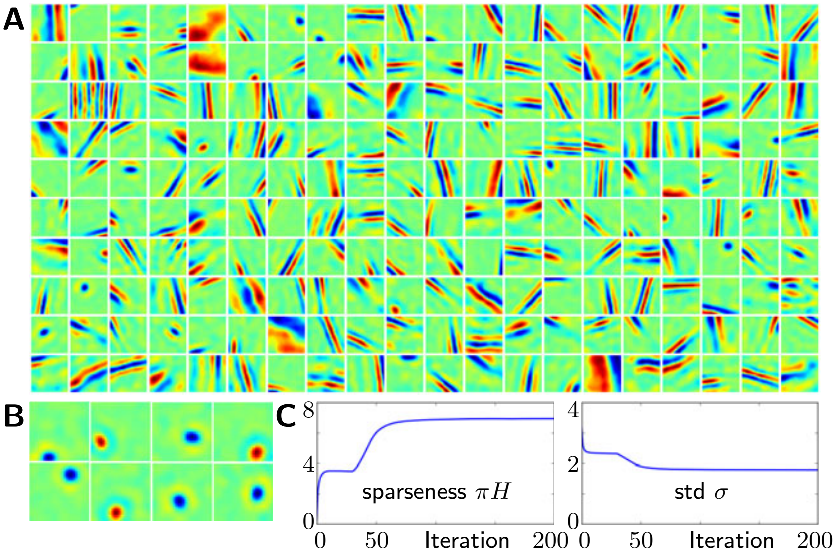


Fig. 5. Numerical experiment on image patches. **A** 200 basis functions randomly selected out of the total $H = 700$. **B** Example of eight of the most globular fields. **C** Time-courses of sparseness (πH) and data noise (standard deviation) σ.

4 Discussion

We have studied a novel sparse coding algorithm using discrete hidden variables. While in most of the former studies on sparse coding the prior distributions are continuous, we, in this work, introduced a symmetric prior defined on the set $\{-1, 0, 1\}$. Such a prior maintains important properties of continuous sparse priors which are, for instance, not preserved for binary hidden variables. As binary variables are a feature of another large class of probabilistic approaches such as deep-belief-networks [12], this work can be regarded as connecting two very successful and active lines of research.

Efficient parameter optimization for our model is made possible by applying Expectation Truncation (ET; [7]), a recent variational EM approach that allowed us to infer all model parameters including data sparsity. In numerical experiments on artificial data, we verified that the resulting algorithm accurately and efficiently recovers the parameters of the generating distribution. Additionally, we found that ET helps in avoiding local optima during learning. While high approximation parameters (H', γ) lead to a very high accuracy of the E-step approximation, learning often converged to local optima corresponding to a solution of relatively low sparsity. Such local solutions were efficiently avoided, however, if the used approximation parameters were lower but still large enough to result in a high approximation quality (Fig. 4). Large-scale applicability of our algorithm was demonstrated using numerical experiments on pre-processed image patches. The obtained basis functions in these experiments are instructive from a neuroscientific perspective. As in previous studies using standard sparse coding ([1] and many others) as well as in studies using binary latents [6,13], we obtained localized Gabor-like basis functions with different orientations and spatial frequencies. Additionally, we obtained globular basis functions as in binary [6] and semi-discrete models [4,5]. Such globular fields are interesting as they have been recorded in neurophysiological experiments but were not obtained with standard SC or ICA before [11]. That globular basis functions also emerge in our study is further evidence for the discreteness of hidden variables facilitating the emergence of globular fields (but also see [14]). As the use of ternary latents is very close to the priors used in [4,5], our approach could be regarded as capturing the essential functional features of these earlier works, and as taking their capabilities a step further by inferring data noise and sparsity.

Acknowledgment. The project was funded by the German Research Foundation (DFG), grant LU 1196/4-1 (GE, JL) and by the Honda Research Institute Europe (MH, JE). Furthermore, we gratefully acknowledge support by the Frankfurt Center for Scientific Computing.

References

1. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
2. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. JMLR 11, 19–60 (2010)
3. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54(11), 4311–4322 (2006)
4. Rehn, M., Sommer, F.T.: A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. J. of Comp. Neurosci. 22(2), 135–146 (2007)
5. Rozell, C.J., Johnson, D.H., Baraniuk, R.G., Olshausen, B.A.: Sparse coding via thresholding and local competition in neural circuits. Neural Computation 20(10), 2526–2563 (2008)
6. Henniges, M., Puertas, G., Bornschein, J., Eggert, J., Lücke, J.: Binary Sparse Coding. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 450–457. Springer, Heidelberg (2010)
7. Lücke, J., Eggert, J.: Expectation truncation and the benefits of preselection in training generative models. JMLR 11, 2855–2900 (2010)
8. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. Neural Networks 11(2), 271–282 (1998)

9. Hateren, J.H., Schaaf, A.: Independent component filters of natural images compared with simple cells in primary visual cortex. Proceedings of the Royal Society of London B 265, 359–366 (1998)
10. Lücke, J.: Receptive field self-organization in a model of the fine-structure in V1 cortical columns. Neural Computation 21(10), 2805–2845 (2009)
11. Ringach, D.L.: Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. Journal of Neurophysiology 88, 455–463 (2002)
12. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554 (2006)
13. Haft, M., Hofman, R., Tresp, V.: Generative binary codes. Pattern Anal. Appl. 6, 269–284 (2004)
14. Puertas, J.G., Bornschein, J., Lücke, J.: The maximal causes of natural scenes are edge filters. NIPS 23, 1939–1947 (2010)

Closed-Form EM for Sparse Coding and Its Application to Source Separation

Jörg Lücke* and Abdul-Saboor Sheikh*

FIAS, Goethe-University Frankfurt, 60438 Frankfurt, Germany
{luecke,sheikh}@fias.ni-frankfurt.de

Abstract. We define and discuss a novel sparse coding algorithm based on closed-form EM updates and continuous latent variables. The underlying generative model consists of a standard 'spike-and-slab' prior and a Gaussian noise model. Closed-form solutions for E- and M-step equations are derived by generalizing probabilistic PCA. The resulting EM algorithm can take all modes of a potentially multimodal posterior into account. The computational cost of the algorithm scales exponentially with the number of hidden dimensions. However, with current computational resources, it is still possible to efficiently learn model parameters for medium-scale problems. Thus, the algorithm can be applied to the typical range of source separation tasks. In numerical experiments on artificial data we verify likelihood maximization and show that the derived algorithm recovers the sparse directions of standard sparse coding distributions. On source separation benchmarks comprised of realistic data we show that the algorithm is competitive with other recent methods.

1 Introduction

Probabilistic generative models are a standard approach to model data distributions and to infer instructive information about the data generating process. Methods like principle component analysis, factor analysis, or sparse coding (SC) [e.g., 15] have all been formulated in the form of probabilistic generative models. Moreover, independent component analysis (ICA), which is a very popular approach to blind source separation, can also be recovered from sparse coding in the limit of zero observation noise [e.g., 4]. A standard procedure to optimize parameters in generative models is the application of Expectation Maximization (EM) [e.g., 14]. However, for many generative models the optimization using EM is analytically intractable. For stationary data only elementary models such as mixture models and factor analysis (which contains probabilistic PCA as special case) have closed-form solutions for E- and M-step equations. EM for more elaborate models requires approximations. In particular, sparse coding models [15, 9, 17, and many more] require approximations because integrals over the latent variables do not have closed-form solutions.

Here, we study a generative model that combines the Gaussian prior of probabilistic PCA (p-PCA) with a binary prior distribution. Distributions combining binary and continuous parts have been discussed and used as priors before [e.g., 12, among many others] and are commonly referred to as 'spike-and-slab' distributions. Also sparse

* Joint first authorship.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 213–221, 2012.

coding variants with spike-and-slab distributions have been studied previously [compare 20, 7, 19, 16, 8, 13]. However, in this work we show that combining binary and Gaussian latents maintains the p-PCA property of having a closed-form solution for EM optimization. We can, therefore, derive an algorithm that uses exact posteriors with potentially many modes to update model parameters.

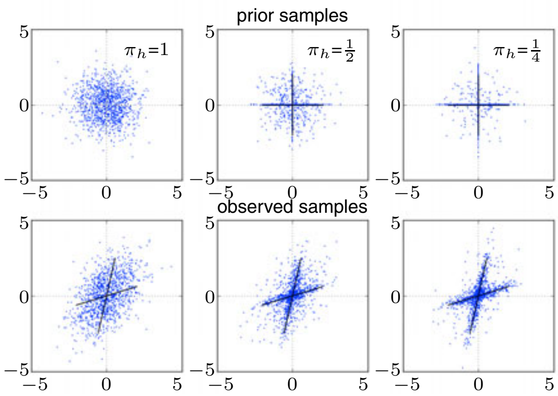


Fig. 1. Distributions generated by the GSC generative model. The left column shows the distributions generated for $\pi_h = 1$ for all h. In this case the model generates p-PCA distributions. The middle column shows an intermediate value of π_h. The generated distributions are not Gaussians anymore but have a slight star shape. The right column shows distributions for small values of π_h. The generated distributions have a salient star shape similar to standard sparse coding distributions.

2 Closed-Form EM for a Spike-and-Slab Sparse Coding Model

Let us first consider a pair of H–dimensional i.i.d. latent vectors, a continuous $\boldsymbol{z}\in\mathbb{R}^H$ and a binary $\boldsymbol{s}\in\{0,1\}^H$ with:

$$p(\boldsymbol{s}\,|\Theta) = \prod_{h=1}^{H} \pi_h^{s_h}\,(1-\pi_h)^{1-s_h} = \text{Bernoulli}(\boldsymbol{s};\boldsymbol{\pi}) \text{ and } p(\boldsymbol{z}\,|\Theta) = \mathcal{N}(\boldsymbol{z};\,\boldsymbol{0},\mathbb{1}_H), \quad (1)$$

where π_h parameterizes the probability of non-zero entries. After generation, the latent vectors are combined using a pointwise multiplication operator: i.e., $(\boldsymbol{s}\odot\boldsymbol{z})_h = s_h\,z_h$ for all h. The resulting hidden random variable is a vector of continuous values and zeroes, and it follows a 'spike-and-slab' distribution. Given a hidden vector (which we will denote by $\boldsymbol{s}\odot\boldsymbol{z}$) , we generate a D–dimensional observation $\boldsymbol{y}\in\mathbb{R}^D$ by linearly combining a set of basis functions W and adding Gaussian noise:

$$p(\boldsymbol{y}\,|\,\boldsymbol{s},\boldsymbol{z},\Theta) = \mathcal{N}(\boldsymbol{y};\,W(\boldsymbol{s}\odot\boldsymbol{z}),\Sigma), \quad (2)$$

where $W\in\mathbb{R}^{D\times H}$ is the matrix containing the basis functions $W_{\cdot h}$ as columns, and $\Sigma\in\mathbb{R}^{D\times D}$ is a covariance matrix parameterizing the data noise. The latents' priors (1), their pointwise combination and the noise distribution (2) define the generative model under consideration. As a special case, the model contains probabilistic PCA (or factor analysis). This can easily be seen by setting all π_h equal to one. The model (1) to (2) is capable of generating a broad range of distributions including sparse coding like distributions. This is illustrated in Fig. 1 where the parameters π_h allow for continuously changing PCA-like to a SC-like distribution.

While the generative model itself has been studied previously [20, 19, 16, 8], we will show that a closed-form EM algorithm can be derived, which can be applied to blind source separation tasks (also see our preliminary work [10]). We will refer to the generative model (1) to (2) as the *Gaussian Sparse Coding* (GSC) model in order to stress that a specific spike-and-slab prior (Gaussian slab) in conjunction with a Gaussian noise model is used. The GSC model is thus an instance of the spike-and-slab sparse coding model (or alternatively known *sparse factor analysis* models [see e.g., 20, 19, 16, 8]).

Expectation Maximization (EM) for Parameter Optimization. Given a set of N independent data points $\{\boldsymbol{y}^{(n)}\}_{n=1,\ldots,N}$, we seek to infer the parameters $\Theta = (W, \Sigma, \boldsymbol{\pi})$ that maximize the data likelihood $\mathcal{L} = \prod_{n=1}^{N} p(\boldsymbol{y}^{(n)} \,|\, \Theta)$ under the GSC generative model. We employ Expectation Maximization (EM) algorithm for parameter optimization. The EM algorithm [see e.g., 14] optimizes the data likelihood w.r.t. the parameters Θ by iteratively maximizing the free-energy given by:

$$\mathcal{F}(\Theta^{\mathrm{old}}, \Theta) = \sum_{n=1}^{N} \sum_{\boldsymbol{s}} \int_{\boldsymbol{z}} p(\boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{y}^{(n)}, \Theta^{\mathrm{old}}) \Big[\log \big(p(\boldsymbol{y}^{(n)} \,|\, \boldsymbol{s}, \boldsymbol{z}, \Theta) \big) + \log \big(p(\boldsymbol{s} \,|\, \Theta) \big) + \log \big(p(\boldsymbol{z} \,|\, \Theta) \big) \Big] \, \mathrm{d}\boldsymbol{z} \, + \, H(\Theta^{\mathrm{old}}) \,, \quad (3)$$

where $H(\Theta^{\mathrm{old}})$ is an entropy term only depending on parameter values held fixed during the optimization of $\mathcal{F}$ w.r.t. Θ. Note that integration over the hidden space involves an integral over the continuous part and a sum over the binary part. Optimizing the free-energy consists of two steps: given the current parameters Θ^{old} the posterior probability is computed in the E-step; and given the posterior, $\mathcal{F}(\Theta^{\mathrm{old}}, \Theta)$ is maximized w.r.t. Θ in the M-step. Iteratively applying E- and M-steps locally maximizes the data likelihood.

M-step parameter updates. Let us first consider the maximization of the free-energy in the M-step before considering expectation values w.r.t. to the posterior in the E-step. Given a generative model, conditions for a maximum free-energy are canonically derived by setting the derivatives of $\mathcal{F}(\Theta^{\mathrm{old}}, \Theta)$ w.r.t. the second argument to zero. For the GSC model we obtain the following parameter updates:

$$W = \Big(\sum_{n=1}^{N} \boldsymbol{y}^{(n)} \big\langle \boldsymbol{s} \odot \boldsymbol{z} \big\rangle_{n}^{\mathrm{T}} \Big) \Big(\sum_{n=1}^{N} \big\langle (\boldsymbol{s} \odot \boldsymbol{z})(\boldsymbol{s} \odot \boldsymbol{z})^{\mathrm{T}} \big\rangle_{n} \Big)^{-1}, \quad (4)$$

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} \Big[\boldsymbol{y}^{(n)} (\boldsymbol{y}^{(n)})^{\mathrm{T}} - 2 W \big\langle \boldsymbol{s} \odot \boldsymbol{z} \big\rangle_{n} (\boldsymbol{y}^{(n)})^{\mathrm{T}} + W \big\langle (\boldsymbol{s} \odot \boldsymbol{z})(\boldsymbol{s} \odot \boldsymbol{z})^{\mathrm{T}} \big\rangle_{n} W^{\mathrm{T}} \Big]$$

$$\text{and} \quad \boldsymbol{\pi} = \frac{1}{N} \sum_{n=1}^{N} \big\langle \boldsymbol{s} \big\rangle_{n}, \quad \text{where} \quad \big\langle f(\boldsymbol{s}, \boldsymbol{z}) \big\rangle_{n} = \sum_{\boldsymbol{s}} \int_{\boldsymbol{z}} p(\boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{y}^{(n)}, \Theta^{\mathrm{old}}) \, f(\boldsymbol{s}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}. \quad (5)$$

Equations (4) to (5) define a new set of parameter values $\Theta = (W, \Sigma, \boldsymbol{\pi})$ given the current values Θ^{old}. These 'old' parameters are only used to compute the sufficient statistics $\langle \boldsymbol{s} \rangle_{n}$, $\langle \boldsymbol{s} \odot \boldsymbol{z} \rangle_{n}$ and $\langle (\boldsymbol{s} \odot \boldsymbol{z})(\boldsymbol{s} \odot \boldsymbol{z})^{\mathrm{T}} \rangle_{n}$ of the model.

Expectation Values. Although the derivation of M-step equations can be analytically intricate, it is the E-step that, for most generative models, poses the major challenge. It usually involves computations of analytically intractable integrals that are required for posterior distributions and for expectation values w.r.t. the posterior. The true posterior is therefore often replaced by an approximate distribution [see e.g., 2, 17] or in the form of factored variational distributions [6, 5]. The most frequently used approximation is the maximum-a-posterior (MAP) estimate [see, e.g., [15], [9]] which replaces the true posterior by a delta-function around the posterior's maximum value. Alternatively, analytically intractable expectation values are often approximated using sampling approaches. Using approximations always implies, however, that many analytical properties of exact EM are not maintained. Approximate EM iterations may, for instance, decrease the likelihood or may not recover (local or global) likelihood optima in many cases. There are nevertheless, a limited number of models with exact EM solutions; e.g., mixture models such as the mixture-of-Gaussians, p-PCA or factor analysis etc. Our work adds a sparse coding model with continuous latents to the set of models with exact EM solution. By following along the same lines as for the p-PCA derivations, we maintain in our E-step the analytical tractability of computing expectation values w.r.t. the posterior of the GSC model (5).

Posterior Probability. First observe that the discrete latent variable $\boldsymbol{s}$ of the GSC model can be directly combined with the basis functions, i.e., $W(\boldsymbol{s} \odot \boldsymbol{z}) = \tilde{W}_{\boldsymbol{s}}\,\boldsymbol{z}$, where $(\tilde{W}_{\boldsymbol{s}})_{dh} = W_{dh} s_h$. Now we apply Bayes' rule to write down the posterior:

$$p(\boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{y}^{(n)}, \Theta) = \frac{\mathcal{N}(\boldsymbol{y}^{(n)};\, \tilde{W}_{\boldsymbol{s}}\,\boldsymbol{z}, \Sigma)\, \mathcal{N}(\boldsymbol{z};\, \boldsymbol{0}, \mathbb{1}_H)\, p(\boldsymbol{s} \,|\, \Theta)}{\sum_{\boldsymbol{s}'} \int \mathcal{N}(\boldsymbol{y}^{(n)};\, \tilde{W}_{\boldsymbol{s}'}\,\boldsymbol{z}', \Sigma)\, \mathcal{N}(\boldsymbol{z}';\, \boldsymbol{0}, \mathbb{1}_H)\, p(\boldsymbol{s}' \,|\, \Theta)\, \mathrm{d}\boldsymbol{z}'}. \quad (6)$$

Note that given a state $\boldsymbol{s}$ in (6), the Gaussian governing the observations $\boldsymbol{y}^{(n)}$ is only dependent on the Gaussian over the continuous latent $\boldsymbol{z}$, which is analytically independent of $\boldsymbol{s}$. We can exploit this joint relation to refactorize the Gaussians. Using Gaussian identities the posterior can be rewritten as:

$$\begin{aligned} p(\boldsymbol{s}, \boldsymbol{z} \,|\, \boldsymbol{y}^{(n)}, \Theta) &= \frac{\mathcal{N}(\boldsymbol{y}^{(n)}; \boldsymbol{0}, C_{\boldsymbol{s}})\, p(\boldsymbol{s} \,|\, \Theta)\, \mathcal{N}(\boldsymbol{z};\, \boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)}, \Lambda_{\boldsymbol{s}})}{\sum_{\boldsymbol{s}'} \mathcal{N}(\boldsymbol{y}^{(n)}; \boldsymbol{0}, C_{\boldsymbol{s}'})\, p(\boldsymbol{s}' \,|\, \Theta)} \\ &= p(\boldsymbol{s} \,|\, \boldsymbol{y}^{(n)}, \Theta)\, \mathcal{N}(\boldsymbol{z};\, \boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)}, \Lambda_{\boldsymbol{s}}), \end{aligned} \quad (7)$$

where $C_{\boldsymbol{s}} = \tilde{W}_{\boldsymbol{s}} \tilde{W}_{\boldsymbol{s}}^{\mathrm{T}} + \Sigma$, $\Lambda_{\boldsymbol{s}} = \big(\tilde{W}_{\boldsymbol{s}}^{T}\, \Sigma^{-1}\, \tilde{W}_{\boldsymbol{s}} + \mathbb{1}_H\big)^{-1}$ and $\boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)} = \Lambda_{\boldsymbol{s}}\, \tilde{W}_{\boldsymbol{s}}^{\mathrm{T}}\, \Sigma^{-1}\, \boldsymbol{y}^{(n)}$. Equation (7) represents the crucial result for the computation of the E-step below because, first, they show that the posterior does not involve analytically intractable integrals and, second, for fixed $\boldsymbol{s}$ and $\boldsymbol{y}^{(n)}$ the dependency on $\boldsymbol{z}$ follows a Gaussian distribution. This special form allows for the derivation of analytical expressions for the expectation values as required for the M-step parameter updates.

E-step Equations. Derived from (7), the expectation values are computed as:

$$\big\langle \boldsymbol{s} \big\rangle_n = \sum_{\boldsymbol{s}} p(\boldsymbol{s} \,|\, \boldsymbol{y}^{(n)}, \Theta)\; \boldsymbol{s}, \qquad \big\langle \boldsymbol{s} \odot \boldsymbol{z} \big\rangle_n = \sum_{\boldsymbol{s}} p(\boldsymbol{s} \,|\, \boldsymbol{y}^{(n)}, \Theta)\; \boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)} \quad (8)$$

$$\text{and} \quad \big\langle (\boldsymbol{s} \odot \boldsymbol{z})(\boldsymbol{s} \odot \boldsymbol{z})^{\mathrm{T}} \big\rangle_n = \sum_{\boldsymbol{s}} p(\boldsymbol{s} \,|\, \boldsymbol{y}^{(n)}, \Theta)\; \big(\Lambda_{\boldsymbol{s}} + \boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)} (\boldsymbol{\kappa}_{\boldsymbol{s}}^{(n)})^{\mathrm{T}}\big). \quad (9)$$

Note that we have to use the current values $\Theta = \Theta^{\mathrm{old}}$ for all parameters on the right-hand-side. The E-step equations (8) to (9) represent a closed-form solution for expectation values required for the closed-form M-step (4) to (5).

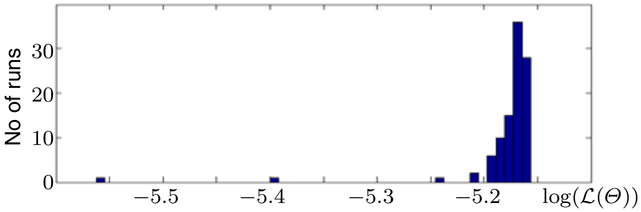


Fig. 2. Histogram of likelihood values for 100 runs of the GSC algorithm on data generated by a SC model with Cauchy prior. Almost all runs converged to high likelihood values.

3 Numerical Experiments

GSC parameter optimization is non-convex. However, as for all algorithms based on closed-form EM, the GSC algorithm always increases the data likelihood at least to a local maxima. We first numerically investigate how frequently local optima are obtained. Later we assess the model's performance on more practical tasks.

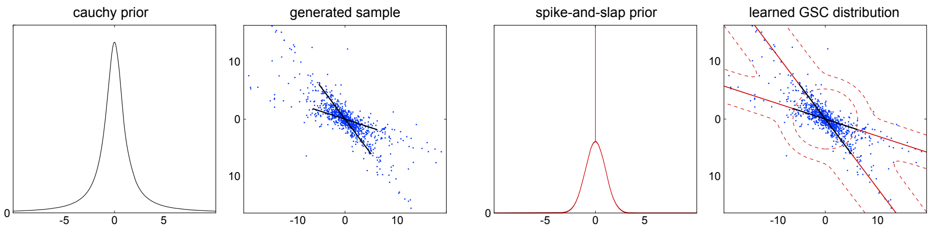
Model Verification: First, we verified on artificial data that the algorithm increases the likelihood and that it can recover the parameters of the generating distribution. For this, we generated $N = 500$ data points $\boldsymbol{y}^{(n)}$ from the GSC generative model (1) to (2) with $D = H = 2$. We used randomly initialized generative parameters[1]. The algorithm was run 250 times on the generated data. For each run we performed 300 EM iterations. For each run, we randomly and uniformly initialized π_h between 0.05 and 1, set Σ to the covariance across the data points, and the elements of W we chose to be independently drawn from a normal distribution with zero mean and unit variance. In all runs the generating parameter values were recovered with high accuracy. Runs with different generating parameters[1] produced essentially the same results.

Recovery of Sparse Directions. To test the model's robustness w.r.t. a relaxation of the GSC assumptions, we applied the GSC algorithm to data generated by standard sparse coding models. We used a standard Cauchy prior and a Gaussian noise model [15] for data generation. Fig. 3 second panel shows data generated by this sparse coding model while the first panel shows the prior density along one of its hidden dimensions. We generated $N = 500$ data points with $H = D = 2$. We then applied the GSC algorithm with the same parameter initialization as in the previous experiment. We performed 100 trials using 300 EM iterations per trial. Again, the algorithm converged to high likelihood values in most runs (see Fig. 2). As a performance measure for this experiment we investigated how well the heavy tails (i.e., the sparse directions) of standard SC were recovered. As a performance metric, we used the Amari index [1]:

$$A(W) = \tfrac{1}{2H(H-1)} \sum_{h,h'=1}^{H} \left(\tfrac{|O_{hh'}|}{\max_{h''} |O_{hh''}|} + \tfrac{|O_{hh'}|}{\max_{h''} |O_{h''h'}|} \right) - \tfrac{1}{H-1} \qquad (10)$$

where $O_{hh'} := \left(W^{-1} W^{\mathrm{gen}} \right)_{hh'}$. The mean Amari index of all runs with high likelihood values was below 10^{-2}, which shows a very accurate recovery of the sparse directions. The right panel in Fig. 3 visualizes the distribution recovered by the GSC

[1] We obtained W^{gen} by independently drawing each matrix entry from a normal distribution with zero mean and standard deviation 3. π_h^{gen} values were drawn from a uniform distribution between 0.05 and 1, $\Sigma = \sigma^{\mathrm{gen}} \mathbb{1}_D$ (where σ^{gen} was uniformly drawn between 0.05 and 10).


cauchy prior

generated sample

spike-and-slap prior

learned GSC distribution

Fig. 3. Comparison of standard sparse coding and GSC. **Left panels**: Cauchy distribution (along one hidden dimension) as a standard SC prior [15] and data generated by it. **Right panels**: Spike-and-slab distribution (one of the hidden dimensions) inferred by the GSC algorithm along with inferred sparse directions (solid red lines) and posterior data density contours (dotted red lines).

algorithm in a typical run. The dotted red lines show the density contours of the learned distribution $p(\boldsymbol{y} \,|\, \Theta)$. High accuracy in the recovery of the generating sparse directions (solid black lines) can be observed by comparison with the recovered directions (solid red lines). The results remain qualitatively the same if we increase the number of hidden and observed dimensions; e.g., for $H = D = 4$ we found the algorithm converged to a high likelihood in 91 (with average Amari index below 10^{-2}) of 100 runs.

Other than standard SC with Cauchy prior, we also ran the algorithm on data generated by SC with Laplace prior [15, 9]. There, for $H = D = 2$, we converged to high likelihood values in 99 of 100 runs with an average Amari index 0.06. In the experiment with $H = D = 4$ the algorithm converged to a high likelihood in 97 of 100 runs. The average Amari index of all runs with high likelihoods was 0.07 in this case.

Source Separation. We applied the GSC algorithm to publicly available benchmarks. We used the non-artificial benchmarks of [18]. The datasets mainly contain acoustic data obtained from [ICALAB; 3]. We generated the observed data by mixing the benchmark sources using randomly generated orthogonal mixing matrices (we followed [18]). Again the Amari index (10) was used as a performance measure.

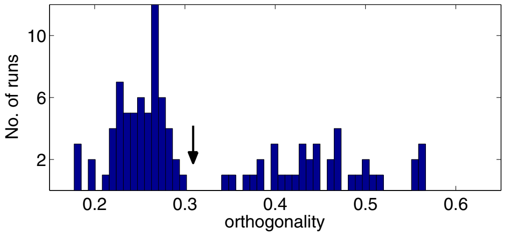


Fig. 4. Histogram of the deviation from orthogonality of the W matrix for 100 runs of the GSC algorithm on the `Speech4` benchmark ($N = 500$). A clear cluster of the most orthogonal runs can automatically be detected: the threshold of runs considered is defined to be the minimum after the cluster (black arrow).

For all the benchmarks we used $N = 200$ and $N = 500$ data points (as selected by [18]). We applied GSC to the data using the same initialization as described before. For each experiment we performed 100 trials with a random new parameter initialization per trial. The first column of Tab. 1 list average Amari indices obtained including

Table 1. Performance of different algorithms on benchmarks for source separation. Data for NG-LICA, KICA, FICA, and JADE are taken from [18]. Performances are compared based on the Amari index [10]. Bold values highlight the best performing algorithm(s).

datasets		Amari index (standard deviation)					
name	N	GSC	GSC$^{\perp}$	NG-LICA	KICA	FICA	JADE
10halo	200	0.34(0.05)	**0.29(0.03)**	**0.29(0.02)**	0.38(0.03)	0.33(0.07)	0.36(0.00)
	500	0.27(0.01)	0.27(0.01)	**0.22(0.02)**	0.37(0.03)	**0.22(0.03)**	0.28(0.00)
Sergio7	200	0.23(0.06)	0.20(0.06)	**0.04(0.01)**	0.38(0.04)	0.05(0.02)	0.07(0.00)
	500	0.18(0.05)	0.17(0.03)	0.05(0.02)	0.37(0.03)	**0.04(0.01)**	**0.04(0.00)**
Speech4	200	0.25(0.05)	**0.17(0.04)**	0.18(0.03)	0.29(0.05)	0.20(0.03)	0.22(0.00)
	500	0.11(0.04)	**0.05(0.01)**	0.07(0.00)	0.10(0.04)	0.10(0.04)	0.06(0.00)
c5signals	200	0.39(0.03)	0.44(0.05)	0.12(0.01)	0.25(0.15)	**0.10(0.02)**	0.12(0.00)
	500	0.41(0.05)	0.44(0.04)	0.06(0.04)	0.07(0.06)	**0.04(0.02)**	0.07(0.00)

all trials per experiment[2]. It is important to note that all the other algorithms listed in the comparison assume orthogonal mixing matrices, while the GSC algorithm does not. Therefore in the column 'GSC$^{\perp}$' in Tab. 1, we report statistics that are only computed over the runs which inferred the most orthogonal bases. As a measure of orthogonality we used the maximal deviation from 90^o between any two axes. Fig. 4 shows as an example a histogram of the maximal deviations of all trials on the `Speech4` data with $N = 500$. As can be observed, we obtained a clear cluster of runs with high orthogonality. We observed worst performance of the GSC algorithm on the `c5signals` dataset. However, the dataset contains sub-Gaussian sources which in general can not be recovered by sparse coding approaches.

4 Discussion

The GSC algorithm is a SC algorithm based on a spike-and-slab prior instead of a standard heavy-tail prior. The algorithm has a distinguishing capability of taking the whole (potentially multimodal) posterior into account for parameter optimization, which is in contrast to the MAP approximation of the posterior (as it is widely used for training SC models [see e.g., 9, 11]). MAP based algorithms can be very efficient but they do not take much of the posterior structure into account and, e.g., require regularization parameters. Other approaches with richer approximations of the posterior are, therefore, actively investigated [e.g., 17, 13]. However, any approximation can introduce learning biases, e.g., through assumptions of monomodal posteriors.

Closed-form EM learning of the GSC algorithm uses no approximations but it comes with a computational cost that is exponential w.r.t. the number of hidden dimensions H. This can be seen considering Eqn. 7 which requires sums over all binary vectors $\boldsymbol{s}$ (similar for expectation values w.r.t. the posterior). Nevertheless, we show in numerical experiments that the approach is well applicable to the typical range of source separation

[2] We obtained the reported results by diagonalizing the updated Σ in the M-step by setting $\Sigma = \sigma^2 \mathbb{1}_D$, where $\sigma^2 = \mathrm{Tr}(\Sigma)/D$.

tasks. As the GSC algorithm takes multimodal posteriors into account and as it infers model parameters including sparsity per latent, it can be considered as more Bayesian than, e.g., SC with MAP estimates [compare 9]. Note, however, that another line of research focuses on a fully Bayesian treatment of SC including approaches using spike-and-slab priors [e.g., 7, 19, 16, 8, 13, etc.]. While these methods emphasize on greater flexibility (estimation of the number of latents, use of different noise models etc.), their great challenge is the procedure of parameter estimation (see e.g., the combination of deterministic and sampling approximations in [13]). In contrast to such more general methodologies, the aim of this work is to generalize sparse coding in a form that still allows for closed-form EM solutions.

To summarize, we have studied a novel sparse coding algorithm and have shown its competitiveness on source separation benchmarks. Along with the reported results on source separation, the main contribution of this work is the derivation and numerical investigation of the (to the knowledge of the authors) first closed-form, exact EM algorithm for spike-and-slab sparse coding.

Acknowledgement. J. Lücke is funded by the German Research Foundation (DFG), grant LU 1196/4-1; A.-S. Sheikh is funded by the German BMBF, grant 01GQ0840.

References

[1] Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: NIPS, pp. 757–763. MIT Press (1996)

[2] Bishop, C.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)

[3] Cichocki, A., Amari, S., Siwek, K., Tanaka, T., Phan, A., Zdunek, R.: ICALAB-MATLAB Toolbox Version 3 (2007)

[4] Dayan, P., Abbott, L.F.: Theoretical Neuroscience. MIT Press, Cambridge (2001)

[5] Jaakkola, T.: Tutorial on variational approximation methods. In: Opper, M., Saad, D. (eds.) Advanced Mean Field Methods: Theory and Practice. MIT Press (2000)

[6] Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Machine Learning 37, 183–233 (1999)

[7] Knowles, D., Ghahramani, Z.: Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 381–388. Springer, Heidelberg (2007)

[8] Knowles, D., Ghahramani, Z.: Nonparametric Bayesian sparse factor models with application to gene expression modelling. CoRR, abs/1011.6293 (2010)

[9] Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: NIPS, vol. 22, pp. 801–808 (2007)

[10] Lücke, J., Sheikh, A.-S.: Closed-form EM for sparse coding and its application to source separation. arXiv:1105.2493v1 [stat.ML]

[11] Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML, p. 87 (2009)

[12] Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. Journal of the American Statistical Association 83(404), 1023–1032 (1988)

[13] Mohamed, S., Heller, K., Ghahramani, Z.: Sparse exponential family latent variable models. In: NIPS Workshop (2010)

[14] Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Jordan, M.I. (ed.) Learning in Graphical Models. Kluwer (1998)

[15] Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)

[16] Paisley, J.W., Carin, L.: Nonparametric factor analysis with beta process priors. In: ICML, p. 98 (2009)

[17] Seeger, M.: Bayesian inference and optimal design for the sparse linear model. JMLR 9, 759–813 (2008)

[18] Suzuki, T., Sugiyama, M.: Least-squares independent component analysis. Neural Computation 23(1), 284–301 (2011)

[19] Teh, Y.W., Görür, D., Ghahramani, Z.: Stick-breaking construction for the indian buffet process. Journal of Machine Learning Research - Proceedings Track 2, 556–563 (2007)

[20] West, M.: Bayesian factor regression models in the ”large p, small n” paradigm. In: Bayesian Statistics, vol. 7, pp. 723–732. Oxford University Press (2003)

Convolutive Underdetermined Source Separation through Weighted Interleaved ICA and Spatio-temporal Source Correlation

Francesco Nesta and Maurizio Omologo

Fondazione Bruno Kessler - Irst, Center of Information Technology, Italy

Abstract. This paper presents a novel method for underdetermined acoustic source separation of convolutive mixtures. Multiple complex-valued Independent Component Analysis adaptations jointly estimate the mixing matrix and the temporal activities of multiple sources in each frequency. A structure based on a recursive temporal weighting of the gradient enforces each ICA adaptation to estimate mixing parameters related to sources having a disjoint temporal activity. Permutation problem is reduced imposing a multiresolution spatio-temporal correlation of the narrow-band components. Finally, aligned mixing parameters are used to recover the sources through L_0-norm minimization and a post-processing based on a single channel Wiener filtering. Promising results obtained over a public dataset show that the proposed method is an effective solution to the underdetermined source separation problem.

1 Introduction

Blind source separation for acoustic sources has been studied for more than ten years and mainly aims to solve the cocktail party problem. In spite of the recent advances in its application to real-world scenarios [1], many issues still remain unsolved. High reverberation and the underdetermined condition are probably the two main obstacles which limit its applicability to more general realistic scenarios. Algorithms for underdetermined source separation in convolutive scenarios have been recently proposed. Some of them exploit spatial models to compute either Wiener filters or binary masks [2][3]. Other methods exploit the narrow-band mixing system formulation to better cope with high reverberation [4]. Finally, other algorithms uses multichannel temporal and spectral redundancies for factorizing different source components [5].

In order to separate sources in the underdetermined case, one of the key tasks that has to be solved is the estimation of the source mixing parameters. If at least the mixing parameters related to a single target source can be estimated, a constrained semi-blind source extraction can be applied in order to recover the source signal [6]. On the other hand, if the complete mixing system is available, binary masking or L_p-norm minimization can be adopted to recover source signals having a high sparse representation [7]. In the frequency-domain BSS based on ICA, mixing parameters for each frequency are estimated from the inverse of the demixing system, but the inversion can be applied only in the determined

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 222–230, 2012.

case, i.e. when the demixing system is a square matrix. On the other hand, it was shown that if one source dominates over the others its mixing parameters can be derived from one column of the inverse of the demixing system [8].

In the traditional formulation of batch-wise ICA, the gradient used in the iterative adaptation is computed by averaging the generalized covariance matrix over different time frames. This is based on the main assumption of stationarity and ergodicity of the random processes. However, acoustic sources are characterized by a prominent high non-stationarity and temporal-spectral sparseness which is naturally in contrast with the above assumptions. In this paper, we show how this cue can be used to our advantage in order to properly define a method based on multiple ICA adaptations, which is able to jointly estimate the complete mixing matrix even in the non-square case. Furthermore, since the adaptation is applied to each frequency independently, a second stage to align mixing parameters of different sources is implemented, which is based on the spatio-temporal correlation of the source signals.

2 Estimation of the Mixing Parameters

2.1 Weighted Natural Gradient

Let's assume to record N acoustic sources by a microphone array of M elements and denote with $s_n(t)$ the time-domain signal generated by the *n-th* source and with $x_m(t)$ the signal sampled at the *m-th* microphone. We move from time-domain to a more sparse representation of the sources by means of a short-time Fourier transform (STFT), applied to frames of N_{bins} samples. Let $S_n(k,l)$ and $X_m(k,l)$ be the *l-th* STFT frame coefficients obtained for the *k-th* frequency bin. Indicating the source signal vector with $\mathbf{S}(k,l) = [S_1(k,l) \cdots S_N(k,l)]^T$, the vector of the transformed observed mixtures $\mathbf{X}(k,l) = [X_1(k,l) \cdots X_M(k,l)]^T$ can be modeled as $\mathbf{X}(k,l) = \mathbf{H}(k)\mathbf{S}(k,l)$, where $\mathbf{H}(k)$ is the $M \times N$ mixing matrix corresponding to the transfer function between sources and microphones at *k-th* frequency bin. If $N = M$, by applying a complex-valued ICA algorithm to the time series at each frequency, one can retrieve the original components by estimating a set of de-mixing matrices $\mathbf{W}(k)$, representing an estimate of $\mathbf{H}(k)^{-1}$ up to scaling and permutation ambiguities. Then, the original signals can be computed as $\mathbf{Y}(k,l) = \mathbf{W}(k)\mathbf{X}(k,l)$. According to the minimization of the Kullback-Liebler divergence with the Natural Gradient (NG) [9], the mixing matrix and the corresponding output signals are estimated as follows

$$\mathbf{Y}_{(i)}(k,l) = \mathbf{W}_{(i)}(k)\mathbf{X}(k,l) = [\mathbf{H}_{(i)}(k)]^{-1}\mathbf{X}(k,l) \tag{1}$$

$$\varDelta\mathbf{H}_{(i)}(k) = \mathbf{H}_{(i)}(k)(\mathbf{I} - E[\varPhi(\mathbf{Y}_{(i)}(k,l))\mathbf{Y}_{(i)}(k,l)^H]) \tag{2}$$

$$\mathbf{H}_{(i+1)}(k) = \mathbf{H}_{(i)}(k) - \eta\varDelta\mathbf{H}_{(i)}(k) \tag{3}$$

where η is the step-size, i is the index of the iteration used to refine the solution, $\Phi(\cdot)$ is a non-linearity and $E[\cdot]$ indicates the expectation. Note, the above Natural Gradient formulation directly updates $\mathbf{H}(k)$ instead of $\mathbf{W}(k)$, as in the most popular NG formulation which does not require any matrix inversion. However (3) is a convenient formulation since it shows that each column of the gradient updates the mixing parameters of a different source.

In a batch approach, the gradient is computed averaging the generalized covariance matrix over the frames, i.e. the expectation operator $E[\cdot]$ is substituted with the time average, implicitly assuming ergodicity and stationarity. However, acoustic signals are highly non-stationary and their spectral representation is sparse over time. In principle, if a prior knowledge on the source activity is available, the expectation can be improved through a time weighted average. In other terms, the gradient part that updates the mixing parameters of a given source is computed only with frames where the source is believed to be dominant. The estimation of (3) can be modified as

$$\Delta\mathbf{H}_{(i)}(k) = \langle[\mathbf{H}_{(i)}(k)(\mathbf{I} - \Phi(\mathbf{Y}_{(i)}(k,l))\mathbf{Y}_{(i)}(k,l)^H)]\mathbf{\Psi}(k,l)\rangle_l \tag{4}$$

where $\langle\cdot\rangle_l$ indicates time average over l, $\mathbf{\Psi}(k,l)$ is a diagonal matrix with element $\psi_{nn}(k,l)$ being a weight (with values ranging from 0 to 1) indicating the dominance of the *n-th* source at each time-frequency point.

2.2 Interleaved Orthogonal Adaptations

In the previous work [10], the weights $\psi_{nn}(k,l)$ were recursively computed from the output power ratios of previously separated frequencies, assuming temporal dependencies across neighbor frequency bins. This approach speeds up the overall convergence and considerably reduces errors due to the statistical bias of a limited amount of observed data. Similarly in [11] the prior weighting was used in a distributed context in order to constrain separate ICA adaptations to estimate mixing parameters with the same source order. In contrast, in this work the weighting is adopted to force multiple ICA adaptations to converge to solutions identifying sources having an orthogonal temporal dominance. Assuming to observe N sources we define N ICA adaptations as

$$\mathbf{Y}^n_{(i)}(k,l) = \mathbf{W}^n_{(i)}(k)\mathbf{X}(k,l) \tag{5}$$

$$\mathbf{H}^n_{(i+1)}(k) \longleftarrow \mathbf{H}^n_{(i)}(k) - \eta\langle[\mathbf{H}^n_{(i)}(k)(\mathbf{I} - \Phi(\mathbf{Y}^n_{(i)}(k,l))\mathbf{Y}^n_{(i)}(k,l)^H)]\mathbf{\Psi}^n(k,l)\rangle_l \tag{6}$$

constrained by the weighting diagonal matrix $\mathbf{\Psi}^n(k,l)$ with diagonal elements equal to $[p^n(k,l), 1-p^n(k,l), \cdots, 1-p^n(k,l)]$, where $p^n(k,l)$ is the posterior probability of observing the n-*th* source in the (k,l) point given the observation $\mathbf{X}(k,l)$. Assuming orthogonality between the temporal evolution of posteriors of different sources, i.e. $\langle p^n(k,l)p^{n'}(k,l)\rangle_l = 0, \forall(n \neq n')$, each ICA adaptation would converge to mixing matrices $[\mathbf{H}^1(k), \cdots, \mathbf{H}^N(k)]$ where their first columns represent the mixing parameters of N different sources. The posteriors $p^n(k,l)$ can be computed through a multichannel statistical model of the observed STFT coefficients

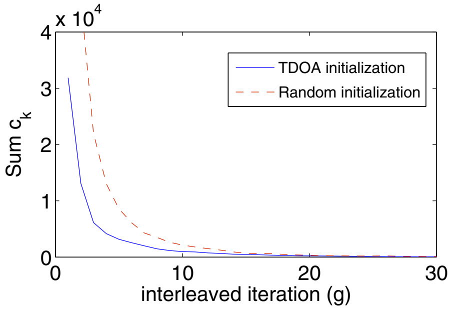


Fig. 1. Typical average convergence curve for mask estimation, starting from a TDOA-based or random initialization

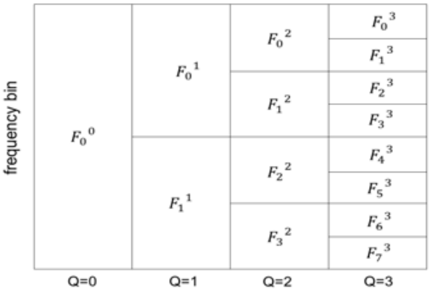


Fig. 2. Dyadic definition of frequency bins subsets for the multi-resolution spatio-temporal correlation analysis

$\mathbf{X}(k,l)$, whose parameters can be estimated a priori (e.g. if the source TDOAs are available beforehand, a spatial model can be used for modeling the inter-channel phase characteristic of $\mathbf{X}(k,l)$). In contrast, here we adopt an EM-like procedure in order to iteratively estimate both $p^n(k,l)$ and $[\mathbf{H}^1(k), \cdots, \mathbf{H}^N(k)]$ only from the observed data, and assuming ideal source sparseness.

Indicating with $\mathbf{h}^n(k)$ the last estimate of the mixing vector of the n-*th* source (i.e. the first column of $\mathbf{H}^n(k)$), the posteriors can be approximated with the ideal binary masks obtained from the estimated mixing vectors as

$$p^n(k,l) = \begin{cases} 1, \quad if \quad \underset{i}{\operatorname{argmax}} \, |[\mathbf{h}^i(k)]^H \mathbf{X}(k,l)| = n \\ \\ 0, \qquad\qquad otherwise \end{cases} \tag{7}$$

where H indicates the Hermitian transpose. In practice, we assume that only one source is observed in each time-frequency point, which is a simplistic assumption but effective to impose the orthogonality constraint. The overall structure of the interleaved weighted ICA adaptation can be summarized as follows

```
for k=1 to $N_k$
g=1; $c_k = \infty$;initialization of $\Psi_0^n(k,l)$;
    while $(c_k > 0)$ and $(g \leq N_g)$,
    for n=1 to N
        for i=1 to $N_i$
            Compute $\mathbf{H}^n_{(i+1)}(k)$ as in (5)-(6)
        end
    end
    Compute $p^n_{(g)}(k,l)$ as in (7) from last $\mathbf{h}^n(k)$, $\forall n$
    $c_k = \sum_{l,n} |p^n_{(g)}(k,l) - p^n_{(g-1)}(k,l)|$
    g=g+1
    end
end.
```

For each frequency, $\mathbf{\Psi}_0^n(k,l)$ is initialized imposing $p^n(k,l) \quad \forall n$ to be a set of orthogonal binary posteriors. If an estimation of the source TDOA vectors is available, $p^n(k,l)$ can be defined as the ideal binary mask obtained with (7) and initially approximating $\mathbf{h}^n(k)$ with the ideal anechoic propagation model. If this information is unknown, mixing parameters can be initialized by using random TDOA vectors. For each ICA adaptation the mixing matrix $\mathbf{H}^n(k)$ is iteratively estimated as in (5)-(6) for N_i iterations. Thus, the posteriors are recomputed as in (7) with the last estimates of $\mathbf{h}^n(k)$ and the procedure is iterated till convergence, i.e. when the estimated binary masks do not change with the iteration. Note, in this work we used the Natural Gradient for the ICA stage but the proposed structure can be used with any algorithm, on condition that the gradient can be formulated in terms of mixing matrix $\mathbf{H}(k)$. Figure 1 shows a typical convergence curve obtained over N_g iterations for both random or TDOA-based initialization of $\mathbf{\Psi}^n(k,l)$.

3 Permutation Alignment

The above procedure estimates the source mixing vectors independently at each frequency. In principle, the adaptation may be run jointly in all the frequencies if a spatial and/or spectral wide-band model for the sources is available. In this case the posteriors may be computed from all the estimated mixing parameters and the adaptations would not be affected by the permutation problem. However, this requires to introduce global constraints in the estimated mixing vectors (or source posteriors) which might reduce the convergence speed and accuracy of the estimated solution. In this work we adopt a two stage approach as for standard ICA based frequency-domain BSS, where the source mixing parameters are independently estimated and the permutations are reduced in a second stage.

3.1 TDOA Vector Estimation

A necessarily preliminary step is the estimation of TDOA vectors related to the multidimensional propagation of the sources over the direct-path. This can be done by seeking for the maxima of the Generalized State Coherence Transform (GSCT) [12], which under ideal source sparseness can be obtained directly from the normalized cross-power spectrum [13].

Note, if there is sufficient sparsity of the TDOA distribution related to different sources and the kernel bandwidth is sufficiently small, the total likelihood may be cumulated over the time-frames l, and the modes can be easily detected from a single density representation. However, if the microphones are too close to each other (e.g. $0.05m$) and the reverberation is high, TDOA distribution of multiple sources highly overlap and the modes detection may become difficult. To circumvent this problem we explicitly exploit the temporal sparsity of the source activity. Multiple maxima are estimated in each frame independently and grouped together through a spatio temporal clustering. Finally, the N TDOA vectors related to different spatial locations and with the highest likelihoods are selected.

3.2 Multi-resolution Spatio-temporal Correlation

Once the TDOA vectors are estimated, we enforce a spatio-temporal correlation between the narrow-band source signals. We define a multichannel narrow-band model of the activity of the n-*th* source as

$$\mathbf{s}^n_Q(k,l) = \mathbf{d}^n(k) s^n_Q(k,l), \qquad \mathbf{d}^n(k) = [1; e^{-2\pi j f_k \tau^n_1}; \cdots; e^{-2\pi j f_k \tau^n_P}]^T \tag{8}$$

where $s^n_Q(k,l) = \langle p^n(k,l) \rangle_{k \in \mathrm{F}^Q_{q(k)}}$ is the average of the posteriors $p^n(k,l)$ related to frequency bins included in the subset $\mathrm{F}^Q_{q(k)}$, at Q-*th* level of spectral resolution, and $[\tau^n_1; \cdots; \tau^n_P]$ is the TDOA vector estimated for the n-*th* source. While the spatial model identified by $\mathbf{d}^n(k)$ is kept constant at each resolution level Q, the spectral resolution of the temporal activity is properly defined through the definition of $\mathrm{F}^Q_{q(k)}$. Here we adopt a dyadic subdivision of the entire set of frequencies as shown in Figure 2. Then, we define the normalized mixing vector as $\overline{\mathbf{h}}^n(k) = \frac{\mathbf{h}^n(k)}{h^n_1(k)} / |\frac{\mathbf{h}^n(k)}{h^n_1(k)}|$, where $h^n_1(k)$ is the first element of $\mathbf{h}^n(k)$ and / refers to the element-wise division. Indicating with $\mathbf{p}^n(k,l) = \overline{\mathbf{h}}^n(k) p^n(k,l)$ the multichannel representation of the source activity, a measure of spatio-temporal correlation between the model $\mathbf{s}^n_Q(k,l)$ and $\mathbf{p}^n(k,l)$ is computed as

$$C[\mathbf{p}^n(k,l), \mathbf{s}^n_Q(k,l)] = \left(\frac{Re\{[\overline{\mathbf{h}}^n(k)]^H \mathbf{d}^n(k)\} + (P-1)}{2P} \right)^{\alpha} \cdot \left(\frac{\tilde{\mathbf{p}}^n(k)[\tilde{\mathbf{s}}^n_Q(k)]^T}{||\tilde{\mathbf{p}}^n(k)|| \times ||\tilde{\mathbf{s}}^n_Q(k)||} \right)^{1-\alpha} \tag{9}$$

where α is a coefficient with values ranging between 0 and 1, $Re\{\cdot\}$ indicates the real part, $\tilde{\mathbf{s}}^n_Q(k) = [s^n_Q(k,1); s^n_Q(k,2); \cdots]$ and $\tilde{\mathbf{p}}^n(k) = [p^n(k,1); p^n(k,2); \cdots]$. Then, given the models $\mathbf{s}^n_Q(k,l)$ $\forall n$, for each k we seek for the permutation matrix Π, with corresponding permutation function $\Psi(n) : N \rightarrow N$ which maximizes

$$\Pi_k = \underset{\Pi}{\mathrm{argmax}} \sum_n C[\mathbf{s}^n_Q(k,l), \mathbf{p}^{\Pi(n)}(k,l)]. \tag{10}$$

For a given resolution stage Q, equations (8)-(10) are iterated till convergence, i.e. when no permutation matrix changes with the iterations. We start from a low spectral resolution, i.e. a single temporal envelope is adopted to align all the frequency bins, till to a high resolution where multiple models are used to locally improve the alignment of neighbor frequencies.

Note, the coefficient α defines the importance of the spatial and temporal correlation. A high value of α increases the robustness of the permutation alignment against temporally correlated source signals, e.g. music signals, or when the sources are observed for a short time. On the other hand, a high value of α would prevent the optimization to converge to an optimal solution in presence of high reverberation and high spatial aliasing, where the propagation over the direct path does not well describe the convolutive mixing system. Here we follow a simple evidence: the estimated TDOA vectors represent a global spatial coherence (related to the direct path propagation) and then have to be used for the optimization at low resolution level. On the other hand, the temporal correlation needs to be locally optimized to better compensate wrong alignments (at lower resolution stages) due to phase discontinuities generated by high reverberation. Therefore, we force $\alpha = 0.5$ for $Q < Q_{max}$ (e.g. with $Q_{max} = 4$), while $\alpha = 0$ for higher resolutions. Nevertheless, we believe that other adaptive strategies to set α worth to be investigated in the future.

4 Source Recovery

Sources are recovered using the L_0-norm minimization and the Minimal Distortion Principle (MDP), in order to estimate the multichannel image of the separated source signals. Since L_0-norm is highly sensitive to the accuracy of the estimated mixing system, it may lead to a poor separation of low frequencies, where the temporal correlation of the sources is high. Therefore, we further apply a Wiener post-processing computing the Wiener gains for the n-*th* source and the m-*th* channel as $g_{mn}(k,l) = \frac{|y_m^n(k,l)|^2}{\sum_n |y_m^n(k,l)|^2}$, where $y_m^n(k,l)$ is the n-*th* source recovered through the L_0-norm minimization. Finally, signals are separated applying the Wiener gains to the input mixtures.

Note, in order to better account for the source sparseness, the Wiener gains are computed from the STFT coefficients obtained with windows of analysis shorter than those required for the ICA stage [6]. That is, the sources images $y_m^n(k,l)$, separated through the L_0-norm minimization, are reconstructed back to time-domain through overlap-and-add (OLA) and after retransformed through the STFT but with shorter windows (e.g. 1024 samples).

5 Experimental Results

The proposed method is validated for the case $M = 2$ with the publicly available "under-determined speech and audio mixtures" development dataset (dev1.zip), used in the Signal Separation Evaluation Campaign (SiSEC) 2011 [14]. Time-domain signals (sampled at $f_s = 16kHz$) were transformed through an STFT analysis with Hanning windows of 2048 or 4096 samples (respectively for the case of $T_{60} = 130$ ms and $T_{60} = 250$ ms) with a shift of 25% of the window. For the Wiener post-filtering Hanning windows of 1024 samples shifted of 128 samples were used. The scaled Natural Gradient was adopted for the ICA, setting the maximum number of iterations to $N_i = 10$, the step-size to $\mu = 0.2$ and the maximum number of interleaved iterations to $N_g = 50$. Performance are measured in terms of average Signal Distortion Ratio (SDR) where the decomposition of the separated signals in target and interference components is determined applying the resulting processing to the separated source image contributions as in [4]. Note that numbers may differ from those provided by the evaluation in SiSEC 2011 since the procedure for the estimation of the signal decomposition is different. For other comparisons with other state-of-art algorithms see the official results of SiSEC 2011 [14]. Furthermore, we compute the Misalignment between the estimated and the optimal Wiener filter gains computed with the true source images as $MIS_{mn} = 10\log_{10}[||\mathbf{g}_{mn} - \mathbf{g}_{mn}^{ideal}||^2/||\mathbf{g}_{mn}^{ideal}||^2]$, where $\mathbf{g}_{mn}$ and $\mathbf{g}_{mn}^{ideal}$

Table 1. Average SDR(Misalignment) for separation results of dev1 of SiSEC 2008

	T_{60} = 130 ms		T_{60} = 250 ms	
mic. spacing	5 cm	1 m	5 cm	1 m
male3	7.68(-5.81) dB	9.82(-5.07) dB	8.72(-5.15) dB	8.81(-4.24) dB
male4	5.44(-4.22) dB	5.89(-3.55) dB	4.65(-3.43) dB	6.64(-3.16) dB
female3	8.06(-5.91) dB	11.88(-5.07) dB	8.73(-5.14) dB	10.69(-4.06) dB
female4	6.80(-4.10) dB	5.92(-3.27) dB	5.96(-3.35) dB	6.92(-2.88) dB
wdrums	-	-	5.88(-2.73) dB	7.21(-2.28) dB
nodrums	-	-	6.67(-3.98) dB	7.61(-3.39) dB

are vectors with elements $g_{mn}(k,l)$ and g_{mn}^{ideal}, for any k and l, computed as in Section 4, and using the estimated and true source signal images, respectively.

Table 1 reports the detailed results and shows that the proposed algorithm is able to recover both speech and music sources with limited distortion. Numerical performance were also confirmed with a subjective evaluation, by human listening, confirming the high perceptual quality of the recovered signals[1].

6 Conclusion

This paper presents a novel method for the estimation of convolutive mixing parameters of multiple acoustic sources, in the underdetermined scenario. The mixing parameters are estimated independently in each frequency with an iterative structure based on interleaved weighted Natural Gradient adaptations, constrained to estimate mixing parameters of sources with a disjoint temporal activity. Narrow-band mixing vectors are then aligned enforcing a multi-resolution spatio-temporal correlation between the estimated narrow-band source signals. The proposed method is evaluated by recovering the source signals through an L_0-norm minimization and Wiener filtering post-processing. Numerical results and a subjective evaluation reveal that the proposed method is a promising robust solution to the underdetermined source separation of convolutive mixtures.

References

1. Christensen, H., Barker, J., Ma, N., Green, P.: The chime corpus: a resource and a challenge for computational hearing in multisource environments. In: Proceedings of Interspeech, Makuhari, Japan (2010)
2. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined convolutive blind source separation using spatial covariance models. In: Proc. ICASSP (2010)
3. Araki, S., Nakatani, T., Sawada, H., Makino, S.: Stereo Source Separation and Source Counting with MAP Estimation with Dirichlet Prior Considering Spatial Aliasing Problem. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 742–750. Springer, Heidelberg (2009)
4. Sawada, H., Araki, S., Mukai, R., Makino, S.: Blind extraction of dominant target sources using ICA and time-frequency masking. IEEE Trans. on Audio, Speech, and Language Processing 14(6), 2165–2173 (2006)

[1] Audio files are at `http://sisec2011.wiki.irisa.fr/tiki-index.php`

5. Ozerov, A., Fevotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. on Audio, Speech and Language Processing 18(3), 550–563 (2010)
6. Nesta, F., Matassoni, M.: Robust automatic speech recognition through on-line semi-blind source extraction. In: Proceedings of CHIME, Florence, Italy (2011)
7. Vincent, E.: Complex Nonconvex l_p Norm Minimization for Underdetermined Source Separation. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 430–437. Springer, Heidelberg (2007)
8. Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., Shikano, K.: Blind spatial subtraction array for speech enhancement in noisy environment. IEEE Trans. on Audio, Speech and Language Processing 17(4), 650–664 (2009)
9. Cichocki, A., Amari, S.-I.: Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. John Wiley & Sons, Inc., New York (2002)
10. Nesta, F., Svaizer, P., Omologo, M.: Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. IEEE Transactions on Audio, Speech, and Language Processing 19(3), 624–639 (2011)
11. Nesta, F., Omologo, M.: Cooperative wiener-ICA for source localization and separation by distributed microphone arrays. In: Proc. of ICASSP (March 2010)
12. Nesta, F., Omologo, M.: Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. IEEE Transactions on Audio, Speech, and Language Processing (2011)
13. Nesta, F., Omologo, M.: Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources. In: Proc. of ICASSP, Kyoto, Japan (to appear, 2012)
14. Araki, S., Nesta, F., Vincent, E., Koldovsky, Z., Nolte, G., Ziehe, A., Benichoux, A.: The 2011 Signal Separation Evaluation Campaign (SiSEC2011):-Audio Source Separation. In: Theis, F., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191. Springer, Heidelberg (2012)

Dictionary Learning with Large Step Gradient Descent for Sparse Representations

Boris Mailhé and Mark D. Plumbley

Queen Mary University of London
School of Electronic Engineering and Computer Science
Centre for Digital Music
Mile End Road, London E1 4NS, United Kingdom
firstname.name@eecs.qmul.ac.uk

Abstract. This work presents a new algorithm for dictionary learning. Existing algorithms such as MOD and K-SVD often fail to find the best dictionary because they get trapped in a local minimum. Olshausen and Field's Sparsenet algorithm relies on a fixed step projected gradient descent. With the right step, it can avoid local minima and converge towards the global minimum. The problem then becomes to find the right step size. In this work we provide the expression of the optimal step for the gradient descent but the step we use is twice as large as the optimal step. That large step allows the descent to bypass local minima and yields significantly better results than existing algorithms. The algorithms are compared on synthetic data. Our method outperforms existing algorithms both in approximation quality and in perfect recovery rate if an oracle support for the sparse representation is provided.

Keywords: Dictionary learning, sparse representations, gradient descent.

1 Introduction

In the method of sparse representations, a signal is expressed as a linear combination of a few vectors named *atoms* taken from a set called a *dictionary*. The sparsity constraint induces that any given dictionary can only represent a small subset of all possible signals, so the dictionary has to be adapted to the data being represented. Good pre-constructed dictionaries are known for common classes of signals, but sometimes it is not the case, for example when the dictionary has to discriminate against perturbations coming from noise [2]. In that case, the dictionary can be learned from examples of the data to be represented.

Several different algorithms have been proposed to learn the dictionary. Many of them iteratively optimize the dictionary and the decomposition [5,3,1]. The difference between those algorithms is the way they update the dictionary to fit a known decomposition. In particular, Olshausen and Field's Sparsenet algorithm [5] uses a fixed step gradient descent. In this work we observe that all those update methods are suboptimal even if the right support for the decomposition is known.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 231–238, 2012.

This work presents a modification to the Sparsenet algorithm that enables it to bypass local minima. We use the fact that the optimal step of the gradient descent can easily be obtained, then multiply it by constant larger than 1. Empirical results show that our method often allows the optimization to reach the global minimum.

2 Dictionary Learning

2.1 Problem

Let $\mathbf{S}$ be a $D \times N$ matrix of N training signals $\{\mathbf{s}_n\}_{n=1}^{N},\ \mathbf{s}_n \in \mathbb{R}^D$. Dictionary learning consists in finding a dictionary $\mathbf{\Phi}$ of size $D \times M$ with $M \geq D$ and sparse coefficients $\mathbf{X}$ such that $\mathbf{S} \approx \mathbf{\Phi X}$. For example, if the exact sparsity level K is known, the problem can be formalized as minimizing the error cost function

$$f(\mathbf{\Phi}, \mathbf{X}) = \|\mathbf{S} - \mathbf{\Phi X}\|_F^2 \tag{1}$$

under the constraints

$$\forall m \in [1, M],\ \|\boldsymbol{\varphi}_m\|_2 = 1 \tag{2}$$

$$\forall n \in [1, N],\ \|\mathbf{x}_n\|_0 \leq K \tag{3}$$

with $\boldsymbol{\varphi}$ an atom (or column) of $\mathbf{\Phi}$ and $\|\mathbf{x}_n\|_0$ the number of non-zero coefficients in the n^{th} column of $\mathbf{X}$.

2.2 Algorithms

Many dictionary learning algorithms follow an alternating optimization method. When the dictionary $\mathbf{\Phi}$ is fixed, estimating the sparse coefficients $\mathbf{X}$ is a sparse representation problem that can be approximately solved by algorithms such as Orthogonal Matching Pursuit (OMP) [6]. Existing algorithms differ in the way they update the dictionary $\mathbf{\Phi}$ once the coefficients $\mathbf{X}$ are fixed:

- Sparsenet [5] uses a projected gradient descent with a fixed step α:

$$\mathbf{R} = \mathbf{S} - \mathbf{\Phi X} \tag{4}$$

$$\nabla f = -\mathbf{R}\mathbf{x}^{mT} \tag{5}$$

$$\boldsymbol{\varphi}_m \leftarrow \boldsymbol{\varphi}_m - \alpha \nabla f \tag{6}$$

$$\boldsymbol{\varphi}_m \leftarrow \frac{\boldsymbol{\varphi}_m}{\|\boldsymbol{\varphi}_m\|_2} \tag{7}$$

  with $\mathbf{x}^m$ the m^{th} line of $\mathbf{X}$.
- MOD [3] directly computes the dictionary that minimizes the error f when the coefficients are fixed. The result is given by a pseudo-inverse:

$$\mathbf{\Phi} \leftarrow \mathbf{S}\mathbf{X}^+ \tag{8}$$

$$\forall m \in [1, M],\ \boldsymbol{\varphi}_m \leftarrow \frac{\boldsymbol{\varphi}_m}{\|\boldsymbol{\varphi}_m\|_2} \tag{9}$$

- K-SVD [1] jointly re-estimates each atom and the amplitude of its non-zero coefficients. For each atom $\boldsymbol{\varphi}_m$, the optimal choice is the principal component of a restricted "error" $\mathbf{E}^{(m)}$ obtained by considering the contribution of $\boldsymbol{\varphi}_m$ alone and removing all other atoms.

$$\mathbf{E}^{(m)} = \mathbf{R} + \boldsymbol{\varphi}_m \mathbf{x}^m \tag{10}$$

$$\boldsymbol{\varphi}_m \leftarrow \underset{\|\boldsymbol{\varphi}\|_2=1}{\operatorname{argmin}} \left\| \mathbf{E}^{(m)} - \boldsymbol{\varphi}\boldsymbol{\varphi}^T \mathbf{E}^{(m)} \right\|_F^2 \tag{11}$$

$$= \underset{\|\boldsymbol{\varphi}\|_2=1}{\operatorname{argmax}} \, \boldsymbol{\varphi}^T \mathbf{E}^{(m)} \mathbf{E}^{(m)^T} \boldsymbol{\varphi} \tag{12}$$

$$\mathbf{x}^m \leftarrow \boldsymbol{\varphi}_m^T \mathbf{E}^{(m)} \tag{13}$$

3 Motivations for an Adaptive Gradient Step Size

This section details an experimental framework used to compare the dictionary update methods presented in Section 2.2. We then show that MOD and K-SVD often get trapped in a local minimum but that with the right step, Sparsenet is more likely to find the global minimum.

3.1 Identifying the Global Optimum: Learning with a Fixed Support

We want to be able to check whether the solution found by an algorithm is the best one. It is easy in the noiseless case: if the training signals are exactly sparse on a dictionary, then there is at least one decomposition that leads to an error of 0: the one used for synthesizing the signals. In that case, a factorization $(\boldsymbol{\Phi}, \mathbf{X})$ is globally optimal if and only if the value of its error cost (1) is 0.

Dictionary learning algorithms often fail at that task because of mistakes done in the sparse representation step: when the dictionary is fixed, tractable sparse approximation algorithms typically fail to recover the best coefficients, although there are particular dictionaries for which the sparse representation is guaranteed to succeed [7]. In order to observe the behavior of the different dictionary update methods, we can simulate a successful sparse representation by using an oracle support: instead of running a sparse representation algorithm, the support used for the synthesis of the training signals is used as an input to the algorithm and only the values of the non-zero coefficients is updated by quadratic optimization. The dictionary learning algorithm is then simplified into Algorithm 1.

3.2 Empirical Observations on Existing Algorithms

We ran a simulation to check whether existing update methods are able to recover the best dictionary once the support is known. Each data set is made of a dictionary containing i.i.d. atoms drawn from a uniform distribution on the

Algorithm 1. $(\mathbf{\Phi}, \mathbf{X}) = \text{dict\_learn}(\mathbf{S}, \sigma)$

```
$\mathbf{\Phi} \leftarrow$ random dictionary
while not converged do
    $\forall n, \mathbf{x}_n^{\boldsymbol{\sigma}_n} \leftarrow \mathbf{\Phi}_{\boldsymbol{\sigma}_n}^{+} \mathbf{s}_n$
    $\mathbf{\Phi} \leftarrow \text{dict\_update}(\mathbf{\Phi}, \mathbf{S}, \mathbf{X})$
end while
```

unit sphere. For each dictionary, 256 8-sparse signals were synthesized by drawing uniform i.i.d. 8-sparse supports and i.i.d. Gaussian amplitudes. Then each algorithm was run for 1000 iterations starting from a random dictionary. The oracle supports of the representations were provided as explained in Section 3.1.

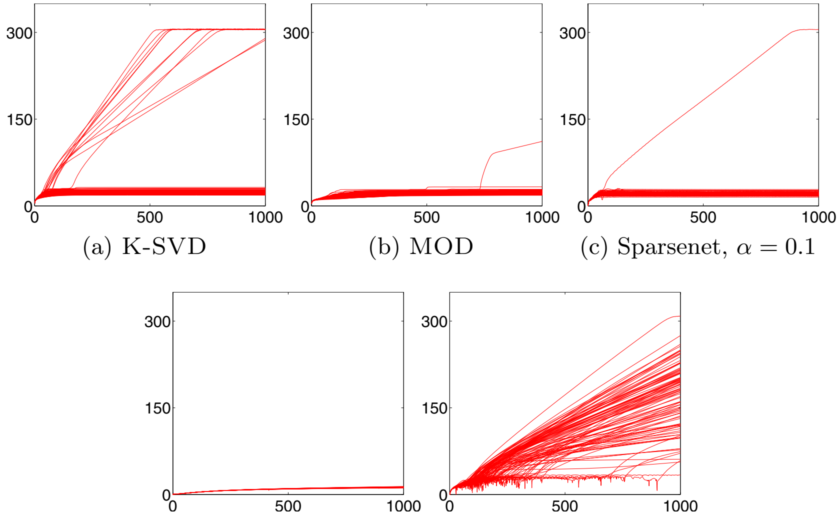
Figure 1 shows the evolution of the SNR $= -10 \log_{10} \frac{\|\mathbf{R}\|_2^2}{\|\mathbf{S}\|_2^2}$ over the execution of the algorithm for each data set. 300dB is the highest SNR that can be reached due to numerical precision. Moreover, we ran some longer simulations and never saw an execution fail to reach 300dB once a threshold of 100dB was passed For each algorithm, the plots show how many runs converged to a global minimum and how fast they did it.

K-SVD found a global minimum in 17 cases and has the best convergence speed of all studied algorithms. MOD only converged to a global minimum in 1 case and shows a tendency to evolve by steps, so even after a large number of iterations it is hard to tell whether the algorithm has converged or not. The best results were obtained when running Sparsenet with a step size $\alpha = 0.05$. In that case most runs converge to a global optimum although the convergence speed is more variable than with K-SVD. The behavior of Sparsenet highly depends on the choice of α. In our case a step of 0.1 is too large and almost always prevented the algorithm to converge, but a step of 0.01 is too small and leads to a very slow convergence.

Moreover, Sparsenet outperforms MOD although they both attempt to solve the same least-square problem. MOD finds that minimum in only one iteration, but if each Sparsenet dictionary update was allowed to iterate on its gradient descent with a well chosen step, it would converge towards the result of the MOD update. So the source of the gain is unlikely to be that the step $\alpha = 0.05$ is well adapted to the descent, but rather that it is larger than what an optimal step would be, thus allowing the descent to jump over local minima. The fact that the SNR sometimes decreases at one iteration for Sparsenet with $\alpha = 0.05$ also hints at a larger than optimal step size.

4 Large Step Gradient Descent

This section presents our method to choose the step size of the gradient descent. Our method is based on optimal step gradient descent, but we purposefully choose a step size that is larger than the optimal one.



(a) K-SVD (b) MOD (c) Sparsenet, $\alpha = 0.1$

(d) Sparsenet, $\alpha = 0.01$ (e) Sparsenet, $\alpha = 0.05$

Fig. 1. Approximation SNR depending on the iteration. K-SVD and MOD often get trapped in a local minimum. With $\alpha = 0.05$, Sparsenet avoids local minima, but $\alpha = 0.1$ is too large and $\alpha = 0.01$ is too small.

4.1 Optimal Step Projected Gradient Descent

When fixing the coefficients and the whole dictionary but one atom $\boldsymbol{\varphi}_m$, there is a closed-form solution for the best atom $\boldsymbol{\varphi}_m^*$ that minimizes the cost function (1) [4].

$$\boldsymbol{\varphi}_m^* = \underset{\|\boldsymbol{\varphi}_m\|_2=1}{\operatorname{argmin}} \ \|\mathbf{S} - \mathbf{\Phi}\mathbf{X}\|_F^2 \tag{14}$$

$$= \underset{\|\boldsymbol{\varphi}_m\|_2=1}{\operatorname{argmin}} \ \left\|\mathbf{E}^{(m)} - \boldsymbol{\varphi}_m \mathbf{x}^m\right\|_F^2 \tag{15}$$

with $\mathbf{E}^{(m)}$ the restricted errors described for K-SVD in Equation (10).

$$\left\|\mathbf{E}^{(m)} - \boldsymbol{\varphi}_m \mathbf{x}^m\right\|_F^2 = \left\|\mathbf{E}_k^{(m)}\right\|_F^2 - 2\left\langle \mathbf{E}_k^{(m)}, \boldsymbol{\varphi}_m \mathbf{x}^m \right\rangle + \|\boldsymbol{\varphi}_m \mathbf{x}^m\|_F^2 \tag{16}$$

$\left\|\mathbf{E}_k^{(m)}\right\|_F^2$ is constant with respect to $\boldsymbol{\varphi}_m$. If $\boldsymbol{\varphi}_m$ is constrained to be unitary, then $\|\boldsymbol{\varphi}_m \mathbf{x}^m\|_F^2 = \|\mathbf{x}^m\|_2^2$ is also constant with respect to $\boldsymbol{\varphi}_m$. So the only variable term is the inner product and the expression of the optimum $\boldsymbol{\varphi}_m^*$ is given by:

$$\boldsymbol{\varphi}_m^* = \underset{\|\boldsymbol{\varphi}_m\|_2=1}{\operatorname{argmax}} \left\langle \mathbf{E}^{(m)}\mathbf{x}^{mT}, \boldsymbol{\varphi}_m \right\rangle \tag{17}$$

$$= \frac{\mathbf{E}^{(m)}\mathbf{x}^{mT}}{\left\|\mathbf{E}^{(m)}\mathbf{x}^{mT}\right\|_2} \quad . \tag{18}$$

The link with the gradient appears when developing the Expression (18):

$$\boldsymbol{\varphi}_m^* \propto (\mathbf{R} + \boldsymbol{\varphi}_m \mathbf{x}^m)\,\mathbf{x}^{mT} \tag{19}$$

$$\propto \boldsymbol{\varphi}_m + \frac{1}{\|\mathbf{x}^m\|_2^2}\mathbf{R}\mathbf{x}^{mT} \quad . \tag{20}$$

Starting from the original atom, the global best atom $\boldsymbol{\varphi}_m^*$ can be obtained with only one iteration of gradient descent and the optimal step α^* of the descent is the inverse of the energy of the amplitude coefficients.

$$\alpha^* = \frac{1}{\|x^m\|_2^2} \tag{21}$$

5 Experimental Validation

This section presents dictionary learning experiments using gradient descent dictionary updates with the step sizes α^* and $2\alpha^*$. The comparison between them shows that the use of a larger than optimal step size improves the results.

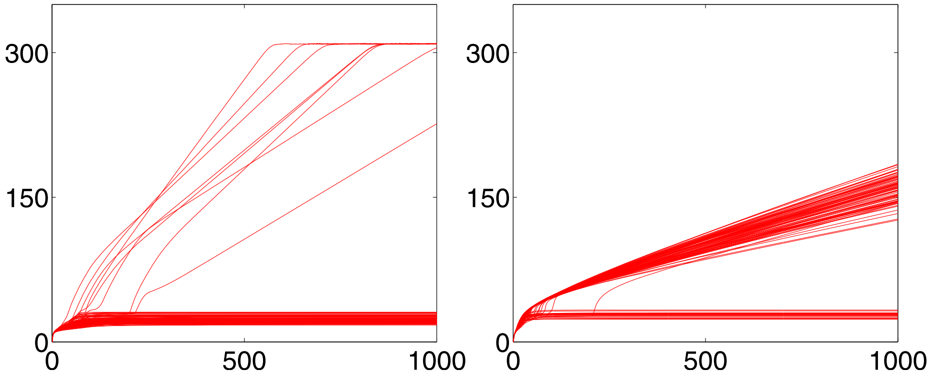
5.1 Learning with a Fixed Support

This experiment uses the same setup as the one presented in Section 3.2. We ran Sparsenet with the optimal step size α^* defined in Equation (21) and a larger step size $2\alpha^*$. As expected, the optimal step gradient descent almost always gets trapped in a local minimum. Doubling that step greatly improves the recovery rate from 8% to 79%.

5.2 Complete Learning

We also compared the different update rules in the context of a complete dictionary learning, i.e. without the use of an oracle support. The sparse decomposition step was performed using OMP.

Figure 3 shows the repartition of the SNR obtained by each algorithm. The different algorithms are sorted by increasing average SNR. For Sparsenet we used the step size $\alpha = 0.05$ which was well suited to the fixed support case. With that choice Sparsenet slightly outperforms K-SVD by 0.01 dB, but in practical cases one might not have access to such previous knowledge to finely tune the step size α. Our large step gradient achieved the best average SNR. It outperforms K-SVD and the fixed step Sparsenet by an average 0.5 dB and converged to a better solution than K-SVD in 98 cases over 100.



(a) Optimal step gradient descent, $\alpha = \alpha^*$ (b) Large step gradient descent, $\alpha = 2\alpha^*$

Fig. 2. Approximation SNR depending on the iteration. The optimal gradient descent only succeeds 8 times whereas using a $2\alpha^*$ step succeeds 79 times.

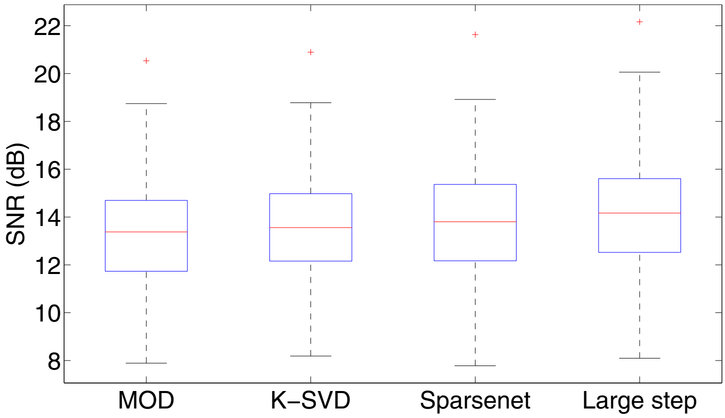


Fig. 3. Repartition of the SNR after learning dictionaries on 100 random data sets with different algorithms. The proposed large step gradient descent results in an average 0.5dB improvement over K-SVD.

6 Conclusion

We have presented a dictionary learning algorithm capable of better approximation quality of the training signals than K-SVD. That algorithm uses a gradient descent with an adaptive step guaranteed to be higher than the optimal step. The large step allows the descent to bypass local minima and converge towards the global minimum.

While our algorithm yields much better recovery rates than the existing ones, it can still be improved. With the step size $2\alpha^*$, the descent still gets trapped in a local minimum in 21% of the cases in our experiments. One could think of using an even larger step, but the algorithm then becomes unstable and fails to converge at all. The solution could be to use a hybrid algorithm that starts with large step gradient descent to find the attraction basin of a global minimum, then switches to one of the fast converging algorithms such as K-SVD to find the minimum itself.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on Signal Processing 54(11), 4311–4322 (2006)
2. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing 15(12), 3736–3745 (2006)
3. Engan, K., Aase, S., Hakon Husoy, J.: Method of optimal directions for frame design. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1999, vol. 5, pp. 2443–2446 (1999)
4. Mailhé, B., Lesage, S., Gribonval, R., Vandergheynst, P., Bimbot, F.: Shift-invariant dictionary learning for sparse representations: extending k-svd. In: Proc. EUSIPCO (2008)
5. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
6. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: 1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 40–44 (November 1993)
7. Tropp, J.: Greed is good: algorithmic results for sparse approximation. IEEE Transactions on Information Theory 50(10), 2231–2242 (2004)

Separation of Sparse Signals in Overdetermined Linear-Quadratic Mixtures

Leonardo T. Duarte[1], Rafael A. Ando[2,⋆], Romis Attux[2],
Yannick Deville[3], and Christian Jutten[4]

[1] School of Applied Sciences - University of Campinas (UNICAMP),
Campinas, Brazil
leonardo.duarte@fca.unicamp.br

[2] School of Electrical and Computer Engineering - University of Campinas (UNICAMP), Campinas, Brazil
{assato,attux}@dca.fee.unicamp.br

[3] IRAP, Université de Toulouse, CNRS, Toulouse, France
ydeville@ast.obs-mip.fr

[4] GIPSA-Lab, CNRS UMR-5216, Grenoble, and Institut Universitaire de France
christian.jutten@gipsa-lab.grenoble-inp.fr

Abstract. In this work, we deal with the problem of nonlinear blind source separation (BSS). We propose a new method for BSS in overdetermined linear-quadratic (LQ) mixtures. By exploiting the assumption that the sources are sparse in a transformed domain, we define a framework for canceling the nonlinear part of the mixing process. After that, separation can be conducted by linear BSS algorithms. Experiments with synthetic data are performed to assess the viability of our proposal.

Keywords: Nonlinear mixtures, sparse signals, blind source separation.

1 Introduction

In blind source separation (BSS), the goal is to retrieve a set of signals (sources) based only on the observation of mixed versions of these original sources [1,2]. Typically, the methods developed to solve this problem work with the assumption that the mixing process can be modeled as a linear system. However, while this framework has been proven successful in many applications, there are some practical examples in which the mixtures are clearly nonlinear — this is the case, for instance, in chemical sensor arrays [3] and hyperspectral imaging [4].

Several works have already pointed out some problems that arise when the mixtures are nonlinear (see [5] for a discussion). In particular, the application of methods based on independent component analysis (ICA) [6], which assumes that the sources are mutually statistically independent random variables, is not valid in a general nonlinear system. In view of this problem, the research on nonlinear BSS has been focused on constrained models, for which the task of

[⋆] R. Assato would like to thank FAPESP (Brazil) for funding his MSc studies.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 239–246, 2012.

source separation can be accomplished by extending the ideas already considered in the linear case. Among the constrained nonlinear models studied so far is the linear-quadratic (LQ) model [7,4]. This model provides a good description of the mixing process in applications such as show-through effect removal in scanned images [8] and design of gas sensor arrays [9]. Besides, the LQ model presents two interesting properties: 1) it can be seen as a first step toward more general polynomial mixtures; 2) it is linear with respect to the mixing coefficients.

When the number of sources is equal to the number of mixtures in an LQ model, the definition of a separating structure is not a simple task, given the difficulty in writing the inverse of the mixing process in an analytical form [7,10]. However, in an overdetermined case, in which the number of mixtures is greater than the number of sources, separation can be achieved by means of a linear structure. This idea has already been exploited in the context of sources belonging to a finite alphabet [11], circular sources [12], non-stationary sources [13] and independent sources [14].

In the present work, we tackle the problem of BSS in overdetermined mixtures assuming that the sources admit a sparse representation in a given basis. By using this property, we propose a strategy to cancel the nonlinear part of the LQ mixing process, so that the resulting problem can be dealt with by linear BSS algorithms. Since our approach for dealing with the nonlinear terms does not rely on the independence assumption, it is possible to tackle problems in which ICA methods fail. Of course, this can be done if the adopted linear BSS method is able to work with dependent mixtures.

2 Mixing Model

Let $\mathbf{s}_j = [s_j(1) \dots s_j(n_d)]^T$ represent j-th source (n_d is the number of samples). In the present work, we consider the case of $n_s = 2$ sources, which is representative in applications such as the design of gas sensor arrays and separation of scanned mixtures. In this case, the i-th mixture can be represented by the vector

$$\mathbf{x}_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 + a_{i3}\mathbf{s}_1 \circ \mathbf{s}_2, \tag{1}$$

where $\circ$ stands for the element-wise product operator (Hadamard product), while the mixing coefficients are denoted by a_{ij}.

An interesting aspect of (1) is that it can be interpreted as a linear mixing process in which the sources are given by $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_1 \circ \mathbf{s}_2$. Therefore, in an overdetermined case, with three mixtures given by

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} s_1(n) \\ s_2(n) \\ s_1(n)s_2(n) \end{bmatrix}, \ \forall n \in \{1, \dots, n_d\}, \tag{2}$$

it is possible to achieve source separation by means of a linear separating system, in which the recovered sources are given by

$$\begin{bmatrix} y_1(n) \\ y_2(n) \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix}, \ \forall n \in \{1, \dots, n_d\}. \tag{3}$$

As will be discussed in the sequel, source separation in this case can be performed by firstly canceling the nonlinear elements of the mixtures, followed by the application of a linear BSS method.

3 Sparsity-Based Cancellation of Quadratic Terms

3.1 The Main Idea

Since we have access to, at least, three mixtures, it is possible to linearly combine them in order to extract the quadratic term expressed in (2). Let us consider the linear combination between the mixtures $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\mathbf{z}_{ij} = \mathbf{x}_i - \alpha_{ij}\mathbf{x}_j, \tag{4}$$

where the index ij corresponds to the mixtures considered in the combination. According to (1), $\mathbf{z}_{ij}$ can be rewritten as follows

$$\mathbf{z}_{ij} = (a_{i1} - \alpha_{ij}a_{j1})\mathbf{s}_1 + (a_{i2} - \alpha_{ij}a_{j2})\mathbf{s}_2 + (a_{i3} - \alpha_{ij}a_{j3})\mathbf{s}_1 \circ \mathbf{s}_2 \tag{5}$$

Therefore, when $\alpha_{ij} = a_{i3}/a_{j3}$, $\mathbf{z}_{ij}$ becomes a linear mixture of $\mathbf{s}_1$ and $\mathbf{s}_2$.

The implementation of the idea described above requires the definition of a criterion to guide the estimation of α_{ij}. A possible idea to accomplish this task can be formulated by rewriting (5) in a transformed domain, which is achieved by multiplying $\mathbf{z}_{ij}$ by the $n_d \times n_d$ orthonormal matrix $\boldsymbol{\Phi}$ that represents such a transformation. In mathematical terms,

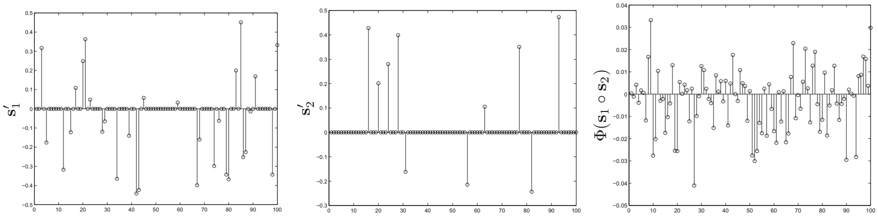
$$\begin{aligned}\mathbf{z}'_{ij} = \boldsymbol{\Phi}\mathbf{z}_{ij} &= (a_{i1} - \alpha_{ij}a_{j1})\boldsymbol{\Phi}\mathbf{s}_1 + (a_{i2} - \alpha_{ij}a_{j2})\boldsymbol{\Phi}\mathbf{s}_2 + (a_{i3} - \alpha_{ij}a_{j3})\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2) \\ &= (a_{i1} - \alpha_{ij}a_{j1})\mathbf{s}'_1 + (a_{i2} - \alpha_{ij}a_{j2})\mathbf{s}'_2 + (a_{i3} - \alpha_{ij}a_{j3})\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2),\end{aligned} \tag{6}$$

where $\mathbf{s}'_i$ is the representation in the transformed domain of the source $\mathbf{s}_i$.

Let us consider for instance that the orthonormal matrix $\boldsymbol{\Phi}$ is related to a frequency transformation, e.g. the discrete cosine transform (DCT). Moreover, let us assume that both $\mathbf{s}'_1$ and $\mathbf{s}'_2$ are sparse vectors, in the sense that not all frequency components of these signals are not null. Note that the term $\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)$ in (6) is related to the convolution of $\mathbf{s}'_1$ and $\mathbf{s}'_2$, since it is given by the DCT transform of a product in time. The key point here is that the convolution of $\mathbf{s}'_1$ and $\mathbf{s}'_2$ tends to produce a signal that is not sparse, or at least less sparse than $\mathbf{s}'_1$ and $\mathbf{s}'_2$, as the nonlinear term causes a spreading in the frequency domain — this feature is illustrated in Figure 1. Based on this observation, our idea is to adjust α_{ij} by maximizing the degree of sparsity of $\mathbf{z}'_{ij}$. If the ℓ_0-norm, which corresponds to the number of non-zero elements of a vector, is adopted as a measure of sparsity [15], then our idea can be formulated as the following optimization problem

$$\min_{\alpha_{ij}} ||\boldsymbol{\Phi}\mathbf{z}_{ij}||_0. \tag{7}$$

It is worth mentioning that the matrix $\boldsymbol{\Phi}$ should not necessarily be related with a frequency transform. The only requirement is that $\boldsymbol{\Phi}$ somehow spreads the representation of $(\mathbf{s}_1 \circ \mathbf{s}_2)$. This point will be further discussed in the sequel.



(a) $\mathbf{s}'_1$. (b) $\mathbf{s}'_2$. (c) $\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)$.

Fig. 1. DCTs of the sources and of the product between these sources

3.2 Theoretical Aspects

We here discuss some theoretical aspects related to the idea expressed in (7). In particular, we provide the guidelines for establishing general conditions for which our proposal is valid. In our analysis, we assume that the mixing matrix is full rank. Moreover, we assume that $||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 \geq \max(||\mathbf{s}'_1||_0, ||\mathbf{s}'_2||_0)$. As our analysis is based on the ℓ_0-norm, it is important to introduce some properties of this measure, which strictly speaking is not a mathematical norm [15]. Yet, the ℓ_0-norm satisfies the triangle inequality, that is, given two vectors $\mathbf{a}$ and $\mathbf{b}$, then $||\mathbf{a}+\mathbf{b}||_0 \leq ||\mathbf{a}||_0 + ||\mathbf{b}||_0$. As a consequence, the ℓ_0-norm also satisfies the reverse triangle inequality, i.e., $||\mathbf{a}-\mathbf{b}||_0 \geq \Big| ||\mathbf{a}||_0 - ||\mathbf{b}||_0 \Big|$. Finally, the ℓ_0-norm is scale invariant, i.e., $||k\mathbf{a}||_0 = ||\mathbf{a}||_0$ for $k \neq 0$.

In order to investigate the cost function $||\boldsymbol{\Phi}\mathbf{z}_{ij}||_0$, let us rewrite (6) as follows:

$$\mathbf{z}'_{ij} = a\mathbf{s}'_1 + b\mathbf{s}'_2 + c\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2). \tag{8}$$

Ideally, to be in accordance with our idea, $||\mathbf{z}'_{ij}||_0$ should attain a minimum if, and only if, $\alpha_{ij} = a_{i3}/a_{j3}$, that is, when $c = 0$. In this case, one only has the linear terms of the mixture, and, thus, by considering the triangle inequality and the scaling invariance property, it turns out that

$$||\mathbf{z}'_{ij}||_0 \leq ||\mathbf{s}'_1||_0 + ||\mathbf{s}'_2||_0. \tag{9}$$

We can also investigate $||\mathbf{z}'_{ij}||_0$ in the cases in which α_{ij} does not lead to the cancellation of the quadratic term, i.e. when $c \neq 0$ in (8). In these situations, our idea will work when $||\mathbf{z}'_{ij}||_0$ is greater than the upper bound (9). When $a = 0$, $b \neq 0$, and $c \neq 0$, one can use the reverse triangle inequality to obtain the following lower bound

$$||\mathbf{z}'_{ij}||_0 = ||b\mathbf{s}'_2 + c\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 \geq ||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 - ||\mathbf{s}'_2||_0. \tag{10}$$

Analogously, when $a \neq 0$, $b = 0$, and $c \neq 0$, one can easily show that

$$||\mathbf{z}'_{ij}||_0 \geq ||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 - ||\mathbf{s}'_1||_0. \tag{11}$$

Finally, when $a \neq 0$, $b \neq 0$, and $c \neq 0$, the following lower bound for $||\mathbf{z}'_{ij}||_0$ can be obtained after the application of the triangle inequality followed by the application of the reversed triangle inequality

$$||\mathbf{z}'_{ij}||_0 \geq ||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 - ||\mathbf{s}'_1||_0 - ||\mathbf{s}'_2||_0. \tag{12}$$

Among the bounds expressed in (10), (11) and (12), and the bound $||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0$ obtained when $a = 0$, $b = 0$, $c \neq 0$, the bound shown in (12) is the smallest one. Therefore, if the lower bound (12) is greater than the higher bound (9), then $||\boldsymbol{\Phi}\mathbf{z}_{ij}||_0$ will necessarily reach the global minimum at $\alpha_{ij} = a_{i3}/a_{j3}$ (i.e. $c = 0$). This observation leads to the following sufficient condition:

$$||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 - ||\mathbf{s}'_1||_0 - ||\mathbf{s}'_2||_0 > ||\mathbf{s}'_1||_0 + ||\mathbf{s}'_2||_0, \tag{13}$$

i.e.

$$||\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)||_0 > 2\left(||\mathbf{s}'_1||_0 + ||\mathbf{s}'_2||_0\right) \tag{14}$$

Some observations can be made on this condition. Firstly, the importance of defining a proper transformation, for which the representation of the quadratic terms is as spread as possible, becomes clear. Note that such a requirement becomes less stringent when the sources have a high degree of sparsity in the transformed domain. For instance, in the situation illustrated in Figure 1, it is clear that the DCT satisfies the sufficient condition expressed in (14). Moreover, when $\boldsymbol{\Phi}$ is given by a random matrix submitted to an orthogonalization process, condition (14) can be satisfied for many different configurations of the sources.

A second point related to (14) is that it provides a sufficient but not necessary condition. Actually, this condition is quite pessimistic, as it considers a very peculiar configuration of the positions for which the signals $\mathbf{s}'_1$ and $\mathbf{s}'_2$ take non-zero values.

3.3 Implementation Issues

In a practical application, the use of the ℓ_0-norm is quite limited, since sparse signals in practice have many elements that are close to zero, but that are not necessarily null. Thus, approximations of the ℓ_0-norm must be considered. A possible choice is the smoothed version of ℓ_0-norm [16], which, for a given signal $\mathbf{y}$, is defined as follows:

$$S_{\ell_0}(\mathbf{y}) = n_d - \sum_{i=1}^{n_d} f(y_{(i)}, \sigma), \tag{15}$$

where $f(\cdot, \sigma)$ corresponds to a zero mean Gaussian kernel of standard deviation σ. As σ approaches to zero, (15) approaches to the ℓ_0-norm. Ideally, the choice of σ depends on how close to zero the low-energy elements of a given signal are.

We can now introduce a general scheme for source separation in overdetermined LQ mixtures of two sources. The proposal, which can be applied when the number of mixtures is greater than 2, is composed of the following steps: 1) *Cancellation of quadratic terms*: considering $n_p > 1$ pairs of mixtures, $\mathbf{x}_i$ and

$\mathbf{x}_j$, find, for each of these pairs, an α_{ij} which minimizes $S_{\ell_0}(\mathbf{z}'_{ij})$. This procedure will provide the set of signals that, ideally, correspond to linear mixtures of the sources. 2) *Source separation*: apply a linear source separation (or extraction) method on the signals obtained in the first stage.

The first stage of the proposed strategy boils down to n_p univariate optimization problems, which, in our work, are carried out by an exhaustive search approach. Note that the first stage can be conducted even when the sources are not statistically independent. Of course, in this case, the second stage should be able to deal with linear mixtures of dependent sources.

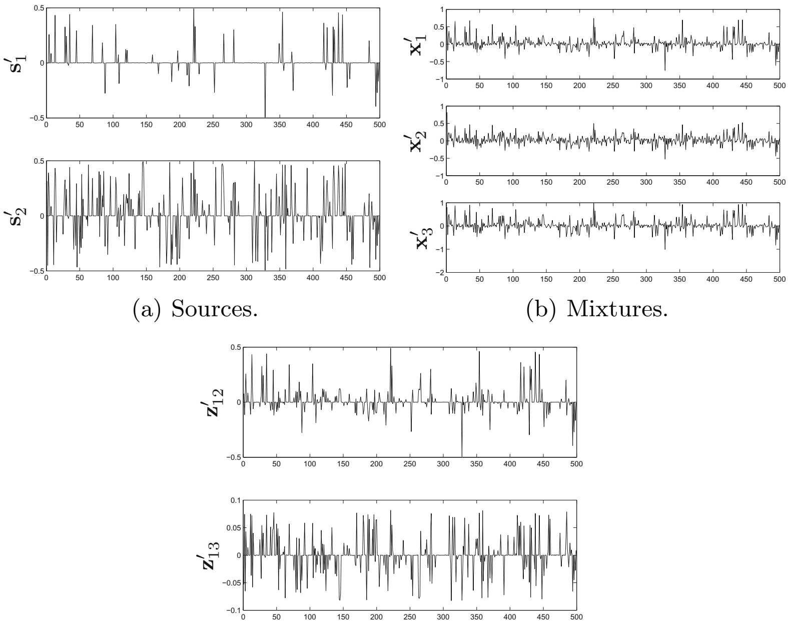
4 Results

Let us consider an overdetermined LQ source separation problem in which the mixing matrix is given by $\mathbf{A} = [1\ 0.5\ 2; 0.5\ 1\ 4; 1\ 1\ 3]$ (see formulation expressed in (2)). The sources here are sparse in the DCT domain. To generate the DCT coefficients, we firstly obtained 500 samples from a distribution uniformly distributed in $[-0.5, 0.5]$. Then, we replaced a given percentage of the generated elements by samples obtained from a zero-mean Gaussian distribution of standard deviation 0.001 (these are the low-energy DCT coefficients). The percentages of these low-energy elements were 90% for the first source and 70% for the second source, and their position were randomly selected. Finally, the sources shared 50 DCT coefficients (these coefficiecients are not the ones with small values), which make them statistically dependent.

In order to remove the quadratic terms in the mixtures, we applied the proposed method to the pairs of mixtures $(\mathbf{x}_1, \mathbf{x}_2)$ and $(\mathbf{x}_1, \mathbf{x}_3)$. After performing 10 runs, each one considering a different set of sources generated according to the procedure described above, our method provided very good solutions in every runs. Indeed, the obtained mean values were $\alpha_{12} = 0.5012$ and $\alpha_{13} = 0.6671$, which are very close to the ideal values $\alpha_{12} = 1/2$ and $\alpha_{13} = 2/3$. To illustrate that the nonlinear terms were removed in this situation, we plot in Figure 2 the DCT coefficients of the sources, mixtures and the provided signals $\mathbf{z}_{12}$ and $\mathbf{z}_{13}$ obtained in a given run. Note that the DCT coefficients of the obtained signals are clearly sparser than those of the mixtures.

With the obtained linear mixtures at hand, we applied the source extraction method proposed in [17] to retrieve the sparsest component. As discussed in [17], this method is able to conduct source separation even when the sources are dependent. Indeed, the sparsest source was estimated with a signal-to-interference ratio[1] (SIR) of 58.7dB. On the other hand, the SIRs obtained after the application of the ICA-based solution proposed in [14] were 4.0dB (first source) and 7.1dB (second source). These low values can be attributed to the fact that the sources were not independent in the considered scenario, thus violating the central assumption of ICA methods.

[1] The SIR is defined as $= 10 \log \left(E\{\hat{y}_i^2\} / E\{(\hat{s}_i - \hat{y}_i)^2\} \right)$, where $\hat{s}_i$ and $\hat{y}_i$ are, respectively, the actual source and its estimate, being both ones obtained after mean, variance and sign normalization.



(a) Sources.

(b) Mixtures.

(c) $\boldsymbol{\Phi}(\mathbf{s}_1 \circ \mathbf{s}_2)$.

Fig. 2. Obtained linear mixtures

5 Conclusions

We proposed a method for suppressing the quadratic terms of overdetermined LQ mixtures. Our approach works with the assumption that the sources are sparse when represented in a proper domain, which should be known in advance, and is based on a ℓ_0-norm minimization procedure. We provided theoretical elements that points out that our proposal is suitable for the cases in which the quadratic terms admit a representation in the considered domain that is less sparse than those of the sources. A numerical experiment illustrated the effectiveness of the obtained method, especially when the sources are dependent.

There are several points to be investigated in future works. For instance, a first one is to extend the theoretical analysis conducted in this paper to the case of the smoothed ℓ_0-norm, paying special attention to the influence of the parameter σ. Another relevant point is to investigate if the two-stage procedure described in Section 3.3 can be merged into a unique step guided by the minimization of the sparsity of the retrieved sources. Finally, we intent to investigate the extension of the idea to the case in which the number of sources is greater than two, and its application to actual problems.

References

1. Comon, P., Jutten, C. (eds.): Handbook of blind source separation, independent component analysis and applications. Academic Press, Elsevier (2010)
2. Romano, J.M.T., Attux, R.R.F., Cavalcante, C.C., Suyama, R.: Unsupervised signal processing: channel equalization and source separation. CRC Press (2011)
3. Duarte, L.T., Jutten, C., Moussaoui, S.: A Bayesian nonlinear source separation method for smart ion-selective electrode arrays. IEEE Sensors Journal 9(12), 1763–1771 (2009)
4. Meganem, I., Deville, Y., Hosseini, S., Déliot, P., Briottet, X., Duarte, L.T.: Linear-quadratic and polynomial non-negative matrix factorization; application to spectral unmixing. In: Proc. of the 19th European Signal Processing Conference, EUSIPCO 2011 (2011)
5. Jutten, C., Karhunen, J.: Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. International Journal of Neural Systems 14, 267–292 (2004)
6. Comon, P.: Independent component analysis, a new concept? Signal Processing 36, 287–314 (1994)
7. Hosseini, S., Deville, Y.: Blind Separation of Linear-Quadratic Mixtures of Real Sources Using a Recurrent Structure. In: Mira, J., Álvarez, J.R. (eds.) IWANN 2003. LNCS, vol. 2687, pp. 241–248. Springer, Heidelberg (2003)
8. Merrikh-Bayat, F., Babaie-Zadeh, M., Jutten, C.: Linear-quadratic blind source separating structure for removing show-through in scanned documents. International Journal on Document Analysis and Recognition, 1–15 (2010)
9. Bedoya, G.: Nonlinear blind signal separation for chemical solid-state sensor arrays. PhD thesis, Universitat Politecnica de Catalunya (2006)
10. Deville, Y., Hosseini, S.: Recurrent networks for separating extractable-target nonlinear mixtures. part i: Non-blind configurations. Signal Processing 89, 378–393 (2009)
11. Castella, M.: Inversion of polynomial systems and separation of nonlinear mixtures of finite-alphabet sources. IEEE Trans. on Sig. Proc. 56(8), 3905–3917 (2008)
12. Abed-Meraim, K., Belouchrani, A., Hua, Y.: Blind identification of a linear-quadratic mixture of independent components based on joint diagonalization procedure. In: Proc. of the IEEE Inter. Conf. on Acous., Spee., and Signal Processing, ICASSP (1996)
13. Deville, Y., Hosseini, S.: Blind identification and separation methods for linear-quadratic mixtures and/or linearly independent non-stationary signals. In: Proc. of the 9th Int. Symp. on Sig. Proc. and its App., ISSPA (2007)
14. Duarte, L.T., Suyama, R., Attux, R., Deville, Y., Romano, J.M.T., Jutten, C.: Blind Source Separation of Overdetermined Linear-Quadratic Mixtures. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 263–270. Springer, Heidelberg (2010)
15. Elad, M.: Sparse and redundant representations from theory to applications in signal and image processing. Springer, Heidelberg (2010)
16. Mohimani, H., Babaie-Zadeh, M., Jutten, C.: A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. IEEE Transactions on Signal Processing 57(1), 289–301 (2009)
17. Duarte, L.T., Suyama, R., Attux, R., Romano, J.M.T., Jutten, C.: Blind extraction of sparse components based on ℓ_0-norm minimization. In: Proc. of the IEEE Statistical Signal Processing Workshop, SSP (2011)

Collaborative Filtering via Group-Structured Dictionary Learning

Zoltán Szabó[1], Barnabás Póczos[2], and András Lőrincz[1]

[1] Faculty of Informatics, Eötvös Loránd University,
Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary
szzoli@cs.elte.hu, andras.lorincz@elte.hu
[2] Carnegie Mellon University, Robotics Institute,
5000 Forbes Ave, Pittsburgh, PA 15213
bapoczos@cs.cmu.edu

Abstract. Structured sparse coding and the related structured dictionary learning problems are novel research areas in machine learning. In this paper we present a new application of structured dictionary learning for collaborative filtering based recommender systems. Our extensive numerical experiments demonstrate that the presented method outperforms its state-of-the-art competitors and has several advantages over approaches that do not put structured constraints on the dictionary elements.

Keywords: collaborative filtering, structured dictionary learning.

1 Introduction

The proliferation of online services and the thriving electronic commerce overwhelms us with alternatives in our daily lives. To handle this information overload and to help users in efficient decision making, recommender systems (RS) have been designed. The goal of RSs is to recommend personalized items for online users when they need to choose among several items. Typical problems include recommendations for which movie to watch, which jokes/books/news to read, which hotel to stay at, or which songs to listen to.

One of the most popular approaches in the field of recommender systems is *collaborative filtering* (CF). The underlying idea of CF is very simple: Users generally express their tastes in an explicit way by rating the items. CF tries to estimate the users' preferences based on the ratings they have already made on items and based on the ratings of other, similar users. For a recent review on recommender systems and collaborative filtering, see e.g., [1].

Novel advances on CF show that *dictionary learning* based approaches can be efficient for making predictions about users' preferences [2]. The dictionary learning based approach assumes that (i) there is a latent, unstructured feature space (hidden representation/code) behind the users' ratings, and (ii) a rating of an item is equal to the product of the item and the user's feature. To increase the generalization capability, usually ℓ_2 regularization is introduced both for the dictionary and for the users' representation.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 247–254, 2012.

Recently it has been shown both theoretically and via numerous applications (e.g., automatic image annotation, feature selection for microarray data, multi-task learning, multiple kernel learning, face recognition, structure learning in graphical models) that it can be advantageous to force different kind of structures (e.g., disjunct groups, trees) on the hidden representation. This regularization approach is called *structured sparsity* [3]. The structured sparse coding problem assumes that the dictionary is already given. A more interesting (and challenging) problem is the combination of these tasks, i.e., learning the best structured dictionary and structured representation. This is the *structured dictionary learning* (SDL) problem. SDL is more difficult than structured sparse coding; one can only find few results in the literature [4–8]. This novel field is appealing for (i) transformation invariant feature extraction [8], (ii) image denoising/inpainting [4,6], (iii) background subtraction [6], (iv) analysis of text corpora [4], and (v) face recognition [5].

Several successful applications show the importance of the SDL problem family. Interestingly, however, to the best of our knowledge, it has not been used for the collaborative filtering problem yet. The *goal of our paper* is to extend the application domain of SDL to CF. In CF further constraints appear for SDL since (i) online learning is desired, and (ii) missing information is typical. There are good reasons for them: novel items/users may appear and user preferences may change over time. Adaptation to users also motivate online methods. Online methods have the additional advantage with respect to offline ones that they can process more instances in the same amount of time, and in many cases this can lead to increased performance. For a theoretical proof of this claim, see [9]. Usually users can evaluate only a small portion of the available items, which leads to incomplete observations, missing rating values. In order to cope with these constraints of the collaborative filtering problem, we will use a novel extension of the structured dictionary learning problem, the so-called online group-structured dictionary learning (OSDL) [10]. OSDL allows (i) overlapping group structures with (ii) non-convex sparsity inducing regularization, (iii) partial observation (iv) in an online framework.

Our paper is structured as follows: We briefly review the OSDL technique in Section 2. We cast the CF problem as an OSDL task in Section 3. Numerical results are presented in Section 4. Conclusions are drawn in Section 5.

Notations. Vectors have bold faces ($\mathbf{a}$), matrices are written by capital letters ($\mathbf{A}$). For a set, $|\cdot|$ denotes the number of elements in the set. For set $O \subseteq \{1,\ldots,d\}$, $\mathbf{a}_O \in \mathbb{R}^{|O|}$ ($\mathbf{A}_O \in \mathbb{R}^{|O| \times D}$) denotes the coordinates (columns) of vector $\mathbf{a} \in \mathbb{R}^d$ (matrix $\mathbf{A} \in \mathbb{R}^{d \times D}$) in O. The ℓ_p (quasi-) norm of vector $\mathbf{a} \in \mathbb{R}^d$ is $\|\mathbf{a}\|_p = (\sum_{i=1}^d |a_i|^p)^{\frac{1}{p}}$ ($p > 0$). $S_p^d = \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_p \leq 1\}$ denotes the ℓ_p unit sphere in $\mathbb{R}^d$. The point-wise product of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ is $\mathbf{a} \circ \mathbf{b} = [a_1 b_1; \ldots; a_d b_d]$. For a set system[1] $\mathcal{G}$, the coordinates of vector $\mathbf{a} \in \mathbb{R}^{|\mathcal{G}|}$ are denoted by a^G ($G \in \mathcal{G}$), that is, $\mathbf{a} = (a^G)_{G \in \mathcal{G}}$.

[1] A set system is also called hypergraph or a family of sets.

2 The OSDL Problem

In this section we formally define the online group-structured dictionary learning problem (OSDL). Let the dimension of the observations be denoted by d_x. Assume that in each time instant ($i = 1, 2, \ldots$) a set $O_i \subseteq \{1, \ldots, d_x\}$ is given, that is, we know which coordinates are observable at time i, and the observation is $\mathbf{x}_{O_i}$. Our goal is to find a dictionary $\mathbf{D} \in \mathbb{R}^{d_x \times d_\alpha}$ that can accurately approximate the observations $\mathbf{x}_{O_i}$ from the linear combinations of the columns of $\mathbf{D}$. These column vectors are assumed to belong to a closed, convex, and bounded set $\boldsymbol{\mathfrak{D}} = \times_{i=1}^{d_\alpha} \boldsymbol{\mathfrak{D}}_i$. To formulate the cost of dictionary $\mathbf{D}$, first a *fixed* time instant i, observation $\mathbf{x}_{O_i}$, and dictionary $\mathbf{D}$ are considered, and the hidden representation $\boldsymbol{\alpha}_i$ associated to this $(\mathbf{x}_{O_i}, \mathbf{D}, O_i)$ triple is defined. Representation $\boldsymbol{\alpha}_i$ is allowed to belong to a closed, convex set $\boldsymbol{\mathcal{A}} \subseteq \mathbb{R}^{d_\alpha}$ ($\boldsymbol{\alpha}_i \in \boldsymbol{\mathcal{A}}$) with certain structural constraints. The structural constraints on $\boldsymbol{\alpha}_i$ are expressed by making use of a given $\mathcal{G}$ group structure, which is a set system on $\{1, \ldots, d_\alpha\}$. Representation $\boldsymbol{\alpha}$ belonging to a triple $(\mathbf{x}_O, \mathbf{D}, O)$ is defined as the solution of the structured sparse coding task

$$l(\mathbf{x}_O, \mathbf{D}_O) = \min_{\boldsymbol{\alpha} \in \boldsymbol{\mathcal{A}}} \left[\frac{1}{2} \left\| \mathbf{x}_O - \mathbf{D}_O \boldsymbol{\alpha} \right\|_2^2 + \kappa \Omega(\boldsymbol{\alpha}) \right], \tag{1}$$

where $l(\mathbf{x}_O, \mathbf{D}_O)$ denotes the loss, $\kappa > 0$, and $\Omega(\mathbf{y}) = \|(\|\mathbf{y}_G\|_2)_{G \in \mathcal{G}}\|_\eta$ is the structured regularizer associated to $\mathcal{G}$ and $\eta \in (0, 1]$. Here, the first term of (1) is responsible for the quality of approximation on the observed coordinates. The second term constrains the solution according to the group structure $\mathcal{G}$ similarly to the sparsity inducing regularizer Ω in [5], i.e., it eliminates the terms $\|\mathbf{y}_G\|_2$ ($G \in \mathcal{G}$) by means of $\|\cdot\|_\eta$. The OSDL problem is defined as the minimization of the cost function:

$$\min_{\mathbf{D} \in \boldsymbol{\mathfrak{D}}} f_t(\mathbf{D}) := \frac{1}{\sum_{j=1}^{t} (j/t)^\rho} \sum_{i=1}^{t} \left(\frac{i}{t} \right)^\rho l(\mathbf{x}_{O_i}, \mathbf{D}_{O_i}). \tag{2}$$

Here the goal is to minimize the average loss belonging to the dictionary, where ρ is a non-negative forgetting factor. If $\rho = 0$, we get the classical average.

As an example, let $\boldsymbol{\mathfrak{D}}_i = S_2^{d_x}$ ($\forall i$), $\boldsymbol{\mathcal{A}} = \mathbb{R}^{d_\alpha}$. In this case, columns of $\mathbf{D}$ are restricted to the Euclidean unit sphere and we have no constraints for $\boldsymbol{\alpha}$. Now, let $|\mathcal{G}| = d_\alpha$ and $\mathcal{G} = \{desc_1, \ldots, desc_{d_\alpha}\}$, where $desc_i$ represents the i^{th} node and its children in a fixed tree. Then the coordinates $\{\alpha_i\}$ are searched in a hierarchical tree structure and the hierarchical dictionary $\mathbf{D}$ is optimized accordingly.

Optimization of cost function (2) is equivalent to the joint optimization:

$$\mathop{\arg\min}_{\mathbf{D} \in \boldsymbol{\mathfrak{D}}, \{\boldsymbol{\alpha}_i \in \boldsymbol{\mathcal{A}}\}_{i=1}^t} f_t(\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^t) = \frac{1}{\sum_{j=1}^{t} (j/t)^\rho} \sum_{i=1}^{t} \left(\frac{i}{t} \right)^\rho \left[\frac{1}{2} \left\| \mathbf{x}_{O_i} - \mathbf{D}_{O_i} \boldsymbol{\alpha}_i \right\|_2^2 + \kappa \Omega(\boldsymbol{\alpha}_i) \right].$$

By using the sequential observations $\mathbf{x}_{O_i}$, one can optimize $\mathbf{D}$ online in an alternating manner: The actual dictionary estimation $\mathbf{D}_{t-1}$ and sample $\mathbf{x}_{O_t}$ are used to optimize (1) for representation $\boldsymbol{\alpha}_t$. After this step, when the estimated representations $\{\boldsymbol{\alpha}_i\}_{i=1}^t$ are given, the dictionary estimation $\mathbf{D}_t$ is derived from the quadratic optimization problem

$$\hat{f}_t(\mathbf{D}_t) = \min_{\mathbf{D}\in\mathfrak{D}} f_t(\mathbf{D}, \{\boldsymbol{\alpha}_i\}_{i=1}^t). \tag{3}$$

These optimization problems can be tackled by making use of the variational property [5] of norm η and using the block-coordinate descent method, which leads to matrix recursions [10].[2]

3 OSDL Based Collaborative Filtering

Below, we transform the CF task into an OSDL problem. Consider the t^{th} user's known ratings as OSDL observations $\mathbf{x}_{O_t}$. Let the optimized group-structured dictionary on these observations be $\mathbf{D}$. Now, assume that we have a test user and his/her ratings, i.e., $\mathbf{x}_O \in \mathbb{R}^{|O|}$. The task is to estimate $\mathbf{x}_{\{1,\ldots,d_x\}\backslash O}$, that is, the missing coordinates of $\mathbf{x}$ (the missing ratings of the user). This can be accomplished by the following steps (Table 1).

Table 1. Solving CF with OSDL

1. Remove the rows of the non-observed $\{1,\ldots,d_x\}\backslash O$ coordinates from $\mathbf{D}$. The obtained $|O| \times d_\alpha$ sized matrix $\mathbf{D}_O$ and $\mathbf{x}_O$ can be used to estimate $\boldsymbol{\alpha}$ by solving the structured sparse coding problem (1).
2. Using the estimated representation $\boldsymbol{\alpha}$, estimate $\mathbf{x}$ as $\hat{\mathbf{x}} = \mathbf{D}\boldsymbol{\alpha}$.

According to the CF literature, *neighbor based correction* schemes may further improve the quality of the estimations [1]. This neighbor correction approach relies on the assumption that similar items (e.g., jokes/movies) are rated similarly. As we will show below, these schemes can be adapted to OSDL-based CF estimation too. Assume that the similarities $s_{ij} \in \mathbb{R}$ $(i, j \in \{1,\ldots,d_x\})$ between individual items are given. We shall provide similarity forms in Section 4. Let $\mathbf{d}_k\boldsymbol{\alpha}_t \in \mathbb{R}$ be the OSDL estimation for the rating of the k^{th} non-observed item of the t^{th} user $(k \notin O_t)$, where $\mathbf{d}_k \in \mathbb{R}^{1\times d_\alpha}$ is the k^{th} row of matrix $\mathbf{D} \in \mathbb{R}^{d_x\times d_\alpha}$, and $\boldsymbol{\alpha}_t \in \mathbb{R}^{d_\alpha}$ is computed as described in Table 1. Let the prediction error on the observable item neighbors (j) of the k^{th} item of the t^{th} user $(j \in O_t\backslash\{k\})$ be $\mathbf{d}_j\boldsymbol{\alpha}_t - x_{jt} \in \mathbb{R}$. These prediction errors can be used for the correction of the OSDL estimation $(\mathbf{d}_k\boldsymbol{\alpha}_t)$ by taking into account the s_{kj} similarities:

$$\hat{x}_{kt} = \gamma_0(\mathbf{d}_k\boldsymbol{\alpha}_t) + \gamma_1\left[\frac{\sum_{j\in O_t\backslash\{k\}} s_{kj}(\mathbf{d}_j\boldsymbol{\alpha}_t - x_{jt})}{\sum_{j\in O_t\backslash\{k\}} s_{kj}}\right], \tag{4}$$

where $\gamma_0, \gamma_1 \in \mathbb{R}$ are weight parameters, and $k \notin O_t$. Equation (4) is a simple modification of the corresponding expression in [2]. It modulates the first term with a separate γ_0 weight, which we found beneficial in our experiments.

[2] The Matlab code of the method is available at http://nipg.inf.elte.hu/szzoli.

4 Numerical Results

We have chosen the Jester dataset [11] for the illustration of the OSDL based CF approach. It is a standard benchmark dataset for CF. It contains $4,136,360$ ratings from $73,421$ users on 100 jokes. The ratings are in the continuous $[-10,10]$ range. The worst and best possible grades are -10 and $+10$, respectively. A fixed 10 element subset of the jokes is called gauge set, and it was evaluated by all users. Two third of the users have rated at least 36 jokes, and the remaining ones have rated between 15 and 35 jokes. The average number of user ratings per joke is 46.

In the neighbor correction step (4), we need the s_{ij} values, which represent the similarities of the i^{th} and j^{th} items. We define this $s_{ij} = s_{ij}(\mathbf{d}_i, \mathbf{d}_j)$ value as the similarity between the i^{th} and j^{th} rows of the optimized OSDL dictionary $\mathbf{D}$. We made experiments with the following two similarities (S_1, S_2):

$$S_1: \quad s_{ij} = \left(\frac{\max(0, \mathbf{d}_i\mathbf{d}_j^T)}{\|\mathbf{d}_i\|_2 \|\mathbf{d}_j\|_2}\right)^{\beta}, \text{ and } S_2: \quad s_{ij} = \left(\frac{\|\mathbf{d}_i - \mathbf{d}_j\|_2^2}{\|\mathbf{d}_i\|_2 \|\mathbf{d}_j\|_2}\right)^{-\beta}. \qquad (5)$$

Here $\beta > 0$ is the parameter of the similarity measure [2]. Quantities s_{ij} are non-negative. If the value of s_{ij} is close to zero (large), then the i^{th} and j^{th} items are very different (very similar).

In our numerical experiments we used the RMSE (root mean square error) measure for the evaluation of the quality of the estimation, since this is the most popular measure in the CF literature. The RMSE is the average squared difference of the true and the estimated rating values:

$$RMSE = \sqrt{\frac{1}{|\mathcal{S}|} \sum_{(i,t)\in\mathcal{S}} (x_{it} - \hat{x}_{it})^2}, \qquad (6)$$

where $\mathcal{S}$ denotes either the validation or the test set. We also performed experiments using the mean absolute error (MAE) and got very similar results.

4.1 Evaluation

We illustrate the efficiency of the OSDL-based CF estimation on the Jester dataset using the RMSE performance measure. To the best of our knowledge, the top results on this database are RMSE = 4.1123 [12] and RMSE = 4.1229 [2]. The method in the first paper is called *item neighbor*, and it makes use of neighbor information only. In [2], the authors used a bridge regression based unstructured dictionary learning model with a neighbor correction scheme. They optimized the dictionary by gradient descent and set d_α to 100.

To study the capability of the OSDL approach in CF, we focused on the following questions:

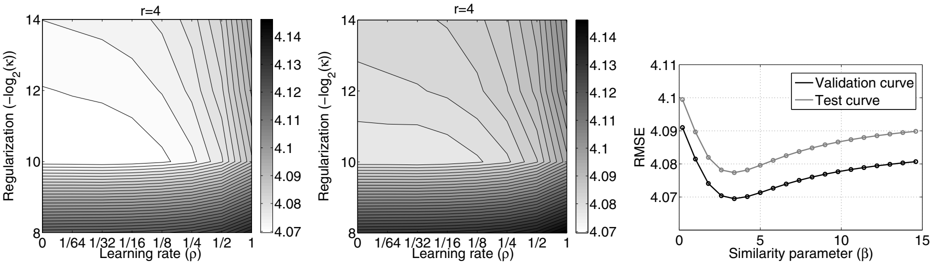
- Is structured dictionary $\mathbf{D}$ beneficial for prediction purposes, and how does it compare to the dictionary of classical (unstructured) sparse dictionary?
- How does the OSDL parameters and the similarity applied affect the efficiency of the prediction?
- How do different group structures $\mathcal{G}$ fit to the CF task?

In our numerical studies we chose the Euclidean unit sphere for $\mathfrak{D}_i = S_2^{d_x}$ $(\forall i)$ and $\boldsymbol{\mathcal{A}} = \mathbb{R}^{d_\alpha}$. We set η of the structure inducing regularizer Ω to 0.5. Group structure $\mathcal{G}$ was realized (i) either on a $\sqrt{d_\alpha} \times \sqrt{d_\alpha}$ toroid with $|\mathcal{G}| = d_\alpha$ applying $r \geq 0$ neighbors to define $\mathcal{G}$,[3] or (ii) on a hierarchy with a complete binary tree structure parameterized by the number of levels l ($|\mathcal{G}| = d_\alpha$, $d_\alpha = 2^l - 1$). The forgetting factor (ρ), the weight of Ω (κ), the size of the mini-batches in $\mathbf{D}$ optimization (R), and the parameter of the S_i similarities (β) were chosen from the sets $\{0, \frac{1}{64}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$, $\{\frac{1}{2^{-1}}, \frac{1}{2^0}, \frac{1}{2^1}, \frac{1}{2^2}, \frac{1}{2^4}, \frac{1}{2^6}, \ldots, \frac{1}{2^{14}}\}$, $\{8, 16\}$, and $\{0.2, 1, 1.8, \ldots, 14.6\}$, respectively. We used a $90\% - 10\%$ (80% training, 10% validation, 10% test) random split for the observable ratings in our experiments, similarly to [2].

First, we provide results using **toroid** group structure. The size of the toroid was 10×10 ($d_\alpha = 100$). In the first experiment we study how the size of neighborhood (r) affects the results. To this end, we set the neighborhood size to $r = 0$ (no structure), and then increased it to 1, 2, 3, 4, and 5. For each (κ, ρ, β), the minimum of the validation/test surface w.r.t. β is illustrated in Fig. 1(a)-(b). According to our experiences, the validation and test surfaces are very similar for a fixed neighborhood parameter r. It implies that the validation surfaces are good indicators for the test errors. For the best r, κ and ρ parameters, we can also observe that the validation and test curves (as functions of β) are very similar [Fig. 1(c)]. Note that (i) both curves have only one local minimum, and (ii) these minimum points are close to each other. The quality of the estimation depends mostly on the κ regularization parameter. The estimation is robust to the different choices of forgetting factor ρ (see Fig. 1(a)-(b)), and this parameter can only help in fine-tuning the results.

From our results (Table 2), we can see that structured dictionaries ($r > 0$) are advantageous over those methods that do not impose structure on the dictionary elements ($r = 0$). Based on this table we can also conclude that the estimation is robust to the selection of the similarity (S) and the mini-batch size (R). We got the best results using similarity S_1 and $R = 8$. Similarly to the role of parameter ρ, adjusting S and R can only be used for fine-tuning. When we increase r up to $r = 4$, the results improve. However, for $r = 5$, the RMSE values do not improve anymore; they are about the same when using $r = 4$. The smallest RMSE we could achieve was 4.0774, and the best known result so far was RMSE = 4.1123 [12]. This proves the efficiency of our OSDL based collaborative filtering algorithm. We note that our RMSE result seems to be significantly better than that of the competitors: we repeated this experiment 5 more times with different randomly selected training, test, and validation sets, and our RMSE results have never been worse than 4.08.

[3] For $r = 0$ ($\mathcal{G} = \{\{1\}, \ldots, \{d_\alpha\}\}$) one gets the classical sparse code based dictionary.



(a) (b) (c)

Fig. 1. (a)-(b): validation and test surface as a function of forgetting factor (ρ) and regularization (κ). For a fixed (κ, ρ) parameter pair, the surfaces show the best RMSE values optimized in the β similarity parameter. (c): validation and test curves for the optimal parameters ($\kappa = \frac{1}{2^{10}}$, $\rho = \frac{1}{2^5}$, mini-batch size $R = 8$). (a)-(c): neighbor size: $r = 4$, group structure ($\mathcal{G}$): toroid, similarity: S_1.

Table 2. Performance of the OSDL prediction using toroid group structure ($\mathcal{G}$) with different neighbor sizes r ($r = 0$: unstructured case). Left: mini-batch size $R = 8$, right: $R = 16$. First row: S_1, second row: S_2 similarity. For fixed R, the best performance is highlighted with boldface typesetting.

	$R = 8$					$R = 16$				
	$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
S_1	4.1594	4.1326	4.1274	4.0792	**4.0774**	4.1611	4.1321	4.1255	4.0804	**4.0777**
S_2	4.1765	4.1496	4.1374	4.0815	4.0802	4.1797	4.1487	4.1367	4.0826	4.0802

In our second experiment, we studied how the **hierarchical** group structure $\mathcal{G}$ affects the results. Our obtained results are similar to that of the toroid structure. We experimented with hierarchy level $l = 3, 4, 5, 6$ (i.e, $d_\alpha = 7, 15, 31, 63$), and achieved the best result for $l = 4$. The RMSE values decrease until $l = 4$, and then increase for $l > 4$. Our best obtained RMSE value is 4.1220, and it was achieved for dimension $d_\alpha = 15$. We note that this small dimensional, hierarchical group structure based result is also better than that of [2], which makes use of unstructured dictionaries with $d_\alpha = 100$ and has RMSE = 4.1229. Our result is also competitive with the RMSE = 4.1123 value of [12].

To sum up, in the studied CF problem on the Jester dataset we found that (i) the application of group structured dictionaries has several advantages and the proposed algorithm can outperform its state-of-the-art competitors. (ii) The toroid structure provides better results than the hierarchical structure, (iii) the quality of the estimation mostly depends on the structure inducing Ω regularization (κ, $\mathcal{G}$, r or l), and (iv) it is robust to the other parameters (ρ forgetting factor, S_i similarity, R mini-batch size).

5 Conclusions

We have proposed an online group-structured dictionary learning (OSDL) approach to solve the collaborative filtering (CF) problem. We casted the CF estimation task as an OSDL problem, and demonstrated the applicability of our novel approach on joke recommendations. Our extensive numerical experiments show that structured dictionaries have several advantages over the state-of-the-art CF methods: more precise estimation can be obtained, and smaller dimensional feature representation can be sufficient by applying group structured dictionaries.

Acknowledgments. The Project is supported by the European Union and co-financed by the European Social Fund (grant agreements no. TÁMOP 4.2.1/B-09/1/KMR-2010-0003 and KMOP-1.1.2-08/1-2008-0002). The research was partly supported by the Department of Energy (grant number DESC0002607).

References

1. Ricci, F., Rokach, L., Shapira, B., Kantor, P.: Recommender Systems Handbook. Springer, Heidelberg (2011)
2. Takács, G., Pilászy, I., Németh, B., Tikk, D.: Scalable collaborative filtering approaches for large recommender systems. J. Mach. Learn. Res. 10, 623–656 (2009)
3. Bach, F., Jenatton, R., Marial, J., Obozinski, G.: Convex optimization with sparsity-inducing norms. In: Optimization for Machine Learning. MIT Press (2011)
4. Jenatton, R., Mairal, J., Obozinski, G., Bach, F.: Proximal methods for sparse hierarchical dictionary learning. In: ICML 2010, pp. 487–494 (2010)
5. Jenatton, R., Obozinski, G., Bach, F.: Structured sparse principal component analysis. AISTATS, J. Mach. Learn. Res.:W&CP 9, 366–373 (2010)
6. Mairal, J., Jenatton, R., Obozinski, G., Bach, F.: Network flow algorithms for structured sparsity. In: NIPS 2010, pp. 1558–1566 (2010)
7. Rosenblum, K., Zelnik-Manor, L., Eldar, Y.: Dictionary optimization for block-sparse representations. In: AAAI Fall 2010 Symposium on Manifold Learning (2010)
8. Kavukcuoglu, K., Ranzato, M., Fergus, R., LeCun, Y.: Learning invariant features through topographic filter maps. In: CVPR 2009, pp. 1605–1612 (2009)
9. Bottou, L., Cun, Y.L.: On-line learning for very large data sets. Appl. Stoch. Model. Bus. - Stat. Learn. 21, 137–151 (2005)
10. Szabó, Z., Póczos, B., Lőrincz, A.: Online group-structured dictionary learning. In: CVPR 2011, pp. 2865–2872 (2011)
11. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. Inform. Retrieval 4, 133–151 (2001)
12. Takács, G., Pilászy, I., Németh, B., Tikk, D.: Matrix factorization and neighbor based algorithms for the Netflix prize problem. In: RecSys 2008, pp. 267–274 (2008)

Group Polytope Faces Pursuit for Recovery of Block-Sparse Signals[⋆]

Aris Gretsistas and Mark D. Plumbley

Queen Mary University of London
Centre for Digital Music
Mile End Road, E1 4NS, London, UK
aris.gretsistas@eecs.qmul.ac.uk

Abstract. Polytope Faces Pursuit is an algorithm that solves the standard sparse recovery problem. In this paper, we consider the case of block structured sparsity, and propose a novel algorithm based on the Polytope Faces Pursuit which incorporates this prior knowledge. The so-called Group Polytope Faces Pursuit is a greedy algorithm that adds one group of dictionary atoms at a time and adopts a path following approach based on the geometry of the polar polytope associated with the dual linear program. The complexity of the algorithm is of similar order to Group Orthogonal Matching Pursuit. Numerical experiments demonstrate the validity of the algorithm and illustrate that in certain cases the proposed algorithm outperforms the Group Orthogonal Matching Pursuit algorithm.

Keywords: block-sparsity, polytopes, sparse representations.

1 Introduction

Over recent years, the study of sparse representations [1] has seen an increasing interest among researchers and its significance has been highlighted in numerous signal processing applications ranging from signal acquisition to de-noising and from coding to source separation. Sparse representations are signal expansions that can accurately represent the signal of interest using a linear combination of a relatively small number of significant coefficients drawn from a basis or a redundant dictionary.

Let $\mathbf{y} \in \mathbb{R}^M$ be the observed vector that we need to decompose and represent in the dictionary $\mathbf{A}$ of size $M \times N$ with $M < N$ using a small number K of significant coefficients corresponding to the columns of the full rank matrix $\mathbf{A}$. The sparse representation problem can then be formulated:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{x} = [x_1, \dots, x_N]^T$ is a K-sparse vector, namely it has only $K = \|\mathbf{x}\|_0$ non-zero entries, with $K \ll N$. The above system of linear equations is said

[⋆] This research is supported by ESPRC Leadership Fellowship EP/G007144/1 and EU FET-Open Project FP7-ICT-225913 "SMALL".

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 255–262, 2012.

to be an underdetermined system, as the number of unknowns is larger than the number of equations. Such a system yields an infinite number of solutions. In *sparse coding* we are interested in obtaining the sparsest solution which has the smallest number of non-zero elements. Two well studied algorithms that can recover under certain conditions the sparse vector $\mathbf{x}$ in equation (1) are *Basis Pursuit* (BP) [2] and *Orthogonal Matching Pursuit* (OMP) [3].

The conventional sparsity model assumes that the non-zero coefficients can be located anywhere in the sparse vector. However, block structures, which imply that the non-zero elements are grouped in blocks (or clusters) instead of being arbitrarily located throughout the vector $\mathbf{x}$, can appear in practical scenarios. More specifically, the sparse coefficients in multi-band signals [4] or harmonic signals [5] can be clustered in groups of dictionary atoms. In that special case of structured sparsity the block-sparse vector $\mathbf{x}$ is treated as a concatenation of blocks of length d:

$$\mathbf{x} = [\underbrace{x_1 \dots x_d}_{\mathbf{x}^T[1]} \underbrace{x_{d+1} \dots x_{2d}}_{\mathbf{x}^T[2]} \dots \underbrace{x_{N-d+1} \dots x_N}_{\mathbf{x}^T[P]}] \tag{2}$$

where $\mathbf{x}[p]$ denotes the p-th block and $N = Pd$. In [6] the block k-sparse vector is defined as the vector $\mathbf{x} \in \mathbb{R}^N$ that has non-zero ℓ_2 norm for at most k indices out of P, namely:

$$\|\mathbf{x}\|_{2,0} = \sum_{p=1}^{P} I(\|\mathbf{x}[p]\|_2 > 0) \leq k \tag{3}$$

where $I(.)$ is the indicator function.

It follows that the redundant dictionary $\mathbf{A}$ can also be represented as a concatenation of P block matrices:

$$\mathbf{A} = [\underbrace{\mathbf{a}_1 \dots \mathbf{a}_d}_{\mathbf{A}^T[1]} \underbrace{\mathbf{a}_{d+1} \dots \mathbf{a}_{2d}}_{\mathbf{A}^T[2]} \dots \underbrace{\mathbf{a}_{N-d+1} \dots \mathbf{a}_N}_{\mathbf{A}^T[P]}] \tag{4}$$

where $\mathbf{A}[p]$ denotes the p-th column block matrix of size $M \times d$.

In order to solve the problem in equation (1) one can attempt the minimization of the mixed ℓ_2/ℓ_1 norm [6]:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_{2,1} \quad \text{such that} \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{5}$$

where $\|\mathbf{x}\|_{2,1} = \sum_{p=1}^{P} \|\mathbf{x}[p]\|_2$. Moreover, greedy algorithms can serve as alternatives to the optimization in equation (5) e.g. *Group Orthogonal Matching Pursuit* (G-OMP) [6].

2 Review of the Polytope Faces Pursuit Algorithm

In this section we will review the original *Polytope Faces Pursuit* (PFP) algorithm, which we will generalize to group form in section 3. The traditional ℓ_1-minimization problem can be converted to its *standard form* using nonnegative coefficients:

$$\min_{\tilde{\mathbf{x}}} \mathbf{1}^T\tilde{\mathbf{x}} \quad \text{such that} \quad \mathbf{y} = \tilde{\mathbf{A}}\tilde{\mathbf{x}},\ \tilde{\mathbf{x}} \geq 0 \tag{6}$$

where $\mathbf{1}$ is a column vector of ones, $\tilde{\mathbf{A}} = [\mathbf{A}, -\mathbf{A}]$ and $\tilde{\mathbf{x}}$ is the $2N$ nonnegative vector:

$$\tilde{\mathbf{x}_\mathbf{i}} = \begin{cases} \max(x_i, 0) & 1 \leq i \leq N \\ \max(-x_{i-N}, 0) & N+1 \leq i \leq 2N. \end{cases} \tag{7}$$

The new linear program has a corresponding *dual linear program*:

$$\max_{\mathbf{c}} \mathbf{y}^T\mathbf{c} \quad \text{such that} \quad \tilde{\mathbf{A}}^T\mathbf{c} \leq \mathbf{1} \tag{8}$$

such that a bounded solution to (8) exists if and only if a bounded solution to (6) exists. Thus, we can initially look for a solution $\mathbf{c}^*$ to (8) and use the Karush-Kuhn-Tucker (KKT) [7] conditions to solve the resulting system for $\mathbf{x}^*$.

The algorithm in an iterative fashion adds one vector at a time, the one with the maximum scaled correlation:

$$\mathbf{a}^k = \arg \max_{\mathbf{a}_i \notin \tilde{\mathbf{A}}^k} \frac{\mathbf{a}_i^T\mathbf{r}^{k-1}}{1 - \mathbf{a}_i^T\mathbf{c}^{k-1}}. \tag{9}$$

After updating the solution vector $\tilde{\mathbf{x}}$ and the corresponding $\mathbf{c}$ the algorithm iterates until the stopping criteria is met. The full PFP algorithm is given in [8].

3 Recovery of Block-Sparse Signals via Group Polytope Faces Pursuit

3.1 Group Selection Criterion

As has been described in [8], the Polytope Faces Pursuit algorithm, based on the conventional sparsity model, starts at $\mathbf{c} = 0$ and adopts a path following approach towards the residual until it hits a face of the *polar polytope* $P^* = \{\mathbf{c} \mid \pm\mathbf{a_i}^T\mathbf{c} \leq 1, \mathbf{a_i} \in \mathbf{A}\}$, which is dual to the *primal polytope* $P = \text{conv}\{\pm\mathbf{a_i}, \mathbf{a_i} \in \mathbf{A}\}$. The next face encountered is the one along the current face towards the projected residual. More specifically, the path of the PFP algorithm at the k-th iteration can be defined as:

$$h^k = \mathbf{a}_i^T(\mathbf{c}^k + \alpha\mathbf{r}^k). \tag{10}$$

The next face will be encountered for the minimum α such that $h^k = 1$. A little manipulation of this condition leads to the maximum scaled correlation of equation (9) as the atom selection criterion of the PFP algorithm.

In order to extend this to the block sparsity case, inspired from the work in [9] which proposes an implementation of the group LARS algorithm, at each step of the algorithm we are looking for a minimum α such that:

$$\|\tilde{\mathbf{A}}[i]^T(\mathbf{c}^k + \alpha\mathbf{r}^k) - \mathbf{1}\|_2^2 = 0 \quad \text{for} \quad i = 1, ..., P \tag{11}$$

where $\tilde{\mathbf{A}}[i] = [\mathbf{A}[i], -\mathbf{A}[i]]$ is the $M \times 2d$ doubled block matrix and in the above expression we consider only d atoms for which the inner product with the residual is nonnegative. After computations we end up with the following second order polynomial:

$$\lambda^2 \|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}\|_2^2 - 2\lambda(\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r} + \|\tilde{\mathbf{A}}[i]^T\mathbf{r}\|_2^2 = 0 \tag{12}$$

where $\lambda = 1/\alpha$. The discriminant of the above quadratic polynomial is given:

$$\Delta = 4(((\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r})^2 - \|\tilde{\mathbf{A}}[i]^T\mathbf{r}\|_2^2\|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}\|_2^2). \tag{13}$$

The discriminant of the polynomial of equation (12) will always be less or equal to zero, and therefore the polynomial will have two complex conjugate solutions. Considering that due to the nonnegative constraint of the solution vector we require that $\tilde{\mathbf{A}}[i]^T\mathbf{r} > \mathbf{0}$ and also that it always holds $\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c} \geq \mathbf{0}$, it is straightforward to show that $(\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r} = \|(\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r}\|_2 \leq \|\tilde{\mathbf{A}}[i]^T\mathbf{r}\|_2\|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}\|_2$, where the Cauchy-Schwarz inequality has been used. Therefore, it follows that $\Delta \leq 0$. Consequently, the Group Polytope Faces Pursuit (G-PFP) algorithm at each iteration will have to choose the group of dictionary atoms with the maximum λ, where:

$$\lambda = \frac{(\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r} \pm j\sqrt{\|\tilde{\mathbf{A}}[i]^T\mathbf{r}\|_2^2\|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}\|_2^2 - ((\mathbf{1}^T - \tilde{\mathbf{A}}[i]\mathbf{c}^T)\tilde{\mathbf{A}}[i]^T\mathbf{r})^2}}{\|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}\|_2^2}. \tag{14}$$

In order to simplify equation (14) we take the squared absolute value of the complex conjugate solution and the group selection criterion of the G-PFP algorithm reduces to:

$$\tilde{\mathbf{A}}[i]^k = \arg \max_{\mathbf{A[i]} \notin \tilde{\mathbf{A}}^\mathbf{k}} \frac{\|\tilde{\mathbf{A}}[i]^T\mathbf{r}^{k-1}\|_2}{\|\mathbf{1} - \tilde{\mathbf{A}}[i]^T\mathbf{c}^{k-1}\|_2}. \tag{15}$$

Note that when the block size is $d = 1$ equation (15) reduces to the maximum scaled correlation of equation (9). In the next section we derive the dual linear program for group sparse signals and show that there exists an optimum primal-dual ($\mathbf{x}^*$,$\mathbf{c}^*$) pair.

3.2 Dual Linear Program of the Group Sparse Recovery Problem

The Lagrangian to the problem of equation (5) is:

$$\mathcal{L}(\mathbf{x}, \mathbf{c}) = \|\mathbf{x}\|_{2,1} - \mathbf{c}^T(\mathbf{A}\mathbf{x} - \mathbf{y}) \tag{16}$$

and subsequently, the differential of $\mathcal{L}$ with respect to $\mathbf{x}$ is:

$$\partial_\mathbf{x}\mathcal{L}(\mathbf{x}, \mathbf{c}) = \partial_\mathbf{x}\|\mathbf{x}\|_{2,1} - \mathbf{A}^T\mathbf{c}. \tag{17}$$

It can easily be shown that the subdifferential $\partial_{\mathbf{x}}\|\mathbf{x}\|_{2,1}$ is given by the expression $\partial_{\mathbf{x}}\|\mathbf{x}\|_{2,1} = \mathbf{x}[p]/\|\mathbf{x}[p]\|_2$ when $\|\mathbf{x}[p]\|_2 > 0$. However, for the zero block-elements of $\mathbf{x}$ the gradient is not defined, but $\partial_{\mathbf{x}}\|\mathbf{x}\|_2$ coincides with the set of unit ℓ_2 norm vectors $\mathcal{B}^r_{\ell_2} = \{\mathbf{u} \in \mathbb{R}^r | \|\mathbf{u}\|_2 \leq 1\}$ [10]. Therefore, for each $p = 1, \dots, P$, we have:

$$\partial_{\mathbf{x}}\|\mathbf{x}\|_{2,1} = \begin{cases} \mathbf{x}[p]/\|\mathbf{x}[p]\|_2 & \|\mathbf{x}[p]\|_2 > 0 \\ \mathcal{B}^r_{\ell_2} & \text{otherwise.} \end{cases} \tag{18}$$

It follows that $\partial_{\mathbf{x}}\|\mathbf{x}\|_{2,1} \leq \mathbf{1}$. The KKT conditions require that $\mathbf{Ax} = \mathbf{y}$ and $\partial_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \mathbf{c}) = 0$. Substituting equations (17) and (18) to the last expression we get the dual to the problem of equation (5):

$$\max_{\mathbf{c}} \mathbf{y}^T\mathbf{c} \quad \text{such that} \quad \|\mathbf{A}^T\mathbf{c}\|_\infty \leq 1. \tag{19}$$

Therefore, for the optimal $\mathbf{x}^*$ exists a corresponding optimal $\mathbf{c}^*$. According to the KKT conditions for the primal-dual optimal $(\mathbf{x}^*,\mathbf{c}^*)$ the necessary and sufficient conditions are $\mathbf{Ax}^* = \mathbf{y}$ and $\|\mathbf{A}^T\mathbf{c}^*\|_\infty \leq 1$.

As already discussed, the G-PFP algorithm is based on the geometry of the polar polytope associated with dual linear program and searches the optimum vertex $\mathbf{c}^*$ using a path following approach. In the following section the proposed algorithm is derived.

3.3 The Proposed Algorithm

Let us now derive the proposed algorithm for recovery of block sparse signals. The G-PFP algorithm is an iterative greedy algorithm that builds the solution vector in a similar way to the G-OMP algorithm. The algorithm at the k-th iteration uses equation (15) to identify the next group of atoms, where we consider only vectors $\tilde{\mathbf{a}}_i$ for which $\tilde{\mathbf{a}}_i^T\mathbf{r}^{k-1} > 0$ within each group of atoms due to the nonnegativity constraint of the solution vector and we exclude the groups that have already been selected in previous iterations. Note that the first iteration will be identical to the G-OMP algorithm as $\mathbf{c}$ is initialized at zero.

Next the algorithm adds the selected group of atoms to the active set and updates the solution vector $\tilde{\mathbf{x}}^k$, the residual $\mathbf{r}^k$ and the corresponding $\mathbf{c}^k$. The algorithm iterates till the stopping criteria are met. The resulting algorithm of G-PFP is given in Algorithm 1.

One of the most expensive computations of the algorithm is the calculation of the Moore-Penrose pseudo-inverse $(\tilde{\mathbf{A}}^k)^\dagger$ required for the update of the solution vector $\tilde{\mathbf{x}}^k$ and the corresponding $\mathbf{c}^k$ at each iteration. As has already been in discussed in [11] for the conventional sparsity PFP algorithm, directional updates could be used (e.g. the method of conjugate gradient) instead of the Cholesky factorization method when dealing with large scale systems.

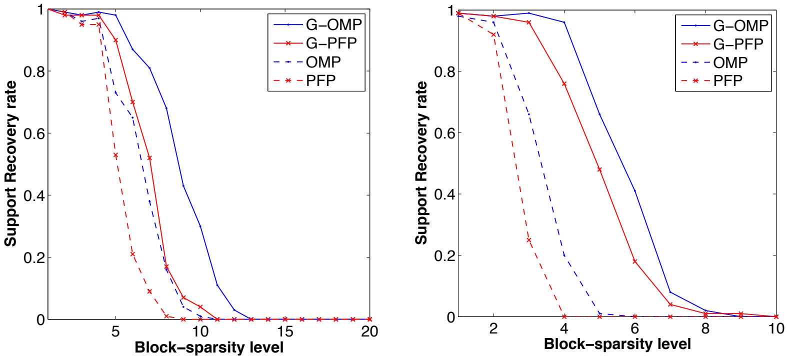
Note also that following LARS, pretty much as we did in [11] we omit the releasing step, which reduces the computational cost but is expected not to lead to a large change to the result.

Algorithm 1. Group-Polytope Faces Pursuit (G-PFP)

```
Input: $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_i]$, $\mathbf{y}$
Set stopping conditions $l_{\max}$ and $\theta_{\min}$
Initialize: $k \leftarrow 0,\ \mathcal{I}^k \leftarrow \emptyset,\ \tilde{\mathbf{A}}^k \leftarrow \emptyset,\ \mathbf{c}^k \leftarrow \mathbf{0},\ \tilde{\mathbf{x}}^k \leftarrow \emptyset,\ \hat{\mathbf{y}}^k \leftarrow \mathbf{0},\ \mathbf{r}^k \leftarrow \mathbf{y}$
while $|\mathcal{I}^k| < l_{\max}$ and $\max_i \tilde{\mathbf{a}}_i^T \mathbf{r}^{k-1} > \theta_{\min}$ do {Find next face}
    $k \leftarrow k + 1$
    Find face:
    $i^k \leftarrow \arg\max_{i \notin \mathcal{I}^{k-1}} \{\|\tilde{\mathbf{A}}[i]^T \mathbf{r}^{k-1}\|_2 / \|1 - \tilde{\mathbf{A}}[i]^T \mathbf{c}^{k-1}\|_2 \mid \tilde{\mathbf{A}}[i]^T \mathbf{r}^{k-1} > 0\}$
    Add constraints:
    $\tilde{\mathbf{A}}^k \leftarrow [\tilde{\mathbf{A}}^{k-1},\ \tilde{\mathbf{A}}[i]^k],\ \mathcal{I}^k \leftarrow \mathcal{I}^{k-1} \cup \{i^k\}$
    $\tilde{\mathbf{x}}^k \leftarrow (\tilde{\mathbf{A}}^k)^\dagger \mathbf{y}, \mathbf{c}^k \leftarrow (\tilde{\mathbf{A}}^k)^{\dagger T} \mathbf{1},\ \hat{\mathbf{y}}^k \leftarrow \tilde{\mathbf{A}}^k \tilde{\mathbf{x}}^k,\ \mathbf{r}^k \leftarrow \mathbf{y} - \hat{\mathbf{y}}^k$
end while
Output: $\mathbf{c}^* = \mathbf{c}^k, \tilde{\mathbf{x}}^* \leftarrow \mathbf{0} +$ corresponding entries from $\tilde{\mathbf{x}}^k$
```

4 Simulation Results

In the first experiment we attempted to quantify the performance of the proposed algorithm and compare against the group sparsity algorithm G-OMP and the standard sparsity algorithms OMP and PFP, using synthetic data. To do so, we randomly generated dictionaries of size 40 × 200 by drawing from i.i.d. Gaussian matrices and normalizing them. The block k-sparse vector $\mathbf{x}$ with block size d was generated by selecting uniformly at random the non-zero groups of atoms.



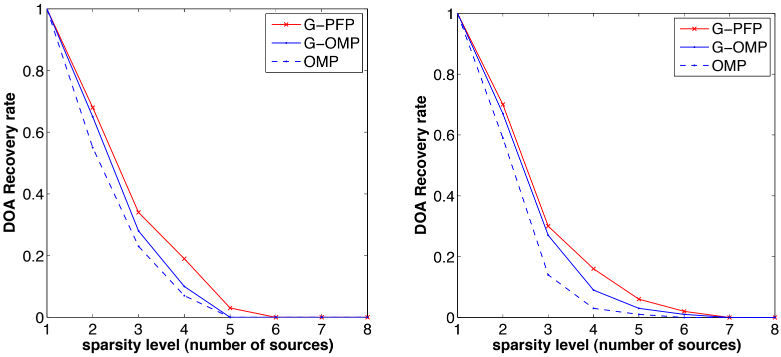
(a)

(b)

Fig. 1. Support recovery rates (over 100 trials) of G-OMP, G-PFP, OMP, PFP vs block-sparsity level k for a dictionary $\mathbf{A} \in \mathbb{R}^{M \times N}$ with $M = 40$, $N = 200$ and block size (a) $d = 2$ and (b) $d = 4$

Fig. 1(a)-(b) illustrates the support recovery rate of all tested algorithms for a variable sparsity level k, where the block size d has chosen equal to 2 and 4, respectively. The results has been averaged over 100 iterations. As can be seen, the greedy group sparsity algorithms perform better in both cases and the performance gain increases with the block size. However, G-OMP shows the best success recovery rates apart from the case when $d = 4$ for high sparsity levels, where G-PFP shows a slightly better performance.

For the second experiment, we chose to apply the algorithms to the problem of direction-of-arrival (DOA) estimation and compare their performance. In this case, we compared G-PFP against G-OMP and OMP. After discretization of the angular space, we formed the redundant dictionary $\mathbf{A}$ of size $M \times N$ containing the impulse responses of $M = 8$ sensors uniformly spaced at half wavelength for all $N = 181$ potential angles of arrival (resolution grid of 1°). Assuming that the $k << N$ plane waves impinge on the array from different angles (which has been chosen randomly) and taking d time-snapshots we formulated the resulting MMV problem as a block sparsity problem by appropriately interleaving the multiple vectors. Therefore, the d snapshots define the number of the size of each block.



(a)

(b)

Fig. 2. DOA recovery rates (over 100 trials) of G-OMP, G-PFP & OMP vs block-sparsity level k (or number of sources). The numbers of sensors is $M = 8$, the angular grid resolution is set at 1° and the number of snapshots (or block size) is (a) $d = 3$ and (b) $d = 4$.

Fig. 2(a)-(b) shows the recovery success rate of the true angles of arrivals averaged over 100 iterations when the number of snapshots and subsequently the block size is 3 and 4, respectively. For the specific setting in both cases G-PFP outperforms the other two algorithms achieving the highest recovery success rates. Considering the fact that the dictionary due to the small number of sensors chosen is quite block-coherent, the results suggest that the G-PFP algorithm can achieve better performance in distinguishing between correlated group of atoms.

5 Conclusions

We have introduced an algorithm for the block sparse recovery problem based on the PFP algorithm. The so-called G-PFP algorithm, which is a greedy algorithm of similar complexity to the G-OMP algorithm, adds one group of atoms at a time and iteratively builds the solution. Experiments on the support recovery of exact sparse block synthetic signals show that the proposed algorithm outperforms the standard PFP algorithm, but performs a little worse than G-OMP. However, on the DOA estimation problem the proposed algorithm showed better performance than G-OMP at all sparsity levels investigated.

Our future work will investigate and attempt to explain this behaviour of G-PFP in the coherent dictionary setting.

References

1. Plumbley, M.D., Blumensath, T., Daudet, L., Gribonval, R., Davies, M.E.: Sparse representations in audio and music: From coding to source separation. Proceedings of the IEEE 98(6), 995–1005 (2010)
2. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing 20(1), 33–61 (1998)
3. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 1-3, pp. 40–44 (1993)
4. Mishali, M., Eldar, Y.C.: Blind multi-band signal reconstruction: Compressed sensing for analog signals. IEEE Transactions on Signal Processing 57(3), 993–1009 (2009)
5. Gribonval, R., Bacry, E.: Harmonic decomposition of audio signals with matching pursuit. IEEE Transactions on Signal Processing 51(1), 101–110 (2003)
6. Eldar, Y.C., Kuppinger, P., Bolcskei, H.: Block-sparse signals: Uncertainty relations and efficient recovery. IEEE Transactions on Signal Processing 58(6), 3042–3054 (2010)
7. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
8. Plumbley, M.D.: On polar polytopes and the recovery of sparse representations. IEEE Transactions on Information Theory 53(9), 3188–3195 (2007)
9. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67 (2006)
10. van den Berg, E., Friedlander, M.P.: Joint-sparse recovery from multiple measurements, Technical Report (2009)
11. Gretsistas, A., Damnjanovic, I., Plumbley, M.D.: Gradient Polytope Faces Pursuit for large scale sparse recovery problems. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2030–2033 (2010)

Nonnegative Matrix Factorization via Generalized Product Rule and Its Application for Classification

Yu Fujimoto[1] and Noboru Murata[2]

[1] Aoyama Gakuin University, Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan
yu.fujimoto@it.aoyama.ac.jp
[2] Waseda University, Ohkubo, Shinjuku, Tokyo 169-8555, Japan
noboru.murata@eb.waseda.ac.jp

Abstract. Nonnegative Matrix Factorization (NMF) is broadly used as a mathematical tool for processing tasks of tabulated data. In this paper, an extension of NMF based on a generalized product rule, defined with a nonlinear one-parameter function and its inverse, is proposed. From a viewpoint of subspace methods, the extended NMF constructs flexible subspaces which plays an important role in classification tasks. Experimental results on benchmark datasets show that the proposed extension improves classification accuracies.

Keywords: Nonnegative matrix factorization, generalized product rule, nonlinear function, subspace method, classification.

1 Introduction

Nonnegative matrix factorization (NMF) is a popular matrix factorization setup [4], and frequently applied to signal processing [15], image processing [8], text mining [14], etc. As a typical application, Benetos et al. [2] have applied NMF to classification tasks based on the idea of subspace methods [16]. In their setup, NMF is applied to obtain the basis matrix $\mathbf{W}$ and the coefficient matrix $\mathbf{H}$ for each class. An essential part of this approach is to classify an unlabeled datum into the class of the nearest subspace defined with the bases. However, this approach will not work when the obtained bases and subspaces are mismatched for given datasets. In such a situation, it is important to select appropriate bases and represent flexible subspaces.

In this paper, an extension of NMF based on a generalized product rule is proposed. Multiplication between two positive values x_1 and x_2 is formally derived with exponential and logarithmic functions, as

$$x_1 \times x_2 = \exp(\log(x_1) + \log(x_2)). \tag{1}$$

The mathematical operator "×" is naturally generalized with a pair of appropriate strictly increasing function $u(\cdot)$ and its inverse function $\xi(\cdot)$ as follows,

$$x_1 \otimes x_2 = u(\xi(x_1) + \xi(x_2)). \tag{2}$$

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 263–271, 2012.

The operator "$\otimes$" is used for the generalized product in this paper. Apparently, Eq. 2 represents various calculation results depending on the function $u(\cdot)$. This type of extension has been proposed from the viewpoint of a class of statistical models related to a divergence minimization problem based on a convex function [11,5]. On the other hand, the role of $u(\cdot)$ is closely related to the link function for the generalized linear model (GLM) [1], or the generator function for the Archimedean copula [12]. For example, the Archimedean copula is a representation of multivariate cumulative distribution functions (cdfs) parametrically specified with given marginal cdfs with strictly decreasing convex function $\psi(\cdot)$; in the bivariate case, the Archimedean copula is given as follows

$$C_\psi(a,b) = \psi^{-1}(\psi(a)+\psi(b)), \tag{3}$$

where $a, b \in [0,1]$.

By using the analogy of the formulation of the Archimedean copula, we introduce a concrete family of $u(\cdot)$ and propose an extension of NMF. The proposed method is expected to represent various factorization results even with the decomposed matrices have the same rank by introducing appropriate nonlinearity denoted with the function $u(\cdot)$. In our method, nonlinear relationship between decomposed matrices is directly and parametrically described, so that this method is different from kernel based approaches [17]. We apply the proposed extension to classification tasks according to a setting of subspace methods. An illustrative example given in this paper shows that the extension achieves "curved" subspaces, so that the extension is expected to provide good classification results when subspaces are appropriately curved.

The paper is organized as follows. At first, in Section 2, a generalized product rule based on a strictly increasing function $u(\cdot)$ is defined for NMF. In Section 3, we briefly introduce a projected gradient descent method to obtain extended NMF. Then, we show an illustrative example of our extension in the context of classification in Section 4. Some experimental results are also shown in this section. At last, concluding remarks are given in Section 5.

2 NMF via Generalized Product Rules

Let $\mathbf{V} = [v_{ij}] = [\mathbf{v}_1, \ldots, \mathbf{v}_J] \in \mathbb{R}_+^{I\times J}$ be a nonnegative matrix. The purpose of NMF is to obtain $\hat{\mathbf{V}} \cong \mathbf{V}$ by using low rank nonnegative matrices, $\mathbf{W} = [w_{ik}] \in \mathbb{R}_+^{I\times K}$ and $\mathbf{H} = [h_{kj}] \in \mathbb{R}_+^{K\times J}$, as $\hat{\mathbf{V}} = \mathbf{WH}$, or equivalently in the elementwise form, as

$$\hat{v}_{ij} = \sum_{k=1}^{K} w_{ik}h_{kj}. \tag{4}$$

We can obtain this type of decomposition under the given cost function; e.g., a typical decomposition is given by minimizing the Frobenius norm as follows,

$$\{\hat{\mathbf{W}}, \hat{\mathbf{H}}\} = \operatorname*{argmin}_{\mathbf{W},\mathbf{H}} \|\mathbf{V}-\mathbf{WH}\|^2 = \operatorname*{argmin}_{\mathbf{W},\mathbf{H}} \sum_{ij}\left\{v_{ij} - \sum_{k=1}^{K} w_{ik}h_{kj}\right\}^2 .$$

In conventional NMF, an element of the nonnegative matrix $\hat{\mathbf{V}}$ is described with the linear combination of elements in two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ as shown in Eq. 4. By introducing the generalized product rule to Eq. 4, an extension of NMF is defined as follows,

$$\hat{v}_{ij} = \sum_{k=1}^{K} w_{ik} \otimes h_{kj} = \sum_{k=1}^{K} u(\breve{w}_{ik} + \breve{h}_{kj}), \tag{5}$$

where $\breve{w}_{ik} = \xi(w_{ik})$, $\breve{h}_{kj} = \xi(h_{kj})$, and $u(\cdot)$ is a strictly increasing function. The formulation given by Eq. 5 is called u-NMF in this paper. Note that w_{ik} and h_{kj} are assumed to be nonnegative, but the transformed elements $\breve{w}_{ik}$ and $\breve{h}_{kj}$ are not. Therefore, if we choose appropriate $u(\cdot)$ such that $u(x) \geq 0$ $(\forall x \in \mathbb{R})$, we only need to obtain nonrestricted elements $\breve{w}_{ik}$ and $\breve{h}_{kj}$ (i.e., they can be negative) for the u-NMF formulation. Intuitively, Eq. 5 represents a nonlinear relation between two low rank nonnegative matrices through the function $u(\cdot)$[1]. For simplification, we denote the relation, Eq. 5, in a matrix algebra form, as $\hat{\mathbf{V}} = \mathbf{W} \otimes \mathbf{H}$.

Before introducing a concrete $u(\cdot)$, let us focus on domain $\mathcal{D}$ and range $\mathcal{R}$ for the function $\exp(\cdot)$ in conventional multiplication Eq. 1. The relations $\mathcal{D}(\exp) = (-\infty, \infty)$ and $\mathcal{R}(\exp) = (0, \infty)$ satisfy the following conditions,

Condition 1. $\mathcal{R}(u) \subseteq \mathbb{R}_+$ (or equivalently, $\mathcal{D}(\xi) \subseteq \mathbb{R}_+$).
Condition 2. $\breve{w}_{ik}, \breve{h}_{kj}, \breve{w}_{ik} + \breve{h}_{kj} \subseteq \mathcal{D}(u) \quad (\forall i, j, k)$.

Condition 1 guarantees nonnegativity of $\hat{v}_{ij}$, w_{ik} and h_{kj}. And Condition 2 is for feasible calculation of generalized multiplication; let $\underline{\xi}$ and $\overline{\xi}$ be the lower and the upper bounds of $\mathcal{D}(u) (= \mathcal{R}(\xi))$, then the condition is equivalently given as

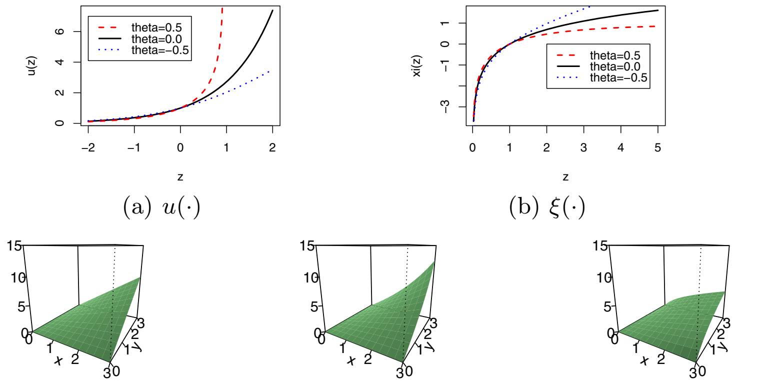
$$\begin{aligned} &(\max\{\underline{\xi}, \underline{\xi} - \min_{j'} \breve{h}_{kj'}\} \leq \breve{w}_{ik} \leq \min\{\overline{\xi}, \overline{\xi} - \max_{j'} \breve{h}_{kj'}\}) \\ &\quad \wedge (\max\{\underline{\xi}, \underline{\xi} - \min_{i'} \breve{w}_{i'k}\} \leq \breve{h}_{kj} \leq \min\{\overline{\xi}, \overline{\xi} - \max_{i'} \breve{w}_{i'k}\}) \qquad (\forall i, j, k), \end{aligned} \tag{6}$$

which plays an important role in implementation. Note that $\mathcal{D}(\exp)$ contains all the real values; therefore, Condition 2 is naturally satisfied in the case of conventional NMF. For generalization of product rule in the context of NMF, we introduce a function $u(\cdot)$ which satisfies Condition 1, and factorize $\mathbf{V}$ not to violate Condition 2.

In this paper, we introduce the following one-parameter function,

$$\begin{aligned} u_\theta(x) &= -\frac{1}{\theta} \log\{\exp(x)(\exp(-\theta) - 1) + 1\} \\ \xi_\theta(x) &= \log \frac{\exp(-\theta x) - 1}{\exp(-\theta) - 1}, \end{aligned} \tag{7}$$

[1] Nonnegative tensor decomposition also can be extended by introducing the generalized product rule. For example, a generalization of the third-order PARAFAC model [4] is given as $\hat{v}_{ijl} = \sum_{k=1}^{K} u(\xi(x_{ik}) + \xi(y_{jk}) + \xi(z_{lk}))$ where $[\hat{v}_{ijl}] \in \mathbb{R}_+^{I \times J \times L}$, $[x_{ik}] \in \mathbb{R}_+^{I \times K}$, $[y_{jk}] \in \mathbb{R}_+^{J \times K}$ and $[z_{lk}] \in \mathbb{R}_+^{L \times K}$.



(a) $u(\cdot)$

(b) $\xi(\cdot)$

(c) $x \times y$

(d) $x \otimes y$ ($\theta = 0.1$)

(e) $x \otimes y$ ($\theta = -0.5$)

Fig. 1. Functions given in Eq. 7 and relations between $x, y \in \mathbb{R}^+$ and $x \otimes y$

where $\xi_\theta(\cdot)$ is the negative generator function of Frank copula [6]. Frank copula is known to be a comprehensive class of the Archimedean copulas; this class of copulas has a flexible representational power (cf. [12]). The analogy between definitions of generalized multiplication given in Eq. 2 and the Archimedean copula C_ψ in Eq. 3 implies that we can represent a kind of dependency between variables by using the generalized product rule. Note that $\mathcal{D}(u_\theta)$ and $\mathcal{R}(u_\theta)$ are given as follows,

$$\mathcal{D}(u_\theta) = \begin{cases} (-\infty, -\log(1-\exp(-\theta))) & (\theta > 0) \\ (-\infty, \infty) & (\theta < 0) \end{cases}, \qquad \mathcal{R}(u_\theta) = (0, \infty).$$

Figure 1(a) and (b) show the forms of these functions. By using such a one-parameter function, the decomposition result derived with u-NMF is different from the conventional NMF result. Figure 1(c)-(e) show simple multiplication results between two values $x, y \in \mathbb{R}^+$. Intuitively, multiplication results $x \otimes y$ for large x and y tend to be larger than $x \times y$ when θ is positive though they tend to be smaller than $x \times y$ when θ is negative. Note that we regard the case of $\theta = 0$ in Eq. 7 as $\lim_{\theta \to 0} u_\theta(x) = \exp(x)$.

3 Implementation of *u*-NMF

The multiplicative algorithm, which is proposed by Lee and Seung [9], is the most popular approach to achieve conventional NMF. However, this type of multiplicative algorithm will not be derived for most of the u-NMF formulations except when the function $u(\cdot)$ is well-defined with no bounded domain. In this paper, we introduce the simple stochastic gradient descent approach [3] with the idea of projected gradient method [10] to satisfy the constraints shown in Eq. 6, in the case that $u(\cdot)$ has the bounded domain. We note that a gradient-based approach for conventional NMF [13] was also proposed before Lee and Seung's method.

Assume that $u(z) \geq 0$ $(\forall z)$ and $\hat{v}_{ij}$ be the approximated matrix which is given by Eq. 5. When the domain of $u(\cdot)$ (i.e., the range of $\xi(\cdot)$) has lower and upper bounds, $\underline{\xi}$ and $\overline{\xi}$, the minimization problem is defined as the following form with inequality constraints,

$$\min_{\breve{\mathbf{W}},\breve{\mathbf{H}}} \sum_{i,j} \left\{ v_{ij} - \sum_{k=1}^{K} u\left(\breve{w}_{ik} + \breve{h}_{kj}\right) \right\}^2$$

$$\text{subject to } \underline{\xi} \leq \breve{w}_{ik} + \breve{h}_{kj} \leq \overline{\xi}, \quad \underline{\xi} \leq \breve{w}_{ik} \leq \overline{\xi}, \quad \underline{\xi} \leq \breve{h}_{kj} \leq \overline{\xi} \quad (\forall i, j, k). \tag{8}$$

This type of optimization is related with the exponentiated gradient method [7] for conventional NMF. Now, let

$$g_{ijk}(\mathbf{W}, \mathbf{H}) = -2 \left(v_{ij} - \sum_{m=1}^{K} u(\breve{w}_{im} + \breve{h}_{mj}) \right) u'(\breve{w}_{ik} + \breve{h}_{kj}), \tag{9}$$

be the partial derivative of the objective function with respect to $\breve{w}_{ik}$, or equivalently, that with respect to $\breve{h}_{kj}$, at $\{\breve{w}_{ik}, \breve{h}_{kj}\}$. Then, for a feasible set $\{\breve{w}_{ik}, \breve{h}_{kj}\}$, we obtain the following update rules by using the projected gradient method,

$$\begin{aligned} \breve{w}_{ik}^{\text{new}} &\leftarrow \phi\left[\breve{w}_{ik} - \gamma g_{ijk}(\mathbf{W}, \mathbf{H}); \max_j \breve{h}_{kj}, \min_j \breve{h}_{kj}\right] \\ \breve{h}_{ik}^{\text{new}} &\leftarrow \phi\left[\breve{h}_{kj} - \gamma g_{ijk}(\mathbf{W}, \mathbf{H}); \max_i \breve{w}_{ik}^{\text{new}}, \min_i \breve{w}_{ik}^{\text{new}}\right], \end{aligned} \tag{10}$$

where

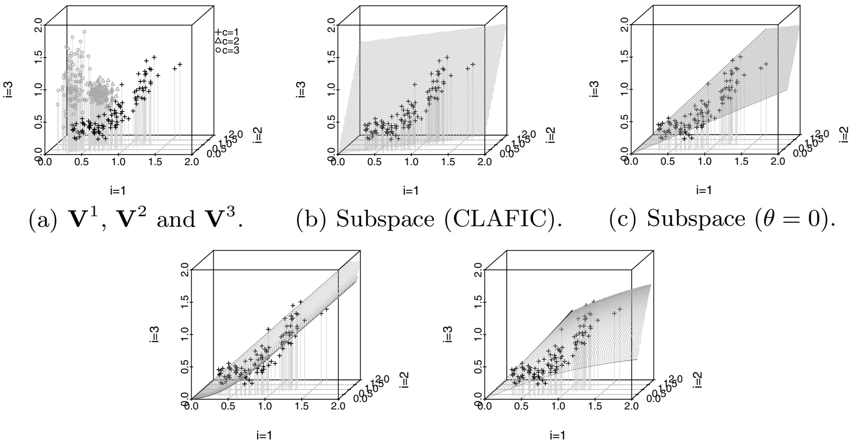
$$\phi[x; z_1, z_2] = \begin{cases} x & \text{if } \max\{\underline{\xi}, \underline{\xi} - z_2\} \leq x \leq \min\{\overline{\xi}, \overline{\xi} - z_1\}, \\ \min\{\overline{\xi}, \overline{\xi} - z_1\} - \varepsilon & \text{if } x \geq \min\{\overline{\xi}, \overline{\xi} - z_1\}, \\ \max\{\underline{\xi}, \underline{\xi} - z_2\} + \varepsilon & \text{if } x \leq \max\{\underline{\xi}, \underline{\xi} - z_2\}, \end{cases} \tag{11}$$

is a function that projects updated $\breve{w}_{ik}$ and $\breve{h}_{kj}$ into the bounded feasible area, ε is a small positive value to keep feasibility and γ is the learning step size. Note that the update rules given by Eq. 10 strictly satisfy feasibility of $\{\breve{w}_{ik}^{\text{new}}, \breve{h}_{kj}^{\text{new}}\}$ for any feasible sets $\{\breve{w}_{ik}, \breve{h}_{kj}\}$; the updated matrices always satisfy Eq. 6. In our experiments, we simply iterated these update rules for all the elements in $\mathbf{V}$ for 5,000 times. We also note that the minimization problem Eq. 8 has local minima, so that we repeated this sequence 10 times from different initial values and adopt the best result in the sense of the Frobenius norm.

4 *u*-NMF for Classification

4.1 Illustrative Example

Now, we show an illustrative example of the proposed formulation from the viewpoint of classification task and compare our setup with a typical subspace method called CLAFIC [16]. Let C be the number of classes and I be the dimension of



(a) $\mathbf{V}^1$, $\mathbf{V}^2$ and $\mathbf{V}^3$. (b) Subspace (CLAFIC). (c) Subspace ($\theta = 0$).

(d) Subspace ($\theta = -5$). (e) Subspace ($\theta = 1$).

Fig. 2. Subspaces in three-dimensional space. (a): Plots of $\{\mathbf{v}_j^1\}$, $\{\mathbf{v}_j^2\}$ and $\{\mathbf{v}_j^3\}$ in $\mathbb{R}_+^3$ ($\{\mathbf{v}_j^1\}$ is highlighted). (b): Subspace S^1 used in CLAFIC. (c)-(e): Subspaces S^1 based on u-NMF with $\theta = \{0, -5, 1\}$.

an observed datum. Figure 2(a) shows a typical distribution of data with $I = 3$ and $C = 3$. For such a dataset, we extract a subset $\mathbf{V}^c = [\mathbf{v}_1^c, \dots, \mathbf{v}_{J^c}^c] \in \mathbb{R}_+^{I \times J^c}$ for each class $c = 1, \dots, C$ where J^c is the number of members in class c. The basic idea of the subspace method is given as follows.

1. Construct a subspace S^c to represent $\mathbf{V}^c$ in the space of $\mathbb{R}^I$ for each c.
2. For an unlabeled datum $\mathbf{v}_* \in \mathbb{R}^I$, a label is given as $\hat{c} = \underset{c}{\operatorname{argmin}} \min_{s \in S^c} \|\mathbf{v}_* - \mathbf{s}\|^2$.

If there is a kind of dependency in some dimensions in I-dimensional vectors $\{\mathbf{v}_j^c\}$, $\mathbf{V}^c$ is expected to be represented in a low-dimensional subspace in the I-dimensional feature space. In such a case, the subspace method will work well by representing appropriate subspaces $\{S^c\}$. A problem in the subspace method is the appropriateness of subspaces for distorted data distributions like $\mathbf{V}^1$ in Figure 2(a).

In CLAFIC, a matrix $\mathbf{V}^c$ is represented with a set of orthonormal basis vectors $\mathbf{U}^c = [\mathbf{u}_1^c, \dots, \mathbf{u}_{K^c}^c] \in \mathbb{R}^{I \times K^c}$ where $\{\mathbf{u}_k^c\}$ is a set of eigenvectors of the matrix $\frac{1}{J^c} \sum_{\mathbf{v} \in \{\mathbf{v}_j^c\}} \mathbf{v}\mathbf{v}^t$ corresponding to the largest K^c eigenvalues for each c. A subspace used in CLAFIC is a linear combination of basis vectors $\{\mathbf{u}_k^c\}$,

$$S^c = \{\mathbf{U}^c \mathbf{a} \mid \forall \mathbf{a} \in \mathbb{R}^{K^c}\}, \tag{12}$$

so that all the subspaces in CLAFIC are interpreted as "flat" (see, Figure 2(b)). Here, we factorize each matrix $\mathbf{V}^c$ into a basis matrix $\mathbf{W}^c = [\mathbf{w}_1^c, \dots, \mathbf{w}_{K^c}^c] \in \mathbb{R}_+^{I \times K^c}$ and a coefficient matrix $\mathbf{H}^c = [\mathbf{h}_1^c, \dots, \mathbf{h}_{J^c}^c] \in \mathbb{R}_+^{K^c \times J^c}$ by using u-NMF respectively. As shown in Figure 2(c), a convex subspace,

$$S^c = \{\mathbf{W}^c \mathbf{a} \mid \forall \mathbf{a} \in \mathbb{R}_+^{K^c}\}, \tag{13}$$

Table 1. Datasets and classification results

dataset	C	I	size of dataset	10-CV lassification errors (mean ± SD) CLAFIC	 NMF	 u-NMF
breast-cancer	2	9	683	**0.161±0.058**	0.348±0.070	0.221±0.074
iris	3	4	150	**0.027±0.047**	0.567±0.141	0.120±0.093
sonar	2	10	990	0.270±0.055	0.226±0.046	**0.150±0.087**
vowel	11	60	208	0.842±0.034	0.846±0.034	**0.794±0.040**
wine	3	13	178	0.315±0.138	0.707±0.074	**0.152±0.066**

which is constructed in conventional NMF ($\theta = 0.0$) is also "flat". However, a subspace based on u-NMF with $u(\cdot) \neq \exp(\cdot)$, which is defined as,

$$S^c = \{\mathbf{W}^c \otimes \mathbf{a} \mid \forall \mathbf{a} \in \mathbb{R}_+^{K^c}\}, \tag{14}$$

is "curved" (and non-convex) according to $u(\cdot)$ in the original data space (see Figure 2(d) and (e)). For this reason, u-NMF can be a flexible description of a given nonnegative dataset by selecting appropriately curved subspaces. To determine $\hat{c}$ for an unlabeled datum $\mathbf{v}_*$ in our classification setup, we obtain $\mathbf{h}_*^c = \operatorname{argmin}_{\mathbf{h}} \|\mathbf{v}_* - \mathbf{W}^c \otimes \mathbf{h}\|^2 \in \mathbb{R}^{K^c}$ for each class c under fixed $\mathbf{W}^c$ for $\mathbf{v}_*$. Then, the corresponding class label $\hat{c}$ is determined as $\hat{c} = \operatorname{argmin}_c \|\mathbf{v}_* - \mathbf{W}^c \otimes \mathbf{h}_*^c\|^2$. And, a value θ is selected for each class respectively; for given learning set $\mathbf{V}^c$, θ is defined as $\operatorname{argmin}_{\theta \in \Theta} \|\mathbf{V}^c - \mathbf{W}^c \otimes \mathbf{H}^c\|$ where Θ is a set of candidates for θ. This procedure implies that a curved subspace derived with u-NMF has a possibility to represent a distorted data distribution more appropriately for each class.

4.2 Benchmark Tests

We compared CLAFIC, subspace methods based on conventional NMF and u-NMF on benchmark datasets called *breast-cancer*, *iris*, *sonar*, *vowel*[2] and *wine*, provided in the UCI Machine Learning Repository[3], from the viewpoint of classification errors based on 10-fold cross validation (10-CV). We divided each learning subset according to class $c = 1, \ldots, C$, thus a set of matrices $\{\mathbf{V}^c \in \mathbb{R}^{I \times J^c}\}$ is generated. We prepared subspaces for each class according to Eqs. 12-14 respectively. The dimension K^c for each class is given as the minimum number of principal components such that the cumulative contribution ratio is more than 0.95 in CLAFIC, and $K^c = \lfloor \frac{IJ^c}{I+J^c} \rfloor$ [2] is used in NMF and u-NMF[4]. In the u-NMF approach, the parameter θ for each class was selected from the set $\Theta = \{-5.0, -4.5, \cdots, 2.0\}$. And in the conventional NMF approach, all the subspaces were constructed under $\theta = 0$. Table 1 also shows 10-CV classification

[2] Log-transformed data values in *vowel* are converted into original nonnegative data by the exponential function.

[3] http://archive.ics.uci.edu/ml/

[4] Obviously, the tuning of K^c is an important topic in classification. Here, we simply compared appropriateness of the subspaces under the fixed K^c for NMF and u-NMF.

errors of three methods. The errors of our method are smaller than those of CLAFIC on *sonar*, *vowel* and *wine*. Moreover, the classification errors of u-NMF are smaller than those of conventional NMF in all the cases. The improvement of classification results indicates that u-NMF flexibly provides appropriate subspaces for various data distributions.

5 Conclusion

In this paper, we proposed an extension of NMF with an idea of generalized multiplication. To formulate this extension, we introduce a strictly increasing function which is derived from the generator function of a comprehensive Archimedean copula. An intuitive interpretation of our proposed factorization is illustrated from the viewpoint of the subspace method. The experimental results show that we can improve the classification accuracy of conventional NMF by selecting a suitably curved subspace for each class.

This type of extension is expected to be valid not only in classification tasks, but also in the other analyses. Although we introduced a concrete form of $u(\cdot)$ from Frank copula in this paper, the appropriateness of $u(\cdot)$ completely depends on the dataset and the task. Further discussions are needed for the appropriate family of $u(\cdot)$. The investigation of the optimization problem and the development of an efficient algorithm for estimation of u-NMF under given $u(\cdot)$ are the other important topics for practical analyses. For the latter issue, if we use the convex function $u(\cdot)$ such that $\mathcal{D}(u) = (-\infty, \infty)$, e.g., Eq. 7 with negative θ, an efficient algorithm like the multiplicative algorithm [9] will be derived by introducing an adequate auxiliary function. This point of view also suggests an importance of the discussion of the convenient family of $u(\cdot)$ for a given task. These topics are remained as future works.

Acknowledgment. This work was supported by JSPS, Grant-in-Aid for Young Scientists(B), 22700292.

References

1. Agresti, A.: Categorical Data Analysis. Wiley Inc. (2002)
2. Benetos, E., Kotropoulos, C., Lidy, T., Rauber, A.: Testing supervised classifiers based on non-negative matrix factorization to musical instrument classification. In: 2006 IEEE International Conference on Multimedia and Expo. (2006)
3. Bottou, L.: Stochastic gradient learning in neural networks. In: Proc. Neuro-Nímes 1991 (1991)
4. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: Nonnegative Matrix and Tensor Factorizations -Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley (2009)
5. Fujimoto, Y., Murata, N.: A Generalization of Independence in Naive Bayes Model. In: Fyfe, C., Tino, P., Charles, D., Garcia-Osorio, C., Yin, H. (eds.) IDEAL 2010. LNCS, vol. 6283, pp. 153–161. Springer, Heidelberg (2010)

6. Genest, C.: Frank's family of bivariate distributions. Biometrika 74(3), 549–555 (1987)
7. Kivinen, J., Warmuth, M.K.: Exponentiated gradient versus gradient descent for linear predictors. Information and Computation 132, 1–63 (1997)
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791 (1999)
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 13, pp. 556–562 (2001)
10. Lin, C.-J.: Projected gradient methods for non-negative matrix factorization. Neural Computation 19(10), 2756–2779 (2007)
11. Murata, N., Fujimoto, Y.: Bregman divergence and density integration. Journal of Math-for-Industry 1, 97–104 (2009)
12. Nelsen, R.B.: An Introduction to Copulas. Springer, Heidelberg (2006)
13. Paatero, P., Tapper, U.: Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics 5, 111–126 (1994)
14. Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing and Management 42(2), 373–386 (2006)
15. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics (2003)
16. Watanabe, S., Pakvasa, N.: Subspace method in pattern recognition. In: Proc. 1st International Joint Conference on Pattern Recognition, pp. 25–32 (1973)
17. Zhang, D., Zhou, Z.-H., Chen, S.: Non-negative matrix factorization on kernels. In: The 9th Pacific Rim International Conference on Artificial Intelligence, pp. 404–412 (2006)

An Algebraic Method for Approximate Rank One Factorization of Rank Deficient Matrices

Franz J. Király, Andreas Ziehe, and Klaus-Robert Müller

Technische Universität Berlin, Machine Learning Group, Franklinstr. 28/29, 10587 Berlin, Germany

Abstract. In this paper we consider the problem of finding approximate common rank one factors for a set of matrices. Instead of jointly diagonalizing the matrices, we perform calculations directly in the problem intrinsic domain: we present an algorithm, AROFAC, which searches the approximate linear span of the matrices using an indicator function for the rank one factors, finding specific single sources. We evaluate the feasibility of this approach by discussing simulations on generated data and a neurophysiological dataset. Note however that our contribution is intended to be mainly conceptual in nature.

1 Introduction

Finding common linear subspaces in data is a classical and well-studied problem in Machine Learning and the applied sciences. Many of these problems can be formulated as finding common eigenspaces resp. singular spaces for a set of matrices, e.g. ICA, CCA, SSA etc. [3,14,1,2]. A standard way to address these problems algorithmically is jointly diagonalizing the matrices. Several highly optimized and efficient algorithms already exist to approximately perform joint diagonalization on a set of matrices [4,12,15,16,17], let them be real or complex, symmetric or non-symmetric, full rank or rank deficient.

However, there are scenarios in signal processing and pattern recognition where finding a complete diagonalization may not be necessary, as the problem often consists intrinsically in finding single common rank one constituents of the matrices and not the whole diagonalization. In the existing approaches the problem is then commonly reduced to joint diagonalization and solved by searching in the space of basis transforms (cf. [5,14,6]).

In this paper, we propose a novel framework to address optimization problems of this sort in its natural and intrinsic formulation without the need to invoke joint diagonalization algorithms. We reformulate the setting as an optimization task on the vector space of matrices and exemplify possible search strategies by a loss function which measures distance to the rank one manifold. The novel algorithm that is computing an approximate rank one factor will be refered to as AROFAC algorithm. We demonstrate its efficacy and noise stability in a signal processing scenario and study inherent phenomena and applicability.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 272–279, 2012.

Let us briefly describe the mathematical underpinnings. Let $A \in \mathbb{C}^{n\times n}$ be a matrix. The rank of $\boldsymbol{A}$ is the number of its non-zero singular values. Equivalently, it is the minimum r such that A can be written as

$$\boldsymbol{A} = \sum_{i=1}^{r} \boldsymbol{a}_i \boldsymbol{b}_i^\top \quad \text{with } \boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{C}^n.$$

In general, this presentation is unique up to numbering and common scaling of the $\boldsymbol{a}_i, \boldsymbol{b}_i$. We will call it the rank one decomposition of $\boldsymbol{A}$. If we are now given matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K \in \mathbb{C}^{n\times n}$ having joint singular vectors, we can similarly write them as

$$\boldsymbol{M}_k = \sum_{i=1}^{r} \sigma_i^{(k)} \boldsymbol{a}_i \boldsymbol{b}_i^\top \quad \text{with } \boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{C}^n \text{ and } \sigma_i^{(k)} \in \mathbb{C}.$$

In practice, the $\boldsymbol{M}_k$ are additionally endowed with noise. Also, if the $\boldsymbol{M}_k$ have additional singular vectors which differ over the k, this may be also modelled as noise if the non-common singular vectors are sufficiently general - so the approach can be also used to model the case where some singular vectors are common and others are not, in contrast to the classical joint diagonalization ansatz.

An recurring problem is now to find a (possibly complex) linear combination $\boldsymbol{M}$ of the $\boldsymbol{M}_k$ which has (approximately) rank one. In the above presentation, $\boldsymbol{M}$ is then a complex multiple of $\boldsymbol{a}_i \boldsymbol{b}_i^\top$ for some i. The standard approach to solve this question would now be to simultaneously diagonalize the $\boldsymbol{M}_k$. However, a more natural and intrinsic approach can be obtained from observing that the vector space spanned by $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$ and the r-dimensional vector space spanned by the rank one matrices $\boldsymbol{a}_i \boldsymbol{b}_i^\top, 1 \leq i \leq r$ (approximately) coincide if $K \geq r$ and the $\boldsymbol{M}_i$ are sufficiently general. Denote this vector space by $\mathcal{V}$. It has a natural, though overdetermined, coordinate parametrization for $\mathcal{V}$ given by the $\lambda_1, \ldots, \lambda_K$. The parametrization can be made unique for example by changing to a principal basis of $\mathcal{V}$ having dimension r. One then can for example search $\mathcal{V}$ in its natural representation for low rank matrices, e.g. rank one matrices, by using rank indicator functions. Also, it is possible to perform dimension reduction procedures such as PCA or deflationary mode computations directly in the vector space $\mathcal{V}$ and its generators.

The AROFAC algorithm presented in this paper searches $\mathcal{V}$ for rank one matrices using an algebraic rank one indicator loss function.

We will explain the mathematical details in the next section.

2 Finding Joint Rank One Factors

In this section, we will present the AROFAC algorithm which finds common rank one factors. Before proceeding to the algorithm, we will recapitulate notation and fix the problem it solves. Start with matrices

$$\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K \in \mathbb{C}^{n \times n},$$

of rank r having a joint rank one decomposition. Assume that $K \geq r$. The $\boldsymbol{M}_i$ can be written as

$$\boldsymbol{M}_k = \sum_{i=1}^{r} \sigma_i^{(k)} \, \boldsymbol{a}_i \boldsymbol{b}_i^\top, \text{ for } 1 \leq k \leq K,$$

where $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{C}^n$ are the singular vectors and $\sigma_i^{(k)}$ the corresponding singular values. The goal is to find $\boldsymbol{a}_i, \boldsymbol{b}_i$, for some i, given the matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$, which are known approximately up to some additive noise.

The central observation for the algorithm is the following: If the $\boldsymbol{M}_i$ are sufficiently different (i.e. generic under the conditions above), they span the r-dimensional vector space

$$\mathcal{V} = \operatorname{span}\langle \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K \rangle = \operatorname{span}\langle \boldsymbol{a}_1 \boldsymbol{b}_1^\top, \ldots, \boldsymbol{a}_r \boldsymbol{b}_r^\top \rangle.$$

Instead of optimizing in the space of coordinate transformations, which is $O(n^2)$-dimensional, we will optimize in this r-dimensional vector space in order to obtain the span vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$. We will explain in the following how:

Due to the above, every element of $\mathcal{V}$ can be written uniquely as

$$\boldsymbol{M} = \sum_{i=1}^{r} \alpha_i \boldsymbol{a}_i \boldsymbol{b}_i^\top$$

for some $\alpha_i \in \mathbb{C}$. If $\boldsymbol{a}_i, \boldsymbol{b}_i$ are general, the rank of $\boldsymbol{M}$ is then exactly the number of nonzero $\alpha_i, 1 \leq i \leq r$ in this expansion. In particular, up to scaling, there are exactly r points in $\mathcal{V}$ corresponding to rank one matrices.

The AROFAC algorithm identifies points of this type in $\mathcal{V}$. For this, AROFAC first performs PCA to obtain the approximate span of matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$ of dimension r; i.e. after this step, we may assume that $K = r$, by replacing $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$ by a principal basis $\boldsymbol{M}'_1, \ldots, \boldsymbol{M}'_r$. Thus, element $\boldsymbol{M}$ of $\mathcal{V}$ can be uniquely written as

$$\boldsymbol{M}(\lambda) = \sum_{i=1}^{r} \lambda_i \boldsymbol{M}'_i,$$

where λ denotes the r-dimensional vector $(\lambda_1, \ldots, \lambda_r)$. Then, gradient descent is performed on the vector space parameterized by the $\lambda_1, \ldots, \lambda_r$ with respect to a loss function measuring the difference from rank one. The optimization is made unconstrained and real by setting $\lambda_1 = 1$ and optimizing with respect to real and complex parts of the λ_i.

To measure distance to rank one, AROFAC uses the following loss function that is zero if and only if $\boldsymbol{M}(\lambda)$ has rank one:

$$L(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \boldsymbol{M}(\lambda)_{ii} \boldsymbol{M}(\lambda)_{jj} - \boldsymbol{M}(\lambda)_{ij} \boldsymbol{M}(\lambda)_{ji} \right\|^2.$$

Here, as usual, $\boldsymbol{M}(\lambda)_{ij}$ denotes the (i,j)-th entry of the matrix $\boldsymbol{M}(\lambda)$. A derivation of the loss function can be found in the appendix. Optimizing the loss function yields an approximate rank one component $\boldsymbol{a}_i\boldsymbol{b}_i^\top$ and is computationally benign if r is small compared with n. In order to obtain a single rank one factor, AROFAC then takes the corresponding linear combination $\boldsymbol{M}(\lambda)$ which is a matrix approximately having rank one, and then factors it approximately in order to obtain $\boldsymbol{a} = \boldsymbol{a}_i$ and $\boldsymbol{b} = \boldsymbol{b}_i$ for some i. A pseudo-code description of AROFAC is given in Algorithm 1.

Algorithm 1. AROFAC.
Input: The matrices $\boldsymbol{M}_1, \dots, \boldsymbol{M}_K$. *Output:* $\boldsymbol{a}, \boldsymbol{b}$.

```
1: Perform PCA on the vector space spanned by the matrices $\boldsymbol{M}_1, \dots, \boldsymbol{M}_K$ and
   identify the $r$ principal components $\boldsymbol{M}'_1, \dots, \boldsymbol{M}'_r$.
2: Minimize $L(\lambda)$, e.g. by gradient descent.
3: Set $\boldsymbol{M} = \boldsymbol{M}'_1 + \sum_{i=2}^r \lambda_i \boldsymbol{M}'_i$.
4: Perform PCA on the row space of $\boldsymbol{M}$, set $\boldsymbol{b} =$ the principal component.
5: Perform PCA on the column space of $\boldsymbol{M}$, set $\boldsymbol{a} =$ the principal component.
6: Return $\boldsymbol{a}, \boldsymbol{b}$.
```

As the coefficient of $\boldsymbol{M}'_1$ is set to one, and $\boldsymbol{M}'_1$ is the principal component, AROFAC finds predominantly the common rank one factors which has the highest weighted occurrence in the decomposition of the $\boldsymbol{M}_1, \dots, \boldsymbol{M}_K$ when the gradient search is initialized with $\lambda = 0$.

The numerical optimization in step 2 is done using the “minFunc” MATLAB routine by Mark Schmidt[1] which is suited for unconstrained optimization of real-valued multivariate functions and can handle problems with large numbers of variables by implementing a limited-memory BFGS method [8].

3 Experiments

In our experiments, we first analyze the convergence behavior of AROFAC on toy data. Then we apply AROFAC to electrophysiological brain data.

The toy data input matrices

$$\boldsymbol{M}_k = \sum_{i=1}^{r} \sigma_i^{(k)} \boldsymbol{a}_i \boldsymbol{b}_i^\top,\ 1 \leq k \leq K$$

are generated as follows: Exact singular vectors $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^n, 1 \leq i \leq r$ are sampled independently and uniformly from the n-sphere. The singular values $\sigma_i^{(k)}$ are sampled independently and uniformly from the standard normal distribution. Then, to each matrix $\boldsymbol{M}_k$, noise is added in the form of a $(n \times n)$ matrix

[1] The “minFunc” webpage is: http://www.di.ens.fr/~mschmidt/Software/minFunc.html

whose entries are independently sampled from a normal distribution with mean 0 and covariance $\varepsilon \in \mathbb{R}^+$. We tested AROFAC for a wide range of parameters with $n \leq 100$, $K \leq 10n$, $r \approx n/5$ and $\varepsilon \leq 1$. Generally, low r and low $\varepsilon \leq 0.5$ increase accuracy. Figure 1 shows the typical convergence of the error for $K = 30, d = 4, r = 4$.

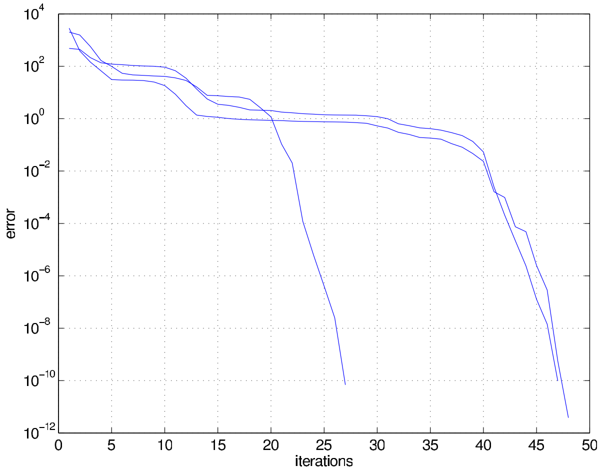


Fig. 1. Convergence of three typical runs on the toy dataset

Finally, we apply our new method to Pairwise Interacting Source Analysis (PISA) [10,11] of real EEG data. As shown by Nolte et al., interaction between source signals results in significant imaginary parts of their cross-spectrum [9]. Therefore, the imaginary parts of the complex-valued cross-spectral matrices are useful quantities for studying brain connectivity. Here we demonstrate that AROFAC can be used to extract a dominating rank one factor of such matrices.

EEG measurements were performed with subjects at rest under the relaxed, eyes-closed condition. EEG data were recorded with 64 Ag/AgCl electrodes, using the *Brain Vision Recorder* system (Brain Products GmbH, Munich, Germany) within the FASOR[2] project [7].

In order to obtain the target matrices $\boldsymbol{M}_k$, 59 channels were selected and cross-spectra were calculated in the frequency range 0 to 25 Hz in 0.4 Hz steps.

Applying AROFAC to the imaginary parts of those 63 matrices yields a complex vector $\boldsymbol{a}$ as the principal rank one component. Figure 2 shows 4 projections of the 2D subspace spanned by the two components $\mathrm{Re}(\boldsymbol{a})$ and $\mathrm{Im}(\boldsymbol{a})$. The four panels correspond to four different directions in this subspace with angles 0 degrees, 45 degrees, 90 degrees and 135 degrees relative to an arbitrary direction.

We note that the pattern is stably reproduced in several runs and bears similarities to typical EOG / visual alpha related components, but also deserves further detailed analysis e.g. by inverse modelling and source localization.

[2] We acknowledge Stefan Haufe and Eike A. Schmidt for providing the data.

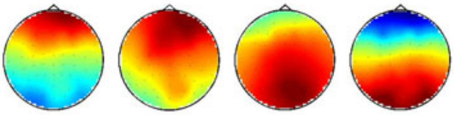


Fig. 2. EEG spatial field patterns of 4 projections of the 2D subspace spanned by the two components $\mathrm{Re}(\boldsymbol{a})$ and $\mathrm{Im}(\boldsymbol{a})$. The pattern has similarities to typical EOG / visual alpha related components.

4 Conclusion

This paper has presented a novel approach to finding common linear subspaces in data which can be modelled by matrices having common singular vectors. The strategy consists of searching in the overdetermined vector space generated by the matrices instead in the space of their transformations.

We have demonstrated the feasibility of the AROFAC algorithm that finds single eigenvectors by identifying common rank one components in this vector space instead of jointly diagonalizing the matrices. The novel approach may be preferable over joint diagonalization if the matrices are rank deficient, and only few eigenvectors which are present in most signals are of interest - in which case searching the space spanned by the matrices is the natural domain for computation. Note that computing only in the span of data is an ubiquitous concept in mathematics in general and signal processing, and kernel methods in particular (e.g.[13]).

We would like to remark that at this point the proposed novel framework and the AROFAC algorithm are mainly conceptual. However, a generalization to tensor factorization seems straightforward. It is also clear that a number of theoretical and practical questions had to remain unanswered in this first contribution. Improvements in numerical efficiency, studies of robustness, an iterative deflation mode of AROFAC and applications in the sciences and industry are still to come.

Concluding, we believe that working directly in the vector space of matrices is worth considering whenever it carries a simple intrinsic structure and thus developing optimization strategies for this domain may be of practical importance for both the signal processing and the machine learning communities.

Appendix

The AROFAC Loss Function

In the following, we will derive the loss function

$$L(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \boldsymbol{M}(\lambda)_{ii} \boldsymbol{M}(\lambda)_{jj} - \boldsymbol{M}(\lambda)_{ij} \boldsymbol{M}(\lambda)_{ji} \right\|^2 .$$

applied in the AROFAC algorithm. This loss function can be seen as a special case of a rank one indicator for arbitrary matrices $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ given by

$$L(\boldsymbol{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \|\boldsymbol{A}_{ii}\boldsymbol{A}_{jj} - \boldsymbol{A}_{ij}\boldsymbol{A}_{ji}\|^2 .$$

We claim that $L(\boldsymbol{A}) = 0$ if and only if $\boldsymbol{A}$ has rank at most one.

Proof: The summands $\|\boldsymbol{A}_{ii}\boldsymbol{A}_{jj} - \boldsymbol{A}_{ij}\boldsymbol{A}_{ji}\|^2$ are real, non-negative integers. Moreover, they are zero if and only if

$$\boldsymbol{A}_{ii}\boldsymbol{A}_{jj} - \boldsymbol{A}_{ij}\boldsymbol{A}_{ji} = 0$$

In particular, $L(\boldsymbol{A})$ is zero if and only if this holds for all i, j. If $\boldsymbol{A}_{jj}$ and $\boldsymbol{A}_{ji}$ are non-zero, the above condition is equivalent to

$$\frac{\boldsymbol{A}_{ii}}{\boldsymbol{A}_{ji}} = \frac{\boldsymbol{A}_{ij}}{\boldsymbol{A}_{jj}}$$

which forces the ratios between elements in fixed rows resp. columns to be constant, thus if $\boldsymbol{A}$ has no zero entries, it is of rank at most one if and only if $L(\boldsymbol{A}) = 0$. The homogenous equation enforces the same condition on the rows resp. columns when we have zero entries, but is always defined, as there are no denominators. So $L(\boldsymbol{A}) = 0$ if and only if $\boldsymbol{A}$ has at most rank one.

In order to reduce the complex to a real optimization problem, one sets $\boldsymbol{A} = \boldsymbol{M}(\lambda)$ and optimizes with respect to the real and imaginary parts of λ. $L(\boldsymbol{A})$ is a polynomial in those, so it is a smooth real function, and its gradient, which can be explicitly obtained by an elementary calculation, also is smooth. Note that in general $L(\boldsymbol{A})$ is not convex, since it already has r absolute minima in the noise-free case, as was stated in the main part of the paper.

Acknowledgement. This work was supported by the "Bernstein Focus: Neurotechnology" Berlin (BFNT-B 01GQ0850), by the German Science Foundation (DFG, contract MU 987/11-1), and by the PASCAL2 European Network of Excellence (ICT-216886). This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein.

References

1. Bießmann, F., Meinecke, F.C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N.K., Müller, K.R.: Temporal kernel CCA and its application in multimodal neuronal data analysis. Machine Learning 79(1-2), 5–27 (2010)
2. von Bünau, P., Meinecke, F.C., Király, F.J., Müller, K.R.: Finding stationary subspaces in multivariate time series. Phys. Rev. Lett. 103(21), 214101 (2009)
3. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non Gaussian signals. IEE - Proceedings -F 140(6), 362–370 (1993)

4. Cardoso, J.F., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM Journal on Matrix Analysis and Applications 17(1), 161–164 (1996)
5. van Der Veen, A.J., Paulraj, A.: An analytical constant modulus algorithm. IEEE Trans. Signal Processing 44(5), 1–19 (1996)
6. Lathauwer, L.D.: A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. SIAM J. Matrix Analysis Applications 28(3), 642–666 (2006)
7. Müller, K.R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., Blankertz, B.: Machine learning for real-time single-trial analysis: From brain-computer interfacing to mental state monitoring. Journal of Neuroscience Methods 167, 82–90 (2008)
8. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Mathematics of Computation 35(151), 773–782 (1980)
9. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology 115(10), 2292–2307 (2004), http://www.ncbi.nlm.nih.gov/pubmed/15351371
10. Nolte, G., Meinecke, F.C., Ziehe, A., Müller, K.R.: Identifying interactions in mixed and noisy complex systems. Phys. Rev. E 73, 051913 (2006), http://link.aps.org/doi/10.1103/PhysRevE.73.051913
11. Nolte, G., Ziehe, A., Meinecke, F., Müller, K.-R.: Analyzing coupled brain sources: Distinguishing true from spurious interaction. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems, vol. 18, pp. 1027–1034. MIT Press, Cambridge (2006)
12. Pham, D.T.: Joint approximate diagonalization of positive definite matrices. SIAM J. on Matrix Anal. and Appl. 22, 1136–1152 (2001)
13. Schölkopf, B., Smola, A.J., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10(5), 1299–1319 (1998)
14. van der Veen, A.: Joint diagonalization via subspace fitting techniques. In: Proc. ICASSP, vol. 5 (2001)
15. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. Signal Processing 50(7), 1545–1553 (2002)
16. Yeredor, A.: On using exact joint diagonalization for noniterative approximate joint diagonalization. IEEE Signal Processing Letters 12(9), 645–648 (2005)
17. Ziehe, A., Laskov, P., Nolte, G., Müller, K.R.: A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. Journal of Machine Learning Research 5, 777–800 (2004)

Bayesian Non-negative Matrix Factorization with Learned Temporal Smoothness Priors

Mathieu Coïc and Juan José Burred

Audionamix
114, avenue de Flandre
75019 Paris, France
{firstname.{middlename.}lastname}@audionamix.com

Abstract. We combine the use of a Bayesian NMF framework to add temporal smoothness priors, with a supervised prior learning of the smoothness parameters on a database of solo musical instruments. The goal is to separate main instruments from realistic mono musical mixtures. The proposed learning step allows a better initialization of the spectral dictionaries and of the smoothness parameters. This approach is shown to outperform the separation results compared to the unsupervised version.

1 Introduction

Non-negative matrix factorization (NMF) is a well-known signal decomposition technique frequently used for sound source separation. NMF decomposes a spectrogram into a set of spectral bases, each one multiplied by a time-varying weight. When dealing with musical mixtures, it is possible to exploit the specific properties of musical instruments, such as the typical temporal evolution of their spectral bases.

One way of integrating such a priori information is by using statistical priors in a Bayesian statistical framework. This was the approach used to force temporal smoothness in [1], and both temporal smoothness and harmonicity in [2]. Another option is to use supervised methods and perform a prior learning based on a database of isolated instrumental sounds. An example of this second approach is the work presented in [3], where NMF is combined with a pre-trained Hidden Markov Model (HMM) to model dynamic behavior.

In this contribution, we use a combination of both Bayesian priors and database learning to model temporal smoothness and improve separation quality. The goal is to extract the lead instrument from realistic mono musical mixtures. In particular, our system is based on a Bayesian NMF model with temporal smoothness priors described by Inverse Gamma (IG) distributions (Sect. 2), as was done in [1,2]. Here, we extend such approach by introducing a learning stage, which is based on performing NMF optimization on isolated instruments with the IG parameters as additional optimization parameters (Sect. 3). We evaluate the performance with 4 different instruments, and for all settings (with or without

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 280–287, 2012.

priors, with or without learning), we compare the performance of two possible implementations of NMF optimization, one based on Multiplicative Updates (NMF-MU), and one based on Expectation-Maximization (NMF-EM).

2 Unsupervised Algorithms

2.1 NMF Framework

The input signal is first transformed into the time-frequency domain by means of a Short Time Fourier Transform (STFT), yielding a matrix $\mathbf{X}$. As in [1], the squared modulus of each element is computed to obtain a matrix of power spectral densities $\mathbf{V} = |\mathbf{X}|^{\circ 2}$. The goal of NMF is to find the non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

$\mathbf{W}$ and $\mathbf{H}$ have dimensions $F \times K$ and $K \times N$, respectively, and it is desirable that $F \times K + K \times N \ll FN$. The rows of $\mathbf{H}$ are usually called *activations* and the columns of $\mathbf{W}$ *atoms* or *bases*.

Such factorization is formulated here as the minimization problem

$$\{\mathbf{W}, \mathbf{H}\} = \underset{\mathbf{W},\mathbf{H} \geq 0}{\operatorname{argmin}}\, D_{IS}(\mathbf{V}|\mathbf{WH}), \tag{2}$$

where D_{IS} is a matrix cost function involving the Itakura-Saito element-wise divergence d_{IS}:

$$D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d_{IS}(\mathbf{V}_{(f,n)}|[\mathbf{WH}]_{(f,n)}). \tag{3}$$

The IS divergence, defined as

$$d_{IS}(x \mid y) = \frac{x}{y} - \log \frac{x}{y} - 1, \tag{4}$$

is a good measure for the perceptual difference between two spectra, which is explained by its scale invariance: $d_{IS}(\gamma x|\gamma y) = d_{IS}(x|y)$, for a given scalar γ.

It can be shown [1] that the above optimization (Eq. 2) is equivalent to a Maximum Likelihood (ML) estimation if the columns of the STFT matrix $\mathbf{X}$, denoted by $\mathbf{x}_n$, are supposed to be generated by a K-component Gaussian Mixture Model (GMM):

$$\mathbf{x}_n = \sum_{k=1}^{K} \mathbf{c}_{kn} \in \mathbb{C}^F, \quad \forall n = 1, ..., N, \tag{5}$$

where latent variables $\mathbf{c}_{kn}$ are independent and follow a zero-mean multivariate normal distribution $\mathbf{c}_{kn} \sim \mathcal{N}(0, h_{kn}\text{diag}(\mathbf{w}_k))$, where h_{kn} are the elements of the activation matrix $\mathbf{H}$ and $\mathbf{w}_k$ are the columns of the dictionary matrix $\mathbf{W}$. The separation process consists in optimizing the criterion $C_{ML}(\boldsymbol{\theta}) \overset{\triangle}{=} \log p(\mathbf{V} \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}\}$ is the parameter vector.

We implement and test two NMF algorithms, one based on Multiplicative Update rules (NMF-MU), and one based on an EM algorithm (NMF-EM). They mainly differ in their speed of convergence to a global solution and in computational performance. The first one was used in [1], and the second one in [2], and both can be adapted to a Bayesian setting.

2.2 NMF-MU Algorithm

Multiplicative Update (MU) rules to iteratively find the optimal $\mathbf{W}$ and $\mathbf{H}$ are given for the IS divergence [4] by

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T \left((\mathbf{WH})^{\circ[-2]} \circ \mathbf{V} \right)}{\mathbf{W}^T (\mathbf{WH})^{\circ[-1]}}, \tag{6}$$

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\left((\mathbf{WH})^{\circ[-2]} \circ \mathbf{V} \right) \circ \mathbf{H}^T}{(\mathbf{WH})^{\circ[-1]} \mathbf{H}^T}, \tag{7}$$

where the $\circ$ symbol denotes element-wise operations, and the division is also element-wise.

2.3 NMF-EM Algorithm

An alternative to MU is to directly perform an ML estimation of the generative model of Eq. 5 via an EM algorithm. In particular, the Space Alternating Generalized EM (SAGE) algorithm [1] is a type of EM algorithm that allows to update large parameter matrices in separate chunks, with fast convergence properties. In particular, we aim at estimating separately the parameters $\mathbf{C}_k = (\mathbf{c}_{k1}, ..., \mathbf{c}_{kN})$. If we partition the parameter space by $\boldsymbol{\theta} = \bigcup_{k=1}^{K} \boldsymbol{\theta}_k$ where $\boldsymbol{\theta}_k = \{\mathbf{w}_k, \mathbf{h}_k\}$, SAGE consists in choosing for each subset $\boldsymbol{\theta}_k$ a *hidden-data space* which is complete for this particular subset, i.e. $\boldsymbol{\theta}_k = \mathbf{C}_k$. The resulting algorithm to estimate $\mathbf{W}$ and $\mathbf{H}$ is defined in detail in [1].

2.4 Bayesian NMF with Temporal Smoothness Prior

The Bayes rule allows to switch from a ML estimation to a Maximum A Posteriori (MAP) estimation. We can thus introduce the prior distributions $p(\mathbf{W})$ and $p(\mathbf{H})$ in this manner:

$$p\left(\mathbf{W}, \mathbf{H} \mid \mathbf{V}\right) = \frac{p\left(\mathbf{V} \mid \mathbf{W}, \mathbf{H}\right) p(\mathbf{W}) p(\mathbf{H})}{p(\mathbf{V})}. \tag{8}$$

In the case of temporal modeling, $p(\mathbf{H})$ is the relevant prior. MAP estimation is obtained by maximizing the following criterion:

$$C_{MAP}(\boldsymbol{\theta}) \triangleq \log p(\boldsymbol{\theta} \mid \mathbf{V}) \overset{c}{=} C_{ML}(\boldsymbol{\theta}) + \log p(\mathbf{H}), \tag{9}$$

where the binary operator $\overset{c}{=}$ denotes equality up to an additive constant.

Based on the MAP estimator, [1] and [2] propose a Markov chain prior structure to model $p(\mathbf{H})$:

$$p(h_k) = p(h_{k1}) \prod_{n=2}^{N} p(h_{kn} \mid h_{k,n-1}). \tag{10}$$

The main objective is to assure smoothness over the rows of $\mathbf{H}$. With an appropriate choice of the Markov transition matrix, we can favor a slow variation of $\mathbf{h}_k$. For example, we can force $p\left(h_{kn} \mid h_{k,n-1}\right)$ reach its maximum at $p(h_{k,n-1})$. The authors propose:

$$p\left(h_{kn} \mid h_{k,n-1}\right) = \mathcal{IG}\left(h_{kn} \mid \alpha_k, (\alpha_k + 1) h_{k,n-1}\right), \tag{11}$$

where $\mathcal{IG}(x \mid \alpha, \beta)$ is the inverse-Gamma distribution[1] with mode $\frac{\beta}{\alpha+1}$ and the initial distribution $p(h_{k1})$ is Jeffrey's non-informative prior: $p(h_{k1}) \propto \frac{1}{h_{k1}}$. Hence, α_k is a parameter that controls the degree of smoothness for the k-th component.

Note that we can have different smoothness parameters for each component. Thus, the smoothness parameter is actually a vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$. In practice, we want to set a smoothness prior only to those components that are supposed to describe the lead instrument. If we assign the first K_s components to the lead instrument, and the remaining ones to the accompaniment, then the α_k priors apply only to $1 \leq k \leq K_s$, and no priors are used for $K_s < k \leq K$.

The priors can be added to both NMF-MU [2] and NMF-EM [1] algorithms, as follows:

- **NMF-MU/IG algorithm.** Eq. (9) gives the following new update rules for $\mathbf{H}$, that replace Eq. 6:

$$h_{k1} \leftarrow h_{k1} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{f1} w_{fk}}{\hat{v}_{f1}^2} + \frac{\alpha_k + 1}{h_{k1}}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{f1}} + \frac{\alpha_k + 1}{h_{k2}}} \right)^{\eta} \tag{12}$$

$$h_{kn} \leftarrow h_{kn} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{fn} w_{fk}}{\hat{v}_{fn}^2} + \frac{(\alpha_k + 1) h_{n-1}}{h_{kn}^2}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fn}} + \frac{1}{h_{kn}} + \frac{\alpha_k + 1}{h_{k,n+1}}} \right)^{\eta} \tag{13}$$

$$h_{kN} \leftarrow h_{kN} \times \left(\frac{\sum_{f=1}^{F} \frac{v_{fN} w_{fk}}{\hat{v}_{fN}^2} + \frac{(\alpha_k + 1) h_{N-1}}{h_{kN}^2}}{\sum_{f=1}^{F} \frac{w_{fk}}{\hat{v}_{fN}} + \frac{\alpha_k + 1}{h_{kN}}} \right)^{\eta}, \tag{14}$$

  where $\eta \in]0, 1]$ plays the role of the step size in gradient descent.

- **NMF-EM/IG algorithm.** To integrate the temporal smoothness prior into NMF-EM, the best way is to add a post estimation after each update, computed as follows:

[1] $\mathcal{IG}\left(x \mid \alpha, \beta\right) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right)$

Table 1. Coefficients for the post estimation of h_{kn} in NMF-EM/IG

	p_2	p_1	p_0
h_{k1}	$\frac{\alpha_k+1}{h_{k2}}$	$F - \alpha_k + 1$	$-F\hat{h}_{k1}$
h_{kn}	$\frac{\alpha_k+1}{h_{kn+1}}$	$F + 1$	$-F\hat{h}_{kn} - (\alpha_k + 1)\, h_{k,n-1}$
h_{kN}	0	$F + \alpha_k + 1$	$-F\hat{h}_{kN} - (\alpha_k + 1)\, h_{k,N-1}$

$$h_{kn} = \frac{\sqrt{p_1^2 - 4p_2p_0} - p_1}{2p_2}, \tag{15}$$

where the coefficients p_0, p_1 and p_2 depend on n and are given in Table 1. In a more recent work [5], a simpler procedure, leading to a better-posed optimization problem and based on Majorization-Minimization (MM), has been proposed as an alternative to a Bayesian EM approach as described above. In the present paper, we use EM as proposed in [1], and will explore the MM alternative in the future.

3 Supervised Algorithms

The smoothness priors α_k defined in the previous section need to be set by hand prior to separation, and remain fixed throughout the optimization process. Furthermore, it would be too cumbersome to find good manual parameters for the individual priors of components with indices $1 \leq k \leq K_s$. Thus, an improvement of separation quality is expected if the α_ks are automatically learned from a training database of isolated instrumental excerpts.

We implement learning by considering the smoothness vector $\boldsymbol{\alpha}$ as an additional parameter to optimize, obtaining the new parameter vector

$$\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\alpha}\}. \tag{16}$$

A MAP estimation (Eq. 9) is performed on an audio file containing concatenated solo excerpts. We keep the estimated dictionary matrix $\hat{\mathbf{W}}$ and the smoothness vector $\hat{\boldsymbol{\alpha}}$ obtained in this way, and use them to initialize the MAP estimation performed on the mixture for actual separation.

The new update rule for the α_k coefficients is derived via ML estimation given the IG Markov chain from Eqs. 10 and 11. The log-likelihood is given by

$$\begin{aligned}
\log\left(p(h_k)\right) &\overset{c}{=} \log\left(\frac{1}{h_{k1}}\right) + \sum_{n=2}^{N} \alpha_k \log((\alpha_k + 1 h_{k,n-1})) - \log(\Gamma(\alpha k)) \qquad (17)\\
&\quad - (\alpha_k + 1)\log(h_{kn}) - \frac{\alpha_k h_{k,n-1}}{h_{kn}} - \frac{h_{k,n-1}}{h_{kn}}\\
&\overset{c}{=} \alpha_k(\log(h_{k1}) - \log(h_{kn})) - \log(h_{k1}) + \sum_{n=2}^{N} \alpha_k \log(\alpha_k + 1) \qquad (18)\\
&\quad - \log(\Gamma(\alpha_k)) - \log(h_{kn}) - \frac{\alpha_k h_{k,n-1}}{h_{k,n}} - \frac{h_{k,n-1}}{h_{kn}}.
\end{aligned}$$

Minimizing the ML criterion gives:

$$
\begin{aligned}
&\frac{\partial \log(p(h_k))}{\partial \alpha_k} = 0 \\
&\Leftrightarrow \ \log\left(\frac{h_{k1}}{h_{kN}}\right) + \sum_{n=2}^{N} \log(\alpha_k + 1) + \frac{\alpha_k}{\alpha_k + 1} - \psi(\alpha_k) - \frac{h_{k,n-1}}{h_{kn}} = 0 \\
&\Leftrightarrow \ \log(\alpha_k + 1) + \frac{\alpha_k}{\alpha_k + 1} - \psi(\alpha_k) = \frac{1}{N-1}\left(\log\left(\frac{h_{kN}}{h_{k1}}\right) + \sum_{n=2}^{N} \frac{h_{k,n-1}}{h_{kn}}\right),
\end{aligned}
\tag{19}
$$

where ψ is the digamma function defined as: $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Since Eq. 19 has no closed-form solution, the estimation of the current α_k is computed numerically.

For separation, we assign again the first K_s components to the main instrument. Thus, learned vector $\hat{\boldsymbol{\alpha}}$ applies only to the first K_s components of the matrix $\mathbf{H}$ is separation, and the first K_s columns of $\mathbf{W}$ are equal to the learned dictionary $\hat{\mathbf{W}}$.

4 Evaluation

For learning, 4 instruments from the RWC musical instrument sound database [6] were used. The instruments chosen were saxophone, trumpet, classical guitar and piano. The saxophone and the trumpet are melodic instruments which usually play only one note at a time. The piano and the guitar are polyphonic instruments that can play several notes at a time (although the guitar will mostly play individual notes when doing a solo). Furthermore, saxophone and trumpet are sustained instruments (the notes can be held as long as the breathing of the player allows), whereas piano and guitar are non-sustained instruments with note energy always decaying after the onset. Thus, the smoothness parameters over the rows of $\mathbf{H}$ are expected to be quite different between both kinds of instruments.

To evaluate separation, 18 mixes were created from songs available in multitrack and featuring solos by those instruments[2]. For each song, one mono track was created for the solo, and one mono track containing all the remaining instruments (accompaniment).

For objective evaluation in terms of Source to Distortion Ratio (SDR), we use the BSS_EVAL toolbox [7]. After separation into K NMF components, it is still necessary to assign the components to one of the sources. For evaluation purposes, SDR is measured between each component and the original tracks. The higher SDR determines if the component represents the solo or the accompaniment.

We evaluate both NMF-MU and NMF-EM algorithms, with or without priors, and in both unsupervised and supervised versions. In the unsupervised version,

[2] Source: `ccmixter.org`

Table 2. Average SDR (in dB)

500 iterations		Unsupervised		Supervised	
		Without priors	IG	Without priors	IG
Saxophone	NMF-MU	10.19	9.12	**10.47**	9.66
	NMF-EM	7.14	6.76	7.01	7.58
Trumpet	NMF-MU	6.16	4.80	6.39	**7.88**
	NMF-EM	3.63	3.98	4.55	5.06
Classic Guitar	NMF-MU	9.84	8.75	8.88	**10.07**
	NMF-EM	7.38	7.52	8.01	6.52
Piano	NMF-MU	**6.99**	4.73	5.44	6.95
	NMF-EM	3.11	3.98	2.97	2.08
Global	NMF-MU	8.30	6.85	7.80	**8.64**
	NMF-EM	5.32	5.56	5.64	5.31

the parameters are initialized randomly except for the smoothness parameters, which are fixed empirically. In that case, we set the same smoothness value for all α_k. In the supervised case, the learned parameters are then used in the separation process to initialize the system, as explained in Sect 3. Note that in the supervised version without smoothness priors, the dictionary $\mathbf{W}$ is learned anyway.

Results are given in Table 2. The following conclusions can be drawn:

- NMF-MU algorithms perform in general better than NMF-EM algorithms.
- In unsupervised algorithms, using the IG smoothness priors is not efficient. This is probably due to the difficulty of manually finding good values for $\hat{\boldsymbol{\alpha}}$.
- Supervised algorithms outperform unsupervised algorithms, except in the case of the piano, in which case the maximum performance is virtually the same. This might indicate that the IG distribution is not well suited to describe the dynamics of the piano spectra.
- In supervised algorithms, using the smoothness priors improves performance, except for the saxophone.

A selection of sound examples can be found online[3].

5 Conclusions and Perspectives

We have proposed a learning stage for the IG temporal smoothness priors within an NMF Bayesian framework for separation of main instruments in mono mixtures. An evaluation of the different system configurations was performed, including supervised and unsupervised versions, with or without priors, and both in MU and EM implementations, for sustained (trumpet and saxophone) and non-sustained (piano and guitar) instruments. Supervised approaches are shown

[3] `http://audionamix.com/BayesianNMF1/`

to perform better than unsupervised ones, except in the case of the piano. Globally, the MU versions of the algorithms perform better.

A refinement of the temporal priors will be subject to further study. In particular, a temporal description will probably benefit from a structured representation considering the attack and sustain parts separately. Also, other prior distributions will be investigated to improve difficult cases, such as the piano. Finally, other instrument-specific priors, such as spectral smoothness or harmonicity, might also be taken into account in order to further improve separation quality.

References

1. Févotte, C., Bertin, N., Durrieu, J.-L.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation 21(3), 793–830 (2009)
2. Bertin, N., Badeau, R., Vincent, E.: Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. IEEE Transactions on Audio, Speech and Language Processing 18(3), 538–549 (2010)
3. Mysore, G.J., Smaragdis, P.: A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In: Proc. IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP), Prague, Czech Republic (May 2011)
4. Cichocki, A., Zdunek, R., Amari, S.-i.: Csiszár's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms. In: Rosca, J.P., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 32–39. Springer, Heidelberg (2006)
5. Févotte, C.: Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization. In: Proc. IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP), Prague, Czech Republic (May 2011)
6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC Music Database: Music genre database and musical instrument sound database. In: Proc. of the 4th International Conference on Music Information Retrieval (ISMIR), pp. 229–230 (2003)
7. Févotte, C., Gribonval, R., Vincent, E.: BSS_EVAL toolbox user guide. IRISA Technical Report 1706, Rennes, France (2005)

On Connection between the Convolutive and Ordinary Nonnegative Matrix Factorizations

Anh Huy Phan[1], Andrzej Cichocki[1,⋆], Petr Tichavský[2,⋆⋆], and Zbyněk Koldovský[3]

[1] Lab for Advanced Brain Signal Processing, Brain Science Institute - RIKEN, Japan
[2] Institute of Information Theory and Automation, Prague, Czech Republic
[3] Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic

Abstract. A connection between the convolutive nonnegative matrix factorization (NMF) and the conventional NMF has been established. As a result, we can convey arbitrary alternating update rules for NMF to update rules for CNMF. In order to illustrate the novel derivation method, a multiplicative algorithm and a new ALS algorithm for CNMF are derived. The experiments confirm validity and high performance of our method and of the proposed algorithm.

Keywords: nonnegative matrix factorization, convolutive nonnegative matrix factorization, nonnegative quadratic programming, ALS, music analysis.

1 Introduction

Expression of a nonnegative data matrix by set of basis patterns (objects) shifting along a direction (horizontal or vertical) of a given data following the convolutive model has recently attracted considerable interest from the view point of applications such as music analysis, image deconvolution [1–5,9,11]. This decomposition model is called the convolutive nonnegative matrix factorization (CNMF), and is considered as an extension of nonnegative matrix factorization (NMF). While there is a vast literature on algorithms for NMF [2], algorithms for CNMF are still very limited in the literature. All existing CNMF algorithms are based on the multiplicative update rules which minimize the least-squares error [1,6,9] or the Kullback-Leiber divergence [1,4], or the generalized alpha- or beta- divergences [2,3]. We note that the multiplicative algorithms have a relatively low complexity of each iteration but they are characterized by rather slow convergence and they sometimes converge to spurious local minima [7,8].

Blind deconvolution of a given nonnegative data $\mathbf{Y} \in \mathbb{R}_+^{I\times J}$ is to find P basis patterns (objects) $\mathbf{A}^{(p)} = [\boldsymbol{a}_1^{(p)}, \boldsymbol{a}_2^{(p)}, \ldots, \boldsymbol{a}_{R_p}^{(p)}] \in \mathbb{R}_+^{I\times R_p}$, $p = 1, 2, \ldots, P$ and a location matrix $\mathbf{X} \in \mathbb{R}_+^{P\times J}$, each p-th row vector $\boldsymbol{x}_{p:}$ representing location and intensity of $\mathbf{A}^{(p)}$. For simplicity, assuming that all basis patterns $\mathbf{A}^{(p)}$ have the same size $R_p = R, \forall p$, otherwise they can be padded with zeros to the right. P basis patterns $\mathbf{A}^{(p)}$ are lateral slices of a

[⋆] Also affiliated with the EE Dept., Warsaw University of Technology and with Systems Research Institute, Polish Academy of Science, Poland.

[⋆⋆] The work of P. Tichavsky was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project 102/09/1278.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 288–296, 2012.

3-D tensor $\mathcal{A} \in \mathbb{R}^{I \times P \times R}$, i.e., $\mathcal{A}(i, p, r) = \mathbf{A}^{(p)}(i, r)$, $i = 1, \ldots, I$, $r = 1, \ldots, R$. Frontal slices $\mathbf{A}_r = [\boldsymbol{a}_r^{(1)}\ \boldsymbol{a}_r^{(2)}\ \cdots\ \boldsymbol{a}_r^{(P)}] \in \mathbb{R}^{I \times P}$ are component matrices, $r = 1, 2, \ldots, R$. The mode-1 matricized version of $\mathcal{A}$ is denoted by $\mathbf{A}_{(1)} = [\mathbf{A}_1\ \mathbf{A}_2\ \cdots\ \mathbf{A}_R] \in \mathbb{R}_+^{I \times RP}$.

We denote a shift matrix $\mathbf{S}_r$ of size $J \times J$ which is a binary matrix with ones only on the r-th superdiagonal for $r > 0$, or on the r-th subdiagonal for $r < 0$, and zeroes elsewhere. $\overset{r\rightarrow}{\mathbf{X}} = \mathbf{X}\mathbf{S}_r$ is an r column shifted version of $\mathbf{X}$ to the right, with the columns shifted in from outside the matrix set to zero. The relation between $\mathbf{Y}$, $\mathbf{A}^{(p)}$ and $\mathbf{X}$ can be expressed as

$$\mathbf{Y} = \sum_{r=1}^{R} \mathbf{A}_r\, \mathbf{X}\, \mathbf{S}_{r-1} + \mathbf{E} = \sum_{r=0}^{R-1} \mathbf{A}_{r+1} \overset{r\rightarrow}{\mathbf{X}} + \mathbf{E}. \tag{1}$$

Most CNMF algorithms were derived by considering (1) as R NMFs [3,4,6,9]

$$\mathbf{Y} = \mathbf{A}_{r+1} \overset{r\rightarrow}{\mathbf{X}} + \left(\sum_{s \neq r} \mathbf{A}_{s+1} \overset{s\rightarrow}{\mathbf{X}} \right) + \mathbf{E} = \mathbf{A}_{r+1} \overset{r\rightarrow}{\mathbf{X}} + \mathbf{E}_{r+1}, \quad r = 0, 1, \ldots, R-1. \tag{2}$$

For example, the multiplicative algorithms [3,9] update $\mathbf{A}_r$ and $\mathbf{X}_r$

$$\mathbf{A}_{r+1} \leftarrow \mathbf{A}_{r+1} \circledast \left(\overset{r\leftarrow}{\mathbf{Y}}\, \mathbf{X}^T \right) \oslash \left(\overset{r\leftarrow}{\widehat{\mathbf{Y}}}\, \mathbf{X}^T \right), \quad r = 0, 1, \ldots, R-1, \tag{3}$$

$$\mathbf{X}_{r+1} \leftarrow \mathbf{X} \circledast \left(\mathbf{A}_{r+1}^T \overset{r\leftarrow}{\mathbf{Y}} \right) \oslash \left(\mathbf{A}_{r+1}^T \overset{r\leftarrow}{\widehat{\mathbf{Y}}} \right), \tag{4}$$

where symbols "$\circledast$" and "$\oslash$" denote the Hadamard element-wise product and division. The coding matrix $\mathbf{X}$ is averaged over R estimations $\mathbf{X}_r$ in (4), i.e., $\mathbf{X} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{X}_r$. Although the approach is simple and quite direct, its average update rule for $\mathbf{X}$ is not optimal. The reason is that the factorization (2) does not consider other shifts $\overset{s\rightarrow}{\mathbf{X}} (s \neq r)$ existing in $\mathbf{E}_{r+1}$. Moreover, practical simulations show that the average rules are not stable and converge slowly.

In the sequel, we present a connection between CNMF and NMF. Based on this, an arbitrary alternating update rule for NMF can be conveyed to CNMF. In order to illustrate the novel derivation method, a multiplicative algorithm and a robust ALS algorithm for CNMF are proposed.

2 A Novel Derivation for CNMF Algorithms

In general, update rules for $\mathcal{A}$ and $\mathbf{X}$ can be derived by minimizing a cost function which can be the Frobenius norm of the approximation error

$$D(\mathbf{Y} \| \widehat{\mathbf{Y}}) = \frac{1}{2} \| \mathbf{Y} - \widehat{\mathbf{Y}} \|_F^2 = \frac{1}{2} \| \mathbf{Y} - \sum_{r=1}^{R} \mathbf{A}_r\, \mathbf{X}\, \mathbf{S}_{r-1} \|_F^2. \tag{5}$$

From (1), the approximation of $\mathbf{Y}$ can be expressed as an NMF with rank PR, that is,

$$\mathbf{Y} = \left[\mathbf{A}_1\ \mathbf{A}_2 \cdots \mathbf{A}_R \right] \begin{bmatrix} \mathbf{X} \\ \overset{1\rightarrow}{\mathbf{X}} \\ \vdots \\ \overset{(R-1)\rightarrow}{\mathbf{X}} \end{bmatrix} + \mathbf{E} = \mathbf{A}_{(1)}\, \mathbf{Z} + \mathbf{E}, \tag{6}$$

or as an approximation of $\mathrm{vec}(\mathbf{Y})$

$$\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}\left(\sum_{r=0}^{R-1} \mathbf{A}_{r+1}\, \mathbf{X}\, \mathbf{S}_r + \mathbf{E}\right) = \mathbf{F}\ \mathrm{vec}(\mathbf{X}) + \mathrm{vec}(\mathbf{E})\,, \tag{7}$$

where $\mathbf{F} = \sum_{r=0}^{R-1}\left(\mathbf{S}_r^T \otimes \mathbf{A}_{r+1}\right) \in \mathbb{R}_+^{IJ\times PJ}$, '$\otimes$' denotes the Kronecker product. From (5), (6) and (7), we can alternatively update $\mathcal{A}$ or $\mathbf{X}$ while fixing the other according to the following procedure

$$\mathbf{X} = \arg\min_{\mathbf{X}}\ \|\,\mathrm{vec}(\mathbf{Y}) - \mathbf{F}\,\mathrm{vec}(\mathbf{X})\,\|_2^2, \quad \text{subject to } \mathbf{X} \geq \mathbf{0} \text{ with fixed } \mathcal{A}, \tag{8}$$

$$\mathcal{A} = \arg\min_{\mathbf{A}_{(1)}}\ \|\mathbf{Y} - \mathbf{A}_{(1)}\mathbf{Z}\|_F^2, \quad \text{subject to } \mathcal{A} \geq \mathbf{0} \text{ with fixed } \mathbf{X}. \tag{9}$$

Note that we can employ any update rules for NMF to update $\mathcal{A}$ and $\mathbf{X}$. For example, by employing the multiplicative update rules [7] we can update $\mathcal{A}$

$$\mathbf{A}_{(1)} \leftarrow \mathbf{A}_{(1)} \circledast \left(\mathbf{Y}\,\mathbf{Z}^T\right) \oslash \left(\widehat{\mathbf{Y}}\mathbf{Z}^T\right) = \mathbf{A}_{(1)} \circledast \left[\mathbf{Y}\,\mathbf{S}_r^T\,\mathbf{X}^T\right]_{r=0}^{R-1} \oslash \left[\widehat{\mathbf{Y}}\,\mathbf{S}_r^T\,\mathbf{X}^T\right]_{r=0}^{R-1}, \tag{10}$$

which can be rewritten for component matrices $\mathbf{A}_r$, $r = 1, 2, \ldots, R$

$$\mathbf{A}_r \leftarrow \mathbf{A}_r \circledast \left(\mathbf{Y}\,\mathbf{S}_{r-1}^T\,\mathbf{X}^T\right) \oslash \left(\widehat{\mathbf{Y}}\,\mathbf{S}_{r-1}^T\,\mathbf{X}^T\right), \quad r = 1, 2, \ldots, R. \tag{11}$$

The multiplicative Least-Squares update rule for $\mathbf{X}$ is given by

$$\begin{aligned}
\mathrm{vec}(\mathbf{X}) &\leftarrow \mathrm{vec}(\mathbf{X}) \circledast \left(\mathbf{F}^T\ \mathrm{vec}(\mathbf{Y})\right) \oslash \left(\mathbf{F}^T\ \mathrm{vec}\big(\widehat{\mathbf{Y}}\big)\right) \\
&= \ \mathrm{vec}(\mathbf{X}) \circledast \mathrm{vec}\left(\sum_{r=0}^{R-1} \mathbf{A}_{r+1}^T\,\mathbf{Y}\,\mathbf{S}_r^T\right) \oslash \mathrm{vec}\left(\sum_{r=0}^{R-1} \mathbf{A}_{r+1}^T\,\widehat{\mathbf{Y}}\,\mathbf{S}_r^T\right)
\end{aligned}$$

or in the matrix form

$$\mathbf{X} \leftarrow \mathbf{X} \circledast \left(\sum_{r=0}^{R-1} \mathbf{A}_{r+1}^T\,\mathbf{Y}\,\mathbf{S}_r^T\right) \oslash \left(\sum_{r=0}^{R-1} \mathbf{A}_{r+1}^T\,\widehat{\mathbf{Y}}\,\mathbf{S}_r^T\right). \tag{12}$$

The update rules in (11) and (12) are particular cases of the multiplicative algorithm for CNMF2D [1]. However, its derivation is much simpler than that in [1]. Similarly, it is straightforward to derive update rules for the multiplicative Kullback-Leiber algorithms, the ALS algorithms. In addition, (6) and (7) also lead to condition on the number of patterns and the number of components $PR \leq min(I, J)$.

3 Alternative Least Squares Algorithm for CNMF

The alternative least squares (ALS) algorithm and its variations are commonly used for nonnegative matrix factorizations (see Chapter 4 [2]). For CNMF, it is straightforward to derive from (8) and (9) two ALS update rules given by

$$\mathbf{A}_{(1)} \leftarrow \left[\mathbf{Y}\,\mathbf{Z}^T\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1}\right]_+\,, \qquad \mathrm{vec}(\mathbf{X}) \leftarrow \left[\mathbf{Q}^{-1}\,\boldsymbol{b}\right]_+\,, \tag{13}$$

where $[x]_+ = \max(x, \varepsilon)$ is the element-wise rectifier which converts negative input to zero or a small enough value, and

$$\mathbf{Q} = \mathbf{F}^T\mathbf{F} = \sum_{r=0}^{R-1}\sum_{s=0}^{R-1}\left(\mathbf{S}_r\,\mathbf{S}_s^T \otimes \mathbf{A}_{r+1}^T\mathbf{A}_{s+1}\right) \quad \in \mathbb{R}_+^{L\times L}, L = JP, \tag{14}$$

$$\boldsymbol{b} = \mathbf{F}^T\,\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}\left(\sum_{r=0}^{R-1}\mathbf{A}_{r+1}^T\,\mathbf{Y}\,\mathbf{S}_r^T\right) \in \mathbb{R}^L. \tag{15}$$

Although the ALS algorithm (13) is simple, it is not stable for sparse nonnegative data as illustrated in Section 5 for decomposition of spectrogram of sound sequences. In the sequel, a robust ALS algorithm is proposed for CNMF. From (5), (8), we consider the least-squares cost function which leads to a nonnegative quadratic programming (NQP) problem

$$D = \frac{1}{2}\|\,\mathrm{vec}(\mathbf{Y}) - \mathbf{F}\;\mathrm{vec}(\mathbf{X})\,\|_2^2 = \frac{1}{2}\|\mathbf{Y}\|_F^2 + \frac{1}{2}\boldsymbol{x}^T\mathbf{Q}\boldsymbol{x} - \boldsymbol{b}^T\,\boldsymbol{x}, \tag{16}$$

where $\boldsymbol{x} = \mathrm{vec}(\mathbf{X})$.

We denote $\tilde{\boldsymbol{x}} = [\tilde{x}_1\;\tilde{x}_2\;\cdots\;\tilde{x}_L]^T = \mathbf{Q}^{-1}\,\boldsymbol{b}$ the solution of the gradient $\nabla D(\boldsymbol{x}) = \mathbf{Q}\,\boldsymbol{x} - \boldsymbol{b}$, and $\mathcal{I}_+ \subset \{1, 2, \ldots, L\}$ a set of $L_1 \le L$ nonnegative entries, i.e. $\tilde{\boldsymbol{x}}_{\mathcal{I}_+} \ge \mathbf{0}$. If $L_1 = L$, then $\tilde{\boldsymbol{x}}$ is solution of (8) as in (13). Otherwise, the $(L - L_1)$ negative entries $\boldsymbol{x}_{\mathcal{I}_-}$, $\mathcal{I}_- = \{1, 2, \ldots, L\}\backslash\mathcal{I}_+$ are set to zeros by the rectifier according to (13). Hence, from (16), the rest L_1 variables $\boldsymbol{x}_{\mathcal{I}_+}$ are solutions of a reduced problem of a lower order L_1, that is

$$D = \frac{1}{2}\|\mathbf{Y}\|_F^2 + \frac{1}{2}\boldsymbol{x}_{\mathcal{I}_+}^T\,\mathbf{Q}_{\mathcal{I}_+}\,\boldsymbol{x}_{\mathcal{I}_+} - \boldsymbol{b}_{\mathcal{I}_+}^T\,\boldsymbol{x}_{\mathcal{I}_+}, \tag{17}$$

where $\mathbf{Q}_{\mathcal{I}_+}$ and $\boldsymbol{b}_{\mathcal{I}_+}$ are parts of $\mathbf{Q}$ and $\boldsymbol{b}$ whose row and column indices are specified by $\mathcal{I}_+$, respectively. If $\tilde{\boldsymbol{x}}_{\mathcal{I}_+} = \mathbf{Q}_{\mathcal{I}_+}^{-1}\,\boldsymbol{b}_{\mathcal{I}_+}$ has $L_2 < L_1$ nonnegative entries, we solve the subproblem of (17) of the lower order L_2. The procedure is recursively applied until there is not any negative entry $\tilde{x}$, i.e. $\mathcal{I}_- = \emptyset$ (see the subfunction nqp in Algorithm 1).

Similarly, the cost function (9) can also be expressed as an NQP problem to update $\mathbf{A}_{(1)}$ or horizontal slices $\mathbf{A}_{i::}$ defined as $\mathbf{A}_{i::}(p, r) = \mathcal{A}(i, p, r)$, $i = 1, 2, \ldots, I$

$$D = \frac{1}{2}\|\mathbf{Y}\|_F^2 + \frac{1}{2}\,\mathrm{vec}\left(\mathbf{A}_{(1)}^T\right)^T\left(\mathbf{I}_I \otimes \left(\mathbf{Z}\mathbf{Z}^T\right)\right)\mathrm{vec}\left(\mathbf{A}_{(1)}^T\right) - \mathrm{vec}\left(\mathbf{Z}\mathbf{Y}^T\right)^T\,\mathrm{vec}\left(\mathbf{A}_{(1)}^T\right) \tag{18}$$

$$= \frac{1}{2}\|\mathbf{Y}\|_F^2 + \sum_{i=1}^{I}\left(\frac{1}{2}\;\mathrm{vec}(\mathbf{A}_{i::})^T\;\left(\mathbf{Z}\mathbf{Z}^T\right)\mathrm{vec}(\mathbf{A}_{i::}) - \left(\boldsymbol{y}_{i:}\;\mathbf{Z}^T\right)\mathrm{vec}(\mathbf{A}_{i::})\right), \tag{19}$$

where $\boldsymbol{y}_{i:}$ denotes the i-th row vector of $\mathbf{Y}$. Finally, pseudo code of the (Q)ALS algorithm is described in Algorithm 1.

4 Initialization for CNMF Algorithms

In general, patterns $\mathbf{A}^{(p)}$ and coding matrix $\mathbf{X}$ can be initialized by nonnegative random values over multiple runs. The final solution can be chosen among them. Practical experiments show that although this simple method can produce acceptable solution, it needs

Algorithm 1. QALS Algorithm

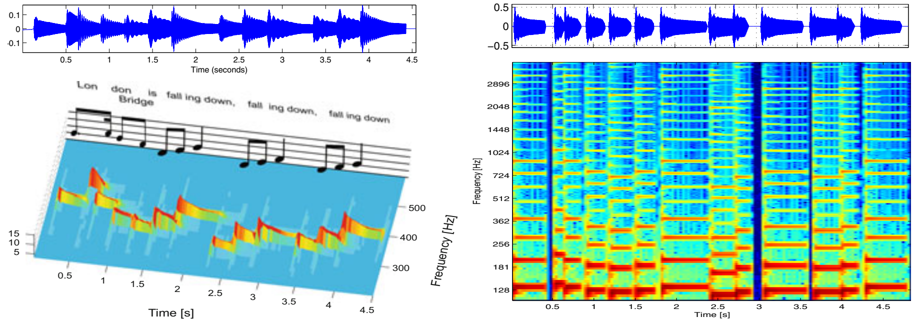
```
Input: $\mathbf{Y}$: nonnegative matrix $I \times J$
        $P, R$: number of patterns and components
Output: $\mathbf{A}^{(p)} \in \mathbb{R}_+^{I \times R}$, $p = 1, 2, \ldots, P$ and $\mathbf{X} \in \mathbb{R}_+^{R \times J}$
begin
    Initialize $\mathcal{A}$ and $\mathbf{X}$
    repeat
        $\mathrm{vec}(\mathbf{X}) = \mathrm{nqp}\,(\mathbf{Q}, \boldsymbol{b})$                                // $\mathbf{Q}$ in (14), $\boldsymbol{b}$ in (15)
        for $i = 1$ to $I$ do $\mathrm{vec}(\mathbf{A}_{i::}) = \mathrm{nqp}\left(\mathbf{Z}\mathbf{Z}^T, \mathbf{Z}\,\boldsymbol{y}_{i:}^T\right)$;        // Update $\mathcal{A}$
    until a stopping criterion is met
end

function $\boldsymbol{x} = \mathrm{nqp}\,(\mathbf{Q}, \boldsymbol{b})$                                    // $\mathbf{Q} \in \mathbb{R}_+^{L \times L}$, $\boldsymbol{b} \in \mathbb{R}^L$
begin
    $\mathcal{I}_+ = \{1, 2, \ldots, L\}$
    repeat
        $\tilde{\boldsymbol{x}}_{\mathcal{I}_+} = \mathbf{Q}_{\mathcal{I}_+}^{-1} \boldsymbol{b}_{\mathcal{I}_+}$; $\mathcal{I}_- = \{l \in \mathcal{I}_+ : \tilde{x}_l < 0\}$; $\mathcal{I}_+ = \mathcal{I}_+ \backslash \mathcal{I}_-$;
    until $\mathcal{I}_- = \emptyset$
    $\boldsymbol{x} = \max\{0, \tilde{\boldsymbol{x}}\}$
end
```

a large number of iterations and several (many) runs from different initial conditions to minimize probability of being stucked in false local minima instead of the global minimum. Noting that from the connection (6), $\mathbf{A}_{(1)}$ in approximation $\|\mathbf{Y} - \mathbf{A}_{(1)}\mathbf{Z}\|_F$ without nonnegativity constraints must comprise PR leading left singular components of $\mathbf{Y}$. Therefore, an SVD-based initialization method is proposed for $\mathbf{A}^{(p)}$ by taking in account that these leading singular components should be distributed among patterns $\mathbf{A}^{(p)}$. That is the first component matrix $\mathbf{A}_1$ takes R leading left singular components, the next R leading left components are for $\mathbf{A}_2$, and so on. Similarly, $\mathbf{X}$ can be initialized by P leading right singular vectors of $\mathbf{Y}$. Moreover, due to nonnegativity constraints, absolute values of singular vectors are used.

5 Simulations

In this section, we compare CNMF algorithms including QALS, the average multiplicative algorithm (aMLS) in (4) [9], the simultaneous multiplicative algorithm (MLS) in (12) [1] through decomposition of two music sequences into basic notes. For the first sequence, the sampled song "London Bridge" composed of five notes D4, E4, F4, G4 and A4 was played on a piano for 4.5 seconds illustrated in Fig. 1(a) (see Chapter 3 [2]). The signal was sampled at 8 kHz and filtered by using a bandpass filter with a bandwidth of $240 - 480$ Hz. The magnitude spectrogram $\mathbf{Y}$ of size 257 frequency bins × 141 time frames is shown in Fig. 1(a), in which each rising part corresponds to the note actually played. It means $\mathbf{Y}$ is very sparse.

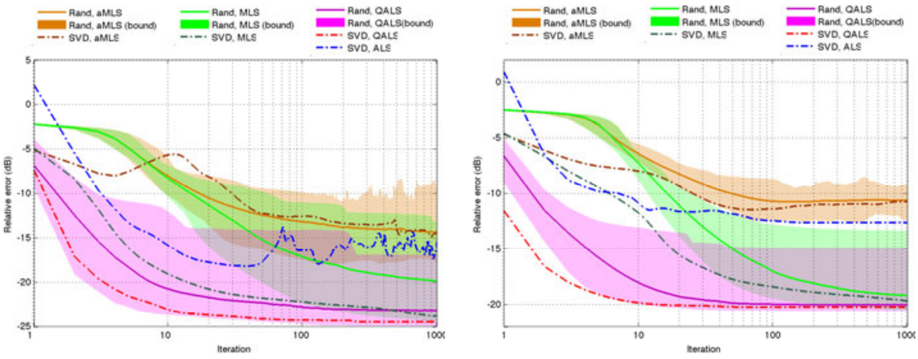
The second sequence was recorded from the same song but composed of five notes A3, G3, F3, E3 and D3 played on a guitar for 5 seconds (see Chapter 3 [2]). The log-frequency spectrogram $\mathbf{Y}$ (364×151) illustrated in Fig. 1(b) was converted from the linear-frequency spectrogram with a quality factor $Q = 100$ and in the frequency range from $f_0 = 109.4$ Hz (bin 8) to $f_I = f_s/2 = 4000$ Hz (bin 257) [10]. The lowest approximation error for this spectrum is 27.56 dB when there was no decomposition.



(a) Sequence 1

(b) Sequence 2

Fig. 1. Waveforms and spectrograms of the two sequences “London Bridge”



(a) Sequence 1

(b) Sequence 2

Fig. 2. Convergence behavior of CNMF algorithms as function of the number of iterations for decomposition of two music sequences. Min-max bounds to the relative errors are shown shaded for random initialization.

Table 1. Performance comparison for various CNMF algorithms

Algo-rithm	Sequence 1			Sequence 2		
	SNR (dB)		RTime (secs)-	SNR (dB)		RTime (secs) -
	Random	SVD-based	SNR (dB)	Random	SVD-based	SNR (dB)
aMLS	15.00 ± 1.98	15.49		11.52 ± 0.54	11.81	
MLS	19.75 ± 4.99	25.08	6.54 - 25.08	19.30 ± 2.18	19.42	13.10 - 19.42
QALS	**24.22± 3.23**	**25.74**	**1.37 - 25.14**	**20.25 ± 0.90**	**20.38**	**3.22 - 20.33**
ALS		18.12			15.07	

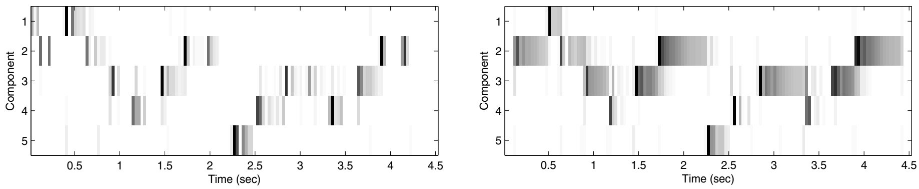
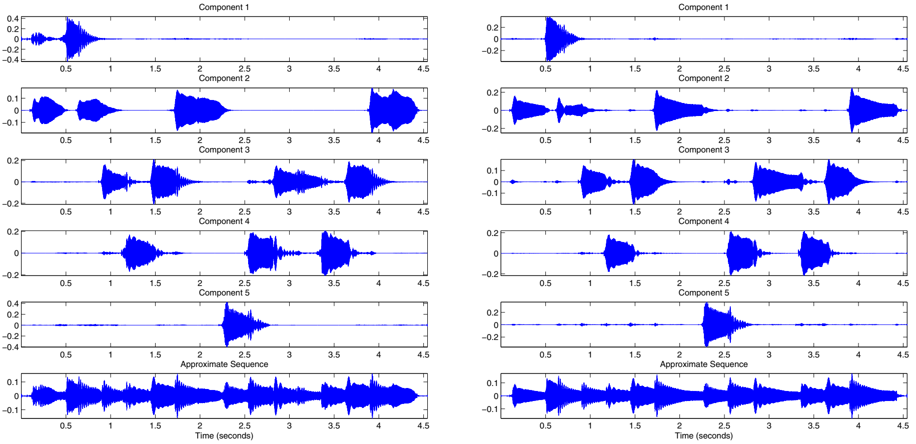


Fig. 3. Coding matrices **X** estimated by aMLS (left) and QALS (right) using SVD-based initialization, and matched with the piano roll for the sequence 1



(a) Basis and reconstructed sequences by aMLS, SNR = 15.49 dB.

(b) Basis and reconstructed sequences by QALS, SNR = 25.74 dB.

Fig. 4. Waveforms of basis spectral patterns $\mathbf{A}^{(p)}$ and the corresponding coding vectors $\boldsymbol{x}_{p:}$ estimated by aMLS and QALS. The reconstructed sequences (in the bottom) are summation of basis sequences.

For both sequences, CNMF algorithms were applied to extract 5 patterns $\mathbf{A}^{(p)} \in \mathbb{R}_+^{I\times 10}$, and to explain the observed sequence through basis audio sequences. The approximate signals were reconstructed from basis patterns, and normalized to have the same energy as the original signal. Algorithms were initialized by the nonnegative random values over 100 times or by absolute values of leading singular vectors extracted from $\mathbf{Y}$. The relative approximation errors $20\log_{10}\left(\frac{\|\mathbf{Y}-\widehat{\mathbf{Y}}\|_F}{\|\mathbf{Y}\|_F}\right)$ (dB) with different initializations are illustrated as function of the number of iterations in Fig. 2. Moreover, Table 1 provides the signal-to-noise (SNR) ratio between the original audio sequence and its approximate signal $-20\ \log\left(\frac{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|_2}{\|\boldsymbol{y}\|_2}\right)$ (dB). Running time (seconds) and SNR as the algorithm converged are given in columns 3 and 5 in Table 1, respectively.

As seen in Fig. 2, aMLS (orange shading) is not stable for both sequences, and often gets stuck in local minima by random initialization. Although SVD-based initialization can improve its performance (brown dash-dot lines), the cost values of aMLS did not always decrease. This was caused by the average rule (4) which is not optimal here. MLS which simultaneously updates $\mathbf{X}$ has better convergence (green shading) than aMLS. Among algorithms with random initialization, QALS mostly achieved the lowest approximation error (magenta shading and magenta dash lines). Moreover, QALS reached the converged values ealier, after 100 iterations, than MLS and aMLS.

Fig. 2 also indicates that SVD-based initialization improved performance compared with random initialization. QALS (dash-dot red lines) converged after 20 iterations in 1.37 seconds and in 3.22 seconds for two sequences, respectively. Whereas MLS (dash-dot green lines) run at least 1000 iterations to achieve similar approximation errors in 6.54 seconds and 13.10 seconds respectively. Running time was measured on a computing server which has 2 quadcore 3.33 GHz processors and 64 GB memory. Therefore, although complexity per iteration of QALS is higher than that of MLS, QALS may converge earlier than MLS due to significantly less computation iterations.

Fig. 3 illustrates two coding matrices $\mathbf{X}$ estimated by QALS and aMLS for the sequence 1 after matching with its piano roll. The coding map $\mathbf{X}$ by QALS is more similar to the piano roll than that of aMLS. The patterns appear continually as the notes played in the piano roll. In addition, waveforms constructed from the basis spectral patterns $\mathbf{A}^{(p)}$ and the corresponding coding row vectors $\boldsymbol{x}_{p:}$, $p = 1, \ldots, 5$, are illustrated in Fig. 4 for aMLS and QALS. aMLS achieved a reconstruction error of 15.49 dB. Whereas QALS obtained much higher performance with an approximation error of 25.74 dB. The standard ALS achieved an error of 18.20 dB. More comparisons between the algorithms are given in Table 1, which confirms the superior performance of QALS.

6 Conclusions

A connection between CNMF and NMFs is presented and allows us to straightforwardly extend arbitrary alternating NMF update rules to CNMF. The novel derivation method has been illustrated by two simple CNMF algorithms. In addition, a novel (Q)ALS algorithm is proposed and has been confirmed to give higher performance than those of the multiplicative algorithms in the sense of convergence, and reconstruction error. Moreover, based on the new connection, an SVD-based initialization method has been proposed for CNMF algorithms.

References

1. Schmidt, M.N., Mørup, M.: Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation. In: Rosca, J.P., Erdogmus, D., Principe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 700–707. Springer, Heidelberg (2006)
2. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley, Chichester (2009)
3. O'Grady, P.D., Pearlmutter, B.A.: Discovering speech phones using convolutive non-negative matrix factorization with a sparseness constraint. Neurocomput. 72, 88–101 (2008)
4. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Transactions on Audio, Speech and Language Processing 15(1), 1–12 (2007)

5. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures with application to blind audio source separation. In: ICASSP 2009, USA, pp. 3137–3140 (2009)
6. Wang, W., Cichocki, A., Chambers, J.A.: A multiplicative algorithm for convolutive nonnegative matrix factorization based on squared Euclidean distance. IEEE Transactions on Signal Processing 57(7), 2858–2864 (2009)
7. Lee, D.D., Seung, H.S.: Algorithms for Nonnegative Matrix Factorization, vol. 13. MIT Press (2001)
8. Berry, M., Browne, M., Langville, A., Pauca, P., Plemmons, R.: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis 52(1), 155–173 (2007)
9. Smaragdis, P.: Non-Negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 494–499. Springer, Heidelberg (2004)
10. Ellis, D.: Spectrograms: Constant-q (log-frequency) and conventional (linear) (May 2004), http://labrosa.ee.columbia.edu/matlab/sgram/
11. FitzGerald, D., Cranitch, M., Coyle, E.: Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation. In: Computational Intelligence and Neuroscience, vol. 2008, Article ID 872425, 15 pages (2008), doi:10.1155/2008/872425

On Revealing Replicating Structures in Multiway Data: A Novel Tensor Decomposition Approach

Anh Huy Phan[1], Andrzej Cichocki[1,⋆], Petr Tichavský[2,⋆⋆],
Danilo P. Mandic[3,⋆⋆⋆], and Kiyotoshi Matsuoka[4]

[1] Brain Science Institute - RIKEN, Japan
[2] Institute of Information Theory and Automation, Czech Republic
[3] Imperial College, London, United Kingdom
[4] Kyushu University of Technology, Japan

Abstract. A novel tensor decomposition is proposed to make it possible to identify replicating structures in complex data, such as textures and patterns in music spectrograms. In order to establish a computational framework for this paradigm, we adopt a multiway (tensor) approach. To this end, a novel tensor product is introduced, and the subsequent analysis of its properties shows a perfect match to the task of identification of recurrent structures present in the data. Out of a whole class of possible algorithms, we illuminate those derived so as to cater for orthogonal and nonnegative patterns. Simulations on texture images and a complex music sequence confirm the benefits of the proposed model and of the associated learning algorithms.

Keywords: tensor decomposition, tensor product, pattern analysis, nonnegative matrix decomposition, structural complexity.

1 Problem Formulation

Estimation problems for data with self-replicating structures, such as images, various textures and music spectrograms require specifically designed approaches to identify, approximate, and retrieve the dynamical structures present in the data. By modeling data via summations of Kronecker products of two matrices (scaling and pattern matrices), Loan and Pitsianis [1] established an approximation to address this problem. Subsequently, Nagy and Kilmer [2] addressed 3-D image reconstruction from real-world imaging systems in which the point spread function was decomposed into a Kronecker product form, Bouhamidi and Jbilou [3] used Kronecker approximation for image restoration, Ford and Tyrtyshnikov focused on sparse matrices in the wavelet domain [4], while the extension to tensor data was addressed in [5].

[⋆] Also affiliated with the EE Dept., Warsaw University of Technology and with Systems Research Institute, Polish Academy of Science, Poland.

[⋆⋆] The work of P. Tichavsky was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project 102/09/1278.

[⋆⋆⋆] The work of Danilo P. Mandic has been supported by a Royal Society International Joint Project with Japan.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 297–305, 2012.

It is important to note that at present, the Kronecker approximation [1] is limited to 2-D structures which are required to have the same dimension. In this paper, we generalize this problem by considering replicas (or similar structures) for multiway data $\boldsymbol{\mathcal{Y}}$. To this end, we explain the tensor $\boldsymbol{\mathcal{Y}}$ by a set of patterns and their locations, while allowing the patterns to have different dimensions. In order to formulate mechanism of data replication, we define a new tensor product which is a generalization of the standard matrix Kronecker product, and is particularly suited for data with recurrent complex structures.

Definition 1 (Kronecker tensor product). *Let $\boldsymbol{\mathcal{A}} = [a_{\boldsymbol{j}}]$ and $\boldsymbol{\mathcal{B}} = [b_{\boldsymbol{k}}]$ be two N-dimensional tensors of size $J_1 \times J_2 \times \cdots \times J_N$ and $K_1 \times K_2 \times \cdots \times K_N$, respectively, $\boldsymbol{j} = [j_1, j_2, \ldots, j_N]$, $1 \le j_n \le J_n$ and $\boldsymbol{k} = [k_1, k_2, \ldots, k_N]$, $1 \le k_n \le K_n$. A Kronecker tensor product of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ is defined as an N-D tensor $\boldsymbol{C} = [c_{\boldsymbol{i}}] \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $\boldsymbol{i} = [i_1, i_2, \ldots, i_N]$, $I_n = J_n K_n$ such that $c_{\boldsymbol{i}} = a_{\boldsymbol{j}}\, b_{\boldsymbol{k}}$, $i_n = k_n + (j_n - 1)K_n$, and is expressed as $\boldsymbol{C} = \boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{B}}$.*

Remark 1. *If $\boldsymbol{C}$ is partitioned into an $J_1 \times J_2 \times \cdots \times J_N$ block tensor, each block-$\boldsymbol{j}$ ($\boldsymbol{j} = [j_1, j_2, \ldots, j_N]$) can be written as $a_{\boldsymbol{j}}\boldsymbol{\mathcal{B}}$.*

In this article, we aim to solve the following problem:

Problem 1 (A New Tensor Decomposition). *Given an N-dimensional tensor $\boldsymbol{\mathcal{Y}}$ of size $I_1 \times I_2 \times \cdots \times I_N$, find smaller scale tensors $\boldsymbol{\mathcal{A}}_p$, $\boldsymbol{\mathcal{X}}_p$, $p = 1, ..., P$ such that*

$$\boldsymbol{\mathcal{Y}} \approx \sum_{p=1}^{P} \boldsymbol{\mathcal{A}}_p \otimes \boldsymbol{\mathcal{X}}_p. \tag{1}$$

We term sub-tensors $\boldsymbol{\mathcal{X}}_p$ of size $K_{p1} \times K_{p2} \times \cdots \times K_{pN}$ as patterns, while $\boldsymbol{\mathcal{A}}_p$ of dimensions $J_{p1} \times J_{p2} \times \cdots \times J_{pN}$ such that $I_n = J_{pn}\, K_{pn}$, are called intensities (see Remark 1).

As such, Problem 1 is a generalization of the 13th problem of Hilbert, which seeks to perform universal function approximation for a function of n variables by a number of functions of $(n-1)$ or fewer variables. This new tensor decomposition is different from other existing tensor/matrix decompositions such as the canonical polyadic decomposition (CP) [6], the Tucker decomposition (TD) [7] and the block component decomposition (BCD) [8], in that it models the relation between latent variables via links between factor matrices and core tensor(s) which can be diagonal (for CP) or dense tensors (for TD). In a particular case when all $\boldsymbol{\mathcal{A}}_p$, $p = 1, \ldots, P$ in (1) become vectors of size I_n or have only one non-singleton dimension, Problem 1 simplifies into BCD which finds only one factor matrix for each core tensor.

In the sequel, we introduce methods to solve Problem 1 with/without nonnegative constraints. Simulations on a music sequence and on complex images containing textures validate the proposed tensor decomposition.

2 Notation and Basic Multilinear Algebra

Throughout the paper, an N-dimensional vector will be denoted by an italic lowercase boldface letters, with its components in squared brackets, for example $\boldsymbol{i} = [i_1, i_2, \ldots, i_N]$ or $\boldsymbol{I} = [I_1, I_2, \ldots, I_N]$.

Definition 2 (Tensor unfolding [9]). *Unfolding a tensor* $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *along modes* $\boldsymbol{r} = [r_1, r_2, \ldots, r_M]$ *and* $\boldsymbol{c} = [c_1, c_2, \ldots, c_{N-M}]$ *where* $[\boldsymbol{r}, \boldsymbol{c}]$ *is a permutation of* $[1, 2, \ldots, N]$ *aims to rearrange this tensor to a matrix* $\mathbf{Y}_{\boldsymbol{r}\times\boldsymbol{c}}$ *of size* $\prod_{k=1}^{M} I_{r_k} \times \prod_{l=1}^{N-M} I_{c_l}$ *whose entries* (j_1, j_2) *are given by* $\mathbf{Y}_{\boldsymbol{r}\times\boldsymbol{c}}(j_1, j_2) = \boldsymbol{\mathcal{Y}}(\boldsymbol{i_r}, \boldsymbol{i_c})$, *where* $\boldsymbol{i_r} = [i_{r_1} \ldots i_{r_M}]$, $\boldsymbol{i_c} = [i_{c_1} \ldots i_{c_{N-M}}]$, $j_1 = \text{ivec}(\boldsymbol{i_r}, \boldsymbol{I_r})$, $j_2 = \text{ivec}(\boldsymbol{i_c}, \boldsymbol{I_c})$, *and* $\text{ivec}(\boldsymbol{i}, \boldsymbol{I}) = i_1 + \sum_{n=2}^{N} (i_n - 1) \prod_{j=1}^{n-1} I_j$.

If $\boldsymbol{c} = [c_1 < c_2 < \cdots < c_{N-M}]$, then $\mathbf{Y}_{\boldsymbol{r}\times\boldsymbol{c}}$ simplifies into $\mathbf{Y}_{(\boldsymbol{r})}$, while for $r = n$ and $\boldsymbol{c} = [1, \ldots, n-1, n+1, \ldots, N]$, we have mode-$n$ matricization $\mathbf{Y}_{\boldsymbol{r}\times\boldsymbol{c}} = \mathbf{Y}_{(n)}$.

Definition 3 (Reshaping). *The reshape operator for a tensor* $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *to a tensor of a size specified by a vector* $\boldsymbol{L} = [L_1, L_2, \ldots, L_M]$ *with* $\prod_{m=1}^{M} L_m = \prod_{n=1}^{N} I_n$ *returns an M-D tensor* $\boldsymbol{\mathcal{X}}$, *such that* $\text{vec}(\boldsymbol{\mathcal{Y}}) = \text{vec}(\boldsymbol{\mathcal{X}})$, *and is expressed as*

$$\boldsymbol{\mathcal{X}} = \texttt{reshape}(\boldsymbol{\mathcal{Y}}, \boldsymbol{L}) \quad \in \mathbb{R}^{L_1 \times L_2 \times \cdots \times L_M}. \tag{2}$$

Definition 4 (Kronecker unfolding). *A* $(\boldsymbol{J} \times \boldsymbol{K})$ *Kronecker unfolding of* $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *with* $I_n = J_n K_n, \forall n$, *is a matrix* $\mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})}$ *of the size* $\prod_{n=1}^{N} J_n \times \prod_{n=1}^{N} K_n$ *whose entries* (j, k) *are given by*

$$\mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})}(j, k) = \boldsymbol{\mathcal{C}}(\boldsymbol{i}),$$

for all $\boldsymbol{j} = [j_1, \ldots, j_N]$, $j_n = 1, \ldots, J_n$, $\boldsymbol{k} = [k_1, \ldots, k_N]$, $k_n = 1, \ldots, K_n$, $n = 1, \ldots, N$ *and* $j = \text{ivec}(\boldsymbol{j}, \boldsymbol{J})$, *and* $k = \text{ivec}(\boldsymbol{k}, \boldsymbol{K})$, $\boldsymbol{i} = [i_1, \ldots, i_N]$, $i_n = k_n + (j_n - 1)K_n$.

Lemma 1 (Rank-1 Factorization). *Consider a tensor product* $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{B}}$ *where* $\boldsymbol{\mathcal{A}}$ *and* $\boldsymbol{\mathcal{B}}$ *have the dimensions as in Definition 1. Then a Kronecker unfolding* $\mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})}$ *is a rank-1 matrix*

$$\mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})} = \text{vec}(\boldsymbol{\mathcal{A}}) \ \text{vec}(\boldsymbol{\mathcal{B}})^T. \tag{3}$$

Lemma 1 also provides a convenient way to compute and update $\boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{B}}$.

Lemma 2 (Implementation of the Kronecker unfolding). *Let* $\widetilde{\boldsymbol{\mathcal{C}}} = \text{reshape}(\boldsymbol{\mathcal{C}}, \boldsymbol{L})$ *of* $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *following* $\boldsymbol{L} = [K_1, J_1, K_2, J_2, \ldots, K_N, J_N]$, $I_n = J_n K_n$, $n = 1, 2, \ldots, N$. *An* $(\boldsymbol{J} \times \boldsymbol{K})$ *Kronecker unfolding of* $\boldsymbol{\mathcal{C}}$ *is equivalent to a tensor unfolding* $\widetilde{\mathbf{C}}_{(\boldsymbol{r})} = \mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})}$ *where* $\boldsymbol{r} = [2, 4, \ldots, 2N]$.

Lemma 3 (Rank-*P* Factorization). *Let a tensor* $\boldsymbol{\mathcal{C}}$ *be expressed as a sum of P Kronecker products* $\boldsymbol{\mathcal{C}} = \sum_{p=1}^{P} \boldsymbol{\mathcal{A}}_p \otimes \boldsymbol{\mathcal{B}}_p$, *where* $\boldsymbol{\mathcal{A}}_p \in \mathbb{R}^{J_1 \times \cdots \times J_N}$ *and* $\boldsymbol{\mathcal{B}}_p \in \mathbb{R}^{K_1 \times \cdots \times K_N}$, $p = 1, 2, \ldots, P$. *Then the Kronecker unfolding of* $\boldsymbol{\mathcal{C}}$ *is a matrix of rank-P, such that*

$$\mathbf{C}_{(\boldsymbol{J}\times\boldsymbol{K})} = \sum_{p=1}^{P} \text{vec}\Big(\boldsymbol{\mathcal{A}}_p\Big) \ \text{vec}\Big(\boldsymbol{\mathcal{B}}_p\Big)^T. \tag{4}$$

Lemmas 1 and 3 give us the necessary insight and physical intuition into methods for solving Problem 1, establishing that the Kronecker tensor decomposition of $\boldsymbol{\mathcal{Y}}$ is equivalent to factorizations of Kronecker unfoldings $\mathbf{Y}_{(\boldsymbol{J}\times\boldsymbol{K})}$. The algorithms for solving Problem 1 are presented in the subsequent section.

3 Decomposition Methods

The desired property of the tensor decomposition (1) is that not all patterns $\boldsymbol{\mathcal{X}}_p$ (and consequently intensities $\boldsymbol{\mathcal{A}}_p$) are required to have the same size. Assume that there are G pattern sizes ($G \leq P$) $K_{g1} \times K_{g2} \times \cdots \times K_{gN}$ ($g = 1, 2, \ldots, G$) corresponding to P patterns $\boldsymbol{\mathcal{X}}_p$ ($p = 1, 2, \ldots, P$). Patterns $\boldsymbol{\mathcal{X}}_p$ which have the same size are classified into the same group. There are G groups of pattern sizes whose indices are specified by $\mathcal{I}_g = \{p : \boldsymbol{\mathcal{X}}_p \in \mathbb{R}^{K_{g1}\times K_{g2}\times\cdots\times K_{gN}}\} = \{p_1^{(g)}, p_2^{(g)}, \ldots, p_{P_g}^{(g)}\}$, $\texttt{card}\{\mathcal{I}_g\} = P_g, \sum_{g=1}^{G} P_g = P$. For simplicity, we assume that the first P_1 patterns $\boldsymbol{\mathcal{X}}_p$ ($p = 1, 2, \ldots, P_1$) belong to group 1, the next P_2 patterns ($p = P_1 + 1, \ldots, P_1 + P_2$) belong to group 2, and so on. The tensor decomposition (1) can now be rewritten as

$$\boldsymbol{\mathcal{Y}} = \sum_{g=1}^{G} \sum_{p_g \in \mathcal{I}_g} \boldsymbol{\mathcal{A}}_{p_g} \otimes \boldsymbol{\mathcal{X}}_{p_g} + \boldsymbol{\mathcal{E}} = \sum_{g=1}^{G} \boldsymbol{\mathcal{Y}}^{(g)} + \boldsymbol{\mathcal{E}} = \widehat{\boldsymbol{\mathcal{Y}}} + \boldsymbol{\mathcal{E}}, \tag{5}$$

where $\boldsymbol{\mathcal{A}}_{p_g} \in \mathbb{R}^{J_{g1}\times J_{g2}\times\cdots\times J_{gN}}$, $\boldsymbol{\mathcal{X}}_{p_g} \in \mathbb{R}^{K_{g1}\times K_{g2}\times\cdots\times K_{gN}}$ and $\boldsymbol{\mathcal{Y}}^{(g)} = \sum_{p_g \in \mathcal{I}_g} \boldsymbol{\mathcal{A}}_{p_g} \otimes \boldsymbol{\mathcal{X}}_{p_g}$. According to Lemma 3, Kronecker unfoldings $\mathbf{Y}^{(g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)}$ with $\boldsymbol{K}_g = [K_{g1}, K_{g2}, \ldots, K_{gN}]$, $\boldsymbol{J}_g = [J_{g1}, J_{g2}, \ldots, J_{gN}]$ are rank-P_g matrices, that is

$$\mathbf{Y}^{(g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)} = \sum_{p_g \in \mathcal{I}_g} \operatorname{vec}\left(\boldsymbol{\mathcal{A}}_{p_g}\right) \operatorname{vec}\left(\boldsymbol{\mathcal{X}}_{p_g}\right)^T . \tag{6}$$

In order to estimate $\boldsymbol{\mathcal{A}}_{p_g}$ and $\boldsymbol{\mathcal{X}}_{p_g}$, $\forall p_g \in \mathcal{I}_g$, we define $\boldsymbol{\mathcal{Y}}^{(-g)} = \boldsymbol{\mathcal{Y}} - \sum_{h \neq g} \boldsymbol{\mathcal{Y}}^{(h)}$, and minimize the cost function

$$\begin{aligned} D(\boldsymbol{\mathcal{Y}} \| \widehat{\boldsymbol{\mathcal{Y}}}) &= \|\boldsymbol{\mathcal{Y}} - \widehat{\boldsymbol{\mathcal{Y}}}\|_F^2 = \|\boldsymbol{\mathcal{Y}}^{(-g)} - \boldsymbol{\mathcal{Y}}^{(g)}\|_F^2 = \|\mathbf{Y}^{(-g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)} - \mathbf{Y}^{(g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)}\|_F^2 \\ &= \|\mathbf{Y}^{(-g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)} - \sum_{p_g \in \mathcal{I}_g} \operatorname{vec}\left(\boldsymbol{\mathcal{A}}_{p_g}\right) \operatorname{vec}\left(\boldsymbol{\mathcal{X}}_{p_g}\right)^T \|_F^2 . \end{aligned} \tag{7}$$

In general, without any constraints, the matrix decomposition in (7) or the tensor decomposition (1) are not unique, since any basis of the columnspace of the matrix $\mathbf{Y}^{(-g)}_{(\boldsymbol{J}_g \times \boldsymbol{K}_g)}$ in (7) can serve as $\operatorname{vec}\left(\boldsymbol{\mathcal{A}}_{p_g}\right)$, $p_g \in I_g$. One possibility to enforce uniqueness is to restrict our attention to orthogonal bases in which the scalar product of two patterns $\boldsymbol{\mathcal{X}}_p$, $\boldsymbol{\mathcal{X}}_q$, defined as a sum of the element-wise products of $\boldsymbol{\mathcal{X}}_p$, $\boldsymbol{\mathcal{X}}_q$, is zero for all $p \neq q$. Alternative constraints for nonnegative data $\boldsymbol{\mathcal{Y}}$, such as nonnegativity, can also be imposed on $\boldsymbol{\mathcal{A}}_p$ and $\boldsymbol{\mathcal{X}}_p$. In other words, by using the background physics to constrain all $\boldsymbol{\mathcal{A}}_q$ and $\boldsymbol{\mathcal{X}}_q$ in the other groups $q \notin \mathcal{I}_g$, we can sequentially minimize (7). These constraints do not have a serious effect on the generality of the proposed solutions as real world nonnegative data often exhibit a degree of orthogonality, and images are nonnegative.

3.1 Orthogonal Patterns

Solving the matrix decomposition in (7) with orthogonal constraints yields vectorizations $\operatorname{vec}\left(\boldsymbol{\mathcal{A}}_{p_g}\right)$ and $\operatorname{vec}\left(\boldsymbol{\mathcal{X}}_{p_g}\right)$ ($p_g \in \mathcal{I}_g$) that are proportional to P_g leading left and right singular vectors of $\mathbf{Y}^{(-g)}_{(\boldsymbol{J}\times\boldsymbol{K})} \approx \mathbf{U}\,\mathrm{diag}\{\boldsymbol{s}\}\,\mathbf{V}^T$, where $\mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{P_g}]$ and $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{P_g}]$, that is,

$$\boldsymbol{\mathcal{A}}_{p_l^{(g)}} = \mathrm{reshape}\Big(s_l\,\boldsymbol{u}_l, \boldsymbol{J}_g\Big), \quad \boldsymbol{J}_g = [J_{g1}, J_{g2}, \ldots, J_{gN}], \tag{8}$$

$$\boldsymbol{\mathcal{X}}_{p_l^{(g)}} = \mathrm{reshape}\Big(\boldsymbol{v}_l, \boldsymbol{K}_g\Big), \quad \boldsymbol{K}_g = [K_{g1}, K_{g2}, \ldots, K_{gN}]. \tag{9}$$

If all the patterns have the same size, then $\boldsymbol{K}_p = \boldsymbol{K}, \forall p$, $\boldsymbol{\mathcal{A}}_p$ and $\boldsymbol{\mathcal{X}}_p$ are reshaped from P leading left and right singular vectors of the Kronecker unfolding $\mathbf{Y}_{(\boldsymbol{J}\times\boldsymbol{K})}$.

3.2 Nonnegative Patterns

We shall now revisit Problem 1 and introduce nonnegative constraints in order to find nonnegative $\boldsymbol{\mathcal{A}}_p$ and $\boldsymbol{\mathcal{X}}_p$ from a nonnegative tensor $\boldsymbol{\mathcal{Y}}$. Such a constrained problem can be solved in a manner similar to the previous problem, that is, $\boldsymbol{\mathcal{A}}_p$ and $\boldsymbol{\mathcal{X}}_p$ are updated by minimizing the cost functions in (7). Note that we can employ straightforwardly update rules for nonnegative least squares approximation: the multiplicative update rules [10] and the ALS algorithms. In the following, we present the multiplicative update rules, which can be directly applied to (7) and have the form

$$\mathrm{vec}\Big(\boldsymbol{\mathcal{A}}_{p_g}\Big) \leftarrow \mathrm{vec}\Big(\boldsymbol{\mathcal{A}}_{p_g}\Big) \circledast \Big(\mathbf{Y}_{(\boldsymbol{J}_g\times\boldsymbol{K}_g)}\ \mathrm{vec}\Big(\boldsymbol{\mathcal{X}}_{p_g}\Big)\Big) \oslash \Big(\widehat{\mathbf{Y}}_{(\boldsymbol{J}_g\times\boldsymbol{K}_g)}\ \mathrm{vec}\Big(\boldsymbol{\mathcal{X}}_{p_g}\Big)\Big), \tag{10}$$

$$\mathrm{vec}\Big(\boldsymbol{\mathcal{X}}_{p_g}\Big) \leftarrow \mathrm{vec}\Big(\boldsymbol{\mathcal{X}}_{p_g}\Big) \circledast \Big(\mathbf{Y}^T_{(\boldsymbol{J}_g\times\boldsymbol{K}_g)}\ \mathrm{vec}\Big(\boldsymbol{\mathcal{A}}_{p_g}\Big)\Big) \oslash \Big(\widehat{\mathbf{Y}}^T_{(\boldsymbol{J}_g\times\boldsymbol{K}_g)}\ \mathrm{vec}\Big(\boldsymbol{\mathcal{A}}_{p_g}\Big)\Big). \tag{11}$$

Note that if all the patterns have the same size, the constrained Problem 1 becomes nonnegative matrix factorization of the Kronecker unfolding $\mathbf{Y}_{(\boldsymbol{J}\times\boldsymbol{K})}$. In a particular case when data $\mathbf{Y}$ is matrix and all patterns have the same size, Problem 1 simplifies into the matrix decomposition proposed in [1].

4 Simulations

The introduced algorithms were verified by comprehensive simulations on synthetic benchmark data and on real-world images with texture and music data.

4.1 Synthetic Data

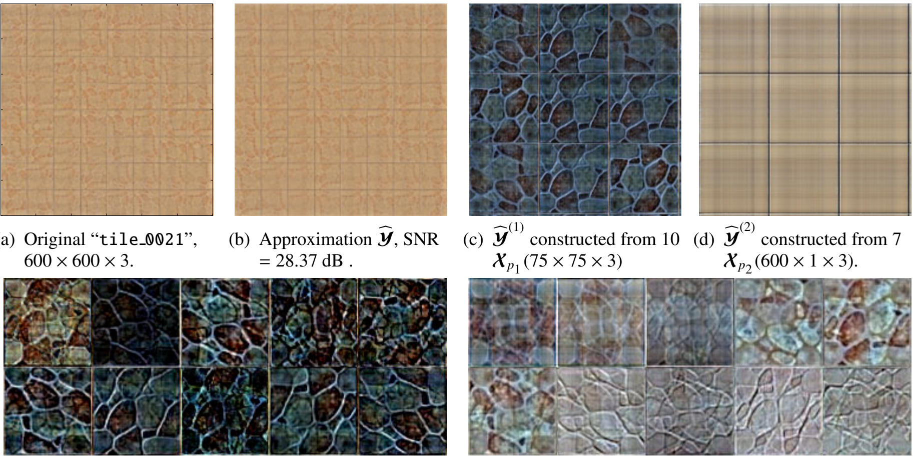
In the first set of simulations, we considered 3-D data of the size 90 × 90 × 12 composed of 12 random nonnegative patterns of different sizes, as given in Table 1 (row 2). Our aim was to extract orthogonal and nonnegative patterns in 50000 iterations or until differences of successive relative errors (SNR) $-20\log_{10}\left(\frac{\|\boldsymbol{\mathcal{Y}}-\widehat{\boldsymbol{\mathcal{Y}}}\|_F}{\|\boldsymbol{\mathcal{Y}}\|_F}\right)$ are lower than 10^{-5}. Results (SNR) in Table 1 (the second row) show an average SNR = 110.16 dB over 100 runs for orthogonal decomposition, and an average SNR = 107.43 dB based on nonnegative patterns. The results confirm the validity of the proposed model and the excellent convergence of the proposed algorithms.

4.2 Analysis of Texture Images

The next set of simulations were performed on RGB textures "`tile_0021`" and "`metal_plate_0020`" taken from `http://texturelib.com`. Textures can be represented by 3-D tensors of pixels, or by 4-D tensors with additional modes for approximation and detail coefficients in the wavelet domain. For example, the image "`tile_0021`" of size 600 × 600 × 3 is tiled by patterns $\boldsymbol{\mathcal{X}}_p$ of size 75 × 75 × 3 as illustrated in Fig. 1(a). Detail coefficients of this image obtained by the biorthogonal wavelet transform formulate a 3-D tensor of size 300 × 300 × 3 × 3. The approximation coefficients can be independently decomposed or combined with the tensor of detail coefficients. Parameters of Kronecker decompositions such as the number of patterns and their dimensions are given in Table 1. Approximation errors (SNR (dB)) and ratio (%) between the number of fitting parameters and the number of data elements are also given in Table 1.

In Fig. 1, the image "`tile_0021`" was approximated by two groups of orthogonal and nonnegative patterns. Two nonnegative basis images corresponding to two groups of patterns are shown in Figs. 1(c), 1(d). The first group consists of 10 patterns $\boldsymbol{\mathcal{X}}_{p_1} \in \mathbb{R}_+^{75\times75\times3}$ (shown in Fig. 1(e)) expressing replicating structures, whereas the second group consists of 7 patterns of size 600 × 1 × 3 representing the background as in Fig. 1(d). In addition, ten orthogonal patterns are shown in Fig. 1(f). For nonnegative patterns, each pattern in Fig. 1(e) represents a replicating structure in the image, whereas the orthogonal patterns in Fig. 1(f) were ranked according to the order of singular values which indicate detail level of patterns. Observe from Fig. 1(f) that the higher the order of the orthogonal patterns $\boldsymbol{\mathcal{X}}_p$, the more details these patterns comprise.

Results for decompositions of the color image "`metal_plate_0012`" are shown in Fig. 2. In the wavelet domain, we formulated a 3-D tensor for the approximation



(a) Original "`tile_0021`", 600 × 600 × 3.

(b) Approximation $\widehat{\boldsymbol{\mathcal{Y}}}$, SNR = 28.37 dB .

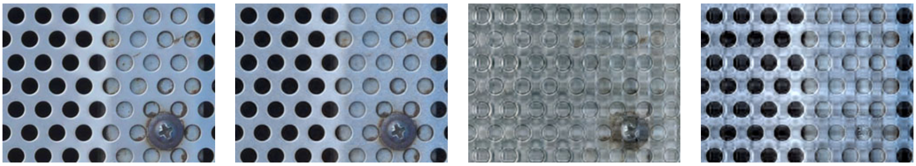
(c) $\widehat{\boldsymbol{\mathcal{Y}}}^{(1)}$ constructed from 10 $\boldsymbol{\mathcal{X}}_{p_1}$ (75 × 75 × 3)

(d) $\widehat{\boldsymbol{\mathcal{Y}}}^{(2)}$ constructed from 7 $\boldsymbol{\mathcal{X}}_{p_2}$ (600 × 1 × 3).

(e) 10 nonnegative $\boldsymbol{\mathcal{X}}_{p_1}$ (75 × 75 × 3).

(f) 10 orthogonal $\boldsymbol{\mathcal{X}}_{p_1}$ (75 × 75 × 3).

Fig. 1. Illustration for orthogonal and nonnegative pattern decompositions of the image "`tile_0021`". (b)-(d) reconstructed images and two basis images by 10 patterns of size 75×75×3 and 7 patterns of size 600 × 1 × 3. (e)-(f) 10 nonnegative and orthogonal patterns.



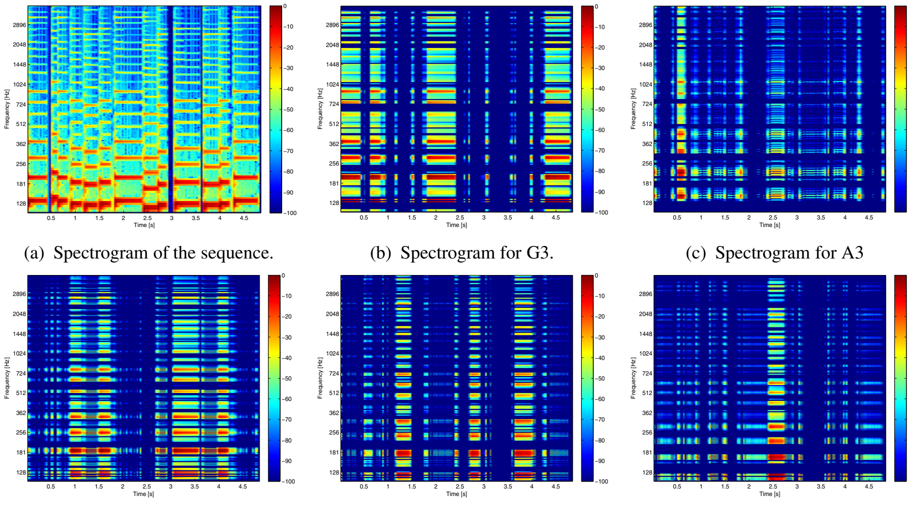
(a) "metal_plate_0012", 180 × 240 × 3.
(b) Approximation $\widehat{\mathcal{Y}}$, SNR = 28.38 dB.
(c) $\widehat{\mathcal{Y}}^{(1)}$ constructed from orthogonal DWT $\mathcal{X}_{p_1}$.
(d) $\widehat{\mathcal{Y}}^{(2)}$ constructed from orthogonal DWT $\mathcal{X}_{p_2}$.

Fig. 2. Approximation of "metal_plate_0012" in the wavelet domain

coefficients and a 4-D tensor comprising the details in the three orientations (horizontal, vertical, and diagonal). The two tensors were independently decomposed to find two groups of patterns whose sizes are given in Table 1 (row 4). The approximate image was then constructed from the basis patterns and achieved an SNR = 28.38 dB using 13.74 % of the number of entries. Figs. 2(c) and 2(d) visualize two basis images, each of which was constructed from one pattern group for the approximation coefficients and all the patterns for the detail coefficients.

4.3 Analysis of Patterns in Music

In this example, we decomposed a sampled song "London Bridge" composed of five notes A3, G3, F3, E3 and D3 played on a guitar for 5 seconds [10]. The log-frequency spectrogram $\mathbf{Y}$ (364 × 151), illustrated in Fig. 3(a), was converted from the



(a) Spectrogram of the sequence.
(b) Spectrogram for G3.
(c) Spectrogram for A3
(d) Spectrogram for F3.
(e) Spectrogram for E3.
(f) Spectrogram for D3.

Fig. 3. Log-frequency spectrograms of the music sequence and 5 basis nonnegative patterns corresponding to 5 notes G3, A3, F3, E3 and D3. The reconstructed signal has SNR = 20.78 dB.

Table 1. Parameters and results for orthogonal (Orho.) and nonnegative (NNG) pattern decompositions

Data	Size	Pattern Size	SNR (dB)		Ratio
		$(K_{g1} \times \cdots \times K_{gN}) \times P_g$	Ortho.	NNG	(%)
random	$90 \times 90 \times 12$	$(5 \times 5 \times 2) \times 2$ & $(6 \times 6 \times 3) \times 4$ & $(9 \times 9 \times 4) \times 6$	110.16	107.43	12.10
tile_0021	$600 \times 600 \times 3$	$(75 \times 75 \times 3) \times 10$ & $(600 \times 1 \times 3) \times 7$	29.69	28.37	17.24
	$300 \times 300 \times 3 \times 3$ (DWT, Detail Coefs.) $300 \times 300 \times 3$ (DWT, Approx. Coefs.)	$(20 \times 15 \times 1 \times 3) \times 20$ & $(300 \times 1 \times 1 \times 3) \times 3$ $(15 \times 15 \times 3) \times 40$ & $(300 \times 1 \times 3) \times 15$	27.84		9.48
metal_plate_0012	$180 \times 240 \times 3$	$(20 \times 20 \times 3) \times 15$ & $(180 \times 1 \times 3) \times 10$	27.58	25.35	21.16
	$90 \times 120 \times 3 \times 3$ (DWT, Detail Coefs.) $90 \times 120 \times 3$ (DWT, Approx. Coefs.)	$(5 \times 20 \times 1 \times 3) \times 3$ & $(90 \times 1 \times 1 \times 3) \times 10$ $(15 \times 15 \times 1) \times 20$ & $(90 \times 1 \times 1 \times 3) \times 5$	28.38		13.74
guitar music sequence	364×151 log-freq. spectrogram	$(4 \times 151) \times 5$ & $(2 \times 151) \times 4$ & $(7 \times 151) \times 2$	22.71	20.78	13.88

linear-frequency spectrogram in the frequency range from $f_0 = 109.4$ Hz (bin 8) to $f_I = f_s/2 = 4000$ Hz (bin 257) with 70 bins per octave. When there was no decomposition, the approximation error was 27.56 dB. The spectrogram was decomposed to find 11 patterns replicating along frequency (see row 5 in Table 1). Among the 11 log-frequency spectrograms $\widehat{\mathbf{Y}}^{(p)}$ constructed from $\mathbf{X}_p$, five spectrograms corresponding to five notes are illustrated in Figs. 3(b)-3(f). The approximate sequences (in the time domain) achieved SNR = 22.71 dB and 20.78 dB using orthogonal and nonnegative patterns, respectively. For this example, we may also apply the nonnegative matrix/tensor deconvolutions to seek for the similar patterns $\mathbf{X}_p$ replicating along frequency [11], however, the new tensor decomposition requires fewer fitting parameters.

5 Conclusions

A new tensor approximation has been proposed to identify and extract replicating structures from multiway data. By imposing a constraint on the replicating structures to be nonnegative or orthogonal, the model has been shown to significantly reduce the number of fitting parameters, compared with existing tensor/matrix factorizations. In a particular case when all the patterns have the same size, the new tensor decomposition simplifies into rank-P matrix factorization. This gives us a new insight and the ability to seek for hidden patterns by employing well-known matrix factorizations such as SVD and NMF. It has also been shown that a low-rank approximation by directly applying SVD or NMF to a data tensor results in common patterns which represent the background of the data, whereas factorization on the rearranged data extracts replicating structures. Simulation results for synthetic data, images and music sequence have shown that the proposed model and algorithms have the ability to extract desired patterns, and explain the data with relatively low approximation errors. Future extensions of the presented of this pattern decomposition will include approximating complex data by several subtensors instead of only two (scaling and pattern) tensors. One interesting implementation would be a multistage approach, in which patterns or scaling tensors are Kronecker products of subtensors.

References

1. Loan, C.V., Pitsianis, N.: Approximation with Kronecker products. In: Linear Algebra for Large Scale and Real Time Applications, pp. 293–314. Kluwer Publications (1993)
2. Nagy, J.G., Kilmer, M.E.: Kronecker product approximation for preconditioning in three-dimensional imaging applications. IEEE Transactions on Image Processing 15(3), 604–613 (2006)
3. Bouhamidi, A., Jbilou, K.: A Kronecker approximation with a convex constrained optimization method for blind image restoration. Optimization Letters, 1–14, doi: 10.1007/s11590-011-0370-7
4. Ford, J.M., Tyrtyshnikov, E.E.: Combining Kronecker product approximation with discrete wavelet transforms to solve dense, function-related linear systems. SIAM J. Sci. Comput. 25, 961–981 (2003)
5. Hackbusch, W., Khoromskij, B.N., Tyrtyshnikov, E.E.: Hierarchical Kronecker tensor-product approximations. Journal of Numerical Mathematics 13(2), 119–156 (2005)
6. Harshman, R.: Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis. UCLA Working Papers in Phonetics 16, 1–84 (1970)
7. Tucker, L.: Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311 (1966)
8. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms – Part I: Lemmas for partitioned matrices. SIAM J. Matrix Anal. Appl. 30(3), 1022–1032 (2008)
9. Bader, B., Kolda, T.: Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. ACM Transactions on Mathematical Software 32(4), 635–653 (2006)
10. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, ch.3. Wiley, Chichester (2009)
11. Phan, A.H., Tichavský, P., Cichocki, A., Koldovský, Z.: Low-rank blind nonnegative matrix deconvolution. In: Proc. IEEE ICASSP (2012)

An On-Line NMF Model for Temporal Pattern Learning: Theory with Application to Automatic Speech Recognition

Hugo Van hamme

University of Leuven, Department ESAT, Leuven, Belgium
hugo.vanhamme@esat.kuleuven.be

Abstract. Convolutional non-negative matrix factorization (CNMF) can be used to discover recurring temporal (sequential) patterns in sequential vector non-negative data such as spectrograms or posteriorgrams. Drawbacks of this approach are the rigidity of the patterns and that it is intrinsically a batch method. However, in speech processing, like in many other applications, the patterns show a great deal of time warping variation and recognition should be on-line (possibly with some processing delay). Therefore, time-coded NMF (TC-NMF) is proposed as an alternative to CNMF to locate temporal patterns in time. TC-NMF is motivated by findings in neuroscience. The sequential data are first processed by a bank of filters such as leaky integrators with different time constants. The responses of these filters are modeled jointly by a constrained NMF. Algorithms for learning, decoding and locating patterns in time are proposed and verified with preliminary ASR experiments.

Keywords: non-negative matrix factorization, temporal patterns, integrate-and-fire neurons, automatic speech recognition.

1 Introduction

Non-negative matrix factorization (NMF) [1] can be used to discover recurring patterns in non-negative data such as histograms, spectrograms or images. Multiple observations are projected to a vector space and stacked as the columns of a matrix **V** and approximated by the product **WH** of reduced rank, each with non-negative entries. The columns of **W** will contain the discovered one-dimensional patterns and **H** indicates their presence in the data. Because the non-negative patterns are combined with non-negative weights, the model can be thought of as a decomposition into parts. NMF and the related [2] PLSA [3] have been applied in many domains including image processing, speech enhancement, speech recognition, document clustering and term clustering.

Though this was not addressed in the original work in [1], the patterns to be discovered may show structure, such as adjacency of the pixels belong to a part. NMF disregards this structure by mapping the data to a one-dimensional vector and not imposing any a priori relation between the entries (the features) in this vector.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 306–313, 2012.

Graph-regularized NMF [4] is a generic extension that takes generic feature relations into account.

When the data have a sequential (two-dimensional) structure, such as spectrograms or, like the examples in this paper, time-varying neuron activations caused by spoken words, more specific formulations can be made. The *frame stacking* approach of stacking successive vectors [5] leads to suboptimal solutions because the repetition of features in successive data vectors or patterns is not exploited. Moreover, an inflation of the required number of model patterns corresponding to multiple time alignments is observed. Another solution is not to stack the data vectors of successive frames but rather to add them over a sliding window long enough to span the patterns, which has shown good results for modeling spoken digits in [6]. The disadvantage of this approach is that the event order within the patterns nor the order of occurrence of the patterns within a window are modeled. A more elegant solution is given by convolutional NMF (CNMF) [7,8], where the data are modeled as a sum of two-dimensional patterns, each convolved with an excitation function. While this model alleviates the objections to stacking, it suffers from the rigidity of the patterns. In speech processing, like in many other applications, the patterns show a great deal of variation. Successful models for word-sized patterns include some time warping mechanisms such as state alignments in hidden Markov models or dynamic time warping in exemplar-based speech recognition. For example, if the speaking rate is reduced, the convolutional patterns will tend to be too short and a good match with slow speech can only be obtained by averaging the patterns over all possible durations or by increasing the number of patterns. CNMF is also intrinsically a batch formulation requiring the complete input data to be known in order to decompose it. To recast it as an online method, block-wise approximations are required.

In this paper, an alternative NMF model, *time coded NMF (TC-NMF)*, for pattern discovery in sequential data and subsequent decoding is proposed. Like CNMF, it is capable of modeling the sequential aspects of the patterns and it is capable of locating these patterns along the time axis. However, it does not model the data by rigid patterns, yet allows for a controllable temporal resolution. Instead, it feeds the non-negative data into a bank of filters with different impulse responses and models their output jointly with an NMF. The relative output magnitudes of the different filters allow to locate patterns in time. Simply put, if the output of a sluggish filter is large, while that of a fast filter is small (large), the pattern that caused it must be far (close) to the current analysis time. This mechanism can be used to locate the patterns as a whole during decoding, but also to locate events within the patterns during pattern discovery.

The proposed model is plausible from the neuroscience perspective. Upstream neurons are tuned to fire on specific (acoustic) events. In early stages of auditory processing, the *receptive field* of a neuron corresponds to a specific time-frequency distribution. Further downstream, they respond to more complex temporal patterns. The firing rates of such neurons are the inputs to the TC-NMF. The neurons feed many *integrate-and-fire* (IaF) neurons [9] with exponentially fading action potentials [10]. Neurons with a long memory or slow decay accumulate spikes form the input neurons over a long period of time, a process known as *temporal summation* in neurophysiology. The firing rate of a IaF neuron will depend on the incoming spike frequency relative to its time constant. We assume IaF neurons come with a wide range of time constants. By this process, neurons further downstream that are connected to these IaF neurons will observe different inputs depending on how much time has elapsed since the occurrence of the (acoustic) event. They will effectively result in time-tuned filter behavior, which is proposed as a mechanism for time perception [13]. In the present work, the typical joint response on all IaF neurons is modeled by an NMF, a model that can be viewed as a neural network showing cognitively plausible properties such as lateral inhibition [11] and which can learn incrementally, i.e. without having to store all past training tokens [12]. An algorithm for computing the perceived time of the detected temporal patterns will be given.

2 Model

Consider non-negative data v_{nt} ($n = 1 \ldots N$) that vary with an index t ($t = 1 \ldots T$) which will be called time. In the experimental section, these will be posteriorgrams, but in principle, the proposed method applies equally well to any vector-valued sequential non-negative data such as magnitude spectrograms. The basic NMF model is based on the following generative model:

$$v_{nt} \approx \sum_r w_{nr}\, h_{rt} \ \text{ with } \ v_{nt} \geq 0,\ w_{nr} \geq 0, h_{rt} \geq 0 \tag{1}$$

At each time instant the observations are explained as a linear combination of R one-dimensional patterns w_{nr}. There is no temporal relation between the pattern weights h_{rt}. The one-dimensional pattern model looses the feature relations that exist in higher dimensional space. In some cases, it is adequate to extend (1) to a tensor decomposition [14], but this is not appropriate for spectrograms and most imaging data.

In convolutional NMF, the r-th pattern is 2-dimensional with a feature index n and a time index m ($m = 1 \ldots M_r$) and is convolved with its time-dependent activation h_{rt}:

$$v_{nt} \approx \sum_r \sum_{m=0}^{M_r} w_{nmr}\, h_{rt}^{m\rightarrow} \tag{2}$$

where $h_{rt}^{m\rightarrow}$ means that the sequence h_{rt} is shifted to the right over m samples of the index t , i.e. $h_{rt}^{m\rightarrow} = h_{r(t+m)}$ except that zeros are shifted in on the left ($t = 1$) and elements are lost on the right ($t = T$). See [7] and [8]. The patterns have received a temporal model of duration M_r and are accurately located in time by h_{rt}.

2.1 Time-Coded NMF

To alleviate the disadvantages of CNMF mentioned in section 1, a model is proposed where the temporal data are first filtered by a bank of K filters with decaying impulse response $f_k(t)$ ($k = 1 \ldots K$):

$$v_{nt}^{k} = \sum_{u=1}^{t} v_{nu} f_k(t-u) \tag{3}$$

The decay functions could be the exponential family

$$f_k(t) = e^{-\alpha_k t} \tag{4}$$

a choice which is grounded in neuroscience (IaF neurons), which can easily be implemented as a bank of first order low-pass filters (leaky integrators) and which will be used in section 3. The constants α_k are known, corresponding to the physical property of a specific neuron in the brain. An alternative choice could be a cascade of first order filters with equal time constant $f_k(t) = t^{k-1} e^{-\alpha t}$ which are reminiscent of Laguerre filters.

The TC-NMF model is derived by assuming v_{nt} is composed of an additive combination of patterns located at times τ_{rt} having activation h_{rt}. Each pattern has an internal structure generating data valued at p_{ns}^{r} at chosen times q_{rs} relative to the end of the pattern:

$$v_{nt}^{k} \approx g_{nt}^{k} + \sum_{r} \sum_{s=1}^{S_r} p_{ns}^{r} f_k(q_{rs} + \tau_{rt}) h_{rt} \tag{5}$$

where S_r controls the level of detail within the r-th pattern (much like M_r in CNMF) and g_{nt}^{k} will be explained in section 2.3. Approximating $f_k(q_{rs} + \tau_{rt}) \approx a_{srk} f_k(\tau_{rt})$ yields

$$v_{nt}^{k} \approx g_{nt}^{k} + \sum_{r} w_{nr}^{k} f_k(\tau_{rt}) h_{rt} \triangleq d_{nt}^{k} \tag{6}$$

where patterns are represented by K vectors w_{nr}^{k} irrespective of their internal structure or duration with:

$$w_{nr}^{k} = \sum_{s=1}^{S_r} a_{srk} p_{ns}^{r} \tag{7}$$

For the first order filter bank (4), no approximation is involved and $a_{srk} = e^{-\alpha_k q_{rs}}$

2.2 Learning

Learning is achieved by estimating the parameters p_{ns}^{r}, τ_{rt} and h_{rt} by minimizing the sum over all k of the Kullback-Leibler divergence (KLD) between the left hand side and the right hand side of (6). In principle, this can be done completely unsupervisedly, which has been successful on toy problems. On real speech data, utterance-level information was exploited to avoid local minima.

The update formulae are obtained by equating the partial derivative of the KLD w.r.t. the parameters to zero. Like with the original NMF, this leads to a fixed point multiplicative update for p_{ns}^{r}:

$$p_{ns}^{r} \leftarrow p_{ns}^{r} \frac{\sum_k \sum_t \frac{v_{nt}^{k}}{d_{nt}^{k}} a_{srk} f_k(\tau_{rt}) h_{rt}}{\sum_k \sum_t a_{srk} h_{rt}} \tag{8}$$

The update for h_{rt} and τ_{rt} requires solving a 2×2 set of nonlinear equations. Form:

$$\varphi_{rt} = \sum_k \sum_n \frac{v_{nt}^{k}}{d_{nt}^{k}} w_{nr}^{k} f_k(\tau_{rt}) h_{rt} \tag{9}$$

$$\psi_{rt} = \sum_k \sum_n \frac{v_{nt}^{k}}{d_{nt}^{k}} w_{nr}^{k} f_k'(\tau_{rt}) h_{rt} \tag{10}$$

where prime denotes derivative. For the case of a leaky integrator filter bank, solve:

$$\sum_k \left(1 + \alpha_k \frac{\varphi_{rt}}{\psi_{rt}}\right)\left(\sum_n w_{nr}^{k}\right) e^{-\alpha_k \tau_{rt}} = 0 \tag{11}$$

for τ_{rt} by a coarse search followed by a few Newton-Raphson updates. Then update

$$h_{rt} \leftarrow \frac{\varphi_{rt}}{\sum_k (\sum_n w_{nr}^{k}) f_k(\tau_{rt})} \tag{12}$$

Updates (8), (11) and (12) are repeated until convergence. After each iteration p_{ns}^{r} is normalized to sum to unity over n and s while h_{rt} is normalized inversely to mitigate scale invariance. The KLD was always observed to decrease under these update rules but the convergence behavior remains to be studied from a theoretical point of view.

2.3 Decoding

Estimation of the activation of a pattern and its position is done with updates (11) and (12) with fixed w_{nr}^{k} (or equivalently fixed p_{ns}^{r}). This will yield a different answer at every t for which this problem is solved. Performing actual recognition requires an integration of these estimates into a global decision. A left-to-right decoding that does not involve dynamic programming is applied here and proceeds as follows. If h_{rt} exceeds a threshold and its current τ_{rt} is positive (i.e. the pattern is observed completely for it ends earlier than the current time) and also places pattern r after the last decoded pattern, pattern r is accepted. To avoid that a pattern accepted earlier would be recognized again in the future, its future effect on v_{nt}^{k} is predicted by the recursion (for leaky integrators):

$$g_{nt}^{k} \leftarrow e^{-\alpha_k} g_{nt}^{k} + \sum_{\substack{accepted \\ r\ at\ t}} h_{rt} e^{-\alpha_k \tau_{rt}} w_{nr}^{k} \tag{13}$$

where g_{nt}^{k} is initialized to 0. Adding g_{nt}^{k} to (5) has the desired inhibitory effect and generates a signal model in which multiple occurrences of a pattern are modeled correctly.

3 Application to Speech Recognition

In this section, preliminary experiments on the adult speakers of the TIDIGITS corpus (strings of up to seven digits) are presented. Each of the 11 digits plus silence is considered as a pattern to be modeled. In order not to make the pattern estimation task trivial, the isolated digits are removed from the training set, leaving 6159 utterances from 112 speakers. For each utterance, a phone lattice is generated using an acoustic model and a bigram phonotactic model which are both trained on the 284 speakers of the Wall Street Journal corpus. The vertices of the lattice are labeled with frame number t (= time at a 10 ms resolution) and the edges are labeled with one of 44 phone units and a posterior probability. The data v_{nt} are composed of phone features ($1 \leq n \leq 44$) and 44^2 phone co-occurrence features ($45 \leq n \leq 1980 = N$). A phone feature is the posterior probability of phone n at time t. The co-occurrence feature of phone pair (A,B) is only non-zero of there is a vertex at time t with an incoming edge with label A and an outgoing edge with label B. The feature value is the product of their posterior probability. The bank of 17 first order filters uses damping coefficients α_k linearly spaced between 0.02 and 0.1.

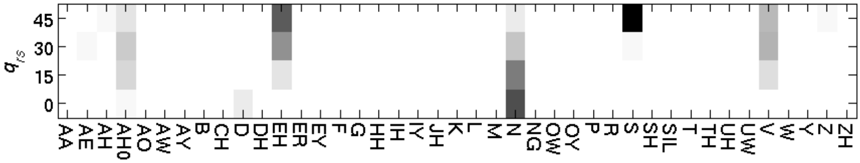


Fig. 1. Phone posterior part of the learned pattern model p^r_{ns} for the digit “seven”. Dark tones are large values. All phone labels were mapped to TIMIT symbols.

The training procedure is organized in several phases resulting in models with increased temporal detail. First $K = 1$ and $\alpha_1 = 0$, which is unable to capture temporal detail in the patterns and $S_r = 1$ (and $q_{r1} = 0$). At this point, utterance-level supervisory information is exploited by setting $h_{rt} = 0$ for all not-occurring patterns (i.e. only word identity but not word order is used) in the estimation of h_{rt} (one per utterance) and p^r_{n1} (τ_{rt} will not affect the cost function and is irrelevant). Then the full $K = 17$ filters are applied and h_{rt} and τ_{rt} are re-estimated to locate the patterns. Finally, the patterns are refined to $S_r = 4$ with $q_{rs} = 15(s\text{-}1)$ by re-estimation (initialization by duplication of p^r_{n1}). Figure 1 shows an estimated phone posterior model in p^r_{ns}. The models correctly reflect the most likely phones from the beginning to the end of the words (e.g. ‘S’ at the beginning $q_{r1} = 45$). On the test set with 6214 strings of at least two digits from 113 speakers (disjoint from the training set), 1.8% substitutions and 1.9% insertions are observed, but also almost 13% deletions. This preliminary result was obtained without much optimization on parameters or feature sets. While the substitution rate is encouraging, the deletion problem may be due to the bottom-up decoding process that does not consider multiple hypotheses.

4 Discussion and Conclusion

Like CNMF, TC-NMF allows to learn, recognize and locate temporal patterns in data. The solutions to the training and recognition problems were presented in this paper. At ach point in time, the pattern activation h_{rt} can be estimated with their perceived times τ_{rt}, using only the instantaneous data v_{nt}^{k} which makes TC-NMF an on-line method. To conclude, some research directions are listed.

- **Filter bank.** The number of filters K is not a critical parameter but should be greater than all S_r. The time constants α_k need to reflect the time scale of the patterns, but more research is required to suggest optimal values. Inappropriate constants or small K lead to poor numerical conditioning of the R matrices a_{srk} (rows r, columns k), which makes it impossible to infer the pattern model from data. The identifiability of patterns for other filter bank choices is not investigated.

- **Representation of patterns.** S_r and q_{rs} should be optimized for ASR. Parameter sharing among the pattern models could be considered.

- **Robustness to timing and warping mismatch**. Consider an observed pattern instance that is atypical in length (e.g. a slowly spoken digit). In CNMF and frame stacking, having a good match at the beginning of the pattern will imply that the observed features will not match at all towards the end. On spectrogram data for instance, a mismatch in the slope of a formant frequency trajectory will result in the different frequency channels (feature index n) holding the formant peak and hence extremely poor matches. In TC-NMF, a timing mismatch between model and data for pattern r will result in a mismatch in the relative size of $f_k(\tau_{rt})w_{nr}^{k}$ versus its contribution in v_{nt}^{k}, reflecting that close timing matches are better. However, it will not result in a mismatch along the feature dimension. Hence, a possible, today unconfirmed, advantage of TC-NMF could be robustness to timing or warping mismatch.

- **Cascading TC-NMF**. Observe that the input and the output to TC-NMF are alike: they are features with a time stamp. The output integrates inputs to larger, more complex units. Architectures that integrate several layers (e.g. spectra to phones to words) could be explored.

- **Search.** The proposed decoding mechanism maintains only one hypothesis, while traditional architectures for ASR examine a (cognitively implausible) huge set of hypotheses simultaneously. Augmenting the model with a search that does not take instantaneous irrevocable decisions is expected to be advantageous to accuracy and might facilitate the integration of a language model.

- **Supervision.** In the experiments, limited supervision information was used in the form of the identity of the keywords present in an utterance. The method is intrinsically unsupervised and could be applied as such. Like in many unsupervised NMF-based models, local minima of the cost could become problematic.

Acknowledgement. This work is sponsored by IWT-SBO project 100049 (ALADIN) and KU Leuven research grant OT/09/028 (VASI).

References

1. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
2. Gaussier, E., Goutte, C.: Relation between PLSA and NMF and Implications. In: ACM Conference on Research and Development in Information Retrieval, SIGIR, pp. 601–602 (2005)
3. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999 (1999)
4. Cai, D., He, X., Han, T., Huang, T.: Graph regularized non-negative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011) (to appear)
5. Gemmeke, J.F., Virtanen, T., Hurmalainen, A.: Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Transactions on Audio, Speech and Language Processing 19(7), 2067–2080 (2011)
6. Stouten, V., Demuynck, K., Van hamme, H.: Discovering Phone Patterns in Spoken Utterances by Non-negative Matrix Factorisation. IEEE Signal Processing Letters 15, 131–134 (2008)
7. Smaragdis, P.: Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs. In: Puntonet, C.G., Prieto, A. (eds.) ICA 2004. LNCS, vol. 3195, pp. 494–499. Springer, Heidelberg (2004)
8. O'Grady, P.D., Pearlmutter, B.A.: Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. Neurocomputing 72, 88–101 (2008)
9. Gerstner, W., Kistler, M.W.: Spiking Neuron Models. Single Neurons, Populations, Plasticity. Cambridge University Press (2002)
10. Gluss, B.: A model for neuron firing with exponential decay of potential resulting in diffusion equations for probability density. Bulletin of Mathematical Biophysics 29, 233–243 (1967)
11. Van hamme, H.: On the relation between perceptrons and non-negative matrix factorization. In: SPARS 2011 Workshop: Signal Processing with Adaptive Sparse Structured Representations, Edinburgh, U.K. (June 2011)
12. Driesen, J., Van hamme, H.: Modelling Vocabulary Acquisition, Adaptation and Generalization in Infants using Adaptive Bayesian PLSA. Neurocomputing 74(11), 1874–1882 (2011)
13. Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N.W., McGraw, P.V., Whitaker, D.: Duration channels mediate human time perception. Proc. R. Soc. B. (2011), doi:10.1098/rspb.2011.1131
14. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative matrix and Tensor Factorizations. Wiley (2009)

Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models*

Ricard Marxer, Jordi Janer, and Jordi Bonada

Universitat Pompeu Fabra,
Music Technology Group,
Roc Boronat 138, Barcelona
{ricard.marxer,jordi.janer,jordi.bonada}@upf.edu

Abstract. This research focuses on the removal of the singing voice in polyphonic audio recordings under real-time constraints. It is based on time-frequency binary masks resulting from the combination of azimuth, phase difference and absolute frequency spectral bin classification and harmonic-derived masks. For the harmonic-derived masks, a pitch likelihood estimation technique based on Tikhonov regularization is proposed. A method for target instrument pitch tracking makes use of supervised timbre models. This approach runs in real-time on off-the-shelf computers with latency below 250ms. The method was compared to a state of the art Non-negative Matrix Factorization (NMF) offline technique and to the ideal binary mask separation. For the evaluation we used a dataset of multi-track versions of professional audio recordings.

Keywords: Source separation, Singing voice, Predominant pitch tracking.

1 Introduction

Audio source separation consists in retrieving one or more audio sources given a set of one or more observed signals in which the sources are mixed. In the field of music processing, it has received special attention the past few decades. A number of methods have been proposed, most of them based on time-frequency masks. We differentiate between two main strategies in the creation of the time-frequency mask depending on the constraints of the solution.

Realtime solutions are often based on binary masks, because of their simple and inexpensive computation. These solutions assume the target sources are orthogonal in the time-frequency domain. The most common binary mask used in stereo music recordings is based on panning information of the sources [15,8,13].

Non-realtime approaches do not make such an orthogonality assumption, and make use of a soft mask based on Wiener filtering [2] which requires estimating all spectrograms of the constitutive sources. For harmonic sources this estimation is often performed in two steps. First the pitch track of the target source is

* This research has been partially funded by Yamaha Corporation (Japan).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 314–321, 2012.

estimated and then the spectrum of that given pitch track is estimated. The first step often relies on melody extraction algorithms [7,6]. Some methods estimate the pitch of the components independently [10], while others perform a joint estimation of the pitches in the spectrum [10,14]. Most joint pitch estimation methods are computationally expensive since they evaluate a large number of possible pitch combinations. NMF approaches to multipitch likelihood estimation [11,5] address this pitfall by factoring the spectrogram into a multiplication of two positive matrices, a set of spectral templates and a set of time-dependent gains. In [4] and [9] the spectral templates are fixed to a set of comb filters representing the spectra generated by each individual pitch spectrum. We propose combining several sources of information for the creation of the binary mask in order to raise the quality of currently existing methods while maintaining low-latency. We propose two main sources of information for the creation of the masks. Spectral bin classification based on measures such as lateralization (panning), phase difference between channels and absolute frequency is used to create a first mask. Information gathered through a pitch-tracking system is used to create a second mask for the harmonic part of the main melody instrument.

2 Spectral Bin Classification Masks

Panning information is one of the features that have been used successfully [15,8] to separate sources in real-time. In [13] the pan and the IPD (inter-channel phase difference) features are used to classify spectral bins. An interesting feature for source separation is the actual frequency of each spectrum bin, which can be a good complement when the panning information is insufficient. Using pan and frequency descriptors we define a filter in the frequency domain using a binary mask to mute a given source:

$$m_i^{pf}[f] = \begin{cases} 0 & \text{if } p_{low} < p_i[f] < p_{high} \text{ and } f_{low} < f < f_{high}, \\ 1 & \text{otherwise.} \end{cases}$$

where $p_i[f]$ is the pan value of the spectral bin f at frame i. The parameters p_{low} and p_{high} are the pan boundaries and f_{low} and f_{high} are the frequency boundaries fixed at -0.25, 0.25 and 60Hz and 6000Hz respectively, to keep the method unsupervised.

The results show that this method produces acceptable results in some situations. The most obvious limitation being that it is not capable of isolating sources that share the same pan/frequency region. This technique is also ineffective in the presence of strong reverberation or in mono recordings which have no pan information.

3 Harmonic Mask

Harmonic mask creation is based on two assumptions: that the vocal component is fully localized in the spectral bins around the position of the singing voice partials and that the singing voice is the only source present in these bins. Under such assumptions an optimal mask to remove the singing voice consists of zeros around the partials positions and ones elsewhere.

These assumptions are often violated. The singing voice is composed of other components than the harmonic components such as consonants, fricatives or breath. Additionally other sources may contribute significantly to the bins where the singing voice is located. This becomes clear in the results where signal decomposition methods such as Instantaneous Mixture Model (IMM) [4] that do not rely on such assumptions perform better than our binary mask proposal. However these assumptions allow us to greatly simplify the problem.

Under these assumptions we define the harmonic mask m^h to mute a given source as:

$$m_i^h[f] = \begin{cases} 0 & \text{for } (f0_i \cdot h) - L/2 < f < (f0_i \cdot h) + L/2, \, \forall h, \\ 1 & \text{otherwise.} \end{cases}$$

where $f0_i$ is the pitch of the i^{th} frame, and L is the width in bins to be removed around the partial position.We may also combine the harmonic and spectral bin classification masks using a logical operation by defining a new mask m_i^{pfh} as:

$$m_i^{pfh}[f] = m_i^{pf}[f] \vee m_i^h[f] \tag{1}$$

Finally, we are also able to produce a *soloing* mask $\bar{m}_i[f]$ by inverting any of the previously presented muting masks $\bar{m}_i[f] = \neg m_i[f]$.

In order to estimate the pitch contour $f0_i$ of the chosen instrument, we follow a three-step procedure: pitch likelihood estimation, timbre classification and pitch tracking.

3.1 Pitch Likelihood Estimation

The pitch likelihood estimation method proposed is a linear signal decomposition model. Similar to NMF, this method allows us to perform a joint pitch likelihood estimation. The main strengths of the presented method are low latency, implementation simplicity and robustness in multiple pitch scenarios with overlapping partials. This technique performed better than a simple harmonic summation method in our preliminary tests.

The main assumption is that the spectrum $X_i \in \mathbb{R}^{N_S \times 1}$ at a given frame i, is a linear combination of N_C elementary spectra, also named basis components. This can be expressed as $X_i = BG_i$, N_S being the size of the spectrum. $B \in \mathbb{R}^{N_S \times N_C}$ is the basis matrix, whose columns are the basis components. $G_i \in \mathbb{R}^{N_C \times 1}$ is a vector of component gains for frame i.

We set the spectra components as filter combs in the following way:

$$\varphi[m,n] = 2\pi f_l H N_P \frac{2^{\frac{iH - F/2 + n}{HN_P}} - 1}{S_r \ln(2)}$$

$$B_m[k] = \sum_{n=0}^{F} w_a[n] \left(\sum_{h=1}^{N_h} sin\left(h\varphi[m,n]\right) \right) e^{-j2\pi nk/N} \tag{2}$$

with $H = (1 - \alpha)F$. Where α is a coefficient to control the frequency overlap between the components, F is the frame size, S_r the sample rate, $w_a[n]$ is the analysis window, N_h is the number of harmonics of our components, B_m is the spectrum of size N of the component of m^{th} pitch. Flat harmonic combs have been used in order to estimate the pitch likelihoods of different types of sources.

The condition number of the basis matrix B defined in Equation 2 is very high ($\kappa(B) \approx 3.3 \cdot 10^{16}$), possibly due to the harmonic structure and correlation between the components in our basis matrix. For this ill-posed problem we propose using the well-known Tikhonov regularization method to find an estimate of the components gains vector $\hat{G}_i$ given the spectrum X_i. This consists in the minimization of the following objective function:

$$\Phi(G_i) = |BG_i - X_i|^2 + \lambda |G_i|^2 \tag{3}$$

where λ is a positive scalar parameter that controls the effect of the regularization on the solution. Under the assumption of gaussian errors, the problem has the closed-form solution $\hat{G}_i = RX_i$ where R is defined as:

$$R = B^t [BB^t + \lambda I_{N_S}]^+ \tag{4}$$

and $[Z]^+$ denotes the MoorePenrose pseudoinverse of Z. The calculation of R is computationally costly, however R only depends on B, which is defined by the parameters of the analysis process, therefore the only operation that is performed at each frame is $\hat{G}_i = RX_i$.

We must note that in contrast to NMF, our gains $\hat{G}_i$ can take negative values. In order to have a proper likelihood we we define the pitch likelihood as:

$$P_i = [\hat{G}_i]_+ / sum([\hat{G}_i]_+) \tag{5}$$

where $[Z]_+$ denotes the operation of setting to 0 all the negative values of a given vector Z.

3.2 Timbre Classification

Estimating the pitch track of the target instrument requires determining when the instrument is not active or not producing a harmonic signal (e.g. in fricative phonemes).

We select a limited number of pitch candidates n_d by finding the largest local maxima of the pitch likelihood function P_i 5. For each candidate a feature vector c is calculated from its harmonic spectral envelope $e_h(f)$ and a classification algorithm predicts the probability of it being a *voiced* envelope of the target instrument. The feature vector c of each of the candidates is classified using Support Vector Machines (SVM). The envelope computation $e_h(f)$ results from the Akima interpolation [1] between the magnitude at harmonic frequencies bins. The timbre features c are a variant of the Mel-Frequency Cepstrum Coefficients (MFCC), where the input spectrum is replaced by an interpolated harmonic spectral envelope $e_h(f)$. This way the spectrum values between the harmonics,

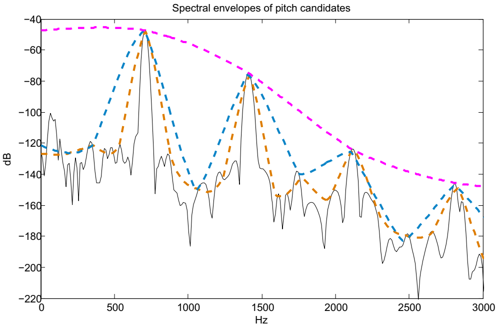


Fig. 1. Spectrum magnitude (solid black line) and the harmonic spectral envelopes (colored dashed lines) of three pitch candidates

where the target instrument is often not predominant, have no influence on the classification task. Figure 1 shows an example of a spectrum $X_i[f]$ (in black) of a singing voice signal, and the interpolated harmonic spectral envelopes $e_{h,1}(f)$, $e_{h,2}(f)$ and $e_{h,3}(f)$ (in magenta, blue and orange respectively), of three different pitch candidates.

The features vector c contains the first 13 coefficients of the Discrete Cosine Transform (DCT), which are computed from the interpolated envelope $e_h(f)$ as:

$$c = DCT\left(10 \cdot \log\left(E[k]\right)\right) \tag{6}$$

where $E[k] = \sum_{f_{k,low}}^{f_{k,high}} e_h(f)^2$, and $f_{k,low}$ and $f_{k,high}$ are the low and high frequencies of the k^{th} band in the Mel scale. We consider 25 Mel bands in a range $[0...5kHz]$. Given an audio signal sampled at $44.1kHz$, we use a window size of 4096 and a hop size of 512 samples. The workflow of our supervised training method is shown in Figure 2. Two classes are defined: *voiced* and *unvoiced* in a frame-based process[1]. *Voiced* frames contain pitched frames from monophonic singing voice recordings (i.e. only a vocal source). Pitched frames have been

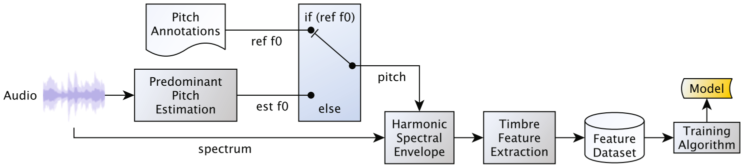


Fig. 2. In the training stage, the $e_h(f)$ is based on the annotated pitch if it exists *if (ref. f0)*, and on the estimated pitch otherwise

[1] The original training and test datasets consist of $384,152$ $(160,779/223,373)$ and $100,047$ $(46,779/53,268)$ instances respectively. Sub-sampled datasets contain $50,000$ and $10,000$ respectively. Values in brackets are given for the voiced and unvoiced instances respectively.

manually annotated. In order to generalize well to real audio mixtures, we also include audio examples composed of an annotated vocal track mixed artificially with background music. *Unvoiced* frames come from three different sources: *a)* non-pitched frames from monophonic singing voice recordings (e.g. fricatives, plosive, aspirations, silences, etc.); *b)* other monophonic instrument recordings (sax, violin, bass, drums); and *c)* polyphonic instrumental recordings not containing vocals. We employ a radial basis function (RBF) kernel for the SVM algorithm [3]. As a pre-process step, we apply standardization to the dataset by subtracting the mean and dividing by the standard deviation. We also perform a random subsampling to reduce model complexity. We obtain an accuracy of 83.54%, when evaluating the model against the test dataset.

3.3 Instrument Pitch Tracking

The instrument pitch tracking step is a dynamic programming algorithm divided into two processes. First a Viterbi is used to find the optimal pitch track in the past C frames, using pitch likelihood P_i for the state probability. Then a second Viterbi allows us to determine the optimal sequence of *voiced* and *unvoiced* frames using the probability found on the timbre classification step for the state. In both cases frequency differences larger than 0.5 semitones between consecutive frames are used to compute transition probabilities. Our implementation works on an online manner with a latency of $C = 20$ frames (232 ms). Due to lack of space the details of the implementation are not presented here.

4 Evaluation

The material used in the evaluation of the source separation method consists of 15 multitrack recordings of song excerpts with vocals, compiled from publicly available resources (MASS[2], SiSEC[3], BSS Oracle[4])

Using the well known BSSEval toolkit [12], we compare the Signal to Distortion Ratio (SDR) error (difference from the ideal binary mask SDR) of several versions of our algorithm and the IMM approach [4]. The evaluation is performed on the "all-minus-vocals" mix versions of the excerpts. Table 1 presents the SDR results averaged over 15 audio files in the dataset. We also plot the results of individual audio examples and the average in Figure 4. *Pan-freq mask* method results in applying the m^{pf} mask from Equation (1). The quality of our low-latency approach to source separation is not as high as for off-line methods such as IMM, which shows an SDR almost 3 dBs higher. However, our LLIS-SVM method shows an increase of 2.2 dBs in the SDR compared to the LLIS-noSVM method. Moreover, adding azimuth information to the multiplicative mask (method *LLIS-SVM-pan*) increases the SDR by 0.7 dBs.

[2] http://www.mtg.upf.edu/static/mass

[3] http://sisec.wiki.irisa.fr/

[4] http://bass-db.gforge.inria.fr/bss_oracle/

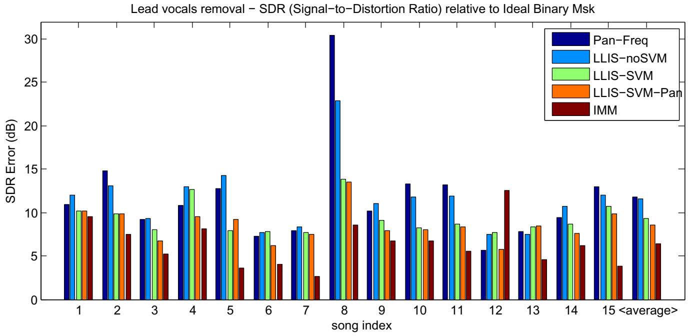


Fig. 3. SDR Error for four methods: pan-frequency mask, LLIS and IMM

Table 1. Signal-To-Distortion Ratio (in dB) for the evaluated methods. The Ideal column shows the results of applying an ideal binary mask with zeros in the bins where the target source is predominant and ones elsewhere.

Method	pan-freq	LLIS-noSVM	LLIS-SVM	LLIS-SVM-pan	IMM	Ideal
SDR-vocals	0.21	0.47	2.70	3.43	6.31	12.00
SDR-accomp	4.79	5.05	7.28	8.01	10.70	16.58

5 Conclusions

We present a source separation approach well suited to low-latency applications. The separation quality of the method is inferior to offline approaches, such as NMF-based algorithms, but it performs significantly better than other existing real-time systems. Maintaining low-latency (232 ms), an implementation of the method runs in real-time on current, consumer-grade computers. The method only targets the harmonic component of a source and therefore does not remove other components such as the unvoiced consonants of the singing voice. Additionally it does not remove the reverberation component of sources. However these are limitations common to other state-of-the-art source separation techniques and are out of the scope of our study.

We propose a method with a simple implementation for low-latency pitch likelihood estimation. It performs joint multipitch estimation, making it well-adapted for polyphonic signals. We also introduce a technique for detecting and tracking a pitched instrument of choice in an online manner by means of a classification algorithm. This study applies the method to the human singing voice, but it is general enough to be extended to other instruments.

Finally, we show how the combination of several sources of information can enhance binary masks in source separation tasks. The results produced by the ideal binary mask show that there are still improvements to be made.

References

1. Akima, H.: A new method of interpolation and smooth curve fitting based on local procedures. JACM 17(4), 589–602 (1970)
2. Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE Transactions on Audio, Speech, and Language Processing 14(1) (2006)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Durrieu, J.L., Richard, G., David, B., Fevotte, C.: Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 564–575 (2010)
5. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Comput. 21, 793–830 (2009)
6. Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.: F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, p. V (May 2006)
7. Goto, M., Hayamizu, S.: A real-time music scene description system: Detecting melody and bass lines in audio signals. Speech Communication (1999)
8. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures. In: Proc (ICASSP) International Conference on Acoustics, Speech, and Signal Processing (2000)
9. Ozerov, A., Vincent, E., Bimbot, F.: A General Modular Framework for Audio Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 33–40. Springer, Heidelberg (2010)
10. Ryynänen, M., Klapuri, A.: Transcription of the singing melody in polyphonic music. In: Proc. 7th International Conference on Music Information Retrieval, Victoria, BC, Canada, pp. 222–227 (October 2006)
11. Sha, F., Saul, L.K.: Real-time pitch determination of one or more voices by nonnegative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1233–1240. MIT Press (2005)
12. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)
13. Vinyes, M., Bonada, J., Loscos, A.: Demixing commercial music productions via human-assisted time-frequency masking. In: Proceedings of Audio Engineering Society 120th Convention (2006)
14. Yeh, C., Roebel, A., Rodet, X.: Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. Trans. Audio, Speech and Lang. Proc. 18, 1116–1126 (2010)
15. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52(7), 1830–1847 (2004)

Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization

Cyril Joder[1], Felix Weninger[1], Florian Eyben[1],
David Virette[2], and Björn Schuller[1,⋆]

[1] Institute for Human-Machine Communication,
Technische Universität München, 80333 München, Germany
[2] HUAWEI Technologies Düsseldorf GmbH, Germany
cyril.joder@tum.de

Abstract. In this paper, we present an on-line semi-supervised algorithm for real-time separation of speech and background noise. The proposed system is based on Nonnegative Matrix Factorization (NMF), where fixed speech bases are learned from training data whereas the noise components are estimated in real-time on the recent past.

Experiments with spontaneous conversational speech and real-life non-stationary noise show that this system performs as well as a supervised NMF algorithm exploiting noise components learned from the same noise environment as the test sample. Furthermore, it outperforms a supervised system trained on different noise conditions.

1 Introduction

Isolating speech from environmental noise remains a challenging problem, especially in the presence of highly non-stationary noise such as background speech or music. On the other hand, a great variety of applications could benefit from a robust separation of speech, such as telephony, automatic speech recognition or hearing aids. In the case of telephony, additional constraints have to be taken into account, since usually only one microphone is available and the separation has to be performed in a real-time, on-line framework with a very small latency between audio input and output, in order to preserve natural communication.

One of the most popular approaches for single-channel source separation is Nonnegative Matrix Factorization (NMF) [3]. It has been shown efficient for speech separation [14,13], when both speech and noise models where learned prior to the separation. In [9], a variant of this algorithm is used, in which only one source is learned, the other being estimated from the mixture. However, this estimation requires off-line processing, where the whole signal is known.

Some studies have considered adapting the NMF algorithm to an incremental, on-line framework. In [11], pattern learning from large amounts of audio data using an on-line version of (convolutive) NMF is discussed. In [1], NMF is used

⋆ The research leading to these results has received funding from the HUAWEI Innovation Research Program 2010 (project GLASS).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 322–329, 2012.

to decompose a sequence formed by the new observation and the basis vectors, which are supposed to encompass the past observations. The approach of [12] and [6] first optimizes the activations for each coming observation, with fixed basis vectors, and then updates the bases based on the past activations. Still, we are not aware of a study on speech separation using on-line NMF algorithms.

In this work, we exploit a simple sliding window approach, where a classic NMF decomposition is performed on the recent past and the noise components are adapted in real-time to the current conditions. We test this semi-supervised on-line NMF method on a speech separation task with realistic data. Results show that the obtained system performs as well as a supervised NMF trained on the same noise environment, with a setting allowing for real-time capabilities.

After presenting the general NMF method in Section 2, we outline the proposed on-line NMF algorithms in Section 3. Then, experiments are detailed in Section 4, before drawing some conclusions.

2 Nonnegative Matrix Factorization (NMF) for Source Separation

Given a matrix of nonnegative data $\mathbf{V} \in \mathbb{R}_+^{m\times n}$, NMF aims at finding the two nonnegative matrices, $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$, which minimize the error $D(\mathbf{V}, \mathbf{WH})$, where D is some divergence measure. In our audio source separation application, $\mathbf{V}$ is the original magnitude spectrogram. The columns of $\mathbf{W}$ then represent characteristic spectra of the recording and $\mathbf{H}$ contains the corresponding 'activation' values of these basis spectra.

Many algorithms for performing this optimization rely on multiplicative update rules, in order to maintain the nonnegativity of the matrices $\mathbf{W}$ and $\mathbf{H}$. For example, with the generalized Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j}, \tag{1}$$

the update rules proposed by [5] are as follows:

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \tag{2}$$

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T}, \tag{3}$$

where $\mathbf{X} \cdot \mathbf{Y}$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ denote element-wise operations and $\mathbf{1}$ is a matrix of ones.

Assuming that each source is described by a set of columns of $\mathbf{W}$ with corresponding rows in $\mathbf{H}$, separated signals can then be reconstructed as follows. Let $\mathbf{W_k}$ be the sub-matrix containing the columns of $\mathbf{W}$ corresponding to a source k, and let $\mathbf{H_k}$ be the according activation sub-matrix. The magnitude spectrogram of the isolated source $\mathbf{V_k}$ is obtained by the Wiener-like equation:

$$\mathbf{V_k} = \mathbf{V} \cdot \frac{\mathbf{W_k H_k}}{\mathbf{WH}}. \tag{4}$$

This spectrogram is then inverted using the phase of the original mixture.

3 On-Line NMF

In this paper, on-line NMF refers to a sliding window method which decomposes the spectrum of the recent past into matrices **W** and **H** as detailed above. The sliding window contains the recent past spectra of the signal. Once a new frame is received, the sliding window is shifted by one frame. The activation matrix **H** is also shifted and the new column is initialized randomly. The matrices are then updated using a fixed number of NMF iterations. By using a sliding window approach, context information (in particular, the activations of the older frames) is available for this update so that a low number of iterations is sufficient.

3.1 On-Line Supervised NMF

In order to exploit the NMF decomposition for a practical source separation task, one needs to determine the source corresponding to each part of the decomposition. For our supervised algorithm, we assume that the sources which are to be separated are known in advance. This can correspond, for example, to the case of a teleconference, where several known people can talk simultaneously. We can then perform a learning of the characteristics of each isolated source. Hence, a spectral basis matrix is created for each considered source, using the (unsupervised) NMF decomposition of the learning data, as in [8].

For the separation phase, the **W** matrix is built by concatenating the basis matrices of the isolated sources. This matrix is kept constant and only the activation matrix **H** is updated using eq. (2). This particular case is straightforward to implement in an on-line system. This is because the update rule for each column $\mathbf{h}_{:,t}$ of **H** can be rewritten as:

$$\mathbf{h}_{:,t} \leftarrow \mathbf{h}_{:,t} \cdot \frac{\mathbf{W}^T \frac{\mathbf{v}_{:,t}}{\mathbf{W}\mathbf{h}_{:,t}}}{\mathbf{W}^T \mathbf{1}_{:,t}}. \tag{5}$$

Thus, each column of the activation matrix can be updated independently of the others, using only the current observation spectrum $\mathbf{v}_{:,t}$. The obtained factorization is then equivalent to an off-line version of supervised NMF.

3.2 On-Line Semi-supervised NMF

In the semi-supervised version of the on-line NMF algorithm, we consider that one source is unknown (modeling for example noise, or a new speaker). Thus, the spectral basis matrix **W** is no longer fully determined in advance. In the separation phase, the columns corresponding to the unknown source are initialized randomly, and updated with each new frame, following eq. (2). The other columns are kept constant.

With this semi-supervised algorithm, it is no longer possible to process each frame independently of the others, since the two matrices **W** and **H** depend on each other. Thus, the length of the sliding window — and thus of the amount of context information considered — does have an impact on the decomposition. Intuitively, a meaningful estimation of the 'noise spectral basis', i.e. the non-constant part of **W** , requires a whole sequence of observations.

3.3 Real-Time Implementation

For a real-time implementation of the on-line semi-supervised NMF, another important parameter is the *delay* parameter. It denotes the position in the sliding window of the frame to be output. If this parameter is equal to 0, only the past context is used for the NMF decomposition. By increasing this value, later observations can be considered. Moreover, the precision of the activations depends not only on the estimation of the matrix $\mathbf{W}$ (which can be controlled by the length of the sliding window), but also on the number of iterations that have been used for computation, which increases with the delay parameter.

The total latency L introduced by the system (neglecting computation time) is then determined by the frame size s and the delay parameter d, thanks to the relation: $L = (d + 1)s$. Note that the delay parameter is not relevant for the supervised algorithm, since the sliding window can be limited to the single current frame. Our implementation of the systems exploits the openSMILE [4] framework, which allows for an efficient incremental processing of audio data.

4 Experimental Evaluation

4.1 Experimental Settings

We evaluate the system on speech that was artificially mixed with real-life noise. Speech was taken from the Buckeye database [7], which contains recordings of interviews. The speech is highly spontaneous and contains a variety of non-linguistic vocalizations. Thus, we believe that this corpus is better suited for evaluation of speech separation in real-life conditions than, e. g., the popular TIMIT corpus of read speech, which is characterized by lower variation. We subdivided the Buckeye recording sessions, each of which is approximately 10 min long, into turns by cutting whenever the subject's speech was interrupted by the interviewer, or by a silence of more than 0.5 s length. Only the subject's speech is used. In these experiments, we only exploit turns of at least 3 s.

The test signals were then corrupted using noise recordings from the official corpus provided for the 2011 PASCAL CHiME Challenge [2]. These contain genuine recordings from a domestic environment obtained over a period of several weeks in a house with two small children. The noise is highly instationary due to abrupt changes such as appliances being turned on/off, impact noises such as banging doors, and interfering speakers [2]. All these data are publicly available[1]. The noise mixed with the speech was randomly drawn from the six hours of noise recordings in the database. We intentionally do not scale speech or noise to attain a distribution of noise levels corresponding to a real-life environment.

The sampling rate of the recordings is 16 kHz and the tested systems employ 32 ms analysis frames, with a 50 % overlap. In our experiments, we used 12 randomly chosen segments of speech, between 3 s and 20 s long. For each speech sample, a training sequence is created by concatenating 20 other speech segments from the same speaker, yielding lengths between 1.5 min and 5.5 min.

[1] http://spandh.dcs.shef.ac.uk/projects/chime/PCC/datasets.html

We constructed two different noise training sequences for supervised NMF. The first was created by concatenating 1024 short segments (0.5 s) drawn from diverse locations in the CHiME noise recordings. Hence, this training sequence contains most of the noise sources that can be found in the database. In order to assess the generalization property of the system to different types of noise, we also constructed another 17 min training sequence, composed of noise recordings from the SiSEC 2010 noisy speech database[2] as well as some extract of the SPIB noise database[3] and some street noise from the *soundcities* website[4]. These sequences are referred to as *matched* and *mismatched* training noise.

Several speech separation systems are tested here. All of them exploit constant basis components for speech, previously learned from the training sequence. The first two systems exploit the on-line supervised NMF algorithm presented in subsection 3.1, with noise components learned respectively from the matched and mismatched training noise. For these systems, the numbers of NMF components for speech and noise are equal to $c_s = c_n = 50$, which has been empirically found satisfactory for the speaker separation task. All the training processes use 256 iterations. The other system uses the on-line semi-supervised NMF algorithm of subsection 3.2, with $c_s = 50$ speech components. The tested values of the different parameters are displayed in Table 1. This values were chosen to maintain a limited computational complexity.

Table 1. Tested values of the parameters for the on-line semi-supervised NMF system

Parameter		Tested Values
c_s	number of speech components	{50}
c_n	number of noise components	{1,2,4,8,12,16}
ℓ	sliding window length	{2,4,6,8,12,16,20,25,30}
d	delay	{0,1,2,3,4,5,6,7}
n	number of optimization iterations	{1,2,4,8,16,32,64}

Several evaluation criteria were computed from the separated speech: the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Source to Artifact Ratio (SAR) [10]. For comparison of the on-line approach, we consider an 'optimal' off-line version of the semi-supervised NMF algorithm, which outperforms supervised NMF on our test data. For this system, 256 iterations are used and the number of noise components was chosen from the set $c_n \in \{1, 2, 4, 8, 12, 16, 20, 25, 30, 35, 40, 45, 50\}$. The value $c_n = 30$ is selected, maximizing the average SDR in the test database. This optimal SDR is equal to 5.2 dB, which represents the best result that can be achieved with basic NMF speech separation algorithms on our test data.

[2] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Source+separation+in+the+presence+of+real-world+background+noise

[3] http://spib.rice.edu/spib/select_noise.html

[4] http://www.soundcities.com

4.2 Results

The results obtained by the supervised NMF systems are displayed in Fig. 1. It can be observed that for both systems, the SIR increases with the number of iterations. However, this reduction of the interferences is at the cost of more artifacts, since the SAR concurrently decreases. The optimal trade-off is here realized for a single iteration, yielding a 4.2 dB SDR, against 0.6 dB for the original corrupted speech. Although the optimal number of iterations may be dependent on the data; this shows that a very small number of iterations is sufficient for a satisfactory separation. Thus, the obtained complexity is very low, achieving a real-time factor of 2 % on a 3.4 GHz, 64 bits CPU.

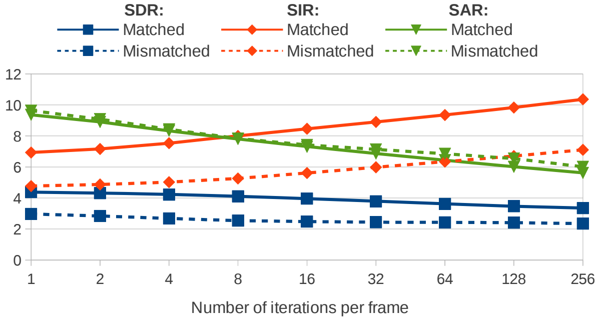


Fig. 1. Average source separation criteria (dB) for the supervised NMF systems, trained on the matched and mismatched noise

Our results also show the importance of an adequate noise model for the separation. Indeed, the supervised NMF systems are outperformed by the off-line semi-supervised algorithm, whose noise spectra seem to fit the observations even better, probably since they are estimated directly on each test sample. Furthermore, whereas the SARs of both supervised systems are roughly equivalent, the 'matched' noise training induces significantly higher SIRs (by over 2 dB) and thus a better separation quality.

Fig. 2 to 4 present a few of the numerous results of the on-line semi-supervised NMF system. The best SDR is equal to 4.4 dB that is slightly better than the result obtained with the supervised NMF, even with the 'matched' noise training. This shows the efficiency of the proposed method to adapt the noise model to the environment in an on-line framework.

The best score is obtained with the parameters $c_n = 8$, $\ell = 20$, $d = 0$ and $i = 1$ (see Table 1). Contrarily to the supervised case, Fig. 2 shows a degradation of the SIR when the number of iterations increases. This can be due to an 'overfitting' phenomenon, where the updated components tend to model speech as well as noise. One can see in Fig. 3 that, with a larger sliding window, the SIR decreases while the SAR is improved. This can be explained by the fact that the adaptation to the environment is then a bit less precise, but it is more robust to

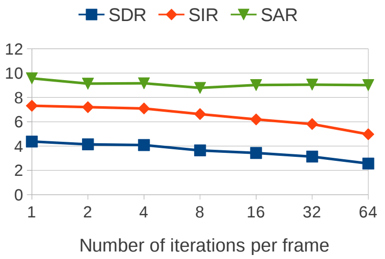


Fig. 2. Criteria (dB) as a function of i for constant $c_n = 8$, $\ell = 20$ and $d = 0$

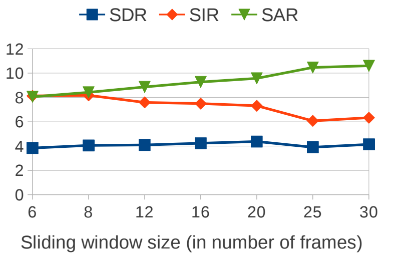


Fig. 3. Criteria (dB) as a function of ℓ for constant $c_n = 8$, $d = 0$ and $i = 1$

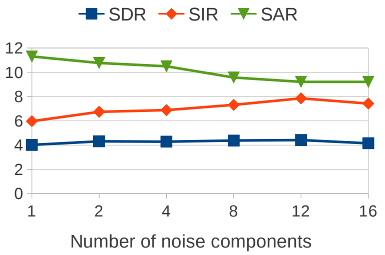


Fig. 4. Criteria (dB) as a function of c_n for constant $l = 20$, $d = 0$ and $i = 1$

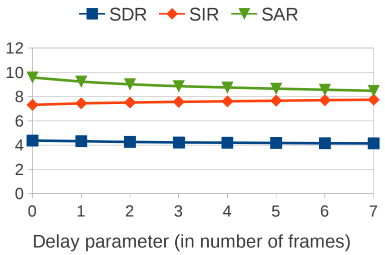


Fig. 5. Criteria (dB) as a function of d for constant $c_n = 8$, $\ell = 20$ and $i = 1$

the overfitting phenomenon. The number of noise components seems to have the opposite influence (Fig. 4). Hence, the values $c_n = 8$ and $\ell = 20$, corresponding to a sliding window of 336 ms, here constitute a reasonable trade-off.

The delay parameter has only a small influence, as shown in Fig. 5. Thus, the value $d = 0$ can be chosen so as to minimize the latency of the system. Furthermore, the best-performing setting has a relatively low complexity, since only one iteration is performed for each frame. The real-time factor then is 20 %, on the aforementioned CPU. Therefore, the system is fully real-time capable.

5 Conclusion

We presented a method for on-line speech separation exploiting a sliding window version of the semi-supervised Nonnegative Matrix Factorization algorithm. An extensive experimental study has been conducted, testing numerous parameter combinations. Our results show that this system performs similarly to (and even slightly better than) a supervised algorithm in which the noise components are learned from the same environment as the test samples. Furthermore, the optimal setting yields a system which is real-time capable on a recent PC.

Among the future works for further improvements of the system can be the introduction of regularization terms such as priors [14] or sparsity and continuity constraints, in order to obtain more meaningful components in both learning and separation phases without considerably affecting the complexity. The use of a small-order Nonnegative Matrix Deconvolution algorithm [8] could also be explored, although at the cost of increased latency and computational complexity. Finally, the observed behavior depending on the number of iterations motivates introduction of relaxation [3] into the multiplicative update algorithm.

References

1. Cao, B., Shen, D., Sun, J.T., Wang, X., Yang, Q., Chen, Z.: Detect and track latent factors with online nonnegative matrix factorization. In: Proceedings of the 20th Intern. Joint Conf. on Artifical Intelligence (IJCAI 2007), pp. 2689–2694 (2007)
2. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In: Proc. of Interspeech, Makuhari, Japan, pp. 1918–1921 (2010)
3. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. Wiley & Sons (2009)
4. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. of ACM Multimedia, pp. 1459–1462 (2010)
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
6. Lefèvre, A., Bach, F., Févotte, C.: Online algorithms for Nonnegative Matrix Factorization with the Itakura-Saito divergence. In: Proc. of IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA), pp. 313–316 (2011)
7. Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye Corpus of Conversational Speech (2nd release). Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA (2007), www.buckeyecorpus.osu.edu
8. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio, Speech and Language Process. 15(1), 1–14 (2007)
9. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 414–421. Springer, Heidelberg (2007)
10. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (2006)
11. Wang, D., Vipperla, R., Evans, N.: Online Pattern Learning for Non-Negative Convolutive Sparse Coding. In: Proc. of Interspeech, pp. 65–68 (2011)
12. Wang, F., Li, P., König, A.C.: Efficient document clustering via online nonnegative matrix factorizations. In: SDM, pp. 908–919. SIAM / Omnipress (2011)
13. Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G.: The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments. In: Proc. Internat. Workshop on Machine Listening in Multisource Environments (CHiME 2011), pp. 24–29 (2011)
14. Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: IEEE Intern. Conf. on Acoustics, Speech and Signal Process. pp. 4029–4032 (2008)

An Audio-Video Based IVA Algorithm for Source Separation and Evaluation on the AV16.3 Corpus

Yanfeng Liang and Jonathon Chambers

School of Electronic, Electrical and Systems Engineering,
Loughborough University, UK
{Y.Liang2,J.A.Chambers}@lboro.ac.uk

Abstract. The machine cocktail party problem has been researched for several decades. Although many blind source separation schemes have been proposed to address this problem, few of them are tested by using a real room audio video recording. In this paper, we propose an audio video based independent vector analysis (AVIVA) method, and test it with other independent vector analysis methods by using a real room recording dataset, i.e. the AV16.3 corpus. Moreover, we also use a new method based on pitch difference detection for objective evaluation of the separation performance of the algorithms when applied on the real dataset which confirms advantages of using the visual modality with IVA.

Keywords: blind source separation, independent vector analysis, real room recording, pitch difference.

1 Introduction

The machine cocktail party problem was first proposed by Colin Cherry in 1953 [1]. For a real room environment, the acoustic sources take multiple paths to the microphone sensors instead of only the direct path. Thus the convolutive model is used to represent the practical situation. In recent years, considerable research has been performed in the field of convolutive blind source separation [2]. Initially, research was aimed at solutions based in the time domain. However, real room impulse responses are typically on the order of thousands of samples in length. Thus the computational cost of the time domain methods will be very expensive. A solution in the frequency domain was proposed to solve this problem [3]. Although the frequency domain blind source separation (FD-BSS) method can reduce the computational cost, it has two indeterminacies, i.e. the scaling problem and permutation problem.

The scaling ambiguity can be managed by matrix normalization. For the permutation ambiguity various methods have been proposed [2]. All of these methods need prior knowledge about the locations of the sources or post-processing exploiting some features of the separated signals. Recently, Kim proposed an exciting new algorithm named independent vector analysis (IVA), which preserves

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 330–337, 2012.

the higher order dependencies and structures of signals across frequencies to solve the permutation problem [4]. This method can solve the permutation problem during the unmixing matrix learning process without any prior knowledge or post-processing. Thus, it is a natural way to solve the permutation problem. Based on the original IVA method, several extended IVA methods have been proposed. An adaptive step size IVA method was proposed to improve the convergence speed by controlling the learning step size [5]. The fast fixed-point IVA method introduces a quadratic Taylor polynomial in the notations of complex variables which is very useful in directly applying Newton's method to a contrast function of complex-valued variables and can achieve a fast and good separation [6]. The first contribution in this paper is to propose an audio and video based independent vector analysis method which uses the video information to initialize the algorithm and thereby improve convergence properties.

Moreover, we compare this with the conventional and fast forms of the audio only IVA algorithms by using a real dataset. For real room recording separation, there is a problem in performing objective evaluation. Traditionally, blind source separation experiments are all simulations, for which we know the mixing matrix and the source signals. Thus, we can evaluate the separation performance by performance index [7], signal to interference rate (SIR) or signal to distortion rate (SDR) [8]. However, a real room recording only provides the mixtures. Objective evaluation of the separation performance becomes a problem. The final contribution in the paper is to use a new evaluation based on pitch information. It detects the pitches of the separated signals respectively, and then calculates the pitch differences between them, and provides an objective evaluation. The paper is organized as follows, in Section 2, different IVA methods are introduced, then the audio and video based IVA method is proposed. The pitch difference based evaluation method is proposed in Section 3. The introduction of the real room recording AV16.3 and the separation performance comparisons are provided in Section 4. Finally, conclusions are drawn in Section 5.

2 Independent Vector Analysis Based Methods

2.1 Model

The basic noise free blind source separation generative model is $\mathbf{x} = \mathbf{H}\mathbf{s}$, which is also adopted by IVA; $\mathbf{x} = [x_1, x_2 \cdots x_m]^T$ is the observed mixed signal vector, $\mathbf{s} = [s_1, s_2 \cdots s_n]^T$ is the source signal vector, and $\mathbf{H}$ is the mixing matrix with $m \times n$ dimension, $(\cdot)^T$ denotes the transpose operator. In this paper, we focus on the exactly determined case, namely $m = n$. Our target is to find the inverse matrix $\mathbf{W}$ of mixing matrix $\mathbf{H}$. Due to the scaling and permutation ambiguities, we can not generally obtain $\mathbf{W}$ uniquely. Actually, $\mathbf{W} = \mathbf{P}\mathbf{D}\mathbf{H}^{-1}$, therefore, $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} = \mathbf{P}\mathbf{D}\mathbf{s}$, where $\mathbf{P}$ is a permutation matrix, $\mathbf{D}$ is a scaling diagonal matrix, and $\hat{\mathbf{s}}$ is the estimation of the source signal $\mathbf{s}$.

In practical situations, due to the reverberation, convolutive methods are more often used which are generally implemented in the frequency domain. Thus, the noise free model in the frequency domain is described as:

$$\mathbf{x}^{(k)} = \mathbf{H}^{(k)}\mathbf{s}^{(k)} \tag{1}$$

$$\hat{\mathbf{s}}^{(k)} = \mathbf{W}^{(k)}\mathbf{s}^{(k)} \tag{2}$$

where $\mathbf{x}^{(k)} = [x_1^{(k)}, x_2^{(k)} \cdots x_m^{(k)}]^T$ is the observed signal vector in the frequency domain, and $\hat{\mathbf{s}}^{(k)} = [\hat{s_1}^{(k)}, \hat{s_2}^{(k)} \cdots \hat{s_n}^{(k)}]^T$ is the estimated signal vector in the frequency domain. The index k denotes the kth frequency bin. It is a multivariate model.

2.2 Independent Vector Analysis

In order to separate multivariate sources from multivariate observations, a cost function for multivariate random variables is needed. The IVA method adopts Kullback-Leibler divergence between the joint probability density function $p(\hat{\mathbf{s}})$ and the product of probability density functions of the individual source vectors $\prod q(\hat{\mathbf{s}})$. This is used as the cost function of the IVA model.

$$J = KL(p(\hat{\mathbf{s}})||\prod q(\hat{\mathbf{s}})) = const - \sum_{k=1}^{K} log|det(W^{(k)})| - \sum_{i=1}^{n} E[log(q(\hat{s}_i))] \tag{3}$$

where $E[\cdot]$ denotes the statistical expectation operator, and $det(\cdot)$ is the matrix determinant operator. The cost function would be minimized when the dependency between the source vectors is removed but the dependency between the components of each vector can be retained. Thus, the cost function preserves the inherent frequency dependency within each source, but it removes the dependency between the sources [4].

The gradient descent method is used to minimize the cost function. By differentiating the cost function J with respect to the coefficients of the separating matrices $w_{ij}^{(k)}$, the gradients for the coefficients can be obtained as follows:

$$\varDelta w_{ij}^{(k)} = -\frac{\partial J}{\partial w_{ij}^{(k)}} = (w_{ij}^{(k)})^{-H} - E[\varphi^{(k)}(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(k)})]x_j^{*(k)} \tag{4}$$

where $(\cdot)^H$ and $(\cdot)^*$ denote the Hermitian transpose and the conjugate operator respectively, and $\varphi^{(k)}(\cdot)$ is the nonlinear function. This nonlinearity represents the core idea of the IVA method, which is the main difference between traditional ICA and IVA. In ICA, this nonlinear function is single-variate. However, for IVA, it becomes multi-variate. As discussed in [4], the Laplacian distribution is used as the source prior, and we can obtain a simple but effective form for the nonlinear function as:

$$\varphi^{(k)}(\hat{s}_i^{(1)} \cdots \hat{s}_i^{(k)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_1^K |\hat{s}_i^{(k)}|^2}} \tag{5}$$

2.3 Fast Fixed-Point Independent Vector Analysis

Fast fixed-point independent vector analysis (FastIVA) employs Newton's method update rules, which converges fast and is free from selecting an efficient learning rate. In order to apply Newton's method update rules, it introduces a quadratic Taylor polynomial in the notations of complex variables which can be used for a contrast function of complex-valued variables [6]. The contrast function used by FastIVA is as follows:

$$J = \sum_i (E[G(\sum_k |\hat{s}_i^{(k)}|^2)] - \sum_k \lambda_i^{(k)} (\mathbf{w}_i^{(k)} (\mathbf{w}_i^{(k)})^H - 1)) \tag{6}$$

where, $\mathbf{w}_i$ is the i-th row of the unmixing matrix $\mathbf{W}$, λ_i is the i-th Lagrange multiplier. $G(\cdot)$ is the nonlinearity function, which has several different types as discussed in [6]. With normalization, the learning rule is:

$$\begin{aligned}(\mathbf{w}_i^{(k)})^H \leftarrow & E[G^{'}(\sum_k |\hat{s}_i^{(k)}|^2) + |\hat{s}_i^{(k)}|^2 G^{''}(\sum_k |\hat{s}_i^{(k)}|^2))](\mathbf{w}_i^{(k)})^H \\ & - E[(\hat{s}_i^{(k)})^* G^{'}(\sum_k |\hat{s}_i^{(k)}|^2) \mathbf{x}^k]\end{aligned} \tag{7}$$

and if this is used for all sources, an unmixing matrix $\mathbf{W}^{(k)}$ can be constructed which must be decorrelated with

$$\mathbf{W}^{(k)} \leftarrow (\mathbf{W}^{(k)} (\mathbf{W}^{(k)})^H)^{-1/2} \mathbf{W}^{(k)} \tag{8}$$

2.4 Audio-Video Based Independent Vector Analysis

For human beings, when we solve the cocktail party problem, we not only use our ears but also our eyes. The video information is potentially helpful to solve the machine cocktail party problem. The positions of the sources can be obtained by the video information. Then a smart initialization of the unmixing matrix can be achieved, which will potentially lead to a faster convergence and better performance. In this paper, the video information is combined with the FastIVA algorithm to propose the audio-video based independent vector analysis (AVIVA) algorithm.

The mixing matrix can be calculated under the plane wave propagation assumption by using the positions of the sources which are captured by video

$$\mathbf{H}^{(k)} = [\mathbf{h}(k, \theta_1, \phi_1) \cdots \mathbf{h}(k, \theta_n, \phi_n)] \tag{9}$$

where

$$\mathbf{h}(k, \theta_i, \phi_i) = \begin{bmatrix} exp(-j\kappa(sin(\theta_i).cos(\phi_i).u_{x_1} + sin(\theta_i). \\ sin(\phi_i).u_{y_1} + cos(\theta_i).u_{z_1})) \\ \vdots \\ exp(-j\kappa(sin(\theta_i).cos(\phi_i).u_{x_m} + sin(\theta_i). \\ sin(\phi_i).u_{y_m} + cos(\theta_i).u_{z_m})) \end{bmatrix} \tag{10}$$

and $\kappa = k/c$ where c is the speed of sound in air at room temperature. The coordinates u_{x_i}, u_{y_i} and u_{z_i} are the 3-D positions of the i-th microphone. The parameters (θ_i) and (ϕ_i) are the elevation angle and azimuth angle of arrival to the center of the microphone array, which can be obtained by the 3D visual tracker as in [9].

Thus, the initialization of the unmixing matrix can be obtained by following the approach in [10].

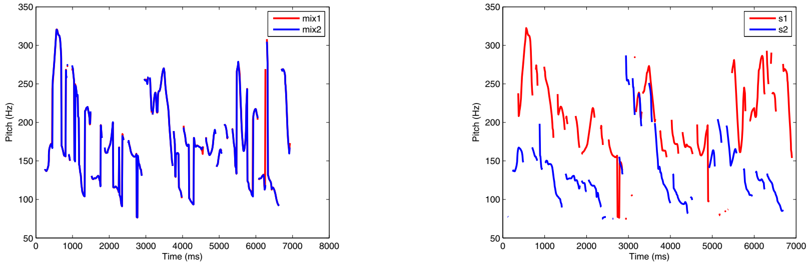
$$\mathbf{W}^{(k)} = \mathbf{Q}^{(k)}\mathbf{H}^{(k)} \tag{11}$$

where Q is the whitening matrix. After that, it can be used as the initialization of the unmixing matrix of FastIVA rather than an identity matrix or random matrix.

3 Pitch Difference Based Evaluation for Real Recordings

For real recording, the only thing we obtain is the mixed signals captured by the microphone array. We can not access either the mixing matrix or the pure source signals. Thus, we can not evaluate the separation performance by traditional methods, such as performance index [7] which is based on the prior knowledge of the mixing matrix, or the SIR or SDR [8] which require prior knowledge about the source signals. It is a tough problem to evaluate objectively real recording separation performance. We can listen to the separated signals, but it is just a kind of subjective evaluation. In order to evaluate the results objectively, the features of the separated signals should be used. Pitch information is one of the features which can help to evaluate the separation performance, because different speech sections at different time slots have different pitches [11]. We adopt the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) method [12], which has better performance compared with traditional pitch estimators.

Fig. 1 shows that the pitches of mixed signals are still mixed, while the pitches of source signals are well separated. It is obvious that good separated pitches can indicate good separation performance provided that the original sources do



(a) (b)

Fig. 1. Comparison of the pitch of a mixture signal and separated signals. (a) the pitches of the mixed signal (b) the pitches of separated signal.

not have substantially overlapping pitch characteristics. In order to evaluate objectively, we calculate the pitch difference:

$$p_{diff}(t) = \sqrt{\sum_{i \neq j}(p_i(t) - p_j(t))^2} \quad i,j = 1, \cdots, m \; t = 1, \cdots, T \tag{12}$$

where T is the number of time slots. Then we set a threshold p_{thr}, if the pitch difference is greater than the threshold at a certain time slot, we can consider that the mixed signals are separated at that time slot and set the separation status equal to 1, otherwise 0.

$$sep\_status(t) = \begin{cases} 1 & if \quad p_{diff}(t) > p_{thr} \\ 0 & otherwise \end{cases} \tag{13}$$

Finally, we can calculate a separation rate to evaluate the separation performance.

$$sep\_rate = \frac{\sum_t sep\_status(t)}{T} \tag{14}$$

The bigger value that the separation rate takes, the better the separation performance. We need to highlight here that it can not evaluate the absolute quality of the separated signal, but it can be used for comparing the separation performance when using different separation methods.

4 Experiments and Results

The real recording used in our experiments is the AV 16.3 corpus [13], which is recorded in a meeting room context. 16.3 stands for 16 microphones and 3 cameras, recorded in a fully synchronized manner. We use the "seq37-3p-0001" recording to perform the experiment, which contains three speakers. Fig.2 shows the room environment, the positions of microphone arrays and the positions of the three speakers. There are two microphone arrays, we choose three microphones (mic3, mic5 and mic7) from microphone array 1 which is in the red circle. The sampling frequency of the recording is 16kHz. The pitch threshold in (13) is set to 5.

We extract the recorded speech from 200s to 220s, during which three speakers are speaking simultaneously. Then, the positions of the speakers are obtained by using the video information. After that, IVA, FastIVA and AVIVA are applied respectively. The experimental results are shown in Fig.3 and Table 1. Fig. 3(a) shows that the pitches of the mixed signals are all mixed. Fig. 3(b),(c),(d) are the separation results by using IVA, FastIVA and AVIVA respectively. It is clear that the pitches are separated, which indicates that the mixed signals are separated. The objective evaluation separation rate and iteration number are shown in Table.1, which confirms that the proposed AVIVA algorithm can achieve the best separation rate by using the least iterations comparing with IVA and FastIVA algorithms.

Further evaluation on the AV16.3 corpus will be presented at the conference.

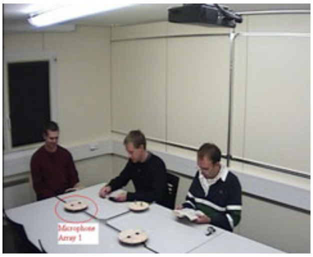
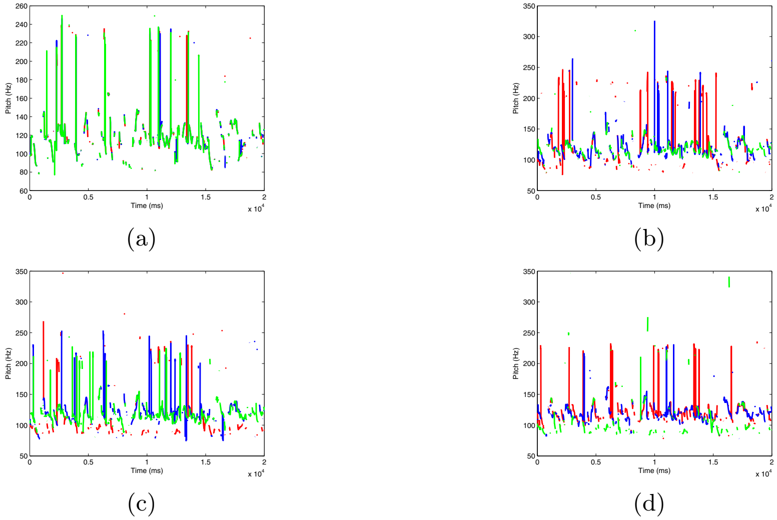


Fig. 2. A single video frame showing the room environment for one of the AV16.3 corpus recordings



(a)

(b)

(c)

(d)

Fig. 3. Separation result comparison by using different IVA methods. (a) mixed signal (b) IVA (c) FastIVA (d)AVIVA.

Table 1. Separation performance comparison for three sources

method	IVA	FastIVA	AVIVA
sep_rate	0.1364	0.1479	0.1604
iter number	100	77	71

5 Conclusion

In this paper, different independent vector analysis methods are introduced, and an audio-video based independent vector analysis method is proposed. Real room recording separation performance evaluation is hard to achieve due to the lack of prior knowledge such as mixing matrix and source signals. A pitch difference based evaluation method is proposed to evaluate objectively the separation performance of the real recording. The experimental results on the real recordings confirm that although all the IVA algorithms can achieve some degree of separation the proposed audio-video based method has the best separation rate with much improved convergence rate as compared to the basic IVA algorithm.

References

1. Cherry, C.: Some experiments on the recognition of speech, with one and with two years. The Journal of The Acoustical Society of America 25, 975–979 (1953)
2. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: A survey of convolutive blind source separation methods. In: Springer Handbook on Speech Processing and Speech Communication, pp. 1–34 (2007)
3. Parra, L.C., Spence, C.: Convolutive blind separation of non-statinary sources. IEEE Transcations on Speech and Audio Processing 8, 320–327 (2000)
4. Kim, T., Attias, H., Lee, S., Lee, T.: Blind Source Separation exploiting higher-order frequency dependencies. IEEE Transcations on Speech and Audio Processing 15, 70–79 (2007)
5. Liang, Y., Naqvi, M., Chambers, J.: Adaptive step size indepndent vector analysis for blind source separation. In: 17th International Conference on Digital Signal Processing, Corfu, Greece (2011)
6. Lee, I., Kim, T., Lee, T.: Fast fixed-point independent vector analysis algorithm for convolutive blind source separation. Signal Processing 87, 1859–1971 (2007)
7. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing: learning algorithms and applications. Wiley (2000)
8. Vincent, E., Fevotte, C., Gribonval, R.: Performance measurement in blind audio source separation. IEEE Transcations on Speech and Audio Processing 14, 1462–1469 (2006)
9. Naqvi, S.M., Yu, M., Chambers, J.A.: A Multimodal Approach to Blind Source Separation of Moving Sources. IEEE Journal of Selected Topics in Signal Processing 4(5), 895–910 (2010)
10. Naqvi, S.M., Zhang, Y., Tsalaile, T., Sanei, S., Chambers, J.A.: A multimodal approach for frequency domain independent component analysis with geometrically-based initialization. In: Proc. EUSIPCO 2008, Lausanne, Switzerland (2008)
11. Shabani, H., Kahaei, M.H.: Missing feature mask generation in BSS outputs using pitch frequency. In: 17th International Conference on Digital Signal Processing, Corfu, Greece (2011)
12. Camacho, A., Harris, J.G.: A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. 124(3), 1638–1652 (2008)
13. Lathoud, G., Odobez, J.-M., Gatica-Perez, D.: AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking. In: Bengio, S., Bourlard, H. (eds.) MLMI 2004. LNCS, vol. 3361, pp. 182–195. Springer, Heidelberg (2005)

Non-negative Matrix Factorization Based Noise Reduction for Noise Robust Automatic Speech Recognition

Seon Man Kim[1], Ji Hun Park[1], Hong Kook Kim[1,*],
Sung Joo Lee[2], and Yun Keun Lee[2]

[1] School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea
{kobem30002,jh_park,hongkook}@gist.ac.kr
[2] Speech/Language Information Research Center
Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea
{lee1862,yklee}@etri.re.kr

Abstract. In this paper, we propose a noise reduction method based on non-negative matrix factorization (NMF) for noise-robust automatic speech recognition (ASR). Most noise reduction methods applied to ASR front-ends have been developed for suppressing background noise that is assumed to be stationary rather than non-stationary. Instead, the proposed method attenuates non-target noise by a hybrid approach that combines a Wiener filtering and an NMF technique. This is motivated by the fact that Wiener filtering and NMF are suitable for reduction of stationary and non-stationary noise, respectively. It is shown from ASR experiments that an ASR system employing the proposed approach improves the average word error rate by 11.9%, 22.4%, and 5.2%, compared to systems employing the two-stage mel-warped Wiener filter, the minimum mean square error log-spectral amplitude estimator, and NMF with a Wiener post-filter , respectively.

Keywords: Automatic speech recognition (ASR), Non-negative matrix factorization (NMF), Noise reduction, Non-stationary background noise, Wiener filter.

1 Introduction

Most automatic speech recognition (ASR) systems often suffer considerably from unexpected background noise [1]. Thus, many noise-robust methods in the frequency domain have been reported such as spectral subtraction [2], minimum mean square error log-spectral amplitude (MMSE-LSA) estimation [3], and Wiener filtering [4][5]. In general, conventional front-ends employing such noise reduction methods perform well in stationary noise environments but not always in non-stationary ones. This is because noise reduction is usually performed by estimating noise components during

[*] This work was supported in part by the ISTD Program, 10035252, and by the Basic Science Research Program through the NRF of Korea funded by the MOST, 2011-0026201.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 338–346, 2012.

the period when target speech is declared inactive under the stationary noise assumption [1][4].

On the other hand, a non-negative matrix factorization (NMF) technique [6] can provide an alternative to estimate target speech from an observed noisy signal. However, the performance of noise reduction methods based on NMF might be degraded when speech and noise have similar distributions in the frequency domain [7]. In other words, there is a large overlap between speech and noise in the frequency domain, thus a certain degree of residual noise remains in the estimated target speech while some speech components are apt to be missed in the target speech. To overcome this problem, we have proposed an NMF-based target speech enhancement method [8], where a Wiener filter was applied to a weighted-sum of speech bases in order to remove the residual noise from the estimated speech. In particular, the temporal continuity constraint technique [9] was also employed so that the characteristics of residual noise remained in the estimated NMF-based target speech became stationary. On the other hand, the target speech was a little damaged after the NMF procedure, even though a regularization technique [7] had been used. Therefore, we need to mitigate such a problem.

In order to mitigate the problem mentioned above, we propose a noise reduction method based on non-negative matrix factorization (NMF) and apply it to noise-robust ASR. The proposed method attenuates non-target noise by a hybrid approach that combines a Wiener filtering and an NMF technique. In addition, stationary noise is estimated from recursively averaging noise components during inactive speech intervals. On the other hand, non-stationary noise is estimated as the difference between the original noise and the estimated noise variance based on recursive averaging. After that, the estimated stationary and non-stationary noises are reduced by Wiener filtering and NMF, respectively. Note here that the NMF bases of the non-stationary noise are trained using a non-stationary noise database (DB), which is generated from an original noise DB.

The rest of this paper is organized as follows. Section 2 proposes an NMF-based noise reduction method. Section 3 demonstrates the effect of the proposed method on ASR performance, and Section 4 concludes this paper.

2 Proposed NMF-Based Noise Reduction Method for ASR

Fig. 1 shows an overall procedure of the proposed noise reduction method which combines NMF with a conventional Wiener filter. As shown in the figure, in the training stage the speech and non-stationary noise bases, $\overline{\mathbf{B}}_S$ and $\overline{\mathbf{B}}_D$, are estimated from speech and non-stationary noise databases (DBs), $\overline{\mathbf{S}}$ and $\overline{\mathbf{D}}$, respectively. In particular, the non-stationary noise DB, $\overline{\mathbf{D}}$, is obtained by applying a recursive averaging method [10] to the original noise DB, $\overline{\mathbf{Y}}$. Note that $\overline{\mathbf{S}}$ or $\overline{\mathbf{Y}}$ is represented in a matrix form by concatenating a sequence of absolute values of speech or noise spectra along the analysis frame, respectively. In the noise reduction stage, activation matrices of target speech and non-stationary (or residual) noise, $\mathbf{A}_S$ and $\mathbf{A}_D$, are estimated by an NMF multiplicative updating rule in order to approximate Wiener

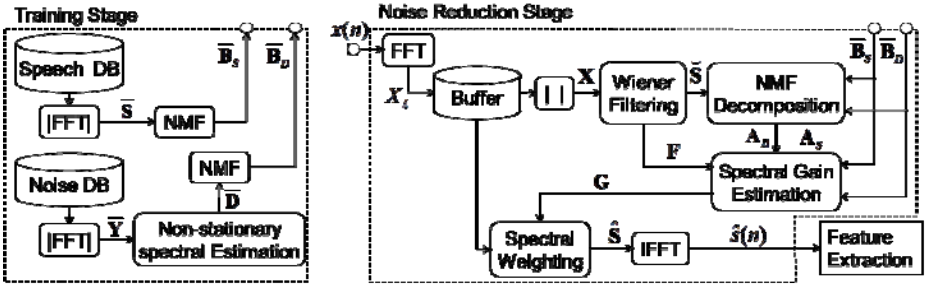
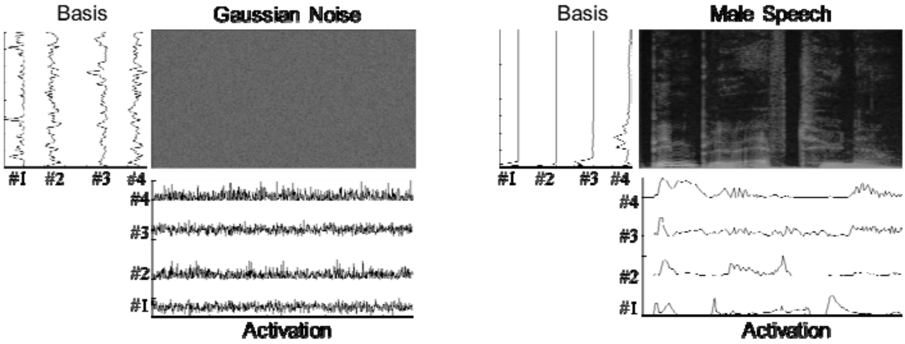


Fig. 1. Block diagram of the proposed NMF-based noise reduction technique applied as a front-end of ASR

filtered outputs, $\breve{\mathbf{S}}$, using known $\bar{\mathbf{B}}_S$ and $\bar{\mathbf{B}}_D$. Then, the weighting value matrix for noise spectral attenuation, $\mathbf{G}$, is obtained from the Wiener filter coefficient matrix, $\mathbf{F}$, and the NMF decomposition outputs, $\mathbf{A}_S$, $\mathbf{A}_D$, $\bar{\mathbf{B}}_S$, and $\bar{\mathbf{B}}_D$. Next, target speech spectral components, $\hat{\mathbf{S}}$, are obtained from $\mathbf{G}$. After that, target speech spectral components are transformed into a time-domain signal, $\hat{s}(n)$, by using an overlap-add method, and $\hat{s}(n)$ is finally used for mel-frequency cepstral coefficient (MFCC) extraction for ASR.

Let $x(n)$, $s(n)$, and $y(n)$ be noisy speech, target speech, and additive noise, respectively, where $x(n)$ and $y(n)$ are assumed to be uncorrelated. In addition, we have $X_k(\ell) = S_k(\ell) + Y_k(\ell)$, where $X_k(\ell)$, $S_k(\ell)$, and $Y_k(\ell)$ denote the spectral components of $x(n)$, $s(n)$, and $y(n)$, respectively, at the k-th frequency bin index $(k = 0,1,\cdots,K-1)$ and ℓ-th segmented frame index $(\ell = 0,1,2,\cdots)$.

As mentioned in Section 1, the performance of an NMF-based noise reduction method could be degraded when speech and noise have similar distributions in the frequency domain [8]. Figs. 2(a) and 2(b) show the spectral distribution and basis



(a) (b)

Fig. 2. Examples of NMF bases and activations for (a) Gaussian noise and (b) male speech

distribution for Gaussian noise and male speech, respectively. As shown in the figure, the spectrum of Gaussian noise is distributed across a wide range of frequencies and overlapped with that of male speech. Similarly, NMF bases of Gaussian noise are also widely distributed and overlapped with those of the male speech. Compared to the male speech, Gaussian noise, which is one of the typical stationary noises, can be removed well by conventional noise reduction methods such as a Wiener filter and an MMSE-LSA. Based on this observation, Wiener filtering and the NMF technique are combined in the proposed method.

Accordingly, it is assumed that $Y_k(\ell)$ is decomposed into stationary noise, $V_k(\ell)$, and non-stationary noise, $D_k(\ell)$; i.e., $Y_k(\ell) = V_k(\ell) + D_k(\ell)$. Assuming that a weight value, $G_{V,k}(\ell)$, for reducing stationary noise gives little damage on target speech, multiplying $G_{V,k}(\ell)$ to $X_k(\ell)$ provides the sum of the estimate of target speech and non-stationary noise such as $X_k(\ell) \cdot G_{V,k}(\ell) = \breve{S}_k(\ell) = \hat{S}_k(\ell) + D_k(\ell)$. As a next step, a weighting value, $G_{D,k}(\ell)$, for the residual noise attenuation is applied to $\breve{S}_k(\ell)$ in order to reduce non-stationary noise from $\breve{S}_k(\ell)$, which results in more enhanced target speech, $\hat{S}_k(\ell)$. That is, $\breve{S}_k(\ell) G_{D,k}(\ell) = \hat{S}_k(\ell)$. By combining these two steps, a weighting value, $G_k(\ell)$, for both the stationary and non-stationary noise reduction can be represented as the product of two weighting values, $G_{V,k}(\ell)$ and $G_{D,k}(\ell)$, such that $\hat{S}_k(\ell) = G_k(\ell) X_k(\ell)$, where $G_k(\ell) = G_{V,k}(\ell) \cdot G_{D,k}(\ell)$.

2.1 Stationary Noise Reduction Based on Wiener Filtering

In this subsection, we explain how to obtain $G_{V,k}(\ell)$ for stationary noise reduction. First, spectral variance of stationary noise, $\hat{\lambda}_{V,k}(\ell)$, is estimated by the recursive averaging method that is executed only when target speech absence is declared [11]. That is, $\hat{\lambda}_{V,k}(\ell) = \zeta_V \hat{\lambda}_{V,k}(\ell - 1) + (1 - \zeta_V) | X_k(\ell) |^2$ if target speech is absent, where ζ_V is a forgetting factor. Then, $G_{V,k}(\ell)$ is represented by employing the *a priori* SNR estimate by the decision-directed (DD) approach [3], $\hat{\xi}_k(\ell)$, as

$$G_{V,k}(\ell) = \frac{\hat{\xi}_k(\ell)}{\hat{\xi}_k(\ell) + 1}. \tag{1}$$

2.2 Non-stationary Noise Reduction Based on NMF

In this subsection, we explain how to estimate $G_{D,k}(\ell)$ for non-stationary (or residual) noise reduction by using the NMF technique. NMF is an algorithm for multivariate data analysis that decomposes a $K \times L$ matrix, $\mathbf{V}_{K \times L}$, into the product of a basis matrix, $\mathbf{B}_{K \times R}$, and an activation matrix, $\mathbf{A}_{R \times L}$; i.e., $\mathbf{V}_{K \times L} \approx \mathbf{B}_{K \times R} \mathbf{A}_{R \times L}$, where K, R, and L correspond to the number of spectral channels, the rank of the basis vector, and the number of frames, respectively. From now on, each matrix is represented in the text without any subscript for the simplicity. To find $\mathbf{B}$ and $\mathbf{A}$, two kinds of cost functions are commonly used [6]: the Euclidean distance and the Kullback–Leibler (KL) divergence. For speech processing, NMF using the KL divergence shows better performance than that using the Euclidean distance [7], thus the KL divergence, $Div(\mathbf{X} \,\|\, \mathbf{BA})$, is used in this paper, defined as [12]

$$Div(\mathbf{X} \,\|\, \mathbf{BA}) = \sum_{i,j} \left(\mathbf{X}_{i,j} \log \frac{\mathbf{X}_{i,j}}{(\mathbf{BA})_{i,j}} - \mathbf{X}_{i,j} + (\mathbf{BA})_{i,j} \right) \tag{2}$$

where *i* and *j* indicate the row and column index of a matrix, respectively. By applying NMF, target speech estimated from Wiener filtering in Section 2.1, $\breve{\mathbf{S}}$, is further decomposed into target speech, $\hat{\mathbf{S}}$, and non-stationary noise, $\mathbf{D}$, by $Div(\breve{\mathbf{S}} \,\|\, \mathbf{BA})$, as

$$\breve{\mathbf{S}}_{K\times L} = \hat{\mathbf{S}}_{K\times L} + \mathbf{D}_{K\times L} \approx \big[\mathbf{B}_{S,K\times R_S}; \mathbf{B}_{D,K\times R_D}\big]\big[\mathbf{A}_{S,R_S\times L}; \mathbf{A}_{D,R_D\times L}\big] = \mathbf{B}_{K\times R}\mathbf{A}_{R\times L} \tag{3}$$

where R_S and R_D are the rank of the basis vectors for speech and non-stationary noise, respectively, and $R = R_S + R_D$. In Eq. 3, the basis matrix $\mathbf{B}(=[\mathbf{B}_S, \mathbf{B}_D])$ is replaced with the pre-trained matrix, $\overline{\mathbf{B}}(=[\overline{\mathbf{B}}_S, \overline{\mathbf{B}}_D])$, assuming that $\overline{\mathbf{B}}_S$ and $\overline{\mathbf{B}}_D$ hold the ability for constructing current speech and noise, respectively. Thus, we have $\mathbf{X} \approx \overline{\mathbf{B}}\mathbf{A}$. To obtain the non-stationary noise basis matrix, $\overline{\mathbf{B}}_D$, the non-stationary noise DB, $\overline{\mathbf{D}}$, is generated from the original noise DB, $\overline{\mathbf{Y}}$. As mentioned earlier, original noise is decomposed into non-stationary noise and stationary noise. Thus, the estimate of the variance, $\overline{\lambda}_{V,k}(\ell)$, is obtained by

$$\overline{D}_k(\ell) = \max(\overline{Y}_k(\ell) - \overline{\lambda}_{V,k}(\ell), 0)_{\forall k,\ell} \tag{4}$$

where $\overline{\lambda}_{V,k}(\ell)$, is the estimate of non-stationary noise by the recursive averaging method. To estimate the activation matrix, $\mathbf{A}(=[\mathbf{A}_S \,; \mathbf{A}_D])$, the activation matrix $\mathbf{A}$ is first randomly initialized. Then, the cost function in Eq. 2 is minimized by iteratively applying an updating rule defined as [12]

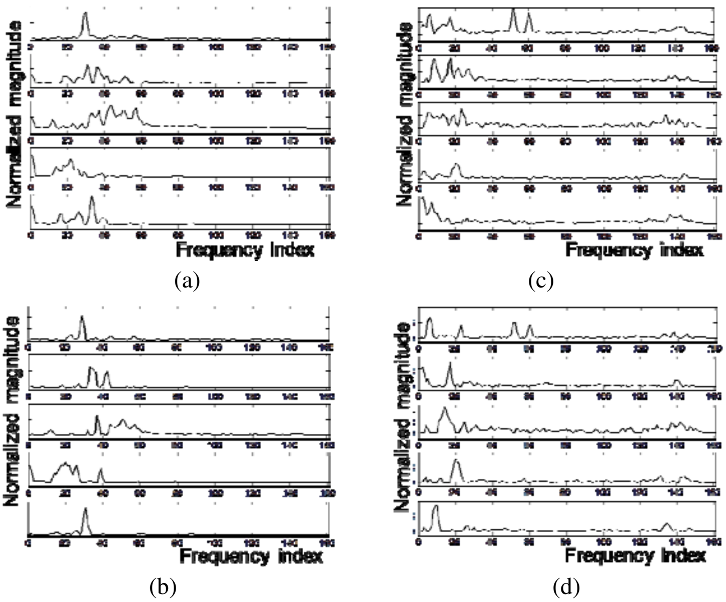
$$\mathbf{A}^{m+1} = \mathbf{A}^m \otimes \frac{\mathbf{B}^{\mathrm{T}} \frac{\mathbf{X}}{\mathbf{BA}}}{\mathbf{B}^{\mathrm{T}}\mathbf{1}} \tag{5}$$

where *m* represents an iteration number, and $\mathbf{1}$ is a matrix with all elements equal to unity. Moreover, both multiplication ⊗ and division denote the element-wise operators. Hence, the weighting value for non-stationary noise reduction is represented as

$$\mathbf{G}_{\mathbf{D}} = \frac{\overline{\mathbf{B}}_{\mathbf{S}}\mathbf{A}_{\mathbf{S}}}{\overline{\mathbf{B}}_{\mathbf{S}}\mathbf{A}_{\mathbf{S}} + \overline{\mathbf{B}}_{\mathbf{D}}\mathbf{A}_{\mathbf{D}}} \tag{6}$$

or

$$G_{D,k}(\ell) = \frac{\sum_{rs=1}^{R_S} \left[\overline{B}_{S,k^i,rs} \; A_{S,rs,\ell^j}\right]}{\sum_{rs=1}^{R_S} \left[\overline{B}_{S,k^i,rs} \; A_{S,rs,\ell^j}\right] + \sum_{rd=1}^{R_D} \left[\overline{B}_{D,k^i,rd} \; A_{D,rd,\ell^j}\right]} \tag{7}$$



(a) (c)

(b) (d)

Fig. 3. Examples of noise bases for two difference noise signals ((a) and (c)) and the estimate of non-stationary noise ((b) and (d))

where k^i and ℓ^j indicate the row and column index of a matrix, respectively, and they correspond to the *k*-th frequency channel and the ℓ-th frame. Eq. 7 implies that the NMF-based noise reduction can also be interpreted as filtering the noisy signal with a time-varying filter, which is similar to the Wiener filtering in Eq. 2.

Fig. 3 shows the bases of the two different noise signals and the estimated non-stationary noise signals obtained from Eq. 7. It is shown from the figure that the bases of the estimated non-stationary noise (Figs. 3(b) and 3(d)) are more localized over frequency than those of the original noise bases (Figs. 3(a) and 3(c)).

2.3 Target Speech Reconstruction

The combined weighting value, $G_k(\ell)$, for noise reduction is represented as the product of $G_{V,k}(\ell)$ and $G_{D,k}(\ell)$ that are described in Eqs. 1 and 7, respectively. That is, $G_k(\ell) = G_{V,k}(\ell) G_{D,k}(\ell)$. Thus, the target speech estimate, $\hat{S}_k(\ell)$, is obtained by multiplying $G_k(\ell)$ to $X_k(\ell)$, and it is transformed into a time-discrete signal, $\hat{s}(n)$, that is finally brought to the MFCC extraction for speech recognition.

3 Speech Recognition Experiments

The performance of the proposed noise reduction method was evaluated in a view of ASR performance. First of all, a word recognition system in several different background noise environments was constructed, where acoustic models were

Table 1. Comparison of average word error rates (WERs) (%)

SNR	Bus Stop					Home TV				
(dB)	No	[3]	[4]	[8]	Proposed	No	[3]	[4]	[8]	Proposed
20	13.0	10.7	10.0	10.7	10.6	19.3	32.3	25.6	18.3	16.1
15	20.5	13.9	12.8	13.6	10.3	28.7	45.5	34.1	26.1	22.2
10	45.4	21.3	20.2	20.6	18.4	42.1	60.1	46.0	36.0	35.2
5	81.4	34.3	32.3	33.5	30.2	65.7	82.0	64.6	51.1	54.0
0	97.9	77.3	74.4	67.3	71.6	88.1	101.4	86.0	77.6	75.8
Avg.	51.6	31.5	30.0	29.1	28.2	48.8	64.3	51.3	41.8	40.7

SNR	Restaurant					Subway				
(dB)	No	[3]	[4]	[8]	Proposed	No	[3]	[4]	[8]	Proposed
20	14.6	11.5	12.7	12.8	11.8	11.9	13.4	12.7	12.9	10.5
15	22.6	17.7	15.3	15.1	15.2	16.3	13.2	13.2	13.2	10.2
10	48.7	24.2	22.2	21.6	20.1	35.1	19.2	16.8	18.8	14.3
5	86.2	43.0	36.1	39.2	34.8	74.5	33.6	31.6	33.3	28.4
0	100	80.8	72.1	70.1	71.4	97.5	67.7	69.2	66.0	64.0
Avg.	54.4	35.5	31.7	31.7	30.7	47.0	29.4	28.7	28.8	25.5

tri-phone based three-state left-to-right hidden Markov models (HMMs). The context-dependent acoustic models were trained from around 170,000 phonetically balanced words [5] recorded from 1,800 persons in quiet environments, where speech signals were sampled at a rate of 16 kHz with 16-bit resolution. As a speech recognition feature, a feature extraction procedure was applied once every 20 ms frame. In other words, 13 mel-frequency cepstral coefficients (MFCCs) including the zeroth order were extracted, and their first two derivatives were added, which resulted in a 39-dimensional feature vector per 20 ms frame.

A speech database was collected using a mobile phone, where there were 20 speakers (10 males and 10 females) and each speaker pronounced 40 utterances in a quiet office. On one hand, two sets of four different environmental noises were recorded such as bus stops, home TV, restaurants, and subways. A noise basis matrix was trained for each noise, whose length was 10 seconds long, from the first noise set. In order to obtain the NMF bases for speech, half of the speech database was used. Note here that each speaker had his/her own NMF bases that were kind of speaker-dependent NMF bases. In this paper, the rank of each basis vector for speech and noise were set at R_S=100 and R_D=50, respectively.

The other noise set was used to generate a test database. That is, each of half of utterances for a speaker was artificially added by each of four different environmental noises, where signal-to-noise ratios (SNRs) varied from 0 to 20 dB with a step of 5 dB. In total, there were 400 noisy speech utterances for the test.

Table 1 compares average word error rates (WERs) of an ASR system employing the proposed method with those employing conventional noise reduction methods such as MMSE-LSA [3], the two-stage mel-warped Wiener filter (Mel-WF) [4], and the NMF-Wiener filter (NMF-WF) [8]. As shown in the table, MMSE-LSA gave the lowest performance in all noise environments under all SNR conditions. On the other hand, the Mel-WF and NMF-WF achieved similar WERs at bus stops, in restaurants, and on subways. However, in the home TV noise environment, NMF-WF outperformed Mel-WF. Note that the non-stationary components in home TV noise environment were more dominant than those in other noise environments. Comparing to NMF-WF, the proposed method provided smaller WER under all different SNRs and noise types. In other words, an ASR system employing the proposed method relatively reduced average WER by 5.2% compared to that using NMF-WF. Moreover, the proposed method provided WER reduction of 11.9% and 22.4% compared to Mel-WF and MMSE-LSA, respectively.

4 Conclusion

In this paper, we proposed an NMF-based noise reduction method for noise-robust ASR. To this end, stationary components in observed noisy speech were reduced by Wiener filtering. Next, an NMF-based decomposition technique was applied to remove the residual non-stationary noise that remained after the Wiener filter processing. In particular, the NMF bases of the residual noise were trained using the non-stationary noise database, estimated from an original noise database. It was shown from the ASR experiments that an ASR system employing the proposed method performed better than those using the conventional two-stage mel-warped Wiener filter, the MMSE-LSA estimator, and the NMF-Wiener filter.

References

1. Wu, J., Droppo, J., Deng, L., Acero, A.: A noise-robust ASR front-end using Wiener filter constructed from MMSE estimation of clean speech and noise. In: IEEE Workshop on ASRU, pp. 321–326 (2003)
2. Choi, H.C.: Noise robust front-end for ASR using spectral subtraction, spectral flooring and cumulative distribution mapping. In: 10th Australian Int. Conf. on Speech Science and Technology, pp. 451–456 (2004)
3. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33(2), 443–445 (1985)
4. Agarwal, A., Cheng, Y.M.: Two-stage mel-warped Wiener filter for robust speech recognition. In: IEEE Workshop on ASRU, pp. 67–70 (1999)
5. Lee, S.J., Kang, B.O., Jung, H.Y., Lee, Y.K., Kim, H.S.: Statistical model-based noise reduction approach for car interior applications to speech recognition. ETRI Journal 32(5), 801–809 (2010)
6. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
7. Wilson, K.K., Raj, B., Smaragdis, R., Divakaran, A.: Speech denoising using non-negative matrix factorization with priors. In: ICASSP, pp. 4029–4032 (2008)

8. Kim, S.M., Kim, H.K., Lee, S.J., Lee, Y.K.: Noise robust speech recognition based on a non-negative matrix factorization. In: Inter-noise 2011 (2011)
9. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Speech Audio Process. 15(3), 1066–1074 (2007)
10. Malah, D., Cox, R., Accardi, A.J.: Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. In: ICASSP, pp. 789–792 (1999)
11. Sohn, J., Kim, N.S., Sung, W.: statistical model-based voice activity detection. IEEE Signal Process. Lett. 6(1), 1–3 (1999)
12. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. In: Adv. Neural Inform. Process. Sys., vol. 13, pp. 556–562 (2000)

Audio Imputation Using the Non-negative Hidden Markov Model

Jinyu Han[1,⋆], Gautham J. Mysore[2], and Bryan Pardo[1]

[1] EECS Department, Northwestern University
[2] Advanced Technology Labs, Adobe Systems Inc.

Abstract. Missing data in corrupted audio recordings poses a challenging problem for audio signal processing. In this paper we present an approach that allows us to estimate missing values in the time-frequency domain of audio signals. The proposed approach, based on the Non-negative Hidden Markov Model, enables more temporally coherent estimation for the missing data by taking into account both the spectral and temporal information of the audio signal. This approach is able to reconstruct highly corrupted audio signals with large parts of the spectrogram missing. We demonstrate this approach on real-world polyphonic music signals. The initial experimental results show that our approach has advantages over a previous missing data imputation method.

1 Introduction

The problem of missing data in an audio spectrogram occurs in many scenarios. For example, the problem is common in signal transmission, where the signal quality is degraded by linear or non-linear filtering operations. In other cases, audio compression and editing techniques often introduce spectral holes to the audio. Missing values also occur frequently in the output of audio source separation algorithms, due to time-frequency component masking [2]. Audio imputation is the task of filling in missing values of the audio signal to improve the perceived quality of the resulting signal. An effective approach for audio imputation could benefit many important applications, such as bandwidth extension, sound restoration, audio declipping, and audio source separation.

Audio imputation from highly corrupted recordings can be a challenging problem. The popular existing generic imputation algorithm [1] is usually ill-suited for use with audio signals and results in audible distortions. Other algorithms such as those in [6] are suitable for imputation of speech, or in the case of musical audio [3] or [7]. However, these algorithms treat individual time frames of the spectrogram as independent of adjacent time frames, disregarding the important temporal dynamics of sound, which makes them less effective for complex audio scenes or severely corrupted audio.

In this paper, we propose an audio imputation algorithm, based on the Non-negative Hidden Markov Model (N-HMM) [4], which takes the temporal dynamics of audio into consideration. The N-HMM jointly learns several small spectral

[⋆] This work was supported in part by National Science Foundation award 0643752.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 347–355, 2012.

dictionaries as well as a Markov chain that describes the structure of transitions between these dictionaries. We extend the N-HMM for missing values imputation by formulating the imputation problem in an Expectation–Maximization (EM) framework. We show promising performance of the proposed algorithm by comparing it to an existing imputation algorithm on real-world polyphonic music audio.

2 Proposed Method

In this section, we describe the proposed audio imputation method. We first give an overview of the modeling strategy. We then briefly describe the probabilistic model that we employ, followed by the actual imputation methodology.

2.1 Overview

The general procedure of supervised audio imputation methods [7] is as follows. We first train a dictionary of spectral vectors from the training data using non-negative spectrogram factorization techniques such as Non-negative Matrix Factorization (NMF) or Probabilistic Latent Component Analysis (PLCA). Each frame of the spectrogram is then modeled as a linear combination of the spectral vectors from the dictionary. Given the spectrogram of a corrupted audio, we estimate the weights for each spectral vector as well as the expected values for the missing entries of the spectrogram using an EM algorithm.

Fig. 1 shows an example of Audio Imputation using PLCA. In this example, a dictionary of 30 spectral vectors is learned from an intact audio spectrogram.

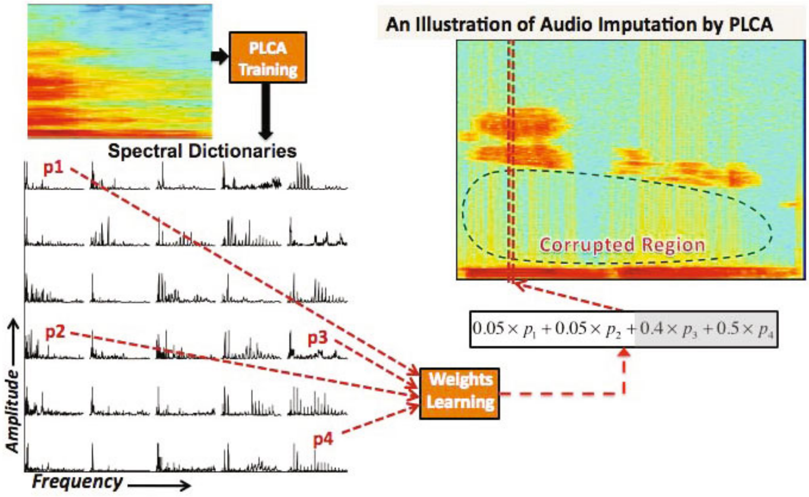

PLCA Training
Spectral Dictionaries
p1
p2
p3
p4
Amplitude
Frequency
Weights Learning
An Illustration of Audio Imputation by PLCA
Corrupted Region
$0.05 \times p_1 + 0.05 \times p_2 + 0.4 \times p_3 + 0.5 \times p_4$

Fig. 1. General Procedure of Supervised Audio Imputation

Given corrupted audio that is similar to the training audio, the original audio spectrogram can be estimated by a linear combination of the spectral vectors from the dictionary.

Previous audio imputation methods [3][7] are based on NMF or PLCA to learn a single dictionary of spectral vectors to represent the entire signal. These approaches treat individual time frame independently, ignoring the temporal dynamics of audio signal. Furthermore, it is not always the case that the training data has exact the same characteristics as the corrupted audio. For example, the corrupted audio may contain a piano solo playing an intro of a song but the training audio from the same song may contain the piano source and the singing voice. In this case, a single dictionary learned from a mixture of the piano and singing voice may be less effective in reconstructing the piano sound from the corrupted audio. This may introduce interference to the reconstructed piano sound from the dictionary elements that are used to explain the singing voice.

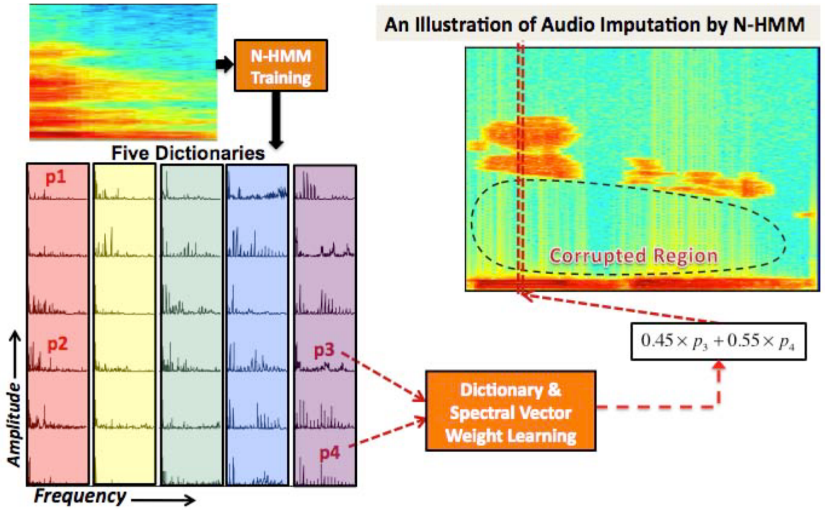

N-HMM Training
Five Dictionaries
p1
p2
p3
p4
Amplitude
Frequency
An Illustration of Audio Imputation by N-HMM
Corrupted Region
$0.45 \times p_3 + 0.55 \times p_4$
Dictionary & Spectral Vector Weight Learning

Fig. 2. Supervised Audio Imputation using N-HMM

As shown in Fig.2, our proposed approach uses a N-HMM to learn several small dictionaries from the training audio. Dictionaries are associated with states in a model that incorporates the dynamic temporal structure of the given audio signal. Several small dictionaries are learned from the training data to explain different aspects of the audio signal. During the imputation process, only spectral vectors from one dictionary are used to reconstruct a certain frame of the corrupted spectrogram. In this way, it is less likely to introduce interference from other sources of the training data.

2.2 Probabilistic Model

Fig.3 shows the graphical model of the N-HMM, an extension of Hidden Markov Model (HMM) by imposing non-negativity on the observation model. The observation alphabet for each state q in the N-HMM is a dictionary of spectral vectors. Each vector z can be thought of a magnitude spectrum. To maintain consistency with prior work [4] we treat it as a probability distribution. For frequency f and time t, we notate the contribution to the magnitude of the spectrogram from a spectral vector z in dictionary q as $P(f_t|z_t, q_t)$. Here, f is one of a set of K frequencies of analysis of a spectrogram. At time t, the observation model is obtained by a linear combination of all spectral vectors z from the current dictionary q:

$$P(f_t|q_t) = \sum_{z_t} P(z_t|q_t)P(f_t|z_t, q_t) \tag{1}$$

where $P(z_t|q_t)$ is the spectral vector mixture weight, given q_t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$.

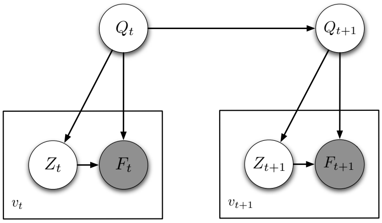


Fig. 3. Graphical Model of the N-HMM. {Q, Z, F} is a set of random variables and {q, z, f} the realization of the random variables. v_t represents the number draws at time t.

In our model, we assume the spectrum V_t at time t is generated by repeated draws from a distribution $P_t(f)$ given by

$$P_t(f) = \sum_{q_t} P(f_t|q_t)\gamma_t(q_t) \tag{2}$$

where $\gamma_t(q_t)$ is the distribution over the states, conditioned on all the observations over all time frames. We can compute $\gamma_t(q_t)$ using the forward-backward algorithm as in traditional HMM. Please refer to [4] for the full formulation. Here, the resulting value $P_t(f)$ can be thought as an estimation of the relative magnitude of the spectrum at frequency f and time t.

A comparison between a N-HMM and a PLCA is illustrated in Fig.4. Compared to most other non-negative spectrogram decomposition techniques, the N-HMM has taken into account the temporal dynamics of the audio signal. Instead of using one large dictionary to explain everything in the audio, the N-HMM learns several small dictionaries, each of which will explain a particular part in the spectrogram. All the parameters of the N-HMM can be learned using the EM algorithm detailed in [4].

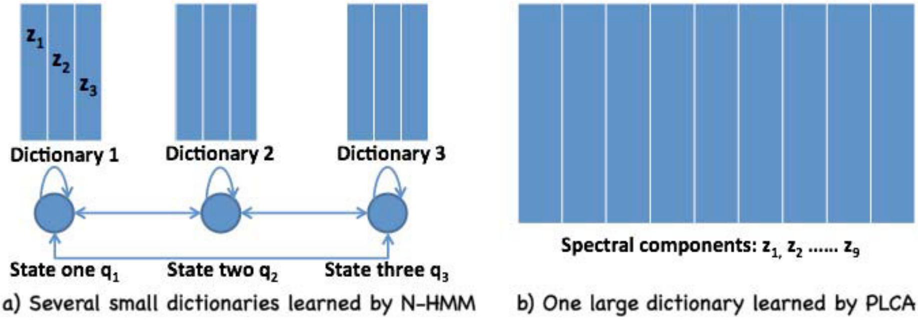

z1
z2
z3
Dictionary 1
Dictionary 2
Dictionary 3
State one q1
State two q2
State three q3
a) Several small dictionaries learned by N-HMM
Spectral components: z1, z2 ...... z9
b) One large dictionary learned by PLCA

Fig. 4. A comparison between PLCA and N-HMM

3 Estimation of Incomplete Data

When the spectrogram is incomplete, a great deal of the entries in the spectrogram could be missing. In this paper, we assume the locations of the corrupted bins are known. Identifying the corrupted region is beyond the scope of this paper. Our objective is to estimate missing values in the magnitude spectrogram of audio signals.

In the rest of the paper we use the following notation: we will denote the observed regions of any spectrogram V as V^o and the missing regions as $V^m = V \setminus V^o$. Within any magnitude spectrum V_t at time t, we will represent the set of observed entries of V_t as V_t^o and the missing entries as V_t^m. $\mathscr{F}_t^o$ will refer to the set of frequencies for which the values of V_t are known, i.e. the set of frequencies in V_t^o. $\mathscr{F}_t^m$ will similarly refer to the set of frequencies for which the values of V_t are missing, i.e. the set of frequencies in V_t^m. $V_t^o(f)$ and $V_t^m(f)$ will refer to specific frequency entries of V_t^o and V_t^m respectively.

To estimate the magnitude of each value in V_t^m we need to scale the value $P_t(f)$ from Eq.2. We do not know the total amplitude at time t because some values are missing. Therefore, we must estimate a scaling factor. We sum the values of the uncorrupted frequencies in the original audio to get $n_t^o = \sum_{f \in \mathscr{F}_t^o} V_t^o(f)$. We sum the values of $P_t(f)$ for $f \in \mathscr{F}_t^o$ to get $p_t^o = \sum_{f \in \mathscr{F}_t^o} P_t(f)$. The expected amplitude at time t is obtained by dividing n_t^o by p_t^o. This gives us a scaling factor. The expected value of any missing term $V_t^m(f)$ can be estimated by:

$$E[V_t^m(f)] = \frac{n_t^o}{p_t^o} P_t(f) \tag{3}$$

The audio imputation process is as follows:

1. Learn the parameters of a N-HMM from the training audio spectrogram, using the EM algorithm.
2. Initialize the missing entries of the corrupted spectrogram to random values.

3. Perform the N-HMM learning on the corrupted spectrogram from step 2. During the learning process,
   - Fix most of the parameters such as $P(f|z,q)$ and $P(q_{t+1}|q_t)$ to the above learned parameters from step 1.
   - Learn the remaining parameters in the N-HMM model using the EM algorithm. Specifically, learn the weights distributions $P(z_t|q_t)$. Then estimate the posterior state distribution $\gamma_t(q_t)$ using the forward-backward algorithm and update $P_t(f)$ using Eq.2.
   - At each iteration, update every missing entry in the spectrogram with its expected value using Eq.3.
4. Reconstruct the corrupted audio spectrogram by:

$$\bar{V}_t(f) = \begin{cases} V_t(f) & \text{if } f \in \mathscr{F}_t^o \\ E[V_t^m(f)] & \text{if } f \in \mathscr{F}_t^m \end{cases} \tag{4}$$

5. Convert the estimated spectrogram to the time domain.

This paper does not address the problem of missing phase recovery. Instead we use the recovered magnitude spectrogram with the original phase to re-synthesize the time domain signal. We found this to be more perceptually pleasing than a standard phase recovery method [5].

4 Experiments

We test the proposed N-HMM audio imputation algorithm on real-world polyphonic musical data. We performed the experiment on 12 real-world pop music songs. The proposed method is compared to a recent audio imputation method using PLCA [7].

For a particular audio clip, both the testing data and training data are taken from the same song. The testing data is about 6-second long, taken from the beginning of a song. The corrupted audio is obtained from the testing data by removing all the frequencies between 800 Hz and 12k Hz in the spectrogram. Another clip (not containing the testing audio) of about 11-second long is taken from the same song as the training data. The details of each audio clip is listed in Table.1. We learn the N-HMM parameters for each song from the training data, and update the N-HMM for the corrupted audio during the imputation process. Specifically, we learned 10 dictionaries of 8 spectral vectors each as well as the transition matrix from the training data. When using PLCA, we learn 1 dictionary of 40 spectral vectors. The values for the parameters are determined by the authors empirically. Signal-to-Noise-Ratio (SNR)[1] is used to measure the outputs of both imputation methods. During the experiments, we find out the existing signal measurements

[1] $SNR = 10log_{10}\frac{\sum_t s(t)^2}{\sum_t (\bar{s}(t)-s(t))^2}$ where $s(t)$ and $\bar{s(t)}$ are the original and the reconstructed signals respectively.

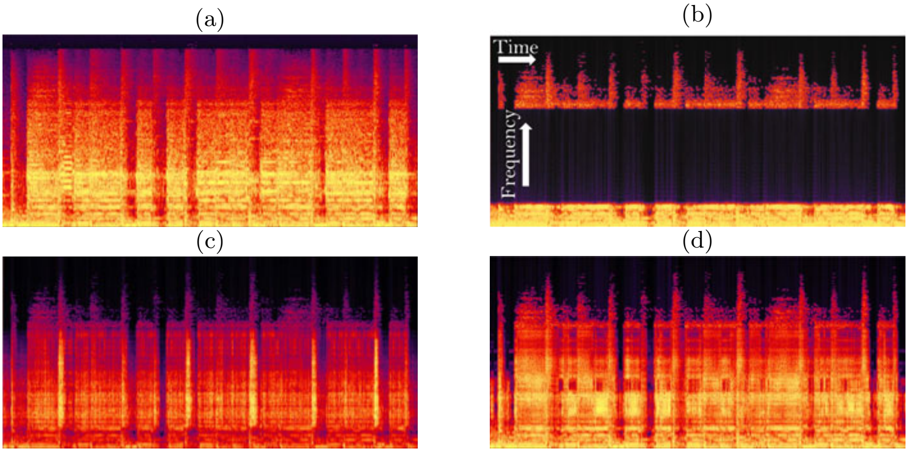


Fig. 5. 5.5-second audio clip from “Born to be wild” by “Steppenwolf”. a)Original audio; b) Corrupted audio input (1.05 dB SNR); c) Imputation result by PLCA (1.41 dB SNR); d) Imputation result by proposed algorithm (4.89 dB SNR).

do not always correspond well to the perceptual quality of the audio. More examples of the experimental results are available at the authors’ website [8] to show the perceptual quality of the reconstructed signals.

We first examine two examples that favor the proposed approach against the PLCA method. The first one is a 5.5-second audio clip from “Born to be wild” by “Steppenwolf”. The spectrogram of the original audio, corrupted audio, output of the proposed method and PLCA are illustrated in Fig.5. The proposed method produces an output with a higher SNR than PLCA.

The next example is a 5.4-second audio clip from “Scar Tissue” by “Red Hot Chili Peppers”. In this example, both PLCA and the proposed method improve the SNR of the corrupted audio by about 7 dB. The proposed method has a lower SNR measurement, however, when listening to the reconstructed audio, the output of the proposed method has better perceptual quality compared to the output of the PLCA method. This difference is also shown in the spectrogram plot in Fig.6. The spectrogram reconstructed by PLCA has more random energy scatted in the high frequency region, while the proposed method only reconstructs the signal in the region where it should have been.

Table 1 presents the performance of PLCA and the proposed algorithm on 12 clips of real-world music recordings using the SNR measurement. The average performance of the proposed method is 15.32 dB SNR, improving 5.67 dB from the corrupted audio and 1.8 dB from the output of the PLCA. The proposed method has better SNR measurement than PLCA on 9 out of 12 song clips. For the audio where the proposed method does not have better SNR measurement, as shown by the example in Fig.6, the proposed method may still produce an audio signal with equivalent or better perceptual quality. We encourage the readers to compare the results of both methods by listening more examples listed at the authors’ website [8].

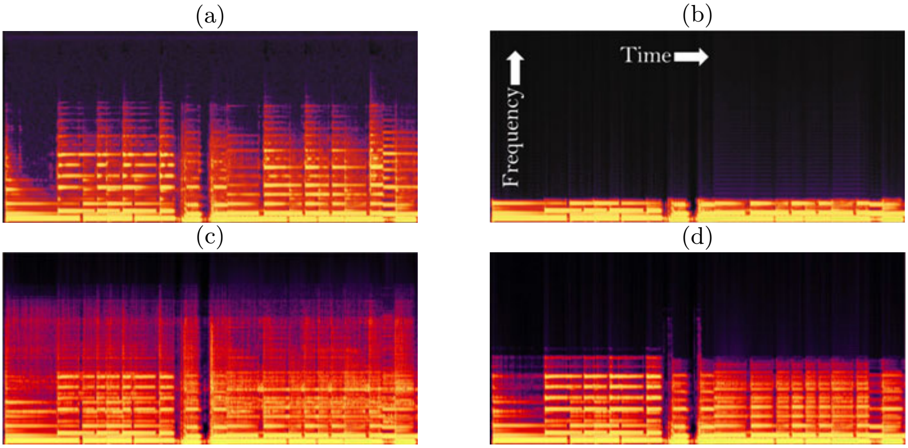


Fig. 6. 5.4-second audio clip from "Scar Tissue" by "Red Hot Chili Peppers". a) Original Audio; b) Corrupted Audio Input (7.46 dB SNR); c) Imputation result by PLCA (15.56 dB SNR); d) Imputation result by proposed algorithm (14.34 dB SNR).

Table 1. Performances of the Imputation results by the proposed method and PLCA

	SNR (dB)			Audio length (Second)	
Song name	Input	Proposed	PLCA	Testing	Training
Better together	11.23	**22.48**	19.5	4.5	10.3
1979	14.43	**19.72**	18.07	5.7	11.3
Born to be wild	1.05	**4.89**	1.41	5.5	20.4
Scar tissue	7.46	14.34	**15.56**	5.4	10
Bad day	6.48	**13.84**	12.55	6.3	11.5
Wonderwall	-2.21	**8.36**	5.28	5.8	5.5
Here I go again	11.49	**15.95**	14.5	5.1	9.5
Every breath you take	7.46	14.34	**15.65**	6.9	10
Viva La Vida	7.6	11.66	**11.77**	6.2	10.1
She will be loved	17.66	**18.46**	15.2	5.7	11.9
Making memories of us	18.06	**21.3**	18.11	9.8	12.8
Daughters	15.11	**18.47**	14.56	8.2	16.2
Average measurement	9.65	**15.32**	13.52	6.29	11.63

5 Conclusions

In this paper we present an approach that allows us to estimate the missing values in the time-frequency domain of audio signals. The proposed approach is based on the N-HMM, which enables us to learn the spectral information as well as the temporal dynamics of the audio signal. Initial experimental results showed that this approach is quite effective in reconstructing missing values from corrupted spectrograms and has advantages over performing imputation using PLCA. Future work includes developping techiniques for missing phase recovery.

References

1. Brand, M.: Incremental Singular Value Decomposition of Uncertain Data with Missing Values. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 707–720. Springer, Heidelberg (2002)
2. Han, J., Pardo, B.: Reconstructing completely overlapped notes from musical mixtures. In: ICASSP (2011)
3. Le Roux, J., Kameoka, H., Ono, N., de Cheveigné, A., Sagayama, S.: Computational auditory induction as a missing-data model-fitting problem with bregman divergence. Speech Communication (2010)
4. Mysore, G.J.: A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures. Ph.d. dissertation, Stanford University (2010)
5. Nawab, S., Quatieri, T., Lim, J.: Signal reconstruction from short-time fourier transform magnitude. IEEE Trans. on Acoustics, Speech & Signal Processing 31, 986–998 (1983)
6. Raj, B.: Reconstruction of Incomplete Spectrograms for Robust Speech Recognition. Ph.d. dissertation, Carnegie Mellon University (2000)
7. Smaragdis, P., Raj, B., Shashanka, M.: Missing data imputation for time-frequency representations of audio signals. J. Signal Processing Systems (2010)
8. www.cs.northwestern.edu/~jha222/imputation

A Non-negative Approach to Language Informed Speech Separation

Gautham J. Mysore[1] and Paris Smaragdis[1,2]

[1] Advanced Technology Labs, Adobe Systems Inc.,
[2] University of Illinois at Urbana-Champaign

Abstract. The use of high level information in source separation algorithms can greatly constrain the problem and lead to improved results by limiting the solution space to semantically plausible results. The automatic speech recognition community has shown that the use of high level information in the form of language models is crucial to obtaining high quality recognition results. In this paper, we apply language models in the context of speech separation. Specifically, we use language models to constrain the recently proposed non-negative factorial hidden Markov model. We compare the proposed method to non-negative spectrogram factorization using standard source separation metrics and show improved results in all metrics.

1 Introduction

The cocktail party problem is a classical source separation problem in which the goal is to separate speech of multiple concurrent speakers. This is a challenging problem, particularly in the single channel case. It would therefore be beneficial to use any high level information that is available to us. Specifically, if it is known that the speakers follow a certain grammar (constrained sequences of words), this information could be useful. We refer to this as a language model. This is routinely used in automatic speech recognition [1] and is in fact crucial to obtaining recognition results with high accuracy.

Non-negative spectrogram factorization algorithms [2] are a major research area in the source separation community and have been quite successful. They provide rich models of the spectral structure of sound sources by representing each time frame of the spectrogram of a given source as a linear combination of non-negative spectral components (analogous to basis vectors) from a dictionary. However, they model each time frame of audio as independent and consequently ignore an important aspect of audio – temporal dynamics. In order to address this issue, we proposed the non-negative hidden Markov model (N-HMM) [3] in which we model a given source using multiple dictionaries of spectral components such that each time frame of audio is explained by a linear combination of spectral components from one of the dictionaries. This gives us the rich spectral modeling capability of non-negative spectrogram factorizations. Additionally, we learn a Markov chain that explains the temporal dynamics between the dictionaries. The dictionaries therefore correspond to states of the Markov chain. We model

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 356–363, 2012.

mixtures by combining N-HMMs of individual sources into the non-negative factorial hidden Markov model (N-FHMM).

There has been some other work [4,5,6] that extends non-negative spectrogram factorizations to model temporal dynamics. Ozerov [4] and Nakano [5] modeled the temporal dynamics between individual spectral components rather than dictionaries. They therefore model each time frame of a given source with a single spectral component rather than a linear combination of spectral components and can thus be too restrictive. Smaragdis [6] introduced a model that does allow linear combinations of spectral components with transitions between dictionaries. However, it also allows all spectral components of all dictionaries to be active at the same time, which is often not restrictive enough.

Since we use a hidden Markov model structure, we can readily use the ideas of language modeling from automatic speech recognition in the context of source separation. That is the context of this paper. Specifically, we constrain the Markov chain of each individual source to explain a valid grammar.

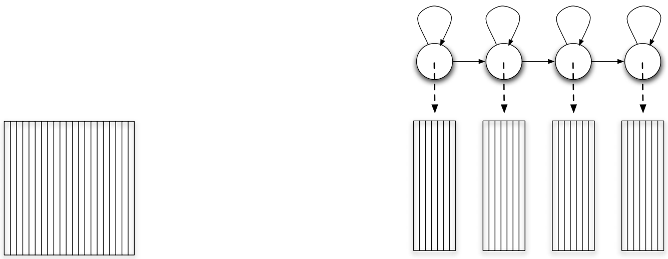
There has been some previous work [6,7,8] on modeling concurrent speakers using hidden Markov models and factorial hidden Markov models with language models. However, the goal has been concurrent speech recognition of multiple speakers. These papers report speech recognition performance and are presumably optimized for this. On the other hand, our goal is high quality source separation and we make design decisions for this goal. Also, to the best of our knowledge, no previous work on using language models for multiple concurrent speakers has reported source separation metrics.

2 Models of Individual Speakers

In this section, we explain how we learn models of individual speakers. We first describe the Non-negative hidden Markov model (N-HMM). We then explain how to learn N-HMMs for individual words of a given speaker. Finally, we explain how to combine these individual word models into a single N-HMM according to the rules of the grammar, as dictated by the language model.

2.1 Non-negative Hidden Markov Model

Non-negative spectrogram factorizations (Fig. 1a) include non-negative matrix factorization (NMF) and their probabilistic counterparts such as probabilistic latent component analysis (PLCA). These models use a single dictionary of non-negative spectral components to model a given sound source. Specifically, they explain each time frame of the spectrogram of a given source with a linear combination of spectral components from the dictionary. These models however ignore two important aspects of audio – non-stationarity and temporal dynamics. To overcome this issue, we proposed the N-HMM (Fig.1b) [3]. This model uses multiple dictionaries such that each time frame is explained by any one of the several dictionaries (accounting for non-stationarity). Additionally it uses a Markov chain to explain the transitions between dictionaries (accounting for temporal dynamics).



(a) Non-negative Spectrogram Factorization

(b) Non-negative Hidden Markov Model with left to right transitions

Fig. 1. Comparison of non-negative models. Non-negative spectrogram factorization uses a single large dictionary to explain a sound source, whereas the N-HMM uses multiple small dictionaries and a Markov chain.

The graphical model of the N-HMM is shown in Fig.2. Each dictionary corresponds to a state q. At time t, the N-HMM is in state q_t. Each spectral component of a given dictionary q is represented by z. A given spectral component is a discrete distribution. Therefore, spectral component z of dictionary q is represented by $P(f|z,q)$. The *non-negativity* in the N-HMM comes from the fact that the parameters of a discrete distribution are non-negative by definition. Since each column of the spectrogram is modeled as a linear combination of spectral components, time frame t (modeled by state q) is given by the following observation model:

$$P(f_t|q_t) = \sum_{z_t} P(f_t|z_t, q_t)P(z_t|q_t), \tag{1}$$

where $P(z_t|q_t)$ is a discrete distribution of mixture weights for time t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$.

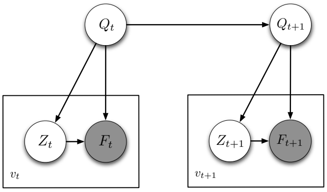


Fig. 2. Graphical model of the N-HMM

2.2 Word Models

Given an instance of a word, we can estimate the parameters of all of the distributions of the N-HMM using the expectation–maximization (EM) algorithm [3]. In this paper, we extend this idea to learn word models from multiple instances of a given word as routinely done in speech recognition [1]. We compute the E step of EM algorithm separately for each instance. The procedure is the same as in [3]. This gives us the marginalized posterior distributions $P_t^{(k)}(z, q|f, \overline{\mathbf{f}})$ and $P_t^{(k)}(q_t, q_{t+1}|\overline{\mathbf{f}})$ for each instance k. We use these in the M step of the EM algorithm. Specifically, we compute a separate weights distribution for each instance k as follows:

$$P_t^{(k)}(z_t|q_t) = \frac{\sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \overline{\mathbf{f}})}{\sum_{z_t} \sum_{f_t} V_{ft}^{(k)} P_t^{(k)}(z_t, q_t|f_t, \overline{\mathbf{f}})}, \tag{2}$$

where $V_{ft}^{(k)}$ is the spectrogram of instance k. However, we estimate a single set of dictionaries of spectral components and a single transition matrix using the marginalized posterior distributions of all instances as follows:

$$P(f|z, q) = \frac{\sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z, q|f, \overline{\mathbf{f}})}{\sum_f \sum_k \sum_t V_{ft}^{(k)} P_t^{(k)}(z, q|f, \overline{\mathbf{f}})}, \tag{3}$$

$$P(q_{t+1}|q_t) = \frac{\sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\overline{\mathbf{f}})}{\sum_{q_{t+1}} \sum_k \sum_{t=1}^{T-1} P_t^{(k)}(q_t, q_{t+1}|\overline{\mathbf{f}})}. \tag{4}$$

We restrict the transition matrix to use only left to right transitions.

2.3 Combining Word Models

Once we learn N-HMMs for each word of a given speaker, we combine them into a single speaker dependent N-HMM. We do this by constructing a large transition matrix that consists of each individual transition matrix. The transition matrix of each individual word stays the same. However, the transitions between words are dictated by a language model. Each state of the speaker dependent N-HMM corresponds to a specific dictionary of that speaker. Therefore, this N-HMM also contains all dictionaries of all words.

3 Model of Mixtures

We first describe how to combine models of individual speakers into a model of speech mixtures. We then explain how to use this model for speech separation. Finally, we describe the pruning that we use to reduce computational complexity.

3.1 Combining Speaker Dependent Models

We model a mixture of two speakers using the non-negative factorial hidden Markov model (N-FHMM) [3]. Given the N-HMM of two speakers, we can combine them into an N-FHMM. We use the dictionaries and the Markov chains of the N-HMMs of the two speakers. A given time frame is then explained using any

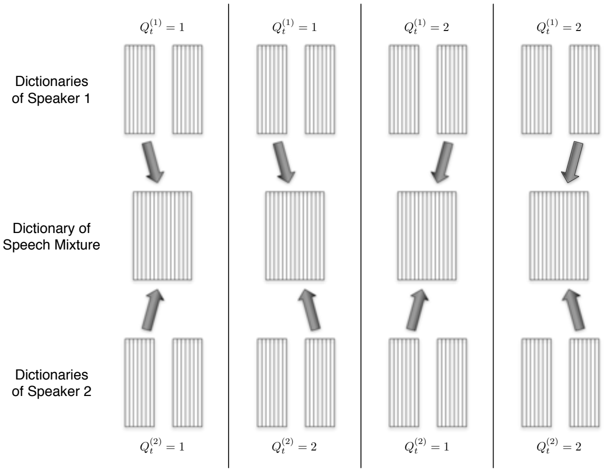


Fig. 3. Combining dictionaries of two sources to model a mixture. In this simple example, each source has two dictionaries so there are a total of four ways of combining them.

one dictionary of the first speaker and any one dictionary of the second speaker. Specifically, the given time frame is modeled using a linear combination of the spectral components of the two appropriate dictionaries. This is illustrated in Fig. 3.

The graphical model of the N-FHMM is shown in Fig. 3. An N-HMM can be seen in the upper half of the graphical model and another one can be seen in the lower half. The interaction model (of the two sources) introduces a new variable s_t that indicates the ratio of the sources at a given time frame. $P(s_t|q_t^{(1)}, q^{(2)})$ is a Bernoulli distribution that depends on the states of the sources at the given time frame. The interaction model is given by:

$$P(f_t|q_t^{(1)}, q_t^{(2)}) = \sum_{s_t} \sum_{z_t} P(f_t|z_t, s_t, q_t^{(s_t)}) P(z_t, s_t|q_t^{(1)}, q_t^{(2)}), \tag{5}$$

where $P(f_t|z_t, s_t, q_t^{(s_t)})$ is spectral component z_t of state $q_t^{(s_t)}$ of source s_t.

3.2 Speech Separation

$P(z_t, s_t|q_t^{(1)}, q_t^{(2)})$ combines the new distribution $P(s_t|q_t^{(1)}, q^{(2)})$ and the weights distributions of each source into a single weights distribution of the mixture. Since the dictionaries and the Markov chain of each source are already specified, if we learn the weights distribution of the mixture, we can estimate soft masks to separate the two sources. This is done using the EM algorithm. Details on how to estimate the masks and then separate the sources can be found in [3].

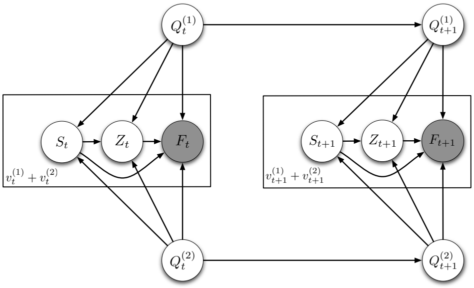


Fig. 4. Graphical model of the N-FHMM

3.3 Pruning

At every time frame, we need to compute the likelihood of every possible state pair (one state from each source). This causes the computational complexity of the N-FHMM to be exponential in the number sources. This can lead to intractable computation. However, we do not need to consider state pairs that have a very small probability. Specifically, we prune out all of the state pairs whose posterior probability $\gamma(q_t^{(1)}, q_t^{(2)})$, is below a pre-determined threshold. In our experiments, we set this threshold to -1000 in the log domain. Even though it is an extremely small number, this pruned out around 99% of the state pairs. This is due to the heavy constraining of the language model.

For each speaker, we used an N-HMM of 127 states (more details are in Sec. 4). Therefore, there are a total of 16129 possible state pairs. With our pruning, we need to consider less than 250 state pairs in most time frames. With the number of states that we used, this corresponds to computation complexity that is linear in the number of sources.

4 Experimental Results and Discussion

We performed experiments on a subset of the data from the speech separation challenge [9]. The test data from the challenge does not contain ground truth, without which we cannot compute source separation metrics. Therefore, we divided the training data into a training set and a test set. We trained N-HMMs for 10 speakers using 450 of the 500 sentences from the training set of each speaker. The remaining 50 sentences were used to construct the test set. We segmented the training sentences into words in order to learn individual word models as described in Section 2.2. We used one dictionary (state) per phoneme. This is less than what is typically used in speech recognition. However, we did not want to excessively constrain the model in order to obtain high quality reconstructions. We used 10 spectral components per dictionary as this number was previously found to give good results in N-HMMs [3].

We then combined the word models of a given speaker into a single N-HMM according to the language model, as described in Section 2.3.

We performed speech separation using the N-FHMM on speakers of different genders and the same gender[1]. For both categories, we constructed 10 test mixtures from our test set. The mixing was done at 0dB. We evaluated the source separation performance using the BSS-EVAL metrics [10]. As a comparison, we performed separation using a non-negative spectrogram factorization technique (PLCA) [2]. When using PLCA, we used the same training and test sets that we used with the proposed model. However, we simply concatenated all of the training data of a given speaker and learned a single dictionary for that speaker, which is customary when using non-negative spectrogram factorizations [2]. We used a dictionary size of 100 spectral components as this gave the best separation results. This is more than used in our previous paper [3] since the database used in this paper has much more training data for each speaker. The proposed method has the advantage (over PLCA) of using language information. However, the point that we are trying to make is that language information can lead to improved speech separation results.

Our results are shown in Table 1. The proposed model outperforms PLCA in all metrics of both categories. Specifically, we see a 7-8dB improvement in source to interference ratio (SIR) while still maintaining a higher source to artifacts ratio (SAR). This means that we are achieving much higher amounts of separation than PLCA and also introducing less artifacts. The source to distortion ratio (SDR), which reflects both of these things is therefore also higher.

Another observation is that when we compare the performance of the N-FHMM in the two categories, we see only a small deterioration in performance from the different gender to the same gender case (0.5-1 dB in each metric). With PLCA, however, we see a greater deterioration in SIR and SDR (2-3 dB). This is because the dictionaries of the two sources are much more similar in the same gender case than in the different gender case. With the N-FHMM, the language model helps disambiguate the sources. However, only the spectral information is used in the case of PLCA.

Table 1. Source separation performance of the N-FHMM and PLCA

Different Gender	SIR	SAR	SDR
N-FHMM	14.91	10.29	8.78
PLCA	7.96	9.08	4.86

Same Gender	SIR	SAR	SDR
N-FHMM	13.88	9.89	8.24
PLCA	5.11	8.77	2.85

The introduction of constraints, priors, and additional structure in non-negative models often leads to improved separation quality (higher SIR), when compared to PLCA or NMF. However, this usually leads to more artifacts (lower SAR). Ozerov [4] noted this with the FS-HMM. We have improved results in both metrics. The reason is that the language model only attempts to determine the correct dictionary to explain each source but not the exact fitting of the spectral components of the given dictionary to the data. Once this dictionary of

[1] Examples at https://ccrma.stanford.edu/~gautham/Site/lva_ica_2012.html

each source is determined for a given time frame, the algorithm fits the corresponding spectral components to the mixture data to obtain the closest possible reconstruction of the mixture. This flexibility after determining the appropriate dictionary avoids excessive artifacts.

5 Conclusions

We presented a method to perform high quality speech separation using language models in the N-HMM framework. We showed that use of the language model greatly boosts source separation performance when compared to non-negative spectrogram factorization. The methodology was shown for speech but it can be used in other contexts in which high level structure information is available such as incorporating music theory into the N-HMM framework for music separation.

References

1. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
2. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 414–421. Springer, Heidelberg (2007)
3. Mysore, G.J., Smaragdis, P., Raj, B.: Non-Negative Hidden Markov Modeling of Audio with Application to Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 140–148. Springer, Heidelberg (2010)
4. Ozerov, A., Fevotte, C., Charbit, M.: Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (October 2009)
5. Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., Sagayama, S.: Non-negative Matrix Factorization with Markov-Chained Bases for Modeling Time-Varying Patterns in Music Spectrograms. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 149–156. Springer, Heidelberg (2010)
6. Smaragdis, P., Raj, B.: The Markov selection model for concurrent speech recognition. In: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (August 2010)
7. Hershey, J.R., Rennie, S.J., Olsen, P.A., Kristjansson, T.T.: Super-human multi-talker speech recognition: A graphical modeling approach. Computer Speech and Language 24(1), 45–46 (2010)
8. Virtanen, T.: Speech recognition using factorial hidden Markov models for separation in the feature space. In: Proceedings of Interspeech, Pittsburgh, PA (September 2006)
9. Cooke, M., Hershey, J.R., Rennie, S.J.: Monaural speech separation and recognition challenge. Computer Speech and Language 24(1), 1–15 (2010)
10. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (2006)

Temporally-Constrained Convolutive Probabilistic Latent Component Analysis for Multi-pitch Detection

Emmanouil Benetos* and Simon Dixon

Centre for Digital Music, Queen Mary University of London
Mile End Road, London E1 4NS, UK
{emmanouilb,simond}@eecs.qmul.ac.uk

Abstract. In this paper, a method for multi-pitch detection which exploits the temporal evolution of musical sounds is presented. The proposed method extends the shift-invariant probabilistic latent component analysis algorithm by introducing temporal constraints using multiple Hidden Markov Models, while supporting multiple-instrument spectral templates. Thus, this model can support the representation of sound states such as attack, sustain, and decay, while the shift-invariance across log-frequency can be utilized for multi-pitch detection in music signals that contain frequency modulations or tuning changes. For note tracking, pitch-specific Hidden Markov Models are also employed in a post-processing step. The proposed system was tested on recordings from the RWC database, the MIREX multi-F0 dataset, and on recordings from a Disklavier piano. Experimental results using a variety of error metrics, show that the proposed system outperforms a non-temporally constrained model. The proposed system also outperforms state-of-the art transcription algorithms for the RWC and Disklavier datasets.

Keywords: Music signal analysis, probabilistic latent component analysis, hidden Markov models.

1 Introduction

Multi-pitch detection is one of the core problems of music signal analysis, having numerous applications in music information retrieval, computational musicology, and interactive music systems [4]. The creation of a robust multi-pitch detection system for multiple instrument sources is considered to be an open problem in the literature. The performance of multi-pitch estimation systems has not yet matched that of a human expert, which can be partly attributed to the non-stationary nature of musical sounds. A produced musical note can be expressed by a sound state sequence (e.g. attack, transient, decay, and sustain states) [1], and can also exhibit frequency modulations such as vibrato.

* E. Benetos is funded by a Westfield Trust research studentship (Queen Mary University of London).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 364–371, 2012.

A method for modeling sound states in music signals was proposed by Nakano et al. in [8], combining the non-negative matrix factorization (NMF) algorithm with Markov-chained constraints. Smaragdis in [11] employed the shift-invariant probabilistic latent component analysis (PLCA) algorithm for pitch tracking, which can model frequency modulations. Mysore proposed a method for sound modeling which combined the PLCA method with temporal constraints using hidden Markov models (HMMs) [7]. In [3], the authors extended the shift-invariant PLCA model for multi-pitch detection, supporting multiple instrument and pitch templates, with time-dependent source contributions. Finally, the authors combined shift-invariant PLCA with HMMs using sound state templates for modeling the temporal evolution of monophonic recordings [2].

Here, we extend the single-instrument single-pitch model of [2] for multi-pitch detection of multiple-instrument recordings. This is accomplished by extracting sound state templates for the complete pitch range of multiple instruments, and utilizing multiple independent HMMs, one for each pitch, for modeling the temporal evolution of produced notes. Experiments performed on excerpts from the RWC database [6], Disklavier recordings [9], and the MIREX multi-F0 dataset showed that the proposed model outperforms the non-temporally constrained model of [3] and also provides accuracy rates that outperform state-of-the-art methods for automatic transcription.

2 Proposed Method

The motivation behind this model is to propose a multi-pitch detection algorithm which supports multiple instrument sources, can express the temporal evolution of a produced note (by modeling sound states), and can support frequency modulations (e.g. vibrati). Frequency modulations can be supported using a shift-invariant model and a log-frequency representation, while modeling the temporal evolution of a sound can be done by utilizing templates for different sound states and constraining the order of appearance of these states using HMMs. This would allow for a rich and informative representation of the music signal, addressing some drawbacks of current polyphonic transcription systems.

2.1 Model

The proposed model extends the single-pitch single-source algorithm proposed in [2], which incorporated temporal constraints into the single-component shift-invariant PLCA algorithm. Here, this method supports multiple concurrent pitches produced by multiple instrument sources, using as an input the log-frequency spectrogram $V_{\omega,t}$, where ω is the log-frequency index and t is the time index. The model approximates the input spectrogram as a probability distribution $P(\omega, t)$:

$$P(\omega,t) = P(t)\sum_{s,p} P_t(p)P_t(s|p)\sum_{q_t^{(p)}} P_t(q_t^{(p)}|p,\bar{\omega})P(\omega|s,p,q_t^{(p)}) *_\omega P_t(f|p) \quad (1)$$

where $p = 1, \ldots, 88$ is the pitch index, s denotes the instrument source, $q^{(p)}$ the sound state for each pitch, and f the pitch shifting. Thus, $P_t(p)$ expresses the piano-roll transcription, $P_t(q_t^{(p)}|p, \bar{\omega})$ is the sound state activation for the p-th pitch, $P_t(s|p)$ the s-th instrument source contribution, $P_t(f|p)$ the pitch impulse distribution, and $P(\omega|s, p, q_t^{(p)})$ the spectral template for the s-th source, p-th pitch, and $q^{(p)}$-th sound state. The convolution of $P(\omega|s, p, q_t^{(p)}) *_\omega P_t(f|p)$ takes place between ω and f using an area spanning one semitone around the ideal position of p, in order to constrain each template for the detection of the pitch it corresponds to. In addition, such formulation allows a greater control over the polyphony level of the signal, as explained in Section 2.2. It should also be noted that $P_t(f|p)$ is not dependent on the instrument source s for computational speed purposes. This design choice might have an effect in the rare case of two instruments producing the same note concurrently. Since 60 bins per octave are used in the input log-frequency spectrogram, f has a length of 5.

Since the sequence of each pitch-specific sound state is temporally constrained, the corresponding HMM for the p-th pitch is:

$$P(\bar{\omega}) = \sum_{\bar{q}^{(p)}} \sum_{\bar{s}} \sum_{\bar{p}} \sum_{\bar{f}} P(q_1^{(p)}) \prod_t P(q_{t+1}^{(p)}|q_t^{(p)}) \prod_t P_t(\omega_t|q_t^{(p)}) \tag{2}$$

where $\bar{\omega}$ refers to all observations, $P(q_1^{(p)})$ is the state prior distribution, $P(q_{t+1}^{(p)}|q_t^{(p)})$ is the transition probability, and $P_t(\omega_t|q_t^{(p)})$ is the observation probability for the pitch sound state. The observation probability is defined as:

$$P_t(\omega_t|q_t^{(p)}) = 1 - \frac{||P(\omega, t|q_t^{(p)}) - V_{\omega,t}||_2}{\sum_{q_t^{(p)}} ||P(\omega, t|q_t^{(p)}) - V_{\omega,t}||_2} \tag{3}$$

where $|| \cdot ||_2$ is the l^2 norm and

$$P(\omega, t|q_t^{(p)}) = P(t) \sum_s P_t(p) P_t(s|p) P_t(q_t^{(p)}|p, \bar{\omega}) \sum_f P(\omega - f|s, p, q_t^{(p)}) P_t(f|p) \tag{4}$$

is the spectrogram reconstruction for the p-th pitch and $q^{(p)}$-th sound state. Thus, for a specific pitch, a greater observation probability is given to the state spectrogram that better approximates the input spectrogram using the Euclidean distance. Again, for computational speed purposes, the HMMs are not dependent on s, which was done in order to avoid using $S \times 88$ HMMs.

2.2 Parameter Estimation

As in the single-pitch model from [2], the aforementioned parameters can be estimated using the Expectation-Maximization algorithm. For the *Expectation* step, the update equations are:

$$P_t(f_t, s, p, q_t^{(1)}, \ldots, q_t^{(88)}|\bar{\omega}) = P_t(q_t^{(1)}, \ldots, q_t^{(88)}|\bar{\omega}) P_t(f_t, s, p|q_t^{(1)}, \ldots, q_t^{(88)}, \omega_t) \tag{5}$$

$$P_t(q_t^{(1)},\dots,q_t^{(88)}|\bar{\omega}) = \prod_{p=1}^{88} P_t(q_t^{(p)}|\bar{\omega}) \tag{6}$$

$$P_t(q_t^{(p)}|\bar{\omega}) = \frac{\alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})}{\sum_{q_t^{(p)}} \alpha_t(q_t^{(p)})\beta_t(q_t^{(p)})} \tag{7}$$

$$P_t(f_t,s,p|\omega_t,q_t^{(1)},\dots,q_t^{(88)}) = \frac{P_t(p)P(\omega_t - f_t|s,p,q_t^{(p)})P_t(f_t|p)P_t(s|p)}{\sum_p P_t(p)\sum_{s,f_t} P(\omega_t - f_t|s,p,q_t^{(p)})P_t(f_t|p)P_t(s|p)} \tag{8}$$

Equation (5) is the model posterior, for the source components, sound state activity, pitch impulse, and pitch activity. In (7), $\alpha_t(q_t)$ and $\beta_t(q_t)$ are the HMM forward and backward variables, respectively, which can be computed using the forward/backward procedure described in [10] and the observation probability from (3). Also, the posterior for the pitch-wise transition matrices is:

$$P(q_{t+1}^{(p)},q_t^{(p)}|\bar{\omega}) = \frac{\alpha_t(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_{t+1}(q_{t+1}^{(p)})P_t(\omega_{t+1}|q_{t+1}^{(p)})}{\sum_{q_t^{(p)}}\sum_{q_{t+1}^{(p)}} \alpha_t(q_t^{(p)})P(q_{t+1}^{(p)}|q_t^{(p)})\beta_{t+1}(q_{t+1}^{(p)})P_t(\omega_{t+1}|q_{t+1}^{(p)})} \tag{9}$$

For the *Maximization* step, the update equations for the unknown parameters are:

$$P(\omega|s,p,q^{(p)}) = \frac{\sum_{f,s,t}\overline{\sum}_{q_t^{(p)}} V_{\omega+f,t}P_t(f,s,p,q^{(1)},\dots,q^{(88)}|\omega+f)}{\sum_{\omega,f,s,t}\overline{\sum}_{q_t^{(p)}} V_{\omega+f,t}P_t(f,s,p,q^{(1)},\dots,q^{(88)}|\omega+f)} \tag{10}$$

where $\overline{\sum}_{q_t^{(p)}} = \sum_{q_t^{(1)}}\cdots\sum_{q_t^{(p-1)}}\sum_{q_t^{(p+1)}}\cdots\sum_{q_t^{(88)}}$,

$$P_t(f_t|p) = \frac{\sum_{\omega,s}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)}{\sum_{f_t,\omega,s}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)} \tag{11}$$

$$P(q_{t+1}^{(p)}|q_t^{(p)}) = \frac{\sum_t P(q_t^{(p)},q_{t+1}^{(p)}|\bar{\omega})}{\sum_{q_{t+1}^{(p)}}\sum_t P(q_t^{(p)},q_{t+1}^{(p)}|\bar{\omega})} \tag{12}$$

$$P_t(s|p) = \frac{\sum_{\omega,f_t}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)}{\sum_{s,\omega,f_t}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)} \tag{13}$$

$$P_t(p) = \frac{\sum_{\omega,f_t,s}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)}{\sum_{p,\omega,f_t,s}\sum_{q_t^{(1)}}\cdots\sum_{q_t^{(88)}} V_{\omega,t}P_t(f_t,s,p,q_t^{(1)},\dots,q_t^{(88)}|\omega_t)} \tag{14}$$

Finally, the pitch-wise initial state probabilities are: $P(q_1^{(p)}) = P_1(q_1^{(p)}|\bar{\omega})$. It should be noted that the spectral template update rule in (10) is not used in this system since we are utilizing pre-extracted templates, but is included for completeness.

Sparsity constraints were also incorporated, in order for the algorithm to provide as meaningful solutions as possible. Using the technique shown in [3], sparsity was enforced on the update rules for the pitch activity matrix $P_t(p)$ and the source contribution matrix $P_t(s|p)$. This means that we would like few notes active in a time frame, and that each note is produced by few instrument sources. The same sparsity parameters that were used in [3] were used. A pitch spectrogram can also be created using $P(f,p,t) = P(t)P_t(p)P_t(f|p)$ and stacking together slices of tensor $P(f,p,t)$ for all pitch values: $P(f,t) = [P(f,1,t) \cdots P(f,88,t)]$.

2.3 Postprocessing

For performing note smoothing and tracking, the resulting pitch activity matrix $P(p,t) = P(t)P_t(p)$ is postprocessed using pitch-wise HMMs, as in [9,3]. Each pitch p is modeled by a 2-state on/off HMM, while the hidden state sequence is $q'_p[t]$ and the observed sequence $o_p[t]$. MIDI files from the RWC database [6] were employed in order to estimate the pitch-wise state priors and the state transition matrices. For estimating the observation probability for each active pitch $P(o_p[t]|q'_p[t] = 1)$, we use a sigmoid curve which has $P(p,t)$ as input:

$$P(o_p[t]|q'_p[t] = 1) = \frac{1}{1 + e^{-P(p,t)}} \tag{15}$$

and use the Viterbi algorithm [10] for extracting the note tracking output for each pitch. The result of the HMM postprocessing step is a binary piano-roll transcription which can be used for evaluation.

3 Evaluation

3.1 Datasets

For training, the spectral templates $P(\omega|s,p,q^{(p)})$ were extracted for various instruments, over their complete pitch range, using $q = 3$ sound states. The extraction process was performed using the unsupervised single-source single-pitch model of [2] and the constant-Q transform with 60 bins/octave as input. Isolated note samples from 3 piano models were used from the MAPS database [5] and templates from cello, clarinet, flute, guitar, harpsichord, oboe, and violin were extracted from the RWC musical instrument sounds dataset [6]. An example of the sound state template extraction process is given in Fig. 1.

For evaluation, we employed 12 excerpts from the RWC classical and jazz datasets which are widely used for transcription (see [3] for comparative results). We also used the woodwind quintet recording from the MIREX multi-F0 development set[1]. Finally, 10 one-minute recordings taken from a Yamaha Disklavier piano which were presented in [9] were also utilized.

[1] http://www.music-ir.org/mirex

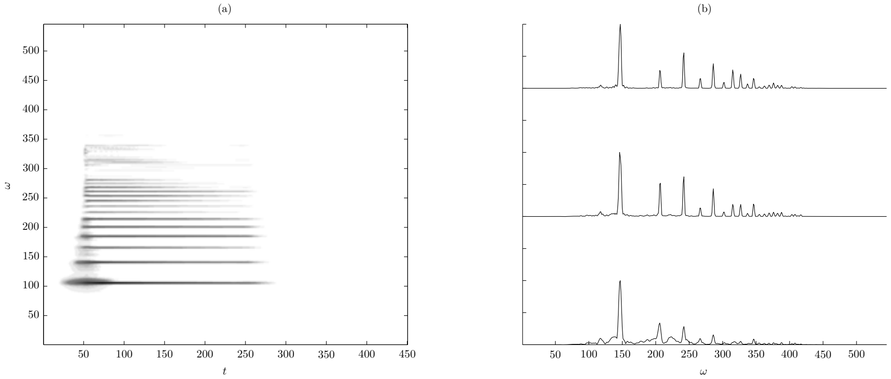


Fig. 1. (a) Spectrogram $V_{\omega,t}$ of a D3 piano note (b) Extracted spectral templates using the method in [2] corresponding to different sound states

3.2 Results

For evaluation, the transcription metrics also used in [3] were utilized, namely the two accuracy measures (Acc_1, Acc_2), the total error (E_{tot}), the substitution error (E_{subs}), missed detection error (E_{fn}), and false alarm error (E_{fp}). Compared to Acc_1, accuracy Acc_2 also takes into account note substitutions. All evaluations take place by comparing the transcribed pitch output and the ground-truth MIDI files at a 10 ms scale.

For comparison, we employed the shift-invariant PLCA-based transcription model of [3] with the same CQT resolution as in the proposed model. The system used for comparison does not support any temporal constraints but uses the same formulation for source contribution, pitch impulse, pitch activity, as well as the same postprocessing step. Experiments were performed using ergodic HMMs (initialized with uniform transition probabilities), as they demonstrated superior performance compared to left-to-right HMMs for the single-pitch detection experiments in [2]. As explained in [2], although left-to-right HMMs might be more suitable for instruments exhibiting a clear temporal structure in note evolution (such as piano), in most instruments a fully connected HMM is more appropriate for expressing the temporal evolution of sound states. An example of the multi-pitch detection process can be seen in Fig. 2 where the pitch spectrogram of a guitar recording can be seen, along with the MIDI ground truth.

Results for the multi-pitch estimation experiments are presented in table 1, comparing the performance of the proposed method with the non-temporally constrained system of [3], over the three datasets. It can be seen that in all cases, the proposed method outperforms the shift-invariant PLCA-based model, with the smallest difference in terms of accuracy occurring for the MIREX recording. It should be noted that for the Disklavier dataset from [9], only piano templates were used in both systems. A common observation for all experiments is that the number of missed pitch detections is higher than the number of false positives.

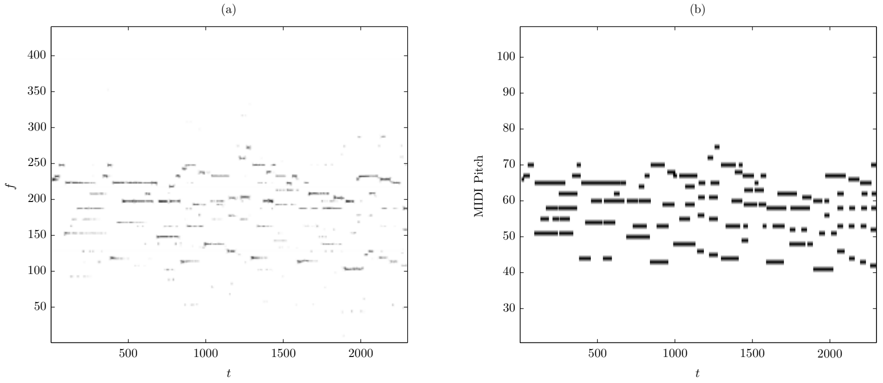


Fig. 2. (a) Pitch spectrogram $P(f, t)$ of an excerpt of "RWC-MDB-J-2001 No. 7" (guitar). (b) The pitch ground truth of the same recording. The abscissa corresponds to 10ms.

Table 1. Multi-pitch detection results using the proposed method compared to the one in [3] using three datasets

Dataset	**Method**	Acc_1	Acc_2	E_{tot}	E_{subs}	E_{fn}	E_{fp}
RWC	Proposed	61.6%	62.8%	37.2%	9.1%	18.3%	9.8%
	[3]	59.5%	60.3%	39.7%	9.2%	20.3%	10.2%
Disklavier	Proposed	58.6%	57.3%	42.7%	9.9%	16.3%	16.5%
	[3]	57.4%	55.5%	44.5%	10.8%	16.3%	17.4%
MIREX	Proposed	41.0%	47.0%	53.0%	25.4%	20.1%	7.5%
	[3]	40.5%	46.3%	53.8%	18.5%	32.3%	3.0%

Also, for the RWC and Disklavier datasets, results outperform state-of-the-art transcription algorithms (see [3] for transcription results using other methods in the literature). It should also be noted that most of the missed detections are located in the decay part of the produced notes. When no sparsity is used, the proposed method reports accuracy metrics $\{Acc_1, Acc_2\}$ of $\{56.3\%, 55.6\%\}$ for the RWC database, $\{56.8\%, 53.1\%\}$ for the Disklavier dataset, and $\{40.8\%, 46.9\%\}$ for the MIREX recording. Selected transcription examples are available online[2], along with the original recordings for comparison.

To the authors' knowledge, no statistical significance tests have been made for multi-pitch detection, apart from the piecewise Friedman tests in the MIREX task. However, given the fact that evaluations actually take place using 10 ms frames, even a small accuracy change can be shown to be statistically significant. Also, it should be noted that although using factorial HMMs (as in the source separation experiments of [7]) for the temporal constraints might in theory produce improved detection results, the model would be intractable, since it would need to compute 3^{88} sound state combinations.

[2] http://www.eecs.qmul.ac.uk/~emmanouilb/transcription.html

4 Conclusions

In this work we proposed a model for multi-pitch detection that extends the shift-invariant PLCA algorithm by introducing temporal constraints using HMMs. The goal was to model the temporal evolution for each produced note using spectral templates for each sound state. Results indicate that the temporal constraints produce improved multi-pitch detection accuracy rates compared to the standard shift-invariant PLCA model. It is also seen that the proposed system outperforms the state-of-the-art methods for the RWC transcription dataset and the Disklavier [9] dataset.

In the future, the proposed model will be tested using different HMM topologies and by incorporating update scheduling procedures for the various parameters to be estimated. Finally, the proposed transcription system will be extended by including an instrument identification step and by jointly performing multi-pitch estimation with note tracking.

References

1. Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.: A tutorial on onset detection of music signals. IEEE Trans. Audio, Speech, and Language Processing 13(5), 1035–1047 (2005)
2. Benetos, E., Dixon, S.: A temporally-constrained convolutive probabilistic model for pitch detection. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, pp. 133–136 (October 2011)
3. Benetos, E., Dixon, S.: Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In: 8th Sound and Music Computing Conf., Padova, Italy, pp. 19–24 (July 2011)
4. de Cheveigné, A.: Multiple F0 estimation. In: Wang, D.L., Brown, G.J. (eds.) Computational Auditory Scene Analysis, Algorithms and Applications, pp. 45–79. IEEE Press/Wiley (2006)
5. Emiya, V., Badeau, R., David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Trans. Audio, Speech, and Language Processing 18(6), 1643–1654 (2010)
6. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: music genre database and musical instrument sound database. In: Int. Conf. Music Information Retrieval, Baltimore, USA (October 2003)
7. Mysore, G.: A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures. Ph.D. thesis, Stanford University, USA (June 2010)
8. Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., Sagayama, S.: Non-negative Matrix Factorization with Markov-Chained Bases for Modeling Time-Varying Patterns in Music Spectrograms. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 149–156. Springer, Heidelberg (2010)
9. Poliner, G., Ellis, D.: A discriminative model for polyphonic piano transcription. EURASIP J. Advances in Signal Processing (8), 154–162 (January 2007)
10. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)
11. Smaragdis, P.: Relative-pitch tracking of multiple arbitary sounds. J. Acoustical Society of America 125(5), 3406–3413 (2009)

A Latently Constrained Mixture Model for Audio Source Separation and Localization

Antoine Deleforge and Radu Horaud

INRIA Grenoble Rhône-Alpes, France

Abstract. We present a method for audio source separation and localization from binaural recordings. The method combines a new generative probabilistic model with time-frequency masking. We suggest that device-dependent relationships between point-source positions and interaural spectral cues may be learnt in order to constrain a mixture model. This allows to capture subtle separation and localization features embedded in the auditory data. We illustrate our method with data composed of two and three mixed speech signals in the presence of reverberations. Using standard evaluation metrics, we compare our method with a recent binaural-based source separation-localization algorithm.

1 Introduction

We address the problem of simultaneous separation and localization of sound sources mixed in an acoustical environment and recorded with two microphones. Time-frequency masking is a technique allowing the separation of an arbitrary number of sources with only two microphones by assuming that a single source is active at every time-frequency point – the *W-disjoint orthogonality* (W-DO). It was shown that this assumption holds, in general, for simultaneous speech signals [8]. The input signal is represented in a time-frequency domain and points corresponding to the target source are weighted with 1 and otherwise with 0. The masked spectrogram is then converted back to a temporal signal. A number of methods combine time-frequency masking with localization-based clustering ([8],[3],[2]), e.g., DUET [8] which allows to separate anechoic mixtures when each source reaches the microphones with a single attenuation coefficient and delay. This mixing model is well suited for "clean" binaural recordings. In practice, more complex filtering effects exist, namely the *head-related transfer function* (HRTF) and the *room impulse response* (RIR). These filters lead to frequency-dependent attenuations and delays between the two microphones, respectively called the *interaural level difference* (ILD) and the *interaural phase difference* (IPD). Some approaches attempted to account for these dependencies by learning a mapping between azimuth, frequencies and interaural cues [7,5,3]. These mappings usually consist in finding a functional relationship that best fits data obtained from an HRTF dataset. To improve robustness to RIR variations, these interaural cues can also be integrated in a mixture model, e.g., [2].

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 372–379, 2012.

In this paper we propose to directly learn a discrete mapping between a set of 3D point sources and IPD/ILD spectral cues. We will refer to such mappings as *Source-Position-to-Interaural-Cues* maps (SPIC maps). Unlike what is done in [7,5,3], the proposed mapping is built point-wise, does not rely on azimuth only and is device-dependent. We explicitly incorporate it into a novel *latently constrained mixture model* for point sound sources. Our model is specifically designed to capture the richness of binaural data recorded with an acoustic dummy head, and this to improve both localization and separation performances. We formally derive an EM algorithm that iteratively performs separation (E-step) followed by localization and source-parameter estimation (M-step). The algorithm is supervised by a training stage consisting in learning a mapping between potential source positions and interaural cues, i.e., SPIC maps. We believe that a number of methods could be used in practice to learn such maps. In particular we propose an *audio-motor mapping* approach. The results obtained with our method compare favorably with the recently proposed MESSL algorithm [2].

2 Binaural Sound Representation

Spectrograms associated with each one of the two microphones are computed using short-term FFT analysis. We use a 64ms time-window with 8ms window overlap, thus yielding $T = 126$ time windows for a 1s signal. Since sounds were recorded at a sample rate of 16,000Hz, each time window contains 1,024 samples. Each window is then transformed via FFT to obtain complex coefficients of $F = 513$ positive frequency channels between 0 and 8,000Hz. We denote with $s^{(k)}_{f,t} \in \mathbb{C}$ the (f, t) point of the spectrogram emitted by sound-source k, and with $s^{(\mathrm{L})}_{f,t}$ and $s^{(\mathrm{R})}_{f,t}$ the spectrogram points perceived by the left- and right-microphone respectively. The W-DO assumption implies that a single sound source k emits at a given point (f, t). The relationships between the emitted and the left and right perceived spectrogram points are:

$$s^{(\mathrm{L})}_{f,t} = h^{(\mathrm{L})}(\boldsymbol{x}_k, f)\, s^{(k)}_{f,t} \;\text{ and }\; s^{(\mathrm{R})}_{f,t} = h^{(\mathrm{R})}(\boldsymbol{x}_k, f)\, s^{(k)}_{f,t} \tag{1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^3$ is the 3D position of sound source k in a listener-centered coordinate frame and $h^{(\mathrm{L})}$ and $h^{(\mathrm{R})}$ denote the left and right HRTFs. The *interaural transfer function* (ITF) is defined by the ratio between the two HRTFs, i.e., $I(\boldsymbol{x}_k, f) = h^{(\mathrm{R})}(\boldsymbol{x}_k, f)/h^{(\mathrm{L})}(\boldsymbol{x}_k, f) \in \mathbb{C}$. The interaural spectrogram is defined by $\hat{I}_{f,t} := s^{(\mathrm{R})}_{f,t}/s^{(\mathrm{L})}_{f,t}$, so that $\hat{I}_{f,t} \approx I(\boldsymbol{x}_k, f)$. Note that the last approximation only holds if there is a source k emitting at frequency-time point (f, t), and if the time delay between microphones (≈ 0.75ms) is much smaller than the Fourier transform time-window that is used (64ms). Under these conditions, at a given frequency-time point, the interaural spectrogram value $\hat{I}_{f,t}$ does not depend on the emitted spectrogram value $s^{(k)}_{f,t}$ but only on the emitting source position $\boldsymbol{x}_k$. We finally define the *ILD spectrogram* α and the *IPD spectrogram* ϕ as the log-amplitude and phase of the complex interaural spectrogram $\hat{I}_{f,t}$:

$$\alpha_{f,t} = 20 \log |\hat{I}_{f,t}| \in \mathbb{R}, \quad \phi_{f,t} = \arg(\hat{I}_{f,t}) \in \,]-\pi, \pi] \tag{2}$$

As already outlined in Section 1 our method makes use of a SPIC map that is learnt during a training stage. Let $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ be a set of 3D sound-source locations in a listener-centered coordinate frame. Let a sound-source n, located at $\boldsymbol{x}_n$ emit white noise and let $\{\alpha_{f,t}^n\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}^n\}_{f=1,t=1}^{F,T}$be the perceived ILD and IPD spectrograms. The *mean ILD* $\boldsymbol{\mu}(\boldsymbol{x}_n) = (\mu_1^n \dots \mu_f^n \dots \mu_F^n)^\top \in \mathbb{R}^F$ and *mean IPD* $\boldsymbol{\xi}(\boldsymbol{x}_n) = (\xi_1^n \dots \xi_f^n \dots \xi_F^n)^\top \in \,]-\pi, \pi]$ *vectors* associated with n are defined by taking the temporal means of α^n and ϕ^n at each frequency channel:

$$\mu_f^n = 1/T \textstyle\sum_{t=1}^T \alpha_{f,t}^n \;\text{ and }\; \xi_f^n = \arg(1/T \textstyle\sum_{t=1}^T e^{j\phi_{f,t}^n}) \tag{3}$$

Vector $\boldsymbol{\xi}$ is estimated in the complex domain in order to avoid problems due to phase circularity [4]. White noise is used because it contains equal power within a fixed bandwidth at any center frequency: The source n is therefore the only source emitting at each point (f, t); μ_f^n and ξ_f^n are thus approximating the log-amplitude and phase of $I(\boldsymbol{x}_k, f)$. The set $\mathcal{X}$ of 3D source locations as well as the mappings $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ will be referred to as the training data to be used in conjunction with the separation-localization algorithm described below.

3 Constrained Mixtures for Separation and Localization

Let's suppose now that there are K simultaneously emitting sounds sources from unknown locations $\{\boldsymbol{x}_k\}_{k=1}^K \subset \mathcal{X}$ and with unknown spectrograms. Using the listener's microphone pair it is possible to build the ILD and IPD observed spectrograms $\{\alpha_{f,t}\}_{f=1,t=1}^{F,T}$ and $\{\phi_{f,t}\}_{f=1,t=1}^{F,T}$. The goal of the sound-source separation and localization algorithm described in this section is to associate each observed point (f, t) with a single source and to estimate the 3D location of each source.

As mentioned in section 2, the observations $\alpha_{f,t}$ (ILD) and $\phi_{f,t}$ (IPD) are significant only if there is a sound source emitting at (f, t). To identify such *significant observations* we estimate the *sound intensity level* (SIL) spectrogram at the two microphones, and retain only those frequency-time points for which the SIL is above some threshold. One empirical way to choose the thresholds (one for each frequency) is to average the SILs at each f in the absence of any emitting source. These thresholds are typically very low compared to SILs of natural sounds, and allow to filter out frequency-time points corresponding to "room silence". Let $M_f \le T$ be the number of significant observations at f and let $\alpha_{f,m}$ and $\phi_{f,m}$ be the m-th significant ILD and IPD observations at f. Let $\mathbf{A} = \{\alpha_{f,m}\}_{f=1,m=1}^{F,M_f}$ and $\boldsymbol{\Phi} = \{\phi_{f,m}\}_{f=1,m=1}^{F,M_f}$ be the *observed data*.

Let $\boldsymbol{z}_{f,m} \in \{0, 1\}^K$ be the *missing data*, i.e., the data-to-source assignment variables, such that $z_{f,m,k} = 1$ if observations $\alpha_{f,m}$ and $\phi_{f,m}$ are generated by source k, and $z_{f,m,k} = 0$ otherwise. The W-DO assumption yields $\sum_{k=1}^{K} z_{f,m,k} = 1$ for all (f, m). $\mathcal{M}_k = \{z_{f,m,k}\}_{f=1,m=1}^{F,M_f}$ is the binary spectral mask of the k-th source. Finally, $\mathbf{Z} = \{\boldsymbol{z}_{f,m}\}_{f=1,m=1}^{F,M_f}$ denotes the set of all missing data. The problem of simultaneous localization and separation amounts to estimate the masking variables $\mathbf{Z}$ and the locations $\{\boldsymbol{x}_k\}_{k=1}^{K}$ conditioned by $\mathbf{A}$ and $\boldsymbol{\Phi}$, given the number of sources K. We assume that observed data are perturbed by Gaussian noise. Hence, the probability of observing $\alpha_{f,m}$ conditioned by source k $(z_{f,m,k} = 1)$ located at $\boldsymbol{x}_k$ is drawn from a normal distribution, and the probability of observing $\phi_{f,m}$ is drawn from a circular normal distribution. The source position $\boldsymbol{x}_k$ acts here as a *latent constraint* on ILD and IPD means:

$$P(\alpha_{f,m}|z_{f,m,k} = 1, \boldsymbol{x}_k, \sigma_{f,k}) = \mathcal{N}(\alpha_{f,m}|\mu_f(\boldsymbol{x}_k), \sigma_{f,k}^2) \text{ and } \quad (4)$$

$$P(\phi_{f,m}|z_{f,m,k} = 1, \boldsymbol{x}_k, \rho_{f,k}) = \mathcal{N}(\Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))|0, \rho_{f,k}^2) \quad (5)$$

where $\sigma_{f,k}^2$ and $\rho_{f,k}^2$ are the ILD and IPD variances associated with source k at frequency f and the Δ function is defined by $\Delta(x, y) = \arg(e^{j(x-y)}) \in]-\pi, \pi]$. As in [2], (5) approximates the normal distribution on the circle $]-\pi, \pi]$ when $\rho_{f,k}$ is small relative to 2π. Preliminary experiments on IPD spectrograms of white noise showed that this assumption holds in the general case. As emphasized in [2], the well known correlation between ILD and IPD does not contradict the assumption that Gaussian noises corrupting the observations are independent. The conditional likelihood of the observed data $(\alpha_{f,m}, \phi_{f,m})$ is therefore given by the product of (4) and (5). We also define the priors $\pi_{f,k} = P(z_{f,m,k})$ which model the proportion of the observed data generated by source k at frequency f. In summary, the model parameters are $\Theta = \{\{\boldsymbol{x}_k\};\ \{\pi_{f,k}\};\ \{\sigma_{f,k}^2\};\ \{\rho_{f,k}^2\}\}_{f=1,k=1}^{F,K}$. The problem can now be expressed as the maximization of the observed-data log-likelihood conditioned by Θ. In order to keep the model as general as possible, there is no assumption on the emitted sounds as well as the way their spectra are spread across the frequency-time points. Therefore, we assume that all the observations are statistically independent, yielding the following expression for the observed-data log-likelihood:

$$\mathcal{L}(\mathbf{A}, \boldsymbol{\Phi}; \Theta) = \log P(\mathbf{A}, \boldsymbol{\Phi}; \Theta) = \textstyle\sum_{f=1}^{F} \sum_{m=1}^{M_f} \log P(\alpha_{f,m}, \phi_{f,m}; \Theta) \quad (6)$$

We address this maximum-likelihood with missing-data problem within the framework of expectation-maximization (EM). In our case, the E-step computes the posterior probabilities of assigning each spectrogram point to a sound source k (separation) while the M-step maximizes the expected complete-data log-likelihood with respect to the model parameters Θ and, most notably, with the source locations $\{\boldsymbol{x}_k\}_{k=1}^{K}$ (localization). The MAP criterion provides binary spectral masks $\mathcal{M}_k$ associated with each source k while the final parameters $\{\boldsymbol{x}_k\}_{k=1}^{K}$ provide estimates for the source locations. The expected complete-data log-likelihood writes ($^{(p)}$ denotes the p-th iteration):

$$Q(\Theta|\Theta^{(p-1)}) = \textstyle\sum_{f=1}^{F} \sum_{m=1}^{M_f} \sum_{k=1}^{K} r_{f,m,k}^{(p)} \log \pi_{f,k} P(\alpha_{f,m}, \phi_{f,m}|\boldsymbol{z}_{f,m}; \Theta) \quad (7)$$

The *E-step* updates the responsibilities according to the standard formula:

$$r_{f,m,k}^{(p)} = \frac{\pi_{f,k} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})}{\sum_{i=1}^{K} \pi_{f,i} P(\alpha_{f,m}, \phi_{f,m} | \boldsymbol{z}_{f,m}; \Theta^{(p-1)})} \tag{8}$$

The *M-step* maximizes (7) with respect to Θ. By combining (4) and (5) with (7) the equivalent minimization criterion writes:

$$\sum_{f=1}^{F} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \left(\log \left(\frac{\sigma_{f,k}^2 \rho_{f,k}^2}{\pi_{f,k}^2} \right) + \frac{(x_{f,m} - \mu_f(\boldsymbol{x}_k))^2}{\sigma_{f,k}^2} + \frac{\Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))^2}{\rho_{f,k}^2} \right) \tag{9}$$

which can be differentiated with respect to $\{\pi_{f,k}\}_f$, $\{\sigma_{f,k}\}_f$ and $\{\rho_{f,k}\}_f$ to obtain closed-form expressions for the optimal parameter values conditioned by $\boldsymbol{x}_k$:

$$\widetilde{\pi}_{f,k} = \frac{\overline{r}_{f,k}}{M_f}, \text{ with } \overline{r}_{f,k} = \sum_{m=1}^{M_f} r_{f,m,k} \tag{10}$$

$$\widetilde{\sigma}_{f,k}^2(\boldsymbol{x}_k) = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} (x_{f,m} - \mu_f(\boldsymbol{x}_k))^2 \tag{11}$$

$$\widetilde{\rho}_{f,k}^2(\boldsymbol{x}_k) = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \Delta(\phi_{f,m}, \xi_f(\boldsymbol{x}_k))^2 \tag{12}$$

By substituting (11) and (12) into (9) the optimal location $\widetilde{\boldsymbol{x}}_k$ is obtained by minimizing the following expression with respect to $\boldsymbol{x}_k$:

$$\sum_{f=1}^{F} \overline{r}_{f,k} \left(\log \left(1 + \frac{(\overline{\alpha}_{f,k} - \mu_f(\boldsymbol{x}_k))^2}{V_{f,k}} \right) + \log \left(1 + \frac{\Delta(\overline{\phi}_{f,k}, \xi_f(\boldsymbol{x}_k))^2}{W_{f,k}} \right) \right) \tag{13}$$

with: $\overline{\alpha}_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \alpha_{f,m}$; $V_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} (\alpha_{f,m} - \overline{\alpha}_{f,k})^2$
$\overline{\phi}_{f,k} = \arg \left(\frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} e^{j\phi_{f,m}} \right)$; $W_{f,k} = \frac{1}{\overline{r}_{f,k}} \sum_{m=1}^{M_f} r_{f,m,k}^{(p)} \Delta(\phi_{f,m}, \overline{\phi}_{f,k})^2$

(13) is evaluated for each source location in the training dataset $\mathcal{X}$ (Section 2) in order to find an optimal 3D location $\widetilde{\boldsymbol{x}}_k$. This is then substituted back in (11) and (12) to estimate $\widetilde{\sigma}_{f,k}$ and $\widetilde{\rho}_{f,k}$ and repeated for each unknown source k.

In general, EM converges to a local maximum of (6). The non-injectivity nature of the interaural functions μ_f and ξ_f and the high cardinality of Θ leads to a very large number of such maxima, especially when the training set $\mathcal{X}$ is large. This makes our algorithm very sensitive to initialization. One way to avoid being trapped in local maxima is to initialize the mixture's parameters at random several times. This cannot be easily applied here since there is no straightforward way to initialize the model's variances. Alternatively, one may randomly initialize the assignment variables $\mathbf{Z}$ and then proceed with the M-step. However, extensive simulated experiments revealed that this solution fails to converge to the ground-truth solution in most of the cases. We therefore propose to combine these strategies by randomly perturbing both the source locations and the source assignments during the first stages of the algorithm. We developed a *stochastic initialization* procedure similar in spirit to SEM [1].

The SEM algorithm includes a stochastic step (S) between the E- and the M-step, during which random samples $R_{f,m,k} \in \{0, 1\}$ are drawn from the responsibilities (8). These samples are then used instead of (8) during the M-step. To initialize our algorithm, we first set $r^{(0)}_{f,m,k} = 1/K$ for all k and then proceed through the sequence S M* E S M, where M* is a variation of M in which the source positions are drawn randomly from $\mathcal{X}$ instead of solving (13). In practice, ten such initializations are used to enforce algorithm convergence, and only the one providing the best log-likelihood after two iterations is iterated twenty more times. A second technique was used to overcome local maxima issues due to the large number of parameters. During the first ten steps of the algorithm only, a unique pair of variances (σ^2_k, ρ^2_k) is estimated for each source. This is done by calculating the means $\overline{\sigma}^2_k(\boldsymbol{x}_k)$ and $\overline{\rho}^2_k(\boldsymbol{x}_k)$ of frequency-dependent variances (11) and (12) weighted by $\overline{r}_{f,k}$. The optimal value $\widetilde{\boldsymbol{x}}_k$ is the one minimizing $\overline{\sigma}^2_k(\boldsymbol{x})\overline{\rho}^2_k(\boldsymbol{x})$ evaluated over all $\boldsymbol{x} \in \mathcal{X}$. Intensive experiments showed that the proposed method converges to a global optimum in most of the cases.

4 Experiments, Results, and Conclusions

In order to evaluate and compare our method, a specific data set of binaural records was built[1] using a Sennheiser MKE 2002 acoustic dummy-head mounted onto a robotic system with two rotational degrees of freedom, namely pan (ψ) and tilt (θ). This device, specifically designed to perform accurate and reproducible motions, allows us to collect both a very dense SPIC map for the training set (section 2) and a large test set of mixed speech point sources. The emitter (a loud speaker) is placed at approximately 2.5 meters in front of the listener. Under these conditions the HRTF mainly depends on the sound-source direction: Hence, the location is parameterized by the angles ψ and θ. All the experiments were carried out in a reverberant room and in the presence of background noise. For recording purposes, the robot is placed in 90 pan angles $\psi \in [-90°, 90°]$ (left-right) and 60 tilt angles $\theta \in [-60°, 60°]$ (top-down), i.e., $N = 5,400$ uniformly distributed *motor states* in front of the *static* emitter, forming the set $\mathcal{X}$. Five binaural recordings are available with each motor state: Sound #0 corresponds to a 1s “room silence” used to estimate the SIL thresholds (section 3). Sound #1 corresponds to 1s white-noise used to build the training set (section 2). Sounds #2, #3 and #4 form the test set and correspond to “They never met you know” by a female (#2), “It was time to go up myself” by a male (#3), and “As we ate we talked” by a male (#4). The three sounds are about 2s long and were randomly chosen from the TIMIT database. Each record was associated to its ground-truth motor-state, thus allowing to create signals of mixed sound sources from different direction with 2° resolution.

We generated 1000 mixtures of two and three speech signals emitted by randomly located sources. 97.7% of the individual sources were correctly mapped to their associated position (i.e. $\leq 2°$ error for both ψ and θ) in the two-source case,

[1] Online at: http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset

Table 1. Comparing the mean source-to-distortion ratio (SDR) and source-to-interference ratio (SIR), in dB, for 1000 mixtures of 2 and 3 sources. Mean separation results with our approach are calculated over all sources (All) and over correctly localized sources only (Loc).

	2 Sources		3 Sources	
	SDR	SIR	SDR	SIR
Oracle Mask	11.73	19.23	9.20	16.16
Our Approach (Loc)	5.28	8.91	2.44	3.92
Our Approach (All)	5.19	8.84	1.72	2.74
MESSL-G	2.83	5.74	1.48	1.47
Original Mixture	0.00	0.45	-3.50	-2.82

and 63.8% in the three-source case. The performance of separation was evaluated with the standard SDR and SIR metrics [6]. We compared our results to those obtained with the original mixture (no mask applied), with the ground truth or *Oracle mask* [8], and with the recently proposed MESSL[2] algorithm [2]. The Oracle mask is set to 1 at every spectrogram point in which the target signal is at least as loud as the combined other signals and 0 everywhere else. The version MESSL-G used includes a garbage component and ILD priors to better account for reverberations and is reported to outperform four methods in reverberant conditions, including [8] and [3]. Table 1 shows that our method yields significantly better results than MESSL-G on an average, although both algorithm require similar computational times. Notice how the localization correctness critically affects the separation performances, and decreases in the three-source case, as the number of observations per source becomes lower and the number of local maxima in (6) becomes higher. Our SDR scores strongly outperform MESSL-G in most cases, while SIR results are only slightly better when sources are more than 70° apart in azimuth (pan angle), e.g., Fig. 1. However, they become much higher when sources are nearby in azimuth, or share the same azimuthal plane with different elevations (tilt angles). This is because MESSL relies on the estimation of a probability density in a discretized ITD space for each source, and thus does not account for more subtle spatial cues induced by the HRTF.

These results clearly demonstrate the efficiency of our method, but they somehow favor our algorithm because of the absence of RIR variations both in the training and the test data sets. The aim of experimenting with these relatively simple data has been to show that our method can conceptually separate and accurately locate both in azimuth and elevation a binaural mixture of 2 to 3 sound sources. The prerequisite is a training stage: the interaural cues associated with source positions need to be learnt in advance using white noise, and we showed that the algorithm performs well even for a very large and dense set of learnt positions. Preliminary results obtained while changing the position of the test sound source in the room suggested that our constrained mixture model coupled with frequency-dependent variances presented some robustness to RIR variations. Alternatively, one could build a training set on different premises such as seat locations in a conference room or musician locations in a concert hall, and thus directly learn the RIR during the training stage.

To conclude, we proposed a novel audio source separation and localization method based on a mixture model constrained by a SPIC map. Experiments

[2] `http://blog.mr-pc.org/2011/09/14/messl-code-online/`.

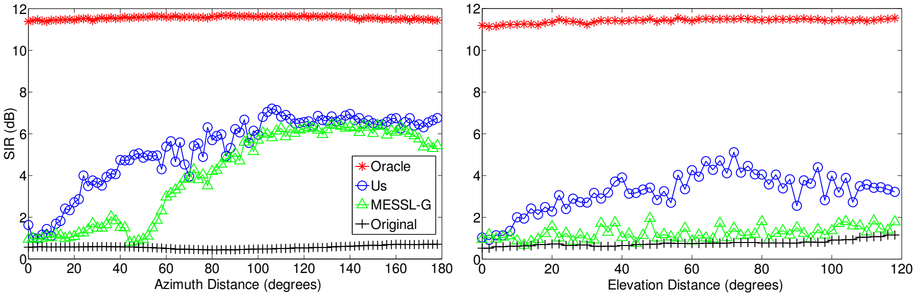


Fig. 1. SIR as a function of azimuth (pan) and elevation (tilt) separation between two sources. Left: one source fixed at $(-90°, 0°)$ while the other takes 90 positions between $(-90°, 0°)$ and $(+90°, 0°)$. Right: one source fixed at $(0°, -60°)$ while the other takes 60 positions between $(0°, -60°)$ and $(0°, +60°)$. SIRs are averaged over 6 mixtures of 2 sources (12 targets). Top-to-down: Oracle ($*$), our method ($\circ$), MESSL-G ($\triangle$), and original mixture ($+$).

and comparisons showed that our algorithm performs better than a recently published probabilistic spectral masking technique in terms of separation and yields very good multi-source localization results. The combination of a SPIC map with a mixture model is a unique feature. In the future, we plan to study more thoroughly the behavior of our algorithm to RIR variations, and improve its robustness by extending our model to a continuous and probabilistic mapping between source positions and interaural parameters. Dynamic models incorporating moving sound sources and head movements could also be included.

References

1. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comp. Stat. & Data An. 14(3), 315–332 (1992)
2. Mandel, M.I., Weiss, R.J., Ellis, D.P.W.: Model-based expectation-maximization source separation and localization. IEEE TASLP 18, 382–394 (2010)
3. Mouba, J., Marchand, S.: A source localization/separation/respatialization system based on unsupervised classification of interaural cues. In: Proceedings of the International Conference on Digital Audio Effects, pp. 233–238 (2006)
4. Nix, J., Hohmann, V.: Sound source localization in real sound fields based on empirical statistics of interaural parameters. JASA 119(1), 463–479 (2006)
5. Roman, N., Wang, D., Brown, G.J.: Speech segregation based on sound localization. JASA 114(4), 2236–2252 (2003)
6. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE TASLP 14(4), 1462–1469 (2006)
7. Viste, H., Evangelista, G.: On the use of spatial cues to improve binaural source separation. In: Proc. Int. Conf. on Digital Audio Effects, pp. 209–213 (2003)
8. Yılmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52, 1830–1847 (2004)

Multiple Instrument Mixtures Source Separation Evaluation Using Instrument-Dependent NMF Models

Francisco J. Rodriguez-Serrano[1], Julio J. Carabias-Orti[1],
Pedro Vera-Candeas[1], Tuomas Virtanen[2], and Nicolas Ruiz-Reyes[1]

[1] Universidad de Jaen, Jaen, Spain
fjrodrig@ujaen.es
[2] Tampere University of Technology, Tampere, Finland

Abstract. This work makes use of instrument-dependent models to separate the different sources of multiple instrument mixtures. Three different models are applied: (a) basic spectral model with harmonic constraint, (b) source-filter model with harmonic-comb excitation and (c) source-filter model with multi-excitation per instrument. The parameters of the models are optimized by an augmented NMF algorithm and learnt in a training stage. The models are presented in [1], here the experimental setting for the application to source separation is explained. The instrument-dependent NMF models are first trained and then a test stage is performed. A comparison with other state-of-the-art software is presented. Results show that source-filter model with multi-excitation per instrument outperforms the other compared models.

Keywords: non-negative matrix factorization (NMF), source-filter model, excitation modeling, spectral analysis, music source separation.

1 Introduction

An audio spectrogram can be decomposed as a linear combination of spectral basis functions. In such a model, the short-term magnitude (or power) spectrum of the signal $x_t(f)$ in frame t and frequency f is modeled as a weighted sum of basis functions as

$$\hat{x}_t(f) = \sum_{n=1}^{N} g_{n,t} b_n(f) \tag{1}$$

where $g_{n,t}$ is the gain of the basis function n in the frame t, and $b_n(f)$, $n = 1, ..., N$ are the bases. In other words, the signal is modeled as a sum of components with fixed basis and time varying amplitudes.

In this paper three models are tested for source separation of musical instruments. The models are derived from eq.(1) and have different constraints to predict note spectrum of musical instruments. All of them are explained in

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 380–387, 2012.

[1] and here are tested and compared with a state-of-the-art software. Despite this, here a brief formulation of each model is shown in order to show a general description of them.

2 NMF Models

2.1 Basic Harmonic Constrained Model

This model is based on the assumption that musical notes spectra have regularly spaced frequency peaks. Because of this, the elements in the basis $b_{n,j}(f)$ are forced to follow this harmonic shape by imposing harmonicity to the spectral basis.

Consequently, in the Basic Harmonic Constrained (BHC) model the short-term magnitude spectrum of the signal $x_t(f)$ at frame t is estimated as

$$\hat{x}_t(f) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} a_{n,m,j} G(f - m f_0(n)) \tag{2}$$

where $m = 1, ..., M$ is the number of harmonics, $a_{n,m,j}$ the amplitude for the m-th partial of the pitch n and instrument j, $f_0(n)$ the fundamental frequency of pitch n, $G(f)$ is the magnitude spectrum of the window function, the spectrum of a harmonic component at frequency $m f_0(n)$ is approximated by translated $G(f - m f_0(n))$ and J is the number of instruments. The parameters to estimate of the BHC model for the NMF iterative algorithm are the time gains $g_{n,t}$ and the pitch amplitudes $a_{n,m,j}$.

2.2 Source-filter Model with Harmonic-Comb Excitation

This model is based on the proposal of Virtanen and Klapuri [2]. Here, each basis is modeled as the product of an excitation $e_n(f)$ and a source-filter $h_j(f)$.

In order not to have a large number of parameter to be fitted by the NMF algorithm, which would have a negative effect in the results, we introduce the excitations $e_n(f)$ as frequency components of unity magnitude at integer multiples of the fundamental frequency of the pitch n as in [3]. This results in modeling the basis using a *harmonic comb* excitation consisting of a sum of harmonic components. When using source-filter model with Harmonic-Comb Excitation (HCE) for the definition of basis functions, the time-frequency representation of the signal can be obtained as

$$\hat{x}_t(f) = \sum_{j=1}^{J} \sum_{n=1}^{N} g_{n,t,j} \sum_{m=1}^{M} h_j(m f_0(n)) G(f - m f_0(n)) \tag{3}$$

The parameters to estimate for the NMF algorithm are the time gains $g_{n,t,j}$ and the source-filter $h_j(f)$ and can be estimated in a NMF framework.

2.3 Source-filter Model with Multi-Excitation Per Instrument

In the third model, we use the multi-excitation model proposed in [1]. Here the excitation is composed of a weighted sum of a set of instrument-dependent excitation basis vectors. For each instrument, the pitch excitation is obtained as the weighted sum of excitation basis functions which are unique for each instrument while the weights vary as function of the pitch. The magnitude spectra of the whole signal is represented as:

$$\hat{x}_t(f) = \sum_{n,j} g_{n,t,j} h_j(f) \sum_{m=1}^{M} \sum_{i=1}^{I} w_{i,n,j} v_{i,m,j} G\left(f - m f_0(n)\right) \qquad (4)$$

where $v_{i,m,j}$ is the i-th excitation basis vector, $w_{i,n,j}$ is the weight of the i-th excitation,$n = 1, ..., N$ (N being the number of pitches), $j = 1, ..., J$ (J being the number of instruments), M represents the number of harmonics and I the number of considered excitations with $I << N$. The parameters to estimate of the model are: the time gains $g_{n,t,j}$, the instrument filter $h_j(f)$, the basis excitation vectors $v_{i,m,j}$ and the excitation weigths $w_{i,n,j}$.

More details of these instrument-dependent NMF models, such as the parameter estimation can be revised in [1].

3 Application to Source Separation

Source separation is a practical application that can be approached even without a priori information [4]. The approach proposed here is based on training NMF models with all the range of notes of some instruments. In the test stage, the parameters of the trained models are fixed except the time gains $g_{n,t,j}$. For the separation task, multiple instrument mixtures are modelled obtaining the estimations of magnitude spectra $\hat{x}_t(f)$ per each instrument.

In this work, we do NMF-based separation based on the three different models explained in Section 2. It must be stressed that these models have been proposed in the literature and are here applied to source separation.

3.1 Experimental Setup

Training Data. For the training stage, the full pitch range of isolated notes for each instrument from RWC musical instrument sound database [5] has been used. Five instruments are considered for the experiments (clarinet, flute, oboe, horn and bassoon). For each instrument, individual sounds are available at semi-tone intervals over the entire range of notes that can be produced by that instrument. From the RWC database we select the files with normal playing style and mezzo dynamic level.

Test Data. To evaluate the separation application, we have used the woodwind database for the Multiple Fundamental Estimation task of the Third Music Information Retrieval Evaluation Exchange (MIREX2007) [6]. This subset is composed of 5 solo instruments (bassoon, clarinet, flute, horn and oboe). Polyphonic signals are generated by mixing the recordings of the individual instrument excerpts. The mixing is performed using just the first 30 seconds of the individual instruments as in [7]. In this way, excerpts range from polyphony 2 to 5 are created, giving 26 individual excerpts as a result.

Time-Frequency Representation. The model parameters are learnt using the training data. Consequently, the training data has to be labelled with the instrument j and pitch n active at each frame t. In our system we use the resolution of a single semitone as in [1]. Taking this decision, the implementation is easier because the database is also annotated with a single semitone resolution.

The most straightforward implementation of this time-frequency representation is the integration of the STFT bins corresponding to the same semitone. However, the frequency representation given by this simple implementation retains energy out of the boundaries of the played MIDI note (and its multiples) especially at low frequency due to the side lobes of the window transform. As a consequence, harmonic constrained NMF models can found energy out of the played MIDI semitone (and its multiples) which limits the separation capabilities.

To avoid this behavior, perceptually most significant sinusoids are extracted at each frame using the Perceptual Matching Pursuit [8]. This approach achieves the cancellation (in a great degree) of each extracted sinusoid minimizing the side lobe effect. After this processing block, all perceptually important peaks of the spectrum are extracted. The frequency resolution of a semitone is achieved retaining in $x_t(f)$ only the most perceptually important sinusoid with the frequency range of each MIDI note f at each frame t. Due to the properties of matching pursuits, the window transform $G(f)$ simplifies to the delta function. The frame size and the hop size are 128 ms and 32 ms, respectively. This sinusoidal model is used to process both the training and test data. More details about the used time-frequency representation can be revised in [1].

NMF-Based Separation Procedure. All the tested NMF-based methods are processed on the same procedure:

- Compute the time-frequency representation of the mixture $x_t(f)$ as explained above.
- Estimate the factorization $\hat{x}_t(f) \approx \sum_n g_{n,t} b_{n,j}(f)$. The bases are fixed from the training stage for all the compared models, while the time gains are optimized using the NMF algorithm for each mixture.
- The factorization can be particularized for each instrument j obtaining $\hat{x}_{t,j}(f)$. Here, we factorize only for the active instruments, in other words, the number and kind of instruments is given to the system as a priori information. Other approaches does not inform to the algorithm about the active instruments [3] in such cases an instrument classification is performed before separation.

- The separation mask is obtained computing the proportional amount of amplitude of each time-frequency cell (frame t and MIDI note interval f) to the instrument j. The use of separation Wiener masks is common in the separation literature [9].
- Once obtained the instrument masks, the spectrogram of the signal is filtered by each mask. Here, as the frequency resolution is different from that of the spectrogram, all frequency bins belonging to the current MIDI note interval are filtered in the same way. The spectrogram is computed with 8192 frequency bins.

State-of-the-Art Comparison. In order to obtain a fair comparison with a state-of-the-art method, we have performed an instrument separation making use of the *Flexible Audio Source Separation Toolbox (FASST)* [10]. This toolbox is a configurable framework for sound source separation. From all the possible scenarios in the toolbox, in the experiments presented at this work, the toolbox has been configured using the most similar as possible scenario than proposed methods. The tested transform are *qerb* and *stft*, both of them are implemented in FASST. The decompositions are always computed using $K = 114$ bases, with this value the range of notes for any instrument is modelled. The excitation-filter decomposition is selected. The source separation is done following a two stage approach: 1) First of all, a training stage is implemented. To do that, a decomposition for each instrument of the training data set is performed. Spectral patterns activations are initialized to zero for all bases k that are not active at frame n. The obtained characteristic spectral patterns are stored for the next stage. 2) Then the test stage is implemented. Now, a decomposition for each excerpt of the test data set (with polyphony from 2 to 5) is computed using the stored characteristic spectral patterns for the active instruments of each excerpt. The obtained spectral patterns activations for each excerpt are utilized for the final separation in FASST. This two stage approach is implemented to make a comparison under the same conditions than the used in the proposed procedure for the instrument-dependent NMF models. In FASST, the sampling frequency is set to $44,100$ Hz, the used transforms are QERB (Quadratic Equivalent Rectangular Bandwidth) and STFT (Short Time Fourier Transform), the length of the time integration function is set to $5,644$ samples (128 ms frames, so the frequency resolution is configured to $F = 5644$) and an overlapping of 50 % between frames is chosen.

Evaluation. For an objective evaluation of the separation performance we use the metrics implemented in [11]. The metrics for each separated signal are the *Source to Distortion Ratio* (SDR), the *Source to Interference Ratio* (SIR), the *Source to Artifacts Ratio* (SAR).

In a NMF framework, the unknown parameters are initialized randomly. Thus, in the test stage, the estimated magnitude spectra are different at each execution giving results to different separation metrics per execution. We have performed a set of 30 executions per algorithm to demonstrate the statistical significance of the metrics. The 95% confidence intervals for the metrics was smaller than 0.7dB for all the algorithms, which means that differences between most of the algorithms are statistically significant.

4 Results

Separation results for the BHC, HCE and MEI models are presented at Table 1. A comparison with the separation performed by ideal Wiener masks and the chosen configuration of the FASST software is also included. These ideal masks are obtained with the original signals, which is an unrealistic situation. Anyway, this particular comparison is interesting to demonstrate the limitations of the system. The separation performance is here limited by the time-frequency resolution. This is the reason why ideal Wiener masks do not perform better.

Table 1. Objective Results for Source Separation of Multiple Instrument Mixtures in dB

Algorithms / Polyphony		2			3			4			5		
(all values in dB)		SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
BHC		8.8	10.9	11.0	6.8	8.6	9.4	5.2	5.9	8.3	3.8	2.4	7.5
HCE		5.5	8.3	11.0	2.9	4.8	9.6	1.7	2.8	8.7	0.9	-8.2	8.2
MEI	I=1	9.0	10.9	11.4	6.5	8.3	9.8	4.2	5.7	8.9	0.9	1.3	7.9
	I=2	9.1	11.1	11.4	6.9	8.9	9.8	5.0	6.7	8.6	2.8	4.1	8.0
	I=4	9.4	11.3	11.5	7.6	9.4	10.0	6.3	7.8	9.0	5.2	6.2	8.2
FASST *stft*		4.9	6.3	6.9	3.3	3.7	4.6	1.3	2.1	3.4	0.4	2.3	3.4
FASST *qerb*		7.1	9.0	8.9	5.2	5.7	5.6	3.8	4.3	5.6	2.1	3.3	4.7
Ideal Wiener masks		12.4	14.0	12.8	11.4	13.2	11.9	10.8	12.7	11.3	10.3	12.4	10.8

In relation to the separation results obtained with FASST, it can be seen that only HCE model performs similarly than FASST software and this happens only for high levels of polyphony. The characteristic spectral patterns trained with FASST are not harmonic constrained. The main differences of the FASST separation procedure and the implemented NMF-based procedure are due to the time-frequency resolution and the sinusoidal model. FASST with *qerb* transform has similar results as the HCE model, however FASST with *stft* transform, which does not use a logarithmic frequency resolution, has worse results. Linear resolution in frequency is not appropriate for the separation of musical instruments as results indicate. When using linear resolution, small variations in the fundamental frequency (smaller than a quarter-tone) can produce variations larger than the main lobe of the window at high frequencies. Consequently, the separation capabilities are limited because trained spectral patterns does not model properly the different spectra produced by the same note (we refer to small variations in the pitch for the same note). Other musical source separation approaches [11] also implement logarithmic resolution. Apart from this, we have experienced that harmonic constraint is particularly beneficial to reduce the interferences between sources (see SIR values in Table 1). For harmonic constrained models, the spectral patterns out of the harmonic excitation is set to zero minimizing, in this way, the interferences.

The Multi-Excitation per Instrument (MEI) model for $I = 4$ excitations obtains the best results. In fact, the results are improved as the number of excitation increases from $I = 1$ to $I = 4$. This behavior is explained by the greater capability of modeling when increasing the number of excitations. The HCE model, which is also based on the source-filter paradigm, achieves worse because it does not provide enough degree of freedom to model accurately the note spectra.

The case of the BHC model is particularly interesting. This method is able to model each instrument note independently from the others. Thus, this model has the highest degree of freedom because it is not restricted by the use of the source-filter. However, this model is comparable with the MEI model for $I = 1, 2$ excitations but does not reach the separation results of the MEI model for $I = 4$ excitations. The differences between the databases for training and test can explain this behavior because the conditions of the music scene produce variations in the note spectra. In other words, BHC model is particularly trained to an implementation of each instrument while the MEI model obtains a more flexible representation thanks to the use of the source-filter. Anyway, the BHC model presents a competitive performance, specially for higher levels of polyphony.

Finally, we conclude that the MEI model proposed in [1] achieves the best results mainly by the two following reasons: 1) The use of a set of excitations provides good modeling capabilities in comparison with HCE model. 2) This method is based on a source-filter model, which provides a better tolerance when the conditions of the music scene vary in opposition to the BHC model.

5 Conclusion and Perspectives

In this work, we have demonstrated the viability of a source separation task of multiple instrument mixtures using instrument-dependent NMF models.This kind of models defines the instrument index j in its parameters, which allows the estimation of the magnitude spectra for each instrument $\hat{x}_{t,j}(f)$. Separation Wiener masks can be defined by assigning the proportional quantity of amplitude to each time-frequency cell to the instrument in relation to the total amplitude of all instruments in this cell. Separation metrics indicate that the Multi-Excitation Model proposed in [1] obtains the better results when including a few set of excitations.

In the future, we plan to work in two directions: 1) The frequency resolution of just a semitone is quite limited. We are planning to extend it interpolating the source filter models. 2) In the source filter models, the basis parameters can be updated in testing to adapt the model to the conditions of the music scene.

Acknowledgment. This work was supported by FEDER, the Spanish Ministry of Science and Innovation under Project TEC2009-14414-C03-02, and the University of Jaen under Project R1/12/2010/64.

References

1. Carabias-Orti, J.J., Virtanen, T., Vera-Candeas, P., Ruiz-Reyes, N., Canadas-Quesada, F.J.: Musical Instrument Sound Multi-Excitation Model for Non-Negative Spectrogram Factorization. IEEE Journal on Selected Topics on Signal Processing 5(6), 1144–1158 (2011)
2. Virtanen, T., Klapuri, A.: Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In: Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop (2006)
3. Heittola, T., Klapuri, A., Virtanen, T.: Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In: Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR), Kobe, Japan (2009)
4. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech, and Language Processing 15(3), 1066–1074 (2007)
5. Goto, M.: Development of the RWC Music Database. In: Proc. of the 18th International Congress on Acoustics (ICA 2004), pp.I-553–I-556 (April 2004) (invited paper)
6. Mirex 2007: Music information retrieval evaluation exchange, http://www.music-ir.org/mirex/wiki/2007:MIREX_HOME
7. Vincent, E., Bertin, N., Badeau, R.: Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 528–537 (2010)
8. Ruiz-Reyes, N., Vera-Candeas, P.: Adaptive Signal Modeling Based on Sparse Approximations for Scalable Parametric Audio Coding. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 447–460 (2010)
9. Every, M.R., Szymanski, J.E.: Separation of synchronous pitched notes by spectral filtering of harmonics. IEEE Trans. Audio, Speech, Lang. Process. 14(5), 1845–1856 (2006)
10. Ozerov, A., Vincent, E.: A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio, Speech, Lang. Process (to appear)
11. Vincent, E.: Musical source separation using time-frequency source priors. IEEE Transactions on Audio, Speech, and Language Processing 14(1), 91–98 (2006)

Complex Extension of Infinite Sparse Factor Analysis for Blind Speech Separation

Kohei Nagira, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{knagira,tall,ogata,okuno}@kuis.kyoto-u.ac.jp

Abstract. We present a method of blind source separation (BSS) for speech signals using a complex extension of infinite sparse factor analysis (ISFA) in the frequency domain. Our method is robust against delayed signals that usually occur in real environments, such as reflections, short-time reverberations, and time lags of signals arriving at microphones. ISFA is a conventional non-parametric Bayesian method of BSS, which has only been applied to time domain signals because it can only deal with real signals. Our method uses complex normal distributions to estimate source signals and mixing matrix. Experimental results indicate that our method outperforms the conventional ISFA in the average signal-to-distortion ratio (SDR).

Keywords: Blind source separation, Infinite sparse factor analysis, Non-parametric Bayes.

1 Introduction

Source separation of speech signals is applicable in many areas including distant speech recognition [1,2] and robot audition systems [3,4], and it has therefore been the focus of intensive research in recent years. The signals captured by the systems from microphones in a real environment consist of a mixture of signals from many talkers, and the captured signals are also contaminated by their reflected signals and reverberations. Source separation is applied to such mixtures of acoustic sounds to recognize the speech signals of each talker.

The main requirements for source separation of speech signals are summarized as follows:

1. Source separation without prior information.
2. Simultaneous separation and source activity detection.
3. Robustness against delayed signals.

Source separation without prior information, such as the locations of the sound sources and microphones, is called *blind source separation* (BSS) [5]. Independent component analysis (ICA) [6] is a method that is frequently used for BSS. Frequency domain ICA can achieve BSS in real environments, but ICA assumes that the number of sources equals that of microphones because it cannot detect their source activities.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 388–396, 2012.

Infinite sparse factor analysis (ISFA) [7] is a BSS method based on a nonparametric Bayesian approach. ISFA simultaneously achieves both source separation and source activity detection. However, with conventional ISFA it is difficult to separate delayed signals that include reflections, reverberations, and time lags of signals arriving at microphones.

The objective of our research is to develop a BSS system that fulfills the three above requirements. This paper presents a method of BSS that meets these three requirements by using a complex extension of ISFA.

2 Blind Source Separation Using ISFA

This section specifies the problem addressed in this paper, explains the conventional ISFA, and presents the matter to be solved.

2.1 Problem Settings of BSS

The blind source separation problem this paper deals with is summarized as follows:

Input: Sound mixtures of K sources captured by D microphones.
Output: Estimated K source signals and their activity
Assumption: $K \leq D$, and
Reverberation time is less than window length of
short-time Fourier transform (STFT).

This system captures the mixed signals from K sound sources by D microphones, and separates them into original K sources without using prior information of mixing processes such as the location of sources or the impulse responses.

2.2 BSS for Speech Signals

Mixed signals captured at the microphones are represented as convoluted mixtures of source signals.

$$\overline{\mathbf{x}}(t) = \sum_{j=0}^{J} \overline{\mathbf{A}}(j)\overline{\mathbf{s}}(t-j) \tag{1}$$

where $\overline{\mathbf{x}}(t)$, $\overline{\mathbf{s}}(t)$, and $\overline{\mathbf{A}}(j)$ are observed signals, source signals, and transfer function coefficients, respectively. In general, when a sound is captured by microphones, it takes a little time for the signal to arrive at the microphone, and this time differs depending on the location of the microphone. In addition, the microphones capture the reflections and reverberations of sound sources. Thus, sounds captured by microphones consist of convoluted mixtures of source signals.

When solving a BSS problem involving convoluted mixtures of signals, STFT is often applied in order to convert a convoluted mixture in a time domain into an instantaneous mixture in a frequency domain, and signals are separated for each frequency bin independently.

2.3 Model of ISFA

Infinite sparse factor analysis [7] is a non-parametric Bayesian BSS method. This subsection presents the ISFA model. Let K, D, and N be the number of sources, the number of microphones, and the length of the source signals, respectively. The instantaneous mixture model is expressed as

$$\mathbf{X} = \mathbf{A}(\mathbf{Z} \odot \mathbf{S}) + \mathbf{E}, \tag{2}$$

where $\mathbf{Z} = [\mathbf{z}_1, \cdots \mathbf{z}_N]$, $\mathbf{X} = [\mathbf{x}_1, \cdots \mathbf{x}_N]$, $\mathbf{S} = [\mathbf{s}_1, \cdots \mathbf{s}_N]$, $\mathbf{E} = [\boldsymbol{\varepsilon}_1, \cdots \boldsymbol{\varepsilon}_N]$, $\mathbf{x}_t = [x_{1t}, x_{2t}, \cdots, x_{Dt}]^{\mathrm{T}}$ is a mixed signal vector at time t, $\mathbf{s}_t = [s_{1t}, s_{2t}, \cdots, s_{Kt}]^{\mathrm{T}}$ is the source signal vector, and $\boldsymbol{\varepsilon}_t = [\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Dt}]^{\mathrm{T}}$ is the Gaussian noise vector. Here, $\mathbf{A}$ is the $D \times K$ mixing matrix, $\mathbf{z}_t = [z_{1t}, z_{2t}, \cdots, z_{Kt}]^{\mathrm{T}}$ is the activity of each source at time t, and source activity z_{kt} is a binary variable: $z_{kt} = 1$ if source k is active at time t, otherwise $z_{kt} = 0$. Operator $\odot$ indicates the element-wise product. ISFA can estimate source signals $\mathbf{S}$, their activity $\mathbf{Z}$, mixing matrix $\mathbf{A}$, and other parameters by using only the observed signal $\mathbf{X}$.

2.4 Problems with Conventional Method

Conventional ISFA [7] cannot separate convoluted mixed signals because it can only handle real signals. This means that it cannot be applied to the complex spectra of source signals that are transformed by STFT. This is one of the main problems to be solved in BSS of speech signals because, as we showed in Section 2.1, the mixing process of speech signals can be considered as a convolution of the transfer function.

3 Complex Extension of ISFA

This section presents the complex extension of ISFA. To separate convoluted mixed signals, this algorithm is applied to observed signals for each frequency bin. First, the inference algorithm of our method is given in Table 1. Our method is based on the Metropolis-Hastings algorithm and Gibbs sampling. Posterior distributions of latent variables are derived from Bayes' theorem by multiplying priors by likelihood function. The following part shows priors and posteriors of each parameter and explains the likelihood functions of this model in detail.

3.1 Priors

The prior distributions of variables are assumed as follows:

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}_C(0, \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}),\ \sigma_{\boldsymbol{\varepsilon}}^2 \sim \mathcal{IG}(p_1, p_2), \tag{3}$$

$$s_{kt} \sim \mathcal{N}_C(0, 1), \tag{4}$$

$$\mathbf{a}_k \sim \mathcal{N}_C(0, \sigma_{\mathbf{A}}^2 \mathbf{I}),\ \sigma_{\mathbf{A}}^2 \sim \mathcal{IG}(p_3, p_4), \text{ and} \tag{5}$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha),\ \alpha \sim \mathcal{G}(p_5, p_6). \tag{6}$$

Table 1. Algorithm for estimating model parameters of complex ISFA

```
Input: Observed signals $\mathbf{X}$, Output: Source signals $\mathbf{S}$ and their activity $\mathbf{Z}$
 1. Initialize mixing matrix$\mathbf{A}$, source activity $\mathbf{Z}$, and source signals $\mathbf{S}$ using their priors.
 2. At each time $t$, carry out the following:
    2-1 In each source $k$, sample $z_{kt}$ from Eq. (14).
    2-2 If $z_{kt} = 1$, sample $s_{kt}$ from Eq. (11); otherwise $s_{kt} = 0$.
    2-3 Determine the number of new classes $\kappa_t$, and initialize the parameters.
 3. In each source $k$, sample mixing matrix $\mathbf{a}_k$ from Eq. (16).
 4. If there is a source that is always inactive, remove it.
 5. Update $\sigma_{\boldsymbol{\varepsilon}}^2$, $\sigma_{\mathbf{A}}^2$, and $\alpha$ from Eqs. (17), (18), and (19), respectively.
 6. Go to 2.
```

Here, $\mathbf{a}_k$ is the kth row of $\mathbf{A}$, and p_1, p_2, p_3, p_4, p_5, and p_6 are hyperparameters. The $\mathcal{N}_C(\mu, \sigma^2)$ is the univariate complex normal distribution with mean μ and variance σ^2. The $\mathcal{G}(b, \theta)$ and $\mathcal{IG}(b, \theta)$ are the gamma distribution and the inverse gamma distribution with shape parameter b and scale parameter θ, respectively. The probability density functions of these distributions are

$$\mathcal{N}_C(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right), \tag{7}$$

$$\mathcal{G}(x; b, \theta) = \frac{x^{b-1}}{\Gamma(b)\,\theta^b} \exp\left(-\frac{x}{\theta}\right), \quad \text{and} \tag{8}$$

$$\mathcal{IG}(x; b, \theta) = \frac{x^{-(b-1)}}{\Gamma(b)\,\theta^b} \exp\left(-\frac{1}{\theta x}\right). \tag{9}$$

IBP(α) means Indian buffet process (IBP) [8] with parameter α. IBP is a stochastic process that can deal with a potentially infinite number of signals. IBP can briefly be explained as follows.

1. Time $t = 1$
   Sample the number of sources from the beginning using Poisson(α).
2. Time $t = i$
   – Source k in the existing sources is active in probability $\frac{m_k}{i}$, where m_k is how many times source k is active from $t = 1$ to $i - 1$.
   – After determining whether existing sources are active or not, sample the number of new sources using Poisson($\frac{\alpha}{i}$).

3.2 Likelihood Function

The likelihood function of complex ISFA is written as follows.

$$\begin{aligned} P(\mathbf{X}|\mathbf{A}, \mathbf{S}, \mathbf{Z}) &= \prod_{t=1}^{N} P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{z}_t) = \prod_{t=1}^{N} \mathcal{N}_C(\mathbf{x}_t; \mathbf{A}(\mathbf{z}_t \odot \mathbf{s}_t), \sigma_{\boldsymbol{\varepsilon}}^2\mathbf{I}) \\ &= \frac{1}{(\pi\sigma_{\boldsymbol{\varepsilon}}^2)^{ND}} \exp\left(-\frac{\mathrm{tr}((\mathbf{X} - \mathbf{A}(\mathbf{Z} \odot \mathbf{S}))^{\mathrm{H}}(\mathbf{X} - \mathbf{A}(\mathbf{Z} \odot \mathbf{S})))}{\sigma_{\boldsymbol{\varepsilon}}^2}\right) \end{aligned} \tag{10}$$

Here, each data point is assumed to be independent and identically distributed.

3.3 Posteriors

This part shows inferences of posteriors based on Bayes' theorem.

Sound Sources. When z_{kt} is active, s_{kt} is sampled by the following posterior.

$$P(s_{kt}|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t \mathbf{z}_t) \propto P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_t, \mathbf{z}_t, \sigma_{\boldsymbol{\varepsilon}}^2) P(s_{kt}) = \mathcal{N}_C \left(s_{kt}; \mu_s, \sigma_s^2 \right), \tag{11}$$

where $\sigma_s^2 = \frac{\sigma_{\boldsymbol{\varepsilon}}^2}{\sigma_{\boldsymbol{\varepsilon}}^2 + \mathbf{a}_k^{\mathrm{H}} \mathbf{a}_k}$, $\mu_s = \frac{\mathbf{a}_k^{\mathrm{H}} \varepsilon_{-kt}}{\sigma_{\boldsymbol{\varepsilon}}^2 + \mathbf{a}_k^{\mathrm{H}} \mathbf{a}_k}$. $\mathbf{s}_{-kt}$ means $\mathbf{s}_t$ except for s_{kt}, and $\boldsymbol{\varepsilon}_{-kt}$ means $\boldsymbol{\varepsilon}|_{z_{kt}=0}$.

Source Activity. The ratio of the probability that z_{kt} becomes active to the probability that z_{kt} becomes inactive is calculated by Eq. (12). This ratio r is divided into two parts, the ratio of prior r_p and the ratio of likelihood r_l.

$$\begin{aligned} r &= \frac{P(z_{kt} = 1|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt})}{P(z_{kt} = 0|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt})} \\ &= \frac{P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt}, z_{kt} = 1, \sigma_{\boldsymbol{\varepsilon}}^2)}{P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt}, z_{kt} = 0, \sigma_{\boldsymbol{\varepsilon}}^2)} \frac{P(z_{kt} = 1|\mathbf{z}_{kt})}{P(z_{kt} = 0|\mathbf{z}_{kt})} = r_l r_p. \end{aligned} \tag{12}$$

The ratio of prior r_p is calculated by $r_p = \frac{P(z_{kt}=1|\mathbf{z}_{-kt})}{P(z_{kt}=0|\mathbf{z}_{-kt})} = \frac{m_{k,-t}}{N - m_{k,-t}}$. This is derived from the priors of source activity based on IBP [8].

The ratio of likelihood r_l is calculated by Eq. (13).

$$r_l = \frac{P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt}, z_{kt} = 1, \sigma_{\boldsymbol{\varepsilon}}^2)}{P(\mathbf{x}_t|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt}, z_{kt} = 0, \sigma_{\boldsymbol{\varepsilon}}^2)} = \sigma^2 \exp \left(\frac{|\mu_s|^2}{\sigma_s^2} \right), \tag{13}$$

The posterior probability of $z_{kt} = 1$ is calculated using ratio r.

$$P(z_{kt} = 1|\mathbf{A}, \mathbf{s}_{-kt}, \mathbf{x}_t, \mathbf{z}_{-kt}) = r/(1 + r) \tag{14}$$

To decide whether or not z_{kt} is active, we sample u from Uniform(0,1) and compare it to $r/(1+r)$. If $u \leq r/(1+r)$, z_{kt} becomes active; otherwise it is not.

Number of New Sources. Some source signals that were not active at the beginning are active at time t for the first time. Let κ_t be the number of these sources. This κ_t is sampled with the Metropolis-Hastings algorithm.

First, the prior distribution of κ_t is $P(\kappa_t|\alpha) = \text{Poisson}\left(\frac{\alpha}{N}\right)$. After sampling κ_t, we initialize new sources and their activities. Next, we decide whether this update is accepted or not. The acceptance probability of transition is $\min(1, r_{\xi \to \xi^*})$. According to Meeds [9] and Knowles [7], $r_{\xi \to \xi^*}$ becomes the ratio of the likelihood of the current state to that of the next state. Then, the ratio can be calculated as follows.

$$r_{\xi \to \xi^*} = (\det \Lambda_\xi)^{-1} \exp \left(\mu_\xi^{\mathrm{H}} \Lambda_\xi \mu_\xi \right), \tag{15}$$

where $\Lambda_\xi = \mathbf{I} + \frac{\mathbf{A}^{*\,\mathrm{H}} \mathbf{A}^*}{\sigma_{\boldsymbol{\varepsilon}}^2}$, $\Lambda_\xi \mu_\xi = \frac{1}{\sigma_{\boldsymbol{\varepsilon}}^2} \mathbf{A}^{*\,\mathrm{H}} \boldsymbol{\varepsilon}_t$. Here, $\mathbf{A}^*$ is the $D \times \kappa_t$ matrix of the additional part of $\mathbf{A}$.

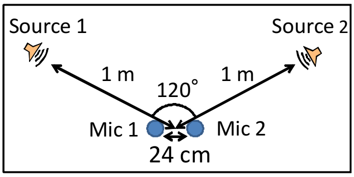


Fig. 1. Locations of microphones and sources

Table 2. Experimental conditions

Number of sources K	2
Number of microphones D	2
Sampling rate	16 [kHz]
STFT window length	64 [msec]
STFT shift length	32 [msec]
Iteration	150 [times]

Mixing matrix. The mixing matrix is estimated in each column. The posterior distribution is

$$P(\mathbf{a}_k|\mathbf{A}_{-k},\mathbf{S},\mathbf{X},\mathbf{Z},\sigma_{\boldsymbol{\varepsilon}}^2,\sigma_{\mathbf{A}}^2) \propto P(\mathbf{X}|\mathbf{A},\mathbf{S},\mathbf{Z},\sigma_{\boldsymbol{\varepsilon}}^2)P(\mathbf{a_k}|\sigma_{\mathbf{A}}^2) = \mathcal{N}_C(\mathbf{a}_k;\mu_{\mathbf{A}},\Lambda_{\mathbf{A}}^{-1}), \tag{16}$$

where $\Lambda_{\mathbf{A}} = \left(\frac{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k}{\sigma_{\boldsymbol{\varepsilon}}^2} + \frac{1}{\sigma_{\mathbf{A}}^2}\right)\mathbf{I}_{D\times D}$, $\mu_{\mathbf{A}} = \frac{\sigma_{\mathbf{A}}^2}{\mathbf{s}_k^{\mathrm{H}}\mathbf{s}_k\sigma_{\mathbf{A}}^2+\sigma_{\boldsymbol{\varepsilon}}^2}\mathbf{E}|_{\mathbf{a}_k=0}\mathbf{s}_k$.

Variance of Noise and Mixing Matrix. The variance of noise corresponds to the noise level of the estimated signals, and the variance of the mixing matrix affects the scale of the estimated signals. Their posteriors are as follows.

$$P(\sigma_{\boldsymbol{\varepsilon}}^2|\mathbf{E}) \propto P(\mathbf{E}|\sigma_{\boldsymbol{\varepsilon}}^2)P(\sigma_{\boldsymbol{\varepsilon}}^2|p_1,p_2) = \mathcal{IG}\left(\sigma_{\boldsymbol{\varepsilon}}^2;p_1+ND,p_2/(1+p_2\,\mathrm{tr}(\mathbf{E}^{\mathrm{H}}\mathbf{E}))\right). \tag{17}$$

$$P(\sigma_{\mathbf{A}}^2|\mathbf{A}) \propto P(\mathbf{A}|\sigma_{\mathbf{A}}^2)P(\sigma_{\mathbf{A}}^2|p_3,p_4) = \mathcal{IG}\left(\sigma_{\mathbf{A}}^2;p_3+DK,p_4/(1+p_4\,\mathrm{tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A}))\right). \tag{18}$$

Parameter of IBP. The posterior distribution of IBP parameter α is

$$p(\alpha|\mathbf{Z}) \propto P(\mathbf{Z}|\alpha)P(\alpha|p_5,p_6) = \mathcal{G}\left(\alpha;K_+ + p_5,p_6/(1+p_6H_N)\right). \tag{19}$$

where K_+ is the active number of sources, and $H_n = \sum_{j-1}^{N}\frac{1}{j}$ is the N-th harmonic number.

3.4 Postprocessing

Just as in frequency domain ICA, our method has problems involving permutation and scaling ambiguity. These problems are caused by a property that prevents our method from determining the amplitude and permutation of output signals at all subbands.

Here, the scaling problem is solved by using the projection back approach [10]. The permutation problem is solved in this paper by using original sources as reference because we want to evaluate the separation performance itself of complex ISFA. Although a solutions to this problem is has been proposed by Sawada *et al.* [11], there has been no exceptional solution to it until now, so this problem is being actively discussed even now.

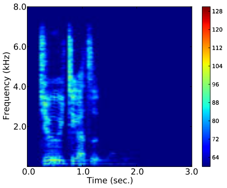


Fig. 2. Spectrogram of source signal

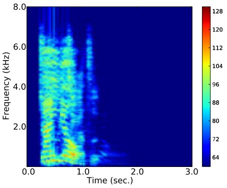


Fig. 3. Spectrogram of mixed signal

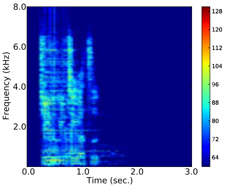


Fig. 4. Separated signal with ours

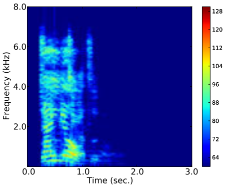


Fig. 5. Baseline separated signal

Table 3. Average separation performance from experimental results [dB]

	Instantaneous			Anechoic chamber			Meeting room		
	Before	Baseline	Proposed	Before	Baseline	Proposed	Before	Baseline	Proposed
SDR	-1.21	**20.54**	2.29	-1.07	-0.87	**2.06**	-2.03	-1.97	**0.55**
ISR	2.37	**26.10**	4.07	1.50	2.64	**3.81**	1.01	1.98	**2.94**
SIR	1.12	**30.34**	10.58	0.94	1.59	**8.84**	1.71	2.36	**4.95**
SAR	75.67	**35.03**	2.83	58.88	**35.89**	2.74	58.68	**36.07**	3.16

4 Experimental Results

We tested our method in a separation experiment using speech signals in order to evaluate the separation performance of our method. In this experiment, our method is compared with the baseline method, real-domain ISFA. We use three kinds of mixed signals for this experiment: instantaneous mixture, convoluted mixture with impulse responses measured in anechoic chamber, and convoluted mixture with impulse responses measured in meeting room ($\mathrm{RT}_{20} = 430$[msec]). Table 2 lists the conditions for this experiment, and Fig. 1 shows the locations of the microphones and sources. We used 214 utterances from ATR phoneme balanced word set.

Figures 2–5 show the spectrograms of a source signal, mixed signal, a signal separated with our method, and a signal separated with the baseline method. We also evaluated our method in terms of the Signal to Distortion Ratio (SDR), the Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR), and the Source to Artifacts Ratio (SAR) [12].

Table 3 summarizes the results. The baseline refers to time-domain ISFA. Our method improves the average SDR by 2.93 dB compared to the baseline on the condition of anechoic chamber, and our method outperformed the baseline on the condition of meeting room. Especially, our method achieves better improvement in SIR than baseline when they are applied to convoluted mixture signals. The result SAR of our method is worse than baseline method, because the separated signals are contaminated by applying STFT and inverse STFT.

5 Conclusion

We presented a method for performing BSS that separates a convoluted mixture of sounds in a real environment, such as reflections, short-time reverberations, and time lags of signals arriving at microphones, with their source activity at the same time. The method was designed by using a non-parametric Bayesian approach. The method separates complex signals in the frequency domain by using a complex extension of ISFA. Our method improves the average SDR by 2.93 dB compared to the baseline based on real-domain ISFA in separating convoluted mixtures in anechoic chamber, and our method also outperforms the conventional ISFA on the condition of meeting room.

In the future, we will evaluate the accuracy of source activity, and we expect to apply source activity to voice activity detection to achieve better speech recognition. Last but not least, the method should be sped up to attain real-time processing so that it can be applied to robot applications.

Acknowledgement. This research was partially supported by Grant-in-Aid for Scientific Research (S), the Global COE Program, and HRI-JP.

References

1. Wölfel, M., McDonough, J.: Distant Speech Recognition. Wiley (2009)
2. Seltzer, M.L., Raj, B., Stern, R.M.: Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans. on Speech and Audio Processing 12(5), 489–498 (2004)
3. Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., Tsujino, H.: Design and Implementation of Robot Audition System "HARK" Open Source Software for Listening to Three Simultaneous Speakers. Advanced Robotics 24(5–6), 739–761 (2010)
4. Valin, J.M., Rouat, J., Michaud, F.: Enhanced robot audition based on microphone array source separation with post-filter. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2004, vol. 3, pp. 2123–2128. IEEE (2004)
5. Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E.: A blind source separation technique using second-order statistics. IEEE Transactions on Signal Processing 45(2), 434–444 (1997)
6. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis. Wiley Interscience (2001)
7. Knowles, D., Ghahramani, Z.: Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 381–388. Springer, Heidelberg (2007)
8. Griffiths, T., Ghahramani, Z.: Infinite latent feature models and the Indian buffet process. Advances in Neural Information Processing Systems 18, 475–482 (2006)
9. Meeds, E., Ghahramani, Z., Neal, R.M., Roweis, S.T.: Modeling dyadic data with binary latent factors. Advances in Neural Information Processing Systems 19, 977–984 (2007)

10. Murata, N., Ikeda, S., Ziehe, A.: An approach to blind source separation based on temporal structure of speech signals. Neurocomputing 41(1-4), 1–24 (2001)
11. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. on Speech and Audio Processing 12(5), 530–538 (2004)
12. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)

A General Framework for Online Audio Source Separation

Laurent S.R. Simon and Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
{laurent.s.simon,emmanuel.vincent}@inria.fr

Abstract. We consider the problem of online audio source separation. Existing algorithms adopt either a sliding block approach or a stochastic gradient approach, which is faster but less accurate. Also, they rely either on spatial cues or on spectral cues and cannot separate certain mixtures. In this paper, we design a general online audio source separation framework that combines both approaches and both types of cues. The model parameters are estimated in the Maximum Likelihood (ML) sense using a Generalised Expectation Maximisation (GEM) algorithm with multiplicative updates. The separation performance is evaluated as a function of the block size and the step size and compared to that of an offline algorithm.

Keywords: Online audio source separation, nonnegative matrix factorisation, sliding block, stochastic gradient.

1 Introduction

Audio source separation is the process of recovering a set of audio signals from a given mixture signal. This can be addressed via established approaches such as Independent Component Analysis (ICA), binary masking and Sparse Component Analysis (SCA) [1] or more recent approaches such as local Gaussian modeling and Nonnegative Matrix Factorisation (NMF) [2]. Most current algorithms are offline algorithms which require the whole signal in order to estimate the sources. In this paper, we focus on online audio source separation, whereby only the past samples of the mixture are available. This constraint arises in particular in real-time scenarios.

A few online implementations have been designed for ICA [3] [4], time-frequency masking [5], local Gaussian modeling [6], spectral continuity-based separation [7] and NMF [8]. However, these algorithms rely either on spatial cues [3] – [6] or on spectral cues [7,8] alone. Such algorithms are not capable of separating mixtures where several sources have the same spatial position and several sources have similar spectral characteristics. For example, in pop music, the voice, the snare drum, the bass drum and the bass are often mixed to the centre and several voices or several guitars are present.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 397–404, 2012.

In order to address this issue, we consider the general flexible source separation framework in [9]. This framework generalises a wide range of algorithms such as certain forms of ICA, local Gaussian modeling and NMF, and enables the specification of additional constraints on the sources such as harmonicity. By jointly exploiting spatial and spectral cues, it makes it possible to robustly separate difficult mixtures such as above.

The two main approaches for online source separation are the sliding block (also known as blockwise) approach, as used in [3] [4] [5] [7], and the stochastic gradient (also known as stepwise) approach, as used in [6] [8]. The sliding block method consists in applying the offline audio source separation algorithm to a block of M time frames. Once this block of signal has been processed, a frame is extracted for each of the J sources before sliding the processing block by one frame. This approach is computationally intensive but accurate. The stepwise method offers to update the model parameters in every frame using only the latest available frame and the model parameters estimated in the previous frame. As it uses only the latest available frame at a given time, this approach is faster than the sliding block approach but can be inaccurate.

In this paper, we propose a general iterative online algorithm for the source separation framework in [9] that combines the sliding block approach and the stepwise approach using two hyper-parameters: the block size M and the step size α. As a by-product, we provide a way of circumventing the annealing procedure in [9], which would require a large number of iterations per block. Moreover, we determine the best trade-off between these two approaches experimentally on a set of real-world music mixtures.

The structure of the rest of the paper is as follows: the flexible framework in [9] is introduced in Section 2. Section 3 presents the online algorithm. Experimental results are shown in Section 4. The conclusion can be found in Section 5.

2 General Audio Source Separation Framework

We operate in the time-frequency (TF) domain by means of the Short-Time Fourier Transform (STFT). In each frequency bin f and each time frame n, the multichannel mixture signal $\mathbf{x}(f,n)$ can be expressed as

$$\mathbf{x}(f,n) = \sum_{j=1}^{J} \mathbf{c}_j(f,n) \tag{1}$$

where J is the number of sources and $\mathbf{c}_j(f,n)$ is the STFT of the spatial image of the j-th source.

2.1 Model

We assume that $\mathbf{c}_j(f,n)$ is a complex-valued Gaussian random vector with zero mean and covariance matrix $\mathbf{R}_{\mathbf{c}_j}(f,n)$

$$\mathbf{c}_j \sim \mathcal{N}_c(\mathbf{0}, \mathbf{R}_{\mathbf{c}_j}) \tag{2}$$

and that $\mathbf{R}_{\mathbf{c}_j}(f,n)$ factors as

$$\mathbf{R}_{\mathbf{c}_j}(f,n) = \mathbf{R}_j(f) v_j(f,n) \tag{3}$$

where $\mathbf{R}_j(f)$ is the spatial covariance matrix of the j-th source and $v_j(f,n)$ is its spectral variance.

In [9], $\mathbf{R}_j(f)$ is expressed as $\mathbf{R}_j(f) = \mathbf{A}_j(f)\mathbf{A}_j^H(f)$, and $\mathbf{A}_j(f)$ is estimated instead. This results in an annealing procedure, which would translate into a large number of iterations within each block in our context. In order to circumvent the annealing, we assume that $\mathbf{R}_j(f)$ is full-rank and directly estimate $\mathbf{R}_j(f)$ instead, similarly to [10].

The spectral variance $v_j(f,n)$ is modeled via a form of hierarchical NMF [9]. The matrix of spectral variances $\mathbf{V}_j \triangleq [v_j(f,n)]_{f,n}$ is first decomposed into the product of an excitation spectral power $\mathbf{V}_j^{\mathrm{x}}$ and a filter spectral power $\mathbf{V}_j^{\mathrm{f}}$

$$\mathbf{V}_j = \mathbf{V}_j^{\mathrm{x}} \odot \mathbf{V}_j^{\mathrm{f}} \tag{4}$$

where $\odot$ denotes entrywise multiplication. $\mathbf{V}_j^{\mathrm{x}}$ is further decomposed into the product of a matrix of narrowband spectral patterns $\mathbf{W}_j^{\mathrm{x}}$, a matrix of spectral envelope weights $\mathbf{U}_j^{\mathrm{x}}$, a matrix of temporal envelope weights $\mathbf{G}_j^{\mathrm{x}}$ and a matrix of time-localised temporal patterns $\mathbf{H}_j^{\mathrm{x}}$, so that

$$\mathbf{V}_j^{\mathrm{x}} = \mathbf{W}_j^{\mathrm{x}}\mathbf{U}_j^{\mathrm{x}}\mathbf{G}_j^{\mathrm{x}}\mathbf{H}_j^{\mathrm{x}}. \tag{5}$$

$\mathbf{V}_j^{\mathrm{f}}$ is decomposed in a similar way.

This factorisation enables the specification of various spectral or temporal constraints over the sources. For example, harmonicity can be enforced by fixing $\mathbf{W}_j^{\mathrm{x}}$ to a set of narrowband harmonic patterns.

2.2 Offline EM-MU Algorithm

In an offline context, the model parameters are estimated in the Maximum Likelihood (ML) sense by a Generalised Expectation-Maximisation (GEM) algorithm combined with Multiplicative Updates (MU) applied to the complete data $\{\mathbf{c}_j(f,n)\}$.

The log-likelihood is defined using the empirical mixture covariance matrix $\widehat{\mathbf{R}}_{\mathbf{x}}(f,n)$ [10] as

$$\log \mathcal{L} = \sum_{f,n} - \operatorname{tr}\left(\mathbf{R}_{\mathbf{x}}^{-1}(f,n)\widehat{\mathbf{R}}_{\mathbf{x}}(f,n)\right) - \log\det(\pi \mathbf{R}_{\mathbf{x}}(f,n)) \tag{6}$$

where

$$\mathbf{R}_{\mathbf{x}}(f,n) = \sum_{j=1}^{J} \mathbf{R}_{\mathbf{c}_j}(f,n) \tag{7}$$

is the covariance of the mixture $\mathbf{x}(f,n)$.

In the E-step, the expectation of the natural statistics is computed via [10]

$$\boldsymbol{\Omega}_j(f,n) = \mathbf{R}_{\mathbf{c}_j}(f,n)\mathbf{R}_{\mathbf{x}}^{\mathbf{-1}}(f,n) \tag{8}$$

$$\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n) = \boldsymbol{\Omega}_j(f,n)\widehat{\mathbf{R}}_{\mathbf{x}}(f,n)\boldsymbol{\Omega}_j^H(f,n) + (\mathbf{I} - \mathbf{W}_j(f,n))\mathbf{R}_{\mathbf{c}_j}(f,n) \tag{9}$$

where $\boldsymbol{\Omega}_j$ is the Wiener filter, $\mathbf{I}$ is the $I \times I$ identity matrix and I is the number of channels of the mixture.

In the M-step, the model parameters are updated as [9,10]

$$\mathbf{R}_j(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_j(f,n)}\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n) \tag{10}$$

$$\mathbf{W}_j^{\mathrm{x}} = \mathbf{W}_j^{\mathrm{x}} \odot \frac{[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}](\mathbf{U}_j^{\mathrm{x}}\mathbf{G}_j^{\mathrm{x}}\mathbf{H}_j^{\mathrm{x}})^T}{\mathbf{V}_j^{\mathrm{x}}.^{-1}(\mathbf{U}_j^{\mathrm{x}}\mathbf{G}_j^{\mathrm{x}}\mathbf{H}_j^{\mathrm{x}})^T} \tag{11}$$

$$\mathbf{U}_j^{\mathrm{x}} = \mathbf{U}_j^{\mathrm{x}} \odot \frac{\mathbf{W}_j^{\mathrm{x}T}[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}](\mathbf{G}_j^{\mathrm{x}}\mathbf{H}_j^{\mathrm{x}})^T}{\mathbf{W}_j^{\mathrm{x}T}\mathbf{V}_j^{\mathrm{x}}.^{-1}(\mathbf{G}_j^{\mathrm{x}}\mathbf{H}_j^{\mathrm{x}})^T} \tag{12}$$

$$\mathbf{G}_j^{\mathrm{x}} = \mathbf{G}_j^{\mathrm{x}} \odot \frac{(\mathbf{W}_j^{\mathrm{x}}\mathbf{U}_j^{\mathrm{x}})^T[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}]\mathbf{H}_j^{\mathrm{x}T}}{(\mathbf{W}_j^{\mathrm{x}}\mathbf{U}_j^{\mathrm{x}})^T\mathbf{V}_j^{\mathrm{x}}.^{-1}\mathbf{H}_j^{\mathrm{x}T}} \tag{13}$$

$$\mathbf{H}_j^{\mathrm{x}} = \mathbf{H}_j^{\mathrm{x}} \odot \frac{(\mathbf{W}_j^{\mathrm{x}}\mathbf{U}_j^{\mathrm{x}}\mathbf{G}_j^{\mathrm{x}})^T[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}]}{(\mathbf{W}_j^{\mathrm{x}}\mathbf{U}_j^{\mathrm{x}}\mathbf{G}_j^{\mathrm{x}})^T\mathbf{V}_j^{\mathrm{x}}.^{-1}} \tag{14}$$

where $.^p$ denotes entrywise raising to the power p, N is the number of time frames in the STFT of the signal and $\widehat{\boldsymbol{\Xi}}_j = [\widehat{\xi}_j(f,n)]_{f,n}$, with

$$\widehat{\xi}_j(f,n) = \frac{1}{I}\operatorname{tr}(\mathbf{R}_j^{-1}(f)\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)). \tag{15}$$

$\mathbf{W}_j^{\mathrm{f}}$, $\mathbf{U}_j^{\mathrm{f}}$, $\mathbf{G}_j^{\mathrm{f}}$ and $\mathbf{H}_j^{\mathrm{f}}$ are updated in a similar way.

After each EM iteration, the model parameters are normalised: the mean of $\mathbf{R}_j$, $\mathbf{W}_j^{\mathrm{x}}$, $\mathbf{U}_j^{\mathrm{x}}$, $\mathbf{G}_j^{\mathrm{x}}$, $\mathbf{H}_j^{\mathrm{x}}$, $\mathbf{W}_j^{\mathrm{f}}$, $\mathbf{U}_j^{\mathrm{f}}$ and $\mathbf{H}_j^{\mathrm{f}}$ are normalised to 1 while $\mathbf{G}_j^{\mathrm{f}}$ is multiplied by the product of the normalisation factors of the other variables.

The separated sources are then obtained via

$$\widehat{\mathbf{c}}_j(f,n) = \boldsymbol{\Omega}_j(f,n)\mathbf{x}(f,n). \tag{16}$$

3 Online EM-MU Algorithm

We now consider an online context where in each time frame t, the data is limited to a block of M STFT frames indexed by n with $t - M + 1 \leq n \leq t$, where $M = 1$ for the stepwise approach and $M = N$ for the full offline approach. We define a step size coefficient $\alpha \in]0;1]$ to stabilise the parameter updates by averaging over time. For each block, the spatial covariance matrices $\mathbf{R}_j^{(t)}(f)$ are initialised to a diffuse spatial covariance spanning a part of the audio space. The temporal weights $\mathbf{G}_j^{\mathrm{x}(t)}$ are randomly initialised and the normalised to the mean spectral power of the signal. Finally, the temporal patterns $\mathbf{H}_j^{\mathrm{x}(t)}$ are initialised to diagonal matrices. The expectation of the natural statistics is computed using (8) and (9) for $t-M+1 \leq n \leq t$, whilst the spatial covariance matrix is updated as follows:

$$\mathbf{R}_j^{(t)}(f) = (1-\alpha)\mathbf{R}_j^{(t-1)}(f) + \alpha\left(\frac{1}{M}\sum_{n=t-M+1}^{t}\frac{1}{v_j(f,n)}\widehat{\mathbf{R}}_{\mathbf{c}_j}(f,n)\right) \tag{17}$$

where the superscript $^{(t)}$ denotes is the value of matrix for the block t.

$\mathbf{G}_j^{\mathrm{x}(t)}$ and $\mathbf{H}_j^{\mathrm{x}(t)}$ are updated using (13) and (14) for $t-M+1 \leq n \leq t$, as they are expected to significantly vary between blocks, whereas the updates of $\mathbf{W}_j^{\mathrm{x}}$ and $\mathbf{U}_j^{\mathrm{x}}$ become

$$\mathbf{W}_j^{\mathrm{x}(t)} = \mathbf{W}_j^{\mathrm{x}(t)} \odot \frac{\mathbf{M}_j^{\mathrm{x}(t)}}{\mathbf{C}_j^{\mathrm{x}(t)}} \tag{18}$$

$$\mathbf{U}_j^{\mathrm{x}(t)} = \mathbf{U}_j^{\mathrm{x}(t)} \odot \frac{\mathbf{N}_j^{\mathrm{x}(t)}}{\mathbf{D}_j^{\mathrm{x}(t)}} \tag{19}$$

where

$$\begin{align} \mathbf{M}_j^{\mathrm{x}(t)} &= (1-\alpha)\mathbf{M}_j^{\mathrm{x}(t-1)} + \alpha[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}](\mathbf{U}_j^{\mathrm{x}(t)}\mathbf{G}_j^{\mathrm{x}(t)}\mathbf{H}_j^{\mathrm{x}(t)})^T \tag{20}\\ \mathbf{C}_j^{\mathrm{x}(t)} &= (1-\alpha)\mathbf{C}_j^{\mathrm{x}(t-1)} + \alpha\mathbf{V}_j^{\mathrm{x}}.^{-1}(\mathbf{U}_j^{\mathrm{x}(t)}\mathbf{G}_j^{\mathrm{x}(t)}\mathbf{H}_j^{\mathrm{x}(t)})^T \tag{21}\\ \mathbf{N}_j^{\mathrm{x}(t)} &= (1-\alpha)\mathbf{N}_j^{\mathrm{x}(t-1)} + \alpha\mathbf{W}_j^{\mathrm{x}(t)T}[\widehat{\boldsymbol{\Xi}}_j \odot \mathbf{V}_j^{\mathrm{x}}.^{-2} \odot \mathbf{V}_j^{\mathrm{f}}.^{-1}](\mathbf{G}_j^{\mathrm{x}(t)}\mathbf{H}_j^{\mathrm{x}(t)})^T \tag{22}\\ \mathbf{D}_j^{\mathrm{x}(t)} &= (1-\alpha)\mathbf{D}_j^{\mathrm{x}(t-1)} + \alpha\mathbf{W}_j^{\mathrm{x}(t)T}\mathbf{V}_j^{\mathrm{x}}.^{-1}(\mathbf{G}_j^{\mathrm{x}(t)}\mathbf{H}_j^{\mathrm{x}(t)})^T \tag{23} \end{align}$$

where $\widehat{\boldsymbol{\Xi}}_j^{(t)}$ is computed as in (16). $\mathbf{M}_j^{\mathrm{f}\,(t)}$, $\mathbf{C}_j^{\mathrm{f}\,(t)}$, $\mathbf{N}_j^{\mathrm{f}\,(t)}$ and $\mathbf{D}_j^{\mathrm{f}\,(t)}$ are updated in a similar way. At each block, several iterations can be performed in order to improve the estimation of the model parameters.

Although equations (17) to (19) look similar to the online update of the local Gaussian model in [6] and [8], there are two crucial differences:

- The framework introduced in the current paper is more general in the sense that it uses hierarchical NMF, enabling the user to apply more specific constraints than when using shallow NMF.
- It is not limited to the sole use of the latest audio frame.

4 Experimental Results

We compared the performance of the online audio source separation framework to the offline framework introduced in section 2.2, as a function of the number of EM iterations, α and M. The project aiming at remixing of recordings for sound engineers, DJs and consumers, we processed five 10 s long stereo commercial pop recordings composed of bass, drums, guitars, strings and voice. All the recordings were recorded at 44100 Hz. The STFT was computed using half-overlapping 2048 sample sine windows. In the offline algorithm as well as in the online algorithm, each of the modeled sources were constrained in a way similar to section V.C in [9]. In the case of an harmonic source, $\mathbf{W}_j^{\mathrm{x}(t)}$ was fixed to a set of narrowband harmonic spectral patterns and the spectral envelope weights in $\mathbf{U}_j^{\mathrm{x}(t)}$ were updated, whereas for bass and percussive sources, $\mathbf{W}_j^{\mathrm{x}(t)}$ was a fixed diagonal matrix and $\mathbf{U}_j^{\mathrm{x}(t)}$ was a fixed matrix of basis spectra learned over a corpus of bass and drum sounds.

Audio samples of the separated sounds of this experiment can be found on http://www.irisa.fr/metiss/lssimon/LVA2012/index.html .

Separation performance was evaluated using the Signal-to-Distortion Ratio (SDR), the Signal-to-Interference Ratio (SIR), the source Image to Spatial distortion Ratio (ISR) and the Source-to-Artifacts Ratio (SAR) defined in [11]. For each set of conditions over the number of iterations, M and α, each of these criteria was averaged over all the mixtures and all the separated sound sources. Over all the results of this experiment, the SDR varied between -1.1 and 0.9 dB, the SIR between -4 and 1 dB, the ISR between 2.3 and 3.9 dB and the SAR between 10 and 19 dB.

Table 1. Separation performance (dB) of the offline and best online algorithms

Algorithm	α	M	number of iterations	SDR	SIR	ISR	SAR
offline	N/A	N/A	100	0.8586	1.2837	3.7989	13.3872
online	1	50	30	**0.8671**	1.0675	**3.9690**	12.3278

As shown in table 1, when $\alpha = 1$, $M = 50$ and 30 GEM iterations are performed, the separation performance of the online algorithm is close to that of the offline algorithm. For smaller block size and smaller number of iterations, the performance decreases. For example, for $M = 10$ and 6 GEM iteration, the SDR is 0.53 dB and the SIR is 3.53 dB. More generally, fig. 1 shows that for $\alpha = 1$, increasing either the block size or the number of iterations increases the SDR, though the block size has less effect on the SDR than the number of iterations. The results also show that increasing the number of iterations from 10 to 30 increases the SDR by 0.2 dB, which can be considered as a significant improvement.

When $\alpha < 1$, the SDR decreases significantly as can be seen in fig. 1. It can also be seen that increasing the number of iterations decreases the SDR and changes of block size have little to no effect on the SDR. This can be explained by an inaccurate estimation of the model parameters of certain sources in the time intervals when these sources are inactive. These inaccurate parameters are then carried over subsequent time frames and may not converge back to accurate values. This undesirable effect is particularly salient for those parameters that are less constrained. For instance, with the considered model, the spatial covariance matrices of all sources gradually diverge towards a diffuse spatial covariance spanning all directions in the mixture, while the effect is more limited for spectral parameters which are fixed or heavily constrained. Potential solutions to this problem are presented in the conclusion.

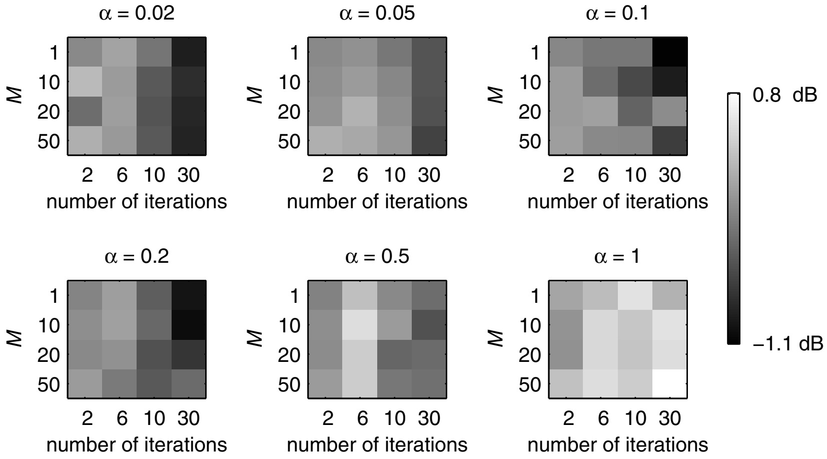


Fig. 1. Mean SDR for all sources and all mixtures, as a function of α, M and step size

5 Conclusion

In this paper, a new framework for online audio source separation was presented. This algorithm offers an increased flexibility both in terms of the range of constraints that can be specified for each source and of the choice of a trade-off between separation accuracy and computational cost. It was shown that the separation accuracy is higher when the block size is large, but that small block sizes nevertheless offer an acceptable separation. However, small step sizes cause the spatial covariance matrices to diverge due to the presence of silence intervals in the sources.

This issue is well-known in the beamforming literature where a voice activity detector is used to restrict the time frames in which the model parameters are updated [12]. While this solution does not readily extend to source separation, we believe that there exist a number of alternative promising solutions, e.g. adding soft constraints over the least constrained parameters by means of probabilistic priors, using different step sizes for the most constrained and the least constrained parameters, and using signal-dependent step sizes related to the power of $\mathbf{R}_{\mathbf{c}_j}(f, n)$ such that the parameters are not updated in the time intervals with low power.

Future work should also include an optimisation of the initialisation of the model parameters for each new block. After these improvements, we expect that the proposed framework will reach its full potential and provide a better trade-off between separation performance and computational cost.

Acknowledgements. This work was supported by the EUREKA Eurostars i3DMusic project funded by Oseo.

References

1. Makino, S., Lee, T.-W., Sawada, H.: Blind Speech Separation. Springer, Heidelberg (2007)
2. Vincent, E., Jafari, M.G., Abdallah, S.A., Plumbley, M.D., Davies, M.E.: Probabilistic modeling paradigms for audio source separation. In: Machine Audition: Principles, Algorithms and Systems, pp. 162–185. IGI Global (2010)
3. Mukai, R., Sawada, H., Araki, S., Makino, S.: Real-time Blind Source Separation For Moving Speakers Using Blockwise ICA and Residual Crosstalk Subtraction. In: 4th Int. Symp. Independent Component Analysis and Blind Signal Separation, pp. 975–980 (2003)
4. Mori, Y., Saruwatari, H., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., Ikeda, Y., Hashimoto, H., Morita, T.: Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking. EURASIP Journal on Advances in Signal Processing 2006(1), 1–17 (2006)
5. Loesch, B., Yang, B.: Online blind source separation based on time-frequency sparseness. In: Proc. 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 117–120 (2009)
6. Togami, M.: Online speech source separation based on maximum likelihood of local Gaussian modeling. In: Proc. 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 213–216 (2011)
7. Ono, N., Miyamoto, K., Sagayama, S.: A real-time equalizer of harmonic and percussive components in music signals. In: Proc. 2008 Int. Conf. on Music Information Retrieval, pp. 139–144 (2008)
8. Wang, D., Vipperla, R., Evans, N.: Online pattern learning for non-negative convolutive sparse coding. In: Proc. Interspeech 2011, pp. 65–68 (2011)
9. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. IEEE Transactions on Audio, Speech, and Language Processing (to appear)
10. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Transactions on Audio, Speech, and Language Processing 18(7), 1830–1840 (2010)
11. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.P.: First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)
12. Brandstein, M.S., Ward, D.B.: Microphone Arrays: Signal Processing Techniques and Applications. Springer, Heidelberg (2001)

Sound Recognition in Mixtures

Juhan Nam[1,⋆], Gautham J. Mysore[2], and Paris Smaragdis[2,3]

[1] Center for Computer Research in Music and Acoustics, Stanford University
[2] Advanced Technology Labs, Adobe Systems Inc.
[3] University of Illinois at Urbana-Champaign

Abstract. In this paper, we describe a method for recognizing sound sources in a mixture. While many audio-based content analysis methods focus on detecting or classifying target sounds in a discriminative manner, we approach this as a regression problem, in which we estimate the relative proportions of sound sources in the given mixture. Using source separation ideas based on probabilistic latent component analysis, we directly estimate these proportions from the mixture without actually separating the sources. We also introduce a method for learning a transition matrix to temporally constrain the problem. We demonstrate the proposed method on a mixture of five classes of sounds and show that it is quite effective in correctly estimating the relative proportions of the sounds in the mixture.

1 Introduction

Nowadays, a huge volume of multimedia content is available and is rapidly increasing over broadband networks. While the content is usually managed or searched using manually annotated text or collaborative information from users, there has been increasing efforts to automatically analyze the content and find relevant information. In particular, some researchers have tried to analyze the content by recognizing sounds in the video because information in the audio domain is crucial for certain tasks, such as sports highlight detection and event detection in surveillance systems [1] and also audio data is generally more efficient to process due to its relatively low bandwidth compared to video data.

The majority of audio-based content analysis methods focus on detecting a target source or classifying sound classes in a discriminative manner [2,3]. Although they are successful in some detection or classification tasks, such discriminative approaches have a limitation in that most real-world sounds are mixtures of multiple sources. It is therefore useful to be able to simultaneously model multiple sources for various applications such as searching for certain scenes in a film soundtrack. For example, if we want to search for a scene with a specific actor in which a car is passing by and background music is present, it would be useful to model each of these sources.

In this paper, we propose a generative approach, which models a mixture sound as multiple single sources and estimates the relative proportion of each

[⋆] This work was performed while interning at Adobe Systems Inc.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 405–413, 2012.

source. Our method is based on probabilistic latent component analysis (PLCA) [4], which is a variant of non-negative matrix factorization (NMF). PLCA has been widely used as a way of modeling sounds in the spectral domain because of the interpretable decomposition and extensible capability as a probabilistic model. We first formalize our problem using a PLCA-based approach and then we propose an improved model which takes temporal characteristics of each source into account. Lastly, we evaluate our method with a dataset and discuss the results.

2 Proposed Method

The basic methodology that we follow is that of supervised source separation using PLCA [5]. For each source, we estimate a dictionary of basis elements from isolated training data of that source. Then, given a mixture, we estimate a set of mixture weights. Using these weights, it is possible to separate the sources (typical PLCA-based supervised source separation). However, without actually separating the sources, we estimate the relative proportion of each source in the mixture. Since we bypass the actual separation process, we can do certain things to improve sound recognition performance even when it does not improve source separation performance. Specifically, we choose the dictionary sizes based on sound recognition performance. Also, we impose a temporal continuity constraint that helps this performance but could introduce fairly heavy artifacts if we were to actually separate the sources. Note that we refer to a source as a general class of sounds, such as speech, music and other environmental sounds in this paper.

2.1 Basic Model

PLCA is an additive latent variable model that is used to decompose audio spectrograms [4]. An asymmetric version of PLCA models each time frame of a spectrogram as a linear combination of dictionary elements as follows:

$$X(f,t) \approx \gamma \sum_{z} P(f|z)P_t(z) \tag{1}$$

where $X(f,t)$ is the audio spectrogram, z is a latent variable, each $P(f|z)$ is a dictionary element, $P_t(z)$ is a distribution of weights at time frame t, and γ is a constant scaling factor. All distributions are discrete. Given $X(f,t)$, we can estimate the parameters of $P(f|z)$ and $P_t(z)$ using the EM algorithm.

We model single sound sources and their mixtures using PLCA. We first compute the spectrogram $X_s(f,t)$ given isolated training data of source s. We then use Eq. 1 to estimate a set of dictionary elements and weights that correspond to that source. In the basic model, we assume that a single source is characterized by the dictionary elements. Therefore, we retain the dictionary elements while discarding the weights. Using the dictionary elements from each single source, we build a larger dictionary to represent a mixture spectrogram. This is formed by simply concatenating the dictionaries of the individual sources. Thus, if we have a spectrogram $X_M(f,t)$ that is a mixture of two sources, we model it as follows[1]:

$$X_M(f,t) \approx \gamma \sum_{z \in \{\mathbf{z_{s_1}}, \mathbf{z_{s_2}}\}} P(f|z) P_t(z) \tag{2}$$

where $\mathbf{z_{s_1}}$ and $\mathbf{z_{s_2}}$ represent the dictionary elements that belong to source 1 and source 2 respectively. Since the dictionary elements of both sources are already known, we keep them fixed and simply estimate the weights $P_t(z)$ at each time frame using the EM algorithm. The weights tell us the relative proportion of each dictionary element in the mixture. It is therefore intuitive that the sum of the weights that correspond to a given source, will give us the proportion of that source present in the mixture. Accordingly, we compute the relative proportions of the sources at each time frame by simply summing the corresponding weights as follows:

$$r_t(s_1) = \sum_{z \in \mathbf{z_{s_1}}} P_t(z) \tag{3}$$

$$r_t(s_2) = \sum_{z \in \mathbf{z_{s_2}}} P_t(z) \tag{4}$$

2.2 Modeling Temporal Dependencies

When we learn a model for a single source from isolated training data of that source, we obtain a dictionary of basis elements and a set of weights. In the previous subsection, we discarded the weights as they simply tell us how to fit the dictionary to that specific instance of training data. This is usually the practice when performing NMF or PLCA based supervised source separation [5].

Although the weights are specific to the training data, they do contain certain information that is more generally applicable. One such piece of information is temporal dependencies amongst dictionary elements. For example, if a dictionary element is quite active in one time frame, it is usually likely to be quite active in the following time frame as well. However, there are usually more such dependencies present such as things like a high presence of dictionary element m in time frame t usually followed by a high presence of dictionary element n in time frame $t+1$. Using the weights of adjacent time frames, we can infer this information. For time frames t and $t+1$ of source s, we can compute this dependency as follows:

$$\phi_s(z_t, z_{t+1}) = P(z_t) P(z_{t+1}), \ \forall z \in \mathbf{z_s}. \tag{5}$$

This gives us the affinity of every dictionary element to every other dictionary element in two adjacent time frames. If we average this value over all time frames

[1] It is straightforward to extend this to more sources.

and normalize, we obtain a set of conditional probability distributions that serve as a transition matrix as follows:

$$P_s(z_{t+1}|z_t) = \frac{\sum_{t=1}^{T-1} \phi_s(z_t, z_{t+1})}{\sum_{z_{t+1}} \sum_{t=1}^{T-1} \phi_s(z_t, z_{t+1})}. \tag{6}$$

When we learn dictionaries from isolated training data, we can compute such a transition matrix for each source. As a result, our model for each source consists of a dictionary and a transition matrix.

Given a mixture, our method of estimating weights should be accordingly changed to make use of the transition matrix. First, we should have a joint transition matrix $P(z_{t+1}|z_t)$ that corresponds to the concatenated dictionaries. Since we assume that the activity of the dictionary elements in one dictionary are independent of those in other dictionaries, we construct the joint transition matrix by diagonalizing individual transition matrices. For example, if we have two sound sources and two corresponding transition matrices $T1$ and $T2$, the joint transition matrix is formed as:

$$T = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}. \tag{7}$$

Once we obtain the concatenated dictionary and transition matrix, we move on to the actual sound recognition stage. Given the mixture, we first estimate the weights $P_t(z)$ as described in the previous subsection. We call this our initial weights estimate $P_t^{(i)}(z)$. Using these estimates, we obtain a new estimate of the weights that is more consistent with the dependencies that are implied by the joint transition matrix[2]. We do this by first computing re-weighting terms in the forward and backward directions to impose the joint transition matrix in both directions:

$$F_{t+1}(z) = \sum_{z_t} P(z_{t+1}|z_t) P_t^{(i)}(z). \tag{8}$$

$$B_t(z) = \sum_{z_{t+1}} P(z_{t+1}|z_t) P_{t+1}^{(i)}(z). \tag{9}$$

Using the above terms, we perform the re-weighting and normalize as follows to get our final estimate of the weights:

$$P_t(z) = \frac{P_t^{(i)}(z)\,(C + F_t(z) + B_t(z))}{\sum_z P_t^{(i)}(z)\,(C + F_t(z) + B_t(z))}, \tag{10}$$

where C is a parameter that controls the influence of the joint transition matrix. As C tends to infinity, the effect of the forward and backward re-weighting terms becomes negligible, whereas as C tends to 0 we tend to modulate the estimated

[2] This is analogous to smoothing an estimated time series with a moving average filter if we believe that the time series is slowly varying.

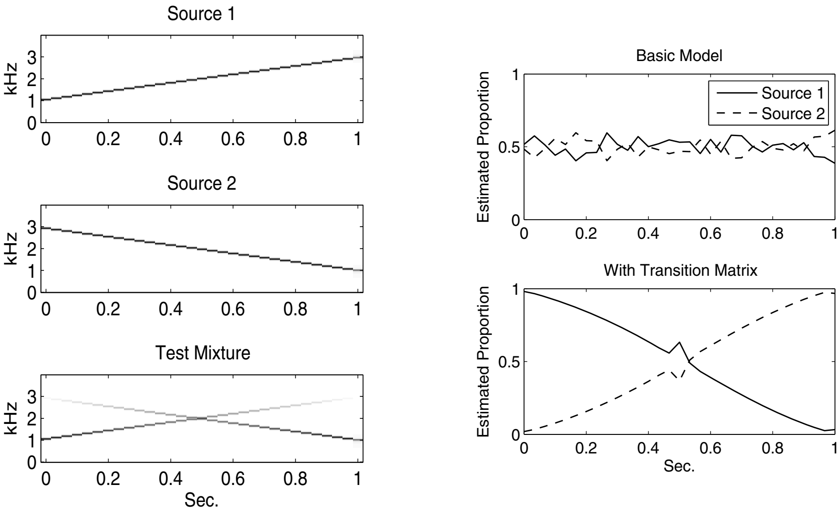


Fig. 1. A toy example: training sources are given as chirps that have frequencies changing in opposite directions and the test mixture is created by linearly cross-fading the two chirps. The basic model fails to discriminate the two sources whereas the model using the transition matrix successfully estimates the cross-fading curves, although there is a little glitch in the intersection.

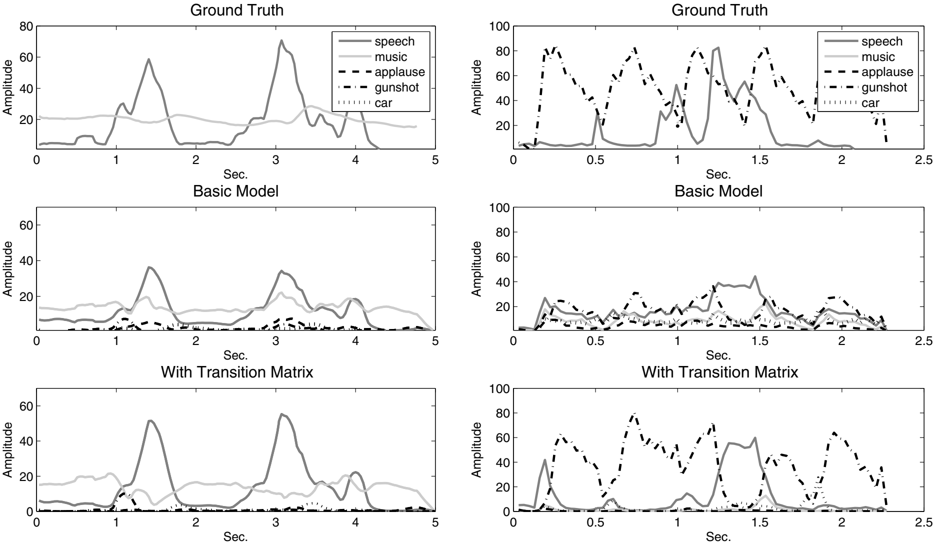
$P_t^{(i)}(z)$ by the predictions of these two terms, thereby imposing the expected structure. This re-weighting is performed after the M step in every EM iteration. Finally, we obtain the relative proportions of single sources at each time frame by simply summing the corresponding weights as in Eq. 3 and 4.

Fig. 1 illustrates the effect of re-weighing by the transition matrix. In the example, two source signals are given as chirps that have frequencies changing in opposite directions and thus they produce the same dictionary but different transition matrices. The test signal is created by cross-fading the two chirps. The basic model estimates approximately the same proportions of the two sources because both dictionaries explain the mixture equally well at every time frame. On the other hand, the re-weighting using the transition matrix successfully estimates the cross-fading curves by filtering out weights inconsistent with temporal dependencies of each source.

3 Experimental Results

We evaluated the proposed method on five classes of sound sources–speech, music, applause, gun shot and car. We collected ten clips of sound files for each class. Speech and music files were extracted from movies, each about 25 seconds long. Other sound files were obtained from a sound effects library.[3] They have different lengths from less than one to five seconds. We resampled all sound

[3] www.sound-ideas.com



(a) speech and music

(b) speech and gun shot

Fig. 2. Estimated relative proportions for mixtures of two sources. For the purpose of visualization, we shows amplitude envelopes of estimated sources instead of the relative proportions. The amplitude envelopes are obtained by multiplying the relative proportions to the sum of the magnitudes in that time frame ($\sum_f X(f,t)$) (an approximation to the envelope of the mixture sound). The top plots are the ground truth computed from individual sources. The middle and bottom plots show the results using the basic model and the improved model with the transition matrix, respectively.

files to 8kHz, and used a 64ms Hann window with 32ms overlap to compute the spectrograms. In the training phase, we obtained a dictionary of elements and a transition matrix separately for each sound source. The size of the dictionary was set to small numbers (less than 15) because we do not need a high-quality reconstruction. In addition, dictionary sizes of speech and music were set to be greater than those of other environmental sounds because speech and music generally have more variations in the training data.

3.1 Examples

Fig. 2 shows examples in which the test sound is given as a mixture of two sources. For the mixture of speech and music sounds, both models recognize the two sources fairly well. However, in the basic model, separation between speech and music is somewhat diluted and loud utterances of speech are partly explained by other sources, which are absent from the test sound. On the other hand, the model with the transition matrix shows better separation between speech and

Table 1. Estimation errors for single test sources

Test sources	speech	music	applause	gun shot	car
basic model	0.37	0.45	0.20	0.76	0.41
with transition matrix	0.26	0.32	0.03	0.42	0.39

music and suppresses other sources more effectively. For the mixture of speech and gunshot sounds, the two models show more apparent difference. The basic model completely fails to estimate the relative proportions as the gunshot sound is represented by many other sources, whereas the model with the transition matrix restores the original envelopes fairly well.

3.2 Evaluation

In order to examine the two models more accurately, we performed a formal evaluation using ten-fold cross-validation. At each validation stage, we split the dataset into nine training files and one test file for each source. From the training files we trained the models with ten sets of dictionary sizes; the maximum numbers of dictionary sizes were 12, 15, 5, 5 and 8 for speech, music, applause, gunshot and car sounds, respectively, and the minimum numbers were 1 for all sources. For the model with transition matrix, we additionally adjusted four reweighting strengths ($C = 0.3$, 0.5, 0.7 and 1.0). For the test files we estimated the relative proportions for single sources and mixtures of two and three sources. The mixtures were created by mixing two or three test files with different relative gains.[4] To quantify the estimation accuracy, we computed the following metric:

$$\text{Estimation error} = \frac{1}{N} \sum_s \sum_t |r_t(s) - g_t(s)|, \tag{11}$$

where $r_t(s)$ is the estimated proportion from Eq. 3 and 4, $g_t(s)$ is the ground truth proportion and N is the number of time frames in the test file. We obtained the ground truth proportion from the ratio of envelope between each single source and the mixture at each time frame. The envelope was computed by summing the magnitudes in that time frame ($\sum_f X(f,t)$). We measured this metric only for active sources, that is, those exist in the test sound. Note that the ground truth proportion is 1 for single test sounds because no other sound is present in that case.

Table 1 shows the best results for the single test source. In the basic model, the significant proportion of the test sound is explained by dictionaries of other sources, particularly for gun shot sounds. However, the model with the transition matrix show significant improvement for most sounds. Table 2 and 3 shows the

[4] For the mixtures of two sources, the relative gains of the two sources were adjusted to be -12, -6, 0, 6 and 12 in dB. For the mixtures of three sources, they were adjusted to be -6, 0 and 6 in dB for each pair.

Table 2. Estimation errors for mixtures of two sources

Test sources	speech/music	speech/gun shot	speech/applause	music/car
basic model	0.17 / 0.27	0.19 / 0.48	0.13 / 0.16	0.26 / 0.25
with transition matrix	0.15 / 0.21	0.15 / 0.34	0.13 / 0.12	0.21 / 0.26

Table 3. Estimation errors for mixtures of three sources

Test sources	speech/music/gun shot	speech/music/car
basic model	0.17 / 0.21 / 0.25	0.16 / 0.20 / 0.20
with transition matrix	0.15 / 0.18 / 0.25	0.15 / 0.17 / 0.21

results for the mixtures of two and three sources. Although the improvements are slightly less than those in the single source case, the model with transition matrix generally outperform the basic model. Note that as we have more sources in the test sound, the estimation errors for individual sources become smaller because the relative proportions of single sources are also smaller.

4 Summary and Discussion

In this paper we presented a method to estimate the relative proportions of single sources in sound mixtures as a way of recognizing real-world sounds which usually contains multiple sources. We first suggested a method of performing this estimation using standard PLCA. We then proposed a method to improve this estimation by accounting for temporal dependencies among dictionary elements. Our experiments on five classes of sound sources and their mixtures showed promising results, particularly with the model that considers temporal dependencies.

A difficulty that we encountered in our experiments was choosing different combinations of dictionary sizes for each single sound source in the training stage because if we consider all possible combinations of dictionary sizes (i.e. grid search), the number of possibilities exponentially grows. Therefore, we had to choose possible combinations of dictionary sizes using some heuristics. For the future works, we need to figure out more algorithmic methods to choose dictionary sizes. In addition, the evaluation metric we used is somewhat rigorous in that it counts accuracy for a very short time. Thus, softer metrics such as mean accuracy over some period or the presence of sound sources (e.g. by checking if the proportion is greater than a certain threshold) could be additionally considered. Finally, the proposed models are desired to be evaluated on a larger dataset.

References

1. Radhakrishnan, R., Xiong, Z., Otsuka, I.: A Content-Adaptive Analysis and Representation Framework for Audio Event Discovery from Unscripted Multimedia. EURASIP Journal on Applied Signal Processing, 1–24 (2006)
2. Li, Y., Dorai, C.: Instructional Video Content Analysis Using Audio Information. IEEE TASLP 14(6) (2006)
3. Tran, H.D., Li, H.: Sound Event Recognition With Probabilistic Distance SVMs. IEEE TASLP 19(6) (2011)
4. Smaragdis, P., Raj, B., Shashanka, M.: A probabilistic latent variable model for acoustic modeling. In: Advances in Models for Acoustic Processing, NIPS (2006)
5. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 414–421. Springer, Heidelberg (2007)

The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio Source Separation -

Shoko Araki[1], Francesco Nesta[2], Emmanuel Vincent[3], Zbyněk Koldovský[4], Guido Nolte[5], Andreas Ziehe[5], and Alexis Benichoux[3]

[1] NTT Communication Science Labs., NTT Corporation, Japan
[2] Fondazione Bruno Kessler - Irst, Center of Information Technology, Italy
[3] INRIA, Centre Inria Rennes - Bretagne Atlantique, France
[4] Technical University of Liberec, Czech Republic
[5] Fraunhofer Institute FIRST IDA, Germany

Abstract. This paper summarizes the audio part of the 2011 community-based Signal Separation Evaluation Campaign (SiSEC2011). Four speech and music datasets were contributed, including datasets recorded in noisy or dynamic environments and a subset of the SiSEC2010 datasets. The participants addressed one or more tasks out of four source separation tasks, and the results for each task were evaluated using different objective performance criteria. We provide an overview of the audio datasets, tasks and criteria. We also report the results achieved with the submitted systems, and discuss organization strategies for future campaigns.

1 Introduction

The Signal Separation Evaluation Campaign (SiSEC) is a regular campaign focused on the evaluation of methods for signal separation. It was built on the experience of previous evaluation campaigns (e.g., the MLSP'05 Data Analysis Competition[1], the PASCAL Speech Separation Challenge [1], and the Stereo Audio Source Separation Evaluation Campaign (SASSEC)) and has been organized since 2008 [2]. SiSEC is not a competition but a community-based scientific evaluation whose aspects are publicly defined. A call for participation precedes the evaluation and aims to define datasets, tasks and evaluation criteria.

This article describes the audio part of SiSEC 2011. In response to the feedback received at SiSEC2008 and SiSEC2010, previous datasets were reorganized as follows:

1. datasets sharing similar scenarios were merged in order to remove some redundancies (e.g. the 2-channel 1-source dataset of the "Source separation in the presence of real-world background noise" task of SiSEC2010 was merged with a new dataset from the PASCAL CHiME Challenge [3]).
2. tasks with little participation in the previous campaign were excluded;

[1] `http://mlsp2005.conwiz.dk/index.php@id=30.html`

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 414–422, 2012.

3. unrealistic data was removed (e.g. the synthetic mixtures of the "Underdetermined speech and music separation" task of SiSEC2010 were eliminated and fresh real-world data was provided).

In general the new campaign was designed so as to better match with real-world scenarios. We believe this data could be of high potential interest for many audio applications in the future years. Specifically, the new datasets embody more realistic features such as a) more reverberant rooms b) real-world diffuse or rapidly varying noise c) source movements.

Datasets and tasks are specified in Section 2 and the obtained outcomes are summarized in Section 3. Due to the variety of the submissions, we focus on the general outcomes of the campaign and ask readers to refer to `http://sisec.wiki.irisa.fr/` for further details.

2 Specifications

This section describes the tasks, datasets and evaluation criteria, which were specified in a collaborative fashion. A few initial specifications were first suggested by the organizers. Potential participants were then invited to provide feedback and contribute additional specifications through the wiki or the mailing list.

2.1 Tasks

For each dataset, audio mixtures spanning a variety of mixing conditions are provided. The channels $x_i(t)$ $(1 \leq i \leq I)$ of each mixture signal were generally obtained as $x_i(t) = \sum_{j=1}^{J} s_{ij}^{\mathrm{img}}(t)$, where $s_{ij}^{\mathrm{img}}(t)$ is the *spatial image* of source j $(1 \leq j \leq J)$ on channel i [2]. For point sources, $s_{ij}^{\mathrm{img}}(t) = \sum_{\tau} a_{ij}(t-\tau, \tau) s_j(t-\tau)$ where $s_j(t)$ are the source signals and $a_{ij}(t, \tau)$ the (possibly time-varying) mixing filters. For these mixtures, we specified the following four tasks:

T1 Source counting
T2 Source signal estimation
T3 Source spatial image estimation
T4 Source DOA estimation

These tasks consist in finding, respectively: (T1) the number of sources J, (T2) the source signals $s_j(t)$, (T3) the spatial images $s_{ij}^{\mathrm{img}}(t)$ of the sources for all channels i, and (T4) the direction of arrival (DOA) of each source. Participants were asked to submit the results of their systems for T2 and/or T3, and optionally for T1 and/or T4.

Two oracle systems were also considered for benchmarking task T3: ideal binary masking over a short-time Fourier transform (STFT) [4] (O1) and over a cochleagram [5] (O2). These systems require the true source spatial images and provide upper bounds on the performance of binary masking-based systems.

2.2 Datasets

Four distinct datasets were provided for SiSEC2011:

D1 Under-Determined Speech and Music Mixtures
This dataset includes the stereo dataset D1 from SiSEC2010 [6], and a fresh dataset containing ten 3-channel mixtures of four audio sources of 10 s duration, sampled at 16 kHz. For 3-channel data we used a linear microphone array. The room reverberation time (RT) for the fresh dataset was 130 ms or 380 ms. Instantaneous mixtures are also included. Tasks T1, T2 and T3 are considered.

D2 Determined Convolutive Mixtures Under Dynamic Conditions
This dataset consists of two kinds of scenarios: (1) random source activity of multiple sources in multiple static locations, and (2) a source continuously moving and overlapped with a source in a fixed or random location. The former aims to simulate a meeting scenario, where multiple talkers utter from fixed locations and their activity is unknown. The latter was specifically designed to evaluate systems able to handle dynamic variations of the mixing parameters. Due to the challenging reverberation conditions, datasets with different difficulty levels were provided (i.e. varying the source-array distance and the angular direction of simultaneously active sources). In the mixtures, two speakers are simultaneously active at most. In these datasets 4-channel mixtures are provided, and participants can decide whether using all the available channels or only a subset of them. The recordings were obtained in a real room of size (6 × 5 × 4 m) with an estimated RT of 700 ms. For both the datasets the signals were recorded by a uniform linear array of four (directional) microphones with a different spacing (of about 2 cm, 8 cm, and 18 cm) and sampled at 16 kHz. T2 and T3 are considered for this dataset.

D3 Professionally Produced Music Recordings
According to many positive requests from the community, we decided to repeat this dataset in SiSEC2011. This dataset contains stereo music signals sampled at 44.1 kHz, including those of the dataset D3 from SiSEC2010 [6]. In addition to the 20-second snips to be separated, full-length recordings are provided as well. The mixtures were created by sound engineers, and the ways of mixing and the mixing effects applied are unknown. Task T3 is imposed on this dataset.

D4 Two-channel Mixtures of Speech and Real-world Background Noise
This task aims to evaluate source separation and denoising techniques in the context of speech enhancement by merging two datasets: the dataset D3 from SiSEC2010 [6] and the CHiME corpus [3]. Both datasets consist of two-channel mixtures of one speech source and real-world background noise sampled at 16 kHz. In both datasets, the spatial image of the background noise was recorded in real-world environments: a subway car, a cafeteria, or a square for the former, and a British family living room for the latter. Tasks T2, T3 and T4 are evaluated for this dataset.

All datasets include both test and development data, and the CHiME corpus in D4 also includes training data. The true source signals and source positions underlying the test data were hidden to the participants, while they were provided for the development data. The true number of speech/music sources was always available.

2.3 Evaluation Criteria

Tasks T2 and T3 were evaluated via the criteria in the BSS Eval toolbox termed signal to distortion ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR) [7,2]. In addition, version 2.0 of the PEASS toolbox [8,9] was used to assess the perceptual quality of the estimated signals for stereo data according to four performance measures akin to SDR, ISR, SIR and SAR: overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifact-related perceptual score (APS).

Task T4 was evaluated by the absolute difference between the true and estimated DOAs.

3 Results

Despite the challenging specifications of each dataset, a remarkable participation was obtained. A total of 32 submissions were received from 18 different research centers. Many participants were involved in SiSEC for the first time, revealing a positive enlargement of the community. Tables 1 to 5 summarize the average performance obtained over the submitted algorithms. The algorithm details and all the results are available at `http://sisec2011.wiki.irisa.fr/tiki-index.php`. It should be noted that the presented values are the absolute values, not the improvements from the values for mixtures.

By comparison with the previous SiSEC, an unexpected high participation was observed for dataset D3. This trend seems to be in line with the recent increasing interest in NMF-based techniques, which have shown to marry well with the task of music recordings separation. The traditional dataset D1 has attracted a satisfactory amount of new participants, although the performance improvement seems to be still limited by the amount of reverberation. The datasets D2 and D4, aimed to simulate more realistic real-world scenarios, have attracted a sufficient but yet limited number of participants, probably due to the intrinsic difficulty of the data. Furthermore, the proposed algorithms do not seem to be equivalently effective in all the scenarios, which reveals that the acoustic source separation is still an open problem for real-world applications.

Note that a close analysis of each table is beyond the scope of this paper and a more detailed investigations will be discussed at the LVA/ICA 2012 conference.

Table 1. Average performance for task T2 or T3 for instantaneous dataset D1. 2 mic: average over test & test2 datasets, 3 mic: average over test3 dataset.

System	2 mic, 3 speech				2 mic, 3 music				2 mic, 4 speech				3 mic, 4 speech			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
S1 [10]	13.4	25.7	21.2	14.5	16.6	27.0	23.1	20.5	8.9	17.2	15.4	9.7	-	-	-	-
	43.9	55.4	61.0	58.6	52.3	58.9	66.6	55.5	42.4	65.1	62.2	47.0	-	-	-	-
S2 [11]	7.9	-	13.6	9.7	6.9	-	12.2	10.2	3.0	-	8.0	5.8	11.7	-	19.1	12.6
	43.2	-	61.7	25.6	40.0	-	62.7	10.6	29.8	-	46.7	10.7	39.7	-	64.0	37.7
S3 [2]	8.8	-	19.8	9.4	5.9	-	13.9	8.5	5.8	-	16.4	6.7	8.0	-	20.5	8.4
	38.5	-	75.3	10.4	35.7	-	68.9	16.2	35.7	-	65.7	12.1	38.6	-	75.5	9.2
O1	10.8	20.1	21.7	11.1	10.4	18.0	18.8	12.5	9.1	17.6	20.0	9.3	-	-	-	-
	38.9	61.8	70.5	37.7	33.3	48.5	64.8	34.2	27.1	57.7	71.8	21.9	-	-	-	-
O2	8.5	15.7	17.4	9.1	9.0	14.1	18.1	11.3	7.5	13.7	16.4	8.1	-	-	-	-
	24.0	29.8	72.4	20.0	30.4	28.3	69.5	21.6	22.0	20.9	70.8	13.1	-	-	-	-

Table 2. Average performance for task T3 for convolutive dataset D1. 2 mic: average over test & test2 datasets, 3 mic: average over test3 dataset. The values are averaged over all the reverberation time.

System	2 mic, 3 speech				2 mic, 3 music				2 mic, 4 speech				3 mic, 4 speech			
	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
S1 [10]	3.4	8.2	6.4	7.8	2.1	7.2	4.4	10.0	2.0	6.1	3.8	5.5	-	-	-	-
	27.9	47.6	38.5	55.9	21.8	33.3	29.7	39.1	31.2	46.9	39.2	48.9	-	-	-	-
S2 [12][3]	1.8	4.1	2.2	4.3	-	-	-	-	1.1	3.3	0.1	2.8	1.6	3.4	1.8	3.4
	21.4	33.9	43.8	38.8	-	-	-	-	19.9	27.4	40.5	35.8	20.1	33.6	53.6	34.0
S3 [13]	5.3	9.3	7.7	10.0	-	-	-	-	-	-	-	-	-	-	-	-
	26.9	51.5	35.3	62.1	-	-	-	-	-	-	-	-	-	-	-	-
S4 [14]	4.3	9.0	6.9	8.8	0.2	4.8	0.6	7.1	1.4	4.7	1.4	6.2	1.2	2.9	2.4	5.6
	25.4	49.9	38.1	56.3	19.7	28.9	19.6	42.0	27.2	41.7	28.4	50.6	29.7	58.8	59.3	30.0
S5 [15][4]	5.4	8.9	8.9	9.1	2.8	6.8	5.0	8.8	3.3	6.3	5.6	6.3	-	-	-	-
	34.4	59.8	52.2	57.7	27.3	43.8	37.8	43.1	35.0	58.3	47.9	49.2	-	-	-	-
S6 [15]	6.1	10.9	10.5	9.1	3.0	7.6	5.4	8.9	3.6	7.4	6.9	6.5	-	-	-	-
	38.3	58.8	53.7	55.0	26.5	39.7	38.0	42.0	35.1	56.0	49.5	48.7	-	-	-	-
S7 [16][5]	5.8	10.8	10.3	8.2	1.7	6.3	3.0	6.7	3.2	7.3	5.9	5.6	5.3	10.0	9.9	7.5
	37.2	61.9	52.3	51.4	22.4	35.9	32.6	38.8	30.3	54.6	48.2	42.6	31.1	63.1	61.6	34.4
O1	10.2	18.7	20.2	10.7	9.9	16.9	18.0	11.0	8.7	16.4	18.5	9.1	-	-	-	-
	43.4	63.1	69.9	45.1	36.0	52.7	64.2	40.6	36.2	63.1	71.9	34.3	-	-	-	-
O2	7.6	13.8	16.8	8.2	7.1	12.3	14.5	8.3	6.1	11.3	15.1	6.3	-	-	-	-
	26.7	41.9	72.7	23.8	25.9	21.8	69.3	19.0	23.7	37.8	72.4	18.8	-	-	-	-

[2] The system details can be found at the SiSEC2011 wiki.

[3] Figure computed by averaging over an incomplete set of mixtures.

[4] The same algorithm as [15] without the Wiener-Filter post-processing.

[5] The values for “2mic.” are from SiSEC2010 submissions.

Table 3. Average performance for dataset D2, ”random source activity of multiple sources in multiple static locations” (top) and ”a continuously moving active source overlapped with a source in a fixed or random location” (bottom). All the signals are evaluated as source signal and spatial source signal estimates. For more details see `http://www.irisa.fr/metiss/SiSEC11/dynamic/main.html`.

System	Source signal estimation						Spatial image estimation							
	SDR	SIR	SAR	OPS	IPS	APS	SDRi	SIRi	SARi	ISRi	OPSi	TPSi	IPSi	APSi
S1 [17]	3.5	9.2	7.0	30.5	69.5	11.9	2.0	6.0	7.5	3.0	29.5	30.3	67.2	27.3
S2 [15,18][6]	3.7	6.2	9.3	35.4	53.2	20.9	2.6	4.1	12.1	4.4	33.1	48.7	51.4	41.3
S3 [15,18][7]	3.5	7.3	7.5	31.8	63.0	7.1	2.3	5.2	10.1	3.6	29.6	41.1	61.6	32.6
S4 [19]	2.2	6.6	6.1	28.5	66.9	4.1	2.1	4.5	7.0	3.6	27.7	41.7	66.5	24.6
S5 [19][8]	2.3	7.4	5.9	26.9	70.6	3.0	1.9	4.8	6.9	3.3	25.9	32.1	70.3	21.1
S6 [20,21]	1.8	7.3	5.1	26.9	71.9	1.6	1.5	5.4	5.7	2.3	26.4	31.1	71.8	23.5
S7 [20,21][9]	3.1	10.6	5.3	27.1	72.6	1.9	1.2	6.7	6.2	1.7	27.0	20.4	72.3	22.5

System	Source signal estimation						Spatial image estimation							
	SDR	SIR	SAR	OPS	IPS	APS	SDRi	SIRi	SARi	ISRi	OPSi	TPSi	IPSi	APSi
S1 [17]	2.5	7.7	6.2	30.9	66.5	6.2	1.3	6.3	8.0	1.9	29.0	31.3	65.1	30.8
S2 [15,18][6]	4.2	7.1	9.1	36.2	55.3	21.3	4.3	5.5	12.8	7.0	33.9	59.5	53.4	40.8
S3 [15,18][7]	4.0	8.5	7.3	32.2	65.9	6.7	4.0	6.9	10.9	6.1	30.2	53.8	64.2	30.7
S4 [19]	3.3	10.5	5.3	26.0	77.1	1.4	2.5	8.0	7.1	3.8	26.9	39.9	76.8	18.1
S5 [19][8]	3.5	11.0	5.4	25.7	78.2	1.3	2.5	8.5	7.2	3.8	27.0	36.5	78.1	18.2
S6 [20,21]	2.4	8.5	5.1	28.1	71.3	2.6	2.0	7.2	6.7	3.0	27.2	36.7	70.8	24.6
S7 [20,21][9]	3.8	12.9	5.4	28.4	72.1	2.9	1.7	9.0	7.6	2.2	28.0	23.8	70.3	24.7

Table 4. Average performance for T2/T3 for testset of D3. The results only for the vocal and drum tracks, which most of the submissions addressed, are summarized. S4, S6 and S8 addressed all the specified tracks. Complete results can be found at SiSEC2011 wiki.

System	Vocal								Drums							
	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS
S1 [22]	3.8	6.2	Inf	3.1	22.4	28.8	59.0	30.8	-	-	-	-	-	-	-	-
S2 [23]	4.5	6.8	Inf	3.8	26.6	29.3	62.7	29.5	-	-	-	-	-	-	-	-
S3 [24]	-2.7	-0.8	Inf	-7.2	22.5	5.0	64.6	10.0	-	-	-	-	-	-	-	-
S4 [25]	-5.5	-1.3	7.0	3.6	15.7	15.7	27.6	15.3	-7.1	-2.9	2.9	2.7	23.3	23.2	50.4	12.8
S5 [26]	2.4	8.5	Inf	0.1	25.2	15.9	70.5	11.6	-0.2	2.2	5.9	-5.4	23.6	44.0	67.8	2.1
S6 [10]	3.1	8.1	7.7	3.7	24.4	37.1	20.8	54.3	2.0	4.3	2.9	2.1	29.3	54.6	28.9	50.4
S7 [27] [3]	4.1	10.7	6.3	7.3	41.6	74.5	61.2	40.0	-	-	-	-	-	-	-	-
S8 [2]	3.0	7.7	9.0	2.4	19.4	28.7	55.2	31.9	1.7	2.1	11.6	1.1	20.7	19.9	58.7	9.0
O1	6.2	22.1	22.3	6.2	28.4	69.1	69.1	16.6	6.3	24.6	23.2	6.2	25.7	73.7	74.5	2.8
O2	4.7	17.0	16.3	4.7	23.6	38.1	61.9	14.8	1.4	2.7	17.3	0.4	18.1	32.2	69.4	4.6

[6] Algorithm derived from the weighted Natural Gradient in [15].

[7] The same algorithm as S2 with additional Binary Masking post-processing.

[8] The same algorithm as S4 with additional TF post-processing.

[9] The same algorithm as S6 with additional Wiener-Filter like post-processing.

Table 5. Average performance for task T2/T3 for test dataset D4. Outdoor and indoor indicates the recordings 2ch-1src in the dataset D3 from SiSEC2010 [6] and the CHiME corpus [3], respectively. Performance of S1 are evaluated on the source signal estimates (i.e. 'src' files), while the remaining systems are evaluated on the spatial source image estimates (i.e. 'sim' files).

System	Outdoor								Indoor							
	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS
S1[2]	-	-	-	-	-	-	-	-	1.8	-	6.1	6.2	-	-	-	-
S2[2]	-1.8	13.3	-0.7	16.0	11.2	46.7	35.8	81.3	-	-	-	-	-	-	-	-
S3[28]	8.8	13.4	15.1	14.4	14.6	50.8	39.3	80.0	-1.7	8.0	0.7	14.1	20.9	50.3	30.0	69.3
S4[28][2]	6.1	13.1	13.4	10.9	43.8	59.8	58.2	57.6	-	-	-	-	-	-	-	-
S5[29,15]	3.5	16.6	6.4	12.2	33.4	59.0	57.5	70.0	6.0	7.3	16.5	11.0	37.3	43.5	68.7	38.9
S6[29,15][10]	3.4	17.6	5.8	12.8	29.6	58.2	55.2	73.3	8.0	11.0	14.7	12.0	38.5	55.3	65.0	49.9
S7[10]	-	-	-	-	-	-	-	-	5.4	7.3	14.0	11.7	35.2	62.2	49.9	51.0
S8[2]	4.0	7.0	8.8	7.6	36.5	51.2	63.4	41.8	-	-	-	-	-	-	-	-
baseline [30]	2.4	8.9	7.2	8.7	22.2	49.9	47.6	64.3	1.7	3.5	5.2	8.6	29.3	34.8	44.1	37.5
O1	15.8	27.1	24.3	16.9	51.3	65.9	75.6	45.5	14.5	20.9	22.7	16.6	53.5	67.2	73.6	57.0

4 Conclusion

This paper presented the specifications of SiSEC2011 and summarized the performance obtained over all the submissions. This time, in accordance with discussions at previous SiSECs, we carefully selected the datasets and tasks in a collaborative fashion. Ultimately, four datasets and tasks were provided which attracted many submissions from 18 research institutions.

Despite some open challenges which still do not allow us to provide an unambiguous evaluation of all the submissions, we hope that SiSEC2011 will continue to represent a common platform for sharing new ideas and perspectives in the source separation research field. We believe SiSEC2011 data could be of high potential interest for many audio applications and encourage the community to use it as a reference for future evaluations.

Following the experience maturated till this campaign, new criteria seem needed for better evaluating more realistic scenarios, such as source separation involving dereverberation or tracking of time-varying mixing conditions. Furthermore, it would be worthwhile to investigate on new objective evaluation criteria more related to the separation filter accuracy rather than to the quality of the signals itself, with the hope of minimizing the presence of outliers. With this regard, we invite all willing participants to join a continuous collaborative discussion on the future of source separation evaluation.

Acknowledgments. We thank all the participants and data providers. We also thank to all the researchers who gave their opinions on the wiki and the mailing-list. Special thanks go to Mr. Antoine Liutkus, who proposed to repeat the dataset D3, and encouraged researches of the field to join in SiSEC2011.

[10] The same algorithm as S5 with different parameter settings.

References

1. Cooke, M.P., Hershey, J., Rennie, S.: Monaural speech separation and recognition challenge. Computer Speech and Language 24, 1–15 (2010)
2. Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., Duong, N.Q.K.: The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges. Signal Processing (to appear)
3. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In: Proc. Interspeech, pp. 1918–1921 (2010)
4. Vincent, E., Gribonval, R., Plumbley, M.D.: Oracle estimators for the benchmarking of source separation algorithms. Signal Processing 87(8), 1933–1950 (2007)
5. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In: Speech Separation by Humans and Machines. Springer, Heidelberg (2005)
6. Araki, S., Ozerov, A., Gowreesunker, V., Sawada, H., Theis, F., Nolte, G., Lutter, D., Duong, N.Q.K.: The 2010 Signal Separation Evaluation Campaign (SiSEC2010): Audio Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 114–122. Springer, Heidelberg (2010)
7. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. on Audio, Speech and Language Processing 14(4), 1462–1469 (2006)
8. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and objective quality assessment of audio source separation. IEEE Trans. on Audio, Speech and Language Processing 19(7), 2046–2057 (2011)
9. Vincent, E.: Improved Perceptual Metrics for the Evaluation of Audio Source Separation. In: Theis, F., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 430–437. Springer, Heidelberg (2012)
10. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. on Audio, Speech and Language Processing PP(99), 1 (2011)
11. Makkiabadi, B., Sanei, S., Marshall, D.: A k-subspace based tensor factorization approach for under-determined blind identification. In: Proc. ASILOMAR 2010 (2010)
12. Hirasawa, Y., Yasuraoka, N., Takahashi, T., Ogata, T., Okuno, H.G.: A GMM Sound Source Model for Blind Speech Separation in Under-determined Conditions. In: Yeredor, A., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 446–453. Springer, Heidelberg (2012)
13. Iso, K., Araki, S., Makino, S., Nakatani, T., Sawada, H., Yamada, T., Nakamura, A.: Blind source separation of mixed speech in a high reverberation environment. In: Proc. HSCMA 2011, pp. 36–39 (2011)
14. Cho, J., Choi, J., Yoo, C.D.: Underdetermined convolutive blind source separation using a novel mixing matrix estimation and MMSE-based source estimation. In: Proc. MLSP 2011 (2011)
15. Nesta, F., Omologo, M.: Convolutive Underdetermined Source Separation through Weighted Interleaved ICA and Spatio-temporal Source Correlation. In: Yeredor, A., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 222–230. Springer, Heidelberg (2012)

16. Sawada, H., Araki, S., Makino, S.: A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In: Proc. WASPAA, pp. 139–142 (2007)
17. Málek, J., Koldovský, Z., Tichavský, P.: Semi-blind Source Separation Based on ICA and Overlapped Speech Detection. In: Yeredor, A., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 462–469. Springer, Heidelberg (2012)
18. Nesta, F., Omologo, M.: Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. IEEE Transactions on Audio, Speech, and Language Processing 20(1), 246–260 (2012)
19. Loesch, B., Yang, B.: Blind Source Separation Based on Time-Frequency Sparseness in the Presence of Spatial Aliasing. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 1–8. Springer, Heidelberg (2010)
20. Loesch, B., Yang, B.: Adaptive Segmentation and Separation of Determined Convolutive Mixtures under Dynamic Conditions. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 41–48. Springer, Heidelberg (2010)
21. Loesch, B., Nesta, F., Yang, B.: On the robustness of the multidimensional state coherence transform for solving the permutation problem of frequency-domain ICA. In: Proc. ICASSP, pp. 225–228 (2010)
22. Durrieu, J.-L., David, B., Richard, G.: A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE Journal of Selected Topics on Signal Processing 5(6), 1180–1191 (2011)
23. Durrieu, J.-L., Thiran, J.-P.: Musical Audio Source Separation Based on User-Selected F0 Track. In: Yeredor, A., et al. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 438–445. Springer, Heidelberg (2012)
24. Cano, E., Dittmar, C., Schuller, G.: Interaction of phase, magnitude and location of harmonic components in the perceived quality of extracted solo signals. In: Proc. AES (2011)
25. Spiertz, M., Gnann, V.: Note clustering based on 2D source-filter modeling for underdetermined blind source separation. In: Proc. AES (2011)
26. Marxer, R., Janer, J.: A Tikhonov regularization method for spectrum decomposition in low latency audio source separation. In: Proc. ICASSP 2012 (to appear, 2012)
27. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. In: Proc. ICASSP 2012 (to appear, 2012)
28. Mustiere, F., Bolic, M., Bouchard, M.: Real-world particle filtering-based speech enhancement. In: Proc. CIP, pp. 75–80 (2010)
29. Nesta, F., Matassoni, M.: Robust automatic speech recognition through on-line semi-blind source extraction. In: Proc. CHIME (2011)
30. Blandin, C., Ozerov, A., Vincent, E.: Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Signal Processing (to appear)

The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Biomedical Data Analysis -

Guido Nolte[1], Dominik Lutter[2], Andreas Ziehe[3], Francesco Nesta[4], Emmanuel Vincent[5], Zbyněk Koldovský[6], Alexis Benichoux[5], and Shoko Araki[7]

[1] Fraunhofer Institute FIRST, Germany
[2] IBIS, Helmholtz Zentrum München, Germany
[3] Technical University Berlin, Germany
[4] Fondazione Bruno Kessler - Irst, Center of Information Technology, Italy
[5] INRIA, Centre Inria Rennes - Bretagne Atlantique, France
[6] Technical University of Liberec, Czech Republic
[7] NTT Communication Science Labs., NTT Corporation, Japan

Abstract. This paper summarizes the bio part of the 2011 community based Signal Separation Evaluation Campaign (SiSEC2011). Two different data sets were given. In the first task, participants were asked to estimate the causal relations of underlying sources from simulated bivariate EEG data. In the second task, participants were asked to reconstruct signaling pathways or parts of it from the microarray expression profiles. The results for each task were evaluated using different objective performance criteria. We provide an overview of the biomedical datasets, tasks and criteria, and we report on the achieved results.

1 Introduction

The Signal Separation Evaluation Campaign (SiSEC) is a regular campaign focused on the evaluation of methods for signal separation. While its main focus is on separation of audio data, after the campaign in 2010 [1] this is now the second time that tasks on biomedical data analysis are proposed. This article describes the bio part of SiSEC 2011.

The standard application of ICA-algorithms in biomedical data analysis are EEG and MEG data. In contrast to signal separation in audio datasets, the respective mixing model is static. The algorithms to solve such a problem are well established and are applied routinely by many researches. It is our opinion that conceptually only minor technical details could be added to present day knowledge. Additionally, a formulation of an ICA challenge for EEG/MEG data is problematic because of two reasons: a) in contrast to audio data, for real EEG/MEG data the ground truth is almost never known, and b), existing ICA algorithms exploit different statistical properties, and the winning method for simulated data will then be the one for which, essentially by coincidence, the simulated statistical properties match the exploited ones.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 423–429, 2012.

We therefore decided to deviate from the 'standard' problem and to propose two different tasks. In the first task, source separation shall be applied to analyze gene expressions, and in the second we simulate EEG data, but the task is not to separate sources but to separate the effect of confounding noise in an estimate of causal relations.

Details of the tasks can be found at http://sisec.wiki.irisa.fr/ and following the link to 'biomedical data analysis'.

2 Estimating Causal Relations

2.1 Task

Noninvasive electrophysiological measurements like EEG/MEG measure to large extent unknown superpositions of very many sources. Any relation observed between channels is dominated by meaningless mixtures of mainly independent sources. The question is how to observe and properly interpret true interactions in the presence of such strong confounders. Since recently, a focus of research are the causal relations between groups of neurons. Many methods have been suggested to address this question for EEG or MEG data [2,3,4,5,6].

In this task contributors are requested to estimate the direction of interaction for simulated unidirectional bivariate dynamical systems. The difficulty is the presence of additive noise which is both non-white and spatially correlated.

The task is to estimate the direction of the interaction of the signal. A submitted result is a vector with 1000 numbers having the values 1, -1, or 0. Here, 1 means direction is from first to second sensor, -1 means direction is from second to first sensor, and 0 means 'I do not know'.

2.2 Dataset

The dataset consists of 1000 examples of bivariate data for 6000 time points. Each example is a superposition of a signal (of interest) and noise. The signal is constructed from a unidirectional bivariate AR-model of order 10 with (otherwise) random AR-parameters and uniformly distributed input. The noise is constructed of three independent sources, generated with 3 univariate AR-models with random parameters and uniformly distributed input, which were instantaneously mixed into the two sensors with a random mixing matrix. The relative strength of noise and signal was set randomly. The Matlab code used to generate the data was provided. Note, that the phrase 'simulated EEG data' is meant loosely. The simulation addresses the conceptual problems of EEG data, but e.g. the actual spectra can be quite different from real EEG data.

The data $\mathbf{z}(t)$ were generated as

$$\mathbf{z}(t) = (1-\gamma)\frac{\mathbf{x}(t)}{||X||} + \gamma\frac{B\mathbf{y}(t)}{||BY||} \tag{1}$$

where $\mathbf{x}$ is a unidirectional linear system and $\mathbf{y}$ are two independent noise sources which are mixed into channels by a random matrix B. The parameter γ was set randomly between 0 and 1, $|| \cdot ||$ denotes Frobenius matrix norm, and X and Y denote the full data as a matrix, e.g. $X = (\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N))$ for N data points. The noise $\mathbf{y}(t)$ was generated with an AR(10)-model with diagonal but otherwise random parameters and uniformly distributed input, i.e.

$$y_i(t) = \sum_{p=1}^{10} A_i(p)(t-p) + \eta_i(t) \tag{2}$$

for $i = 1, 2, 3$. For each data set the parameters A_{ik} were selected randomly according to a Gaussian distribution with a standard deviation 0.25. Nonstationary, i.e. diverging, systems were excluded. If the standard is substantially larger, almost all systems are nonstationary. If it is chosen substantially smaller, the spectra are nearly white. The 'innovation' $\eta_i(t)$ was uniformly distributed in the range $[-.5, .5]$. This takes into account that some algorithms require non-Gaussian data or, especially, non-Gaussian innovations.

The signal $\mathbf{x}(t)$ was generated in the following way. If, e.g., the first channel was the sender, then $x_1(t)$ was generated with a random AR-model of order 10 in the same way as the noise term, and $x_2(t)$ was generated as

$$x_2(t) = \sum_p A_{22}(p)x_2(t-p) + A_{21}(p)x_1(t-p) + \epsilon_2(t) \tag{3}$$

where, again, $\epsilon_2(t)$ was uniformly distributed in the range $[-.5, .5]$. The construction for the other direction is analogous.

2.3 Evaluation Criterion

For all examples either 1 or -1 is correct. The most important point here is the way it is counted: you get +1 point for each correct answer; you get -10 points for each wrong answer; and you get 0 points for each 0 in the result vector. With this counting confidence about the result is added into the evaluation. It is strongly recommended that for each example the evidence for a specific finding is assessed. To our knowledge, this causality challenge is the first time that such an evaluation scheme is proposed.

2.4 Results

We received a total of 5 submissions. Results are shown in table 1 Another submission arrived after the deadline and after announcement of the results and was not counted. All participants were among the list of people who were contacted personally and were encouraged to submit.

This kind of challenge is new within the SiSEC campaign and can therefore not be compared to previous challenges.

Submission	Total Points	Correct Detections	False Detection
S1	-2289	701	299
S2 [7]	252	352	10
S3 [8]	-357	773	113
S4 [9,10,11,12,13]	218	278	6
S5 [14,15]	-247	163	41

Table 1. Results of causality challenge. The total points can be calculated as the number of correct detections minus ten times the number of false detections.

3 Cancer Pathway Reconstruction

3.1 The Task

Cellular signaling pathways are the key transducers from extracellular signals to cellular reaction. Dysfunction of signaling pathways is often involved in the formation of cancer [16]. Thus, understanding the biology of cell signaling helps to understand cancer and to develop new therapies. The regulation of these signaling pathways takes place on multiple layers, from extracellular receptors to intracellular transduction, ending with the transcriptional activation of target genes. Single genes can take part in more than one pathway and the expression profiles can be regarded as linear superpositions of different signaling pathways or more generally biological processes. All gene expression levels are represented by an $M \times N$ data matrix $\mathbf{X} = [\mathbf{x}_i^\top \dots \mathbf{x}_M^\top]$ with each row-vector $\mathbf{x}_m^\top$ representing the gene expression levels off all N genes measured in one experiment, or microarray. Assuming a linear mixture model, each vector $\mathbf{x}_m^\top$ represents a mixture of K unknown source signals $\mathbf{s}_k^\top$, each representing a pathway related gene expression profile with the corresponding mixing coefficients represented as a column-vector $\mathbf{a}_m$. Thus, using blind source separation (BSS) techniques, the data-matrix $\mathbf{X}$ can be decomposed into $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A}$ is the $M \times K$ mixing matrix and $\mathbf{S}$ the $K \times N$ matrix of source signals. These source signals can now be used as a basis to identify distinct signaling pathways in terms of cellular responses [17]. A more detailed discussion of the linear factor model can be found in [18,19].

Here, the task is to reconstruct these signaling pathways or parts of it from the microarray expression profiles using BSS techniques. In a first approximation we consider a signaling pathways as gene lists. These pathway gene lists were taken from NETPATH (www.netpath.org).

3.2 Dataset

The microarray technology a method for mRNA profiling has become one of the most popular approaches in the field of gene expression analysis. Based on the complexity of gene expression profiles, a variety of statistical methods have been developed to provide insights into the biological mechanisms of gene expression regulation [20,21,22]. The dataset consists of the i gene expression profiles. Each expression profile $\mathbf{x}_i$ mirrors the expression of N genes via measuring the level of the corresponding mRNA under a specific condition. In our case, mRNA was extracted from $i = 189$ invasive breast carcinomes [23] and measured using Affymetrix U133A Gene-chips. The Affymetrix raw data was normalized using the RMA algorithm [24] from the R Bioconductor package *simpleaffy*. Non-expressed genes were filtered out and Affymetrix probe sets were mapped to Gene Symbols. This resulted in a total of $N = 11815$ expressed genes.

3.3 Evaluation

Evaluation of the reconstructed pathways was performed by testing for the significance of enriched genes that can be mapped to the distinct pathways. For each source signal or estimated pathway we identify the number of genes that map to the distinct pathways and calculate p-values using Fisher's exact test. To correct for multiple testing we use the Benjamini-Hochberg procedure to estimate false positive rates (FDR). Now, after Benjamini-Hochberg correction a reconstructed pathway was declared as enriched if the p-value was below 0.05. Finally, the number of all different significantly reconstructed pathways were counted.

3.4 Results

There were no submissions.

4 Conclusion

In this paper we presented the specifications of the biomedical data analysis part of SiSEC2011 and summarized the performance obtained over all the submissions. Two different tasks of very different nature we given. The 'Cancer pathway reconstruction' received no submission which could be due to the fact that the mathematical details were unclear to people not familiar with the biology.

For the EEG/MEG data analysis it might appear natural that ICA challenges were proposed. However, the ICA model for these data is not convolutive, which, from an algorithmic viewpoint, is a much simpler case than acoustic data. For instantaneous mixtures the algorithms have become standard. Probably everything which could be said , apart from minor details, was said already, and such a challenge does not attract researchers working on the technical aspects.

It was therefore decided to propose a different kind of challenge, in which causal direction in the presence of noise were to be estimated and in which evidence had to assessed for a successful submission. The large variation across final scores that it is largely unclear how to optimally solve this problem. Although the data were, strictly speaking, nonlinear (i.e. non-Gaussian), the nonlinearity was small, and people working on nonlinear methods were effectively left out. For the future we intend to expand the simulations such that both linear and nonlinear methods can reasonably be applied.

Acknowledgements. The authors gratefully acknowledge financial support by EU, BMBF and DFG. We would like to thank our co-authors for their permission to use material from joint publications.

References

1. Araki, S., Theis, F., Nolte, G., Lutter, D., Ozerov, A., Gowreesunker, V., Sawada, H., Duong, N.Q.K.: The 2010 Signal Separation Evaluation Campaign (SiSEC 2010): Biomedical Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 123–130. Springer, Heidelberg (2010)
2. Chen, Y., Bressler, S.L., Knuth, K.H., Truccolo, W.A., Ding, M.: Stochastic modeling of neurobiological time series: power, coherence, Granger causality, and separation of evoked responses from ongoing activity. Chaos 16(2), 026113 (2006)
3. Kaminski, M., Ding, M., Truccolo, W.A., Bressler, S.L.: Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. Biol. Cybern. 85(2), 145–157 (2001)
4. Baccala, L.A., Sameshima, K.: Partial directed coherence: a new concept in neural structure determination. Biol. Cybern. 84(6), 463–474 (2001)
5. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. 85(2), 461–464 (2000)
6. Nolte, G., Ziehe, A., Nikulin, V.V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.R.: Robustly estimating the flow direction of information in complex physical systems. Phys. Rev. Lett. 100, 234101 (2008)
7. Hu, S., Dai, G., Dai, Q., Worrell, G., Liang, H.: Causality analysis of neural connectivity: critical examination of existing methods and advances of new methods. IEEE Transactions on Neural Networks (Regular Paper) 22(6), 829–844 (2011)
8. Leistritz, L., Hesse, W., Arnold, M., Witte, H.: Development of interaction measures based on adaptive non-linear time series analysis of biomedical signals. Biomedizinische Technik 51, 64–69 (2006)
9. Chavez, M., Martinerie, J., Le Van Quyen, M.: Statistical assessment of nonlinear causality: application to epileptic EEG signals. J. Neurosci. Methods 124(2), 113–128 (2003)
10. Palus, M., Komarek, V., Hrncir, Z., Sterbova, K.: Synchronization as adjustment of infomation rates: Detection from bivariate time series. Phys. Rev. E 63, 046211 (2001)
11. Prichard, D., Theiler, J.: Generalized redundancies for time series analysis. Physica D 84, 476–493 (1995)
12. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D.: Testing for Nonlinearity in Time Series: The Method of Surrogate Data. Physica D 58, 77–94 (1992)
13. Vakorin, V.A., Krakovska, O.A., McIntosh, A.R.: Confounding effects of indirect connections on causality estimation. Journal of Neuroscience Methods 184(1), 152–160 (2009)
14. Vicente, R., Wibral, M., Lindner, M., Pipa, G.: Transfer entropy-a model-free measure of effective connectivity for the neurosciences. RID e-1566-2011. Journal of Computational Neuro- Science 30(1), 45–67 (2011)

15. Wibral, M., Rahm, B., Rieder, M., Lindner, M., Vicente, R., Kaiser, J.: Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. RID e-1566-2011. Progress In Biophysics & Molecular Biology 105(1-2), 80–97 (2011)
16. Hoffman, B.D., Grashoff, C., Schwartz, M.A.: Dynamic molecular processes mediate cellular mechanotransduction. Nature 475(7356), 316–323 (2011)
17. Lutter, D., Langmann, T., Ugocsai, P., Moehle, C., Seibold, E., Splettstoesser, W.D., Gruber, P., Lang, E.W., Schmitz, G.: Analyzing time-dependent microarray data using independent component analysis derived expression modes from human macrophages infected with F. tularensis holartica. J. Biomed. Inform. 42(4), 605–611 (2009)
18. Lutter, D., Ugocsai, P., Grandl, M., Orso, E., Theis, F., Lang, E., Schmitz, G.: Analyzing m-csf dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. BMC Bioinformatics 9(100) (2008)
19. Teschendorff, A.E., Journée, M., Absil, P.-A., Sepulchre, R., Caldas, C.: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Computational Biology 3(8) (2007)
20. Quackenbush, J.: Computational approaches to analysis of DNA microarray data. Yearb Med. Inform., 91–103 (2006)
21. Schachtner, R., Lutter, D., Knollmüller, P., Tomé, A.M., Theis, F.J., Schmitz, G., Stetter, M., Vilda, P.G., Lang, E.W.: Knowledge-based gene expression classification via matrix factorization. Bioinformatics 24(15), 1688–1697 (2008)
22. Kowarsch, A., Blöchl, F., Bohl, S., Saile, M., Gretz, N., Klingmüller, U., Theis, F.J.: Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation. BMC Bioinformatics 11, 585 (2010)
23. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J. Natl. Cancer Inst. 98(4), 262–272 (2006)
24. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P.: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31(4), e15 (2003)

Improved Perceptual Metrics for the Evaluation of Audio Source Separation

Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@inria.fr

Abstract. We aim to predict the perceived quality of estimated source signals in the context of audio source separation. Recently, we proposed a set of metrics called PEASS that consist of three computation steps: decomposition of the estimation error into three components, measurement of the salience of each component via the PEMO-Q auditory-motivated measure, and combination of these saliences via a nonlinear mapping trained on subjective opinion scores. The parameters of the decomposition were shown to have little influence on the prediction performance. In this paper, we evaluate the impact of the parameters of PEMO-Q and the nonlinear mapping on the prediction performance. By selecting the optimal parameters, we improve the average correlation with mean opinion scores (MOS) from 0.738 to 0.909 in a cross-validation setting. The resulting improved metrics are used in the context of the 2011 Signal Separation Evaluation Campaign (SiSEC).

Keywords: audio source separation, objective evaluation, PEASS.

1 Introduction

Audio source separation is the task of extracting the signal of each sound source from a given mixture. In a number of applications such as speech enhancement for hearing aids or denoising of old music recordings, the separation performance amounts to the subjective judgment of listeners.

A popular set of performance metrics can be obtained by decomposing the estimation error into three components, namely *target distortion*, *interference* and *artifacts*, and measuring the salience of these components via energy ratios termed signal to distortion ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR) [11]. Despite the wide use of the associated BSS Eval toolbox[1], *e.g.* within the annual Signal Separation Evaluation Campaign (SiSEC) [11], these metrics are known to poorly correlate with subjective performance for certain mixtures involving *e.g.* low-frequency sounds or time-varying distortion. Two different routes have

[1] http://bass-db.gforge.inria.fr/bss_eval/

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 430–437, 2012.

been taken to increase correlation: assessing the overall distortion via auditory-motivated measures such as PEAQ [9] or PEMO-Q [8], or combining energy ratios via linear or nonlinear mappings trained on subjective opinion scores [4].

In [3], we combined these two routes via a three-step procedure consisting of

1. decomposing the estimation error into target distortion, interference and artifacts components,
2. assessing the salience of each component via PEMO-Q,
3. combining these saliences via trained nonlinear mappings.

We distributed the resulting metrics termed overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS), as the version 1.0 of a toolkit called PEASS[2]. Each of the above three steps involves one or more design parameters. In [3], we showed that the parameters of the first step have little influence on the prediction performance. In this paper, we evaluate the impact of the parameters of the two latter steps and select the optimal parameters maximizing the correlation with mean opinion scores (MOS). The resulting improved metrics are distributed as the version 2.0 of PEASS and used among others for the evaluation of the algorithms submitted to SiSEC 2011.

The structure of the rest of the paper is as follows. In Section 2, we summarize the computation of the PEASS metrics and highlight the parameters involved in each step. In Section 3, we describe the evaluation protocol and show the effect of each parameter on the prediction performance. We conclude in Section 4.

2 The PEASS Metrics

For a given set of separated sources, we aim to predict the perceived quality of the estimated multichannel *spatial image* $\widehat{\mathbf{s}}_j(t)$ of each source j, *i.e.* its contribution to all mixture channels, relatively to the true spatial image $\mathbf{s}_j(t)$ [11]. The PEASS metrics [3] involve three computation steps outlined in the introduction. In the following, we summarize each step with a focus on the two latter steps, including the internal computations of PEMO-Q which were not detailed in [3].

2.1 Distortion Decomposition

In the first step, the estimation error $\widehat{\mathbf{s}}_j(t) - \mathbf{s}_j(t)$ is split into three components: target distortion $\mathbf{e}_j^{\text{target}}(t)$, interference $\mathbf{e}_j^{\text{interf}}(t)$ and artifacts $\mathbf{e}_j^{\text{artif}}(t)$ such that

$$\widehat{\mathbf{s}}_j(t) - \mathbf{s}_{ij}(t) = \mathbf{e}_j^{\text{target}}(t) + \mathbf{e}_j^{\text{interf}}(t) + \mathbf{e}_j^{\text{artif}}(t). \quad (1)$$

This is achieved by passing the signals through a bank of gammatone filters [6], partitioning the output into overlapping time frames, performing decomposition (1) in each subband and each time frame by least-squares projection onto the subspaces spanned by delayed versions of the true source spatial image signals, and reconstructing time-domain signals by filterbank inversion. Compared with BSS Eval, this step aims to improve the handling of time-varying distortion.

[2] http://bass-db.gforge.inria.fr/peass/

2.2 PEMO-Q Component Saliences

In the second step, the perceptual salience of these components is assessed as

$$q_j^{\mathrm{o}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \mathbf{s}_j) \tag{2}$$

$$q_j^{\mathrm{t}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{target}}) \tag{3}$$

$$q_j^{\mathrm{i}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{interf}}) \tag{4}$$

$$q_j^{\mathrm{a}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{artif}}) \tag{5}$$

where PEMO-Q$(\widehat{\mathbf{x}}, \mathbf{x}) \in [-1, 1]$ is the *perceptual similarity* measured by PEMO-Q between a test signal $\widehat{\mathbf{x}}$ and a reference signal $\mathbf{x}$. Compared with BSS Eval, this step accounts for auditory masking and dynamic compression phenomena.

PEMO-Q first computes *internal auditory representations* $\widehat{X}_i$ and X_i of each channel i of $\widehat{\mathbf{x}}$ and $\mathbf{x}$ via the computational auditory model in [2,1]. This model comes in two versions and consists of:

R1 subband decomposition via a bank of gammatone filters linearly spaced on the equivalent rectangular bandwidth (ERB) scale between $f_{\min}$ and $f_{\max}$,

R2 for each subband, halfwave rectification, first-order autoregressive (AR) lowpass filtering with 1 kHz cutoff, and summation with a threshold a_{thresh},

R3 amplitude compression by five consecutive nonlinear feedback loops emphasizing rapid changes up to a maximum amplitude ratio of $r_{\max}$ for each loop,

R4 either first-order AR lowpass filtering with 8 Hz cutoff (*lowpass version* [2]) or decomposition via a bank of eight first-order AR bandpass filters with center frequencies ranging from 0 to 129 Hz (*modulation version* [1]).

This mimics the effect of haircells in the inner ear and modulation processing in the auditory cortex. The outputs $\widehat{X}_i$ and X_i are either two-dimensional time-frequency representations for the lowpass version or three-dimensional time-frequency-rate representations for the modulation version.

The perceptual similarity between $\widehat{X}_i$ and X_i is then measured by [7,8]

S1 partial assimilation of the two representations in each time-frequency-rate bin (t, f, m) as $\widehat{X}_{itfm} \leftarrow \alpha X_{itfm} + (1-\alpha)\widehat{X}_{itfm}$ if $|\widehat{X}_{itfm}| < |X_{itfm}|$,

S2 computation of the time-varying linear cross-correlation between $\widehat{X}_i$ and X_i over time frames of length l_{corr}[3],

S3 computation of the time-varying root mean square (RMS) amplitude of X_i over time frames of length l_{amp},

S4 computation of the p-th percentile of the cross-correlation series weighted by the RMS amplitude.

This attempts to model the perception of global similarity based on the local similarities between the signals. Finally, the overall scalar similarity PEMO-Q$(\widehat{\mathbf{x}}, \mathbf{x})$ is selected as the minimum of the channel-wise similarity over all channels i.

[3] A slightly distinct processing is applied in the modulation version. See [8] for details.

2.3 Trained Nonlinear Mapping

In the third step, the saliences in (2)–(5) are combined by [3]

M1 optional log-mapping from $[-1, 1]$ to $\mathbb{R}$ via $q_j^k \leftarrow \log((1+q_j^k)/(1-q_j^k))$ [7],
M2 selection of one or more saliences forming a *feature vector* $\mathbf{q}_j$,
M3 transformation into a scalar objective score via a feedforward neural network (NN) [5] composed of n_{lay} layers of n_{neur} neurons trained on subjective scores.

Compared with BSS Eval, this accounts for the different perceptual importance of each distortion component by which artifacts may be heard as more disturbing than interference for instance.

Four different perceptual assessment *tasks* were considered in [3]: global quality, preservation of the target source, suppression of other sources, and absence of additional artificial noise. For each task, a different feature vector was selected and a different NN was trained by minimizing the RMS error between the predicted and the actual subjective opinion scores. This resulted in four metrics called OPS, TPS, IPS and APS, respectively.

3 Effect of the Design Parameters

3.1 Data and Evaluation Procedure

Each processing block from R1 to M3 involves some design parameters listed above. In order to evaluate their effect on the prediction performance, we consider the set of 6400 subjective scores collected in [3] using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) protocol [10]. For each of 10 mixtures and each of the four tasks listed in Section 2.3, 20 subjects were asked to score 8 test sounds, including 4 real-world sounds produced by actual source separation algorithms, one *hidden reference* and 3 *anchors*. The scoring scale ranges from 0 to 100, where larger means better. The anchors are artificial sounds with low quality ensuring that the whole scale is used. For information about the variance of subjective scores and outliers, see [3]. In order to avoid overfitting, a 200-fold cross-validation procedure is used. For each fold, the scores of 19 subjects over 9 mixtures are used for training while testing is performed on the scores of the remaining subject over the remaining mixture. The prediction *accuracy* is assessed via the linear correlation between the predicted scores and the MOS.

3.2 Main Results

The version 1.0 of PEASS relies on the following default parameters of PEMO-Q: modulation version, $f_{\min} = 235$ Hz, $f_{\max} = 14500$ Hz, $a_{\text{thresh}} = 10^{-5}$, $r_{\max} = +\infty$, $\alpha = 0.5$, $l_{\text{corr}} = +\infty$ and $l_{\text{amp}} = +\infty$[4]. The mapping consists of a 1.5-layer NN[5] without input log-mapping. For each mixture and subject, all 8 test sounds

[4] p is irrelevant here due to the use of global correlation ($l_{\text{corr}} = +\infty$).
[5] This term refers to a 2-layer NN with linear output layer.

Table 1. Accuracy after successive parameter optimization stages

Optimization stage	OPS	TPS	IPS	APS	Average
Baseline (version 1.0)	0.799	0.396	0.860	0.896	0.738
Optimal mapping and PEMO-Q version	0.909	0.815	**0.934**	0.870	0.882
Optimal PEMO-Q similarity measure	**0.925**	0.812	0.931	0.924	0.898
Optimal PEMO-Q internal representation	0.922	**0.864**	0.926	**0.925**	**0.909**

were used for training but only the 4 real-world sounds for testing. The best feature vector among 3 or 4 candidates and the best number of neurons were then selected so as to maximize accuracy over the test set [3].

In subsequent experiments, we found this approach to be unsuitable for two reasons. First, the absence of references and anchors in the test set resulted in objective metrics that do not span the whole range from 0 to 100 and thus fail to handle better or poorer sounds than those in that set. Second, the 10 references in the training set drew the NN to better fit scores close to 100 instead of uniformly fitting all scores. In order to avoid these drawbacks, we adopt a consistent approach from now on, whereby all real-world sounds and anchors but only one reference are employed in each training and testing fold. The resulting baseline performance of version 1.0 is displayed in the top row of Table 1.

Due to the large number of design parameters, we optimize these parameters in three successive stages, from higher-level to lower-level ones. For simplicity and computational efficiency, the same parameters are used for all four metrics, except the optimal feature vector and number of neurons which depend on the metric. The resulting performance after each stage is shown in the bottom three lines of Table 1. On average, the accuracy improves from 0.738 to 0.909 when combining all three stages. This huge improvement is mostly due to the optimization of higher-level parameters in the first stage, while the two other stages have less impact. We analyze each stage in more details in the following.

3.3 Detailed Impact of the Mapping and the Version of PEMO-Q

The top half of Table 2 describes the effect of the number of neurons n_{neur} and the feature vectors. By simply selecting the optimal n_{neur} (first row) and features (second row), we greatly improve the performance of the TPS and significantly improve that of the three other metrics, resulting in an average accuracy of 0.868. This is a direct consequence of the consistent training approach discussed above, but also of the fact that all possible feature vectors are tested here. Indeed, none of the optimal feature vectors belongs to the list of candidate vectors previously tested in [3].

Table 3 describes the effect of the other parameters of the nonlinear mapping and the version of PEMO-Q. The use of a 2-layer NN with input log-mapping along with the lowpass version of PEMO-Q appears optimal for all metrics except the APS and yields an optimal average accuracy of 0.882. The corresponding feature vectors are shown in the bottom line of Table 2.

Table 2. Accuracy as a function of the feature vectors and of the baseline or the optimal mapping and version of PEMO-Q, assuming optimal number of neurons and default PEMO-Q parameters

Mapping and version	Feature vector	OPS	TPS	IPS	APS	Average
baseline	baseline	0.799	0.710	0.860	0.905	0.819
	optimal	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{a}}]$ 0.871	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{i}}]$ 0.747	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.935	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{t}}\ q_j^{\mathrm{a}}]$ 0.920	0.868
optimal	baseline	0.901	0.801	0.865	0.834	0.850
	optimal	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{a}}]$ 0.909	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.815	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.934	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{a}}]$ 0.870	**0.882**

Table 3. Accuracy as a function of the version of PEMO-Q, the optional log-mapping and the number of NN layers n_{lay}, assuming optimal feature vectors and numbers of neurons in each case and default PEMO-Q parameters

Version	Log-mapping	n_{lay}	OPS	TPS	IPS	APS	Average
filterbank	no	1.5	0.871	0.747	0.935	0.920	0.868
		2	0.877	0.759	0.912	**0.924**	0.868
	yes	1.5	0.884	0.784	0.928	0.916	0.878
		2	0.884	0.761	0.926	0.909	0.870
lowpass	no	1.5	0.886	0.794	0.940	0.869	0.872
		2	0.877	0.788	0.919	0.878	0.866
	yes	1.5	0.903	0.775	**0.939**	0.839	0.864
		2	**0.909**	**0.815**	0.934	0.870	**0.882**

3.4 Detailed Impact of the PEMO-Q Similarity Measure

After fixing the optimal mapping and version of PEMO-Q, we consider the parameters of the PEMO-Q similarity measure in a second stage. The effect of each parameter is illustrated in Figure 1. Among the tested values, the average accuracy appears to increase with p and decrease with α and l_{corr}. This effect is particularly significant for the APS, which may be due to the nonstationary nature of artifacts calling for local rather than global correlation between the reference and the test representation. The optimal values are $\alpha = 0.25$, $l_{\mathrm{corr}} = 100$ ms, $l_{\mathrm{amp}} = 1$ s and $p = 0.5$, yielding an average accuracy of 0.898.

3.5 Detailed Impact of the PEMO-Q Internal Representation

After fixing the optimal parameters of the similarity measure, we consider the parameters of the internal representation in a last stage. The effect of each parameter is illustrated in Figure 2. Among the tested values, the average accuracy appears to increase with f_{min} and r_{max} and decrease with f_{max} and a_{thresh}. This effect is significant for all metrics except the IPS. The optimal parameters are the default f_{min}, f_{max} and r_{max} along with $a_{\mathrm{thresh}} = 10^{-6}$, yielding an average accuracy of 0.909.

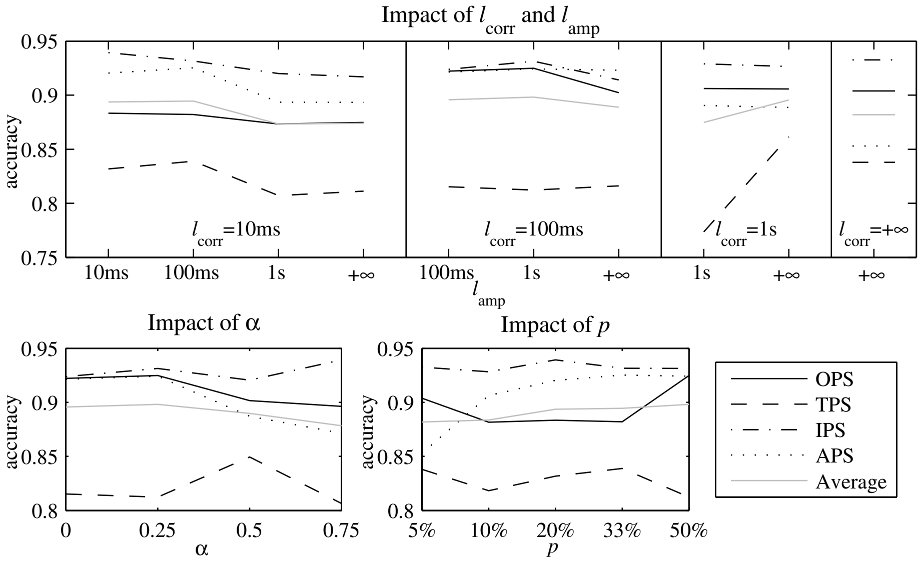


Fig. 1. Accuracy as a function of one of the three parameters (l_{corr}, l_{amp}), α and p given the optimal values of the two other parameters, assuming optimal mapping and version of PEMO-Q and default PEMO-Q internal representation. Note that infinite durations are equivalent to 10 s here, since the duration of the test signals is 5 s.

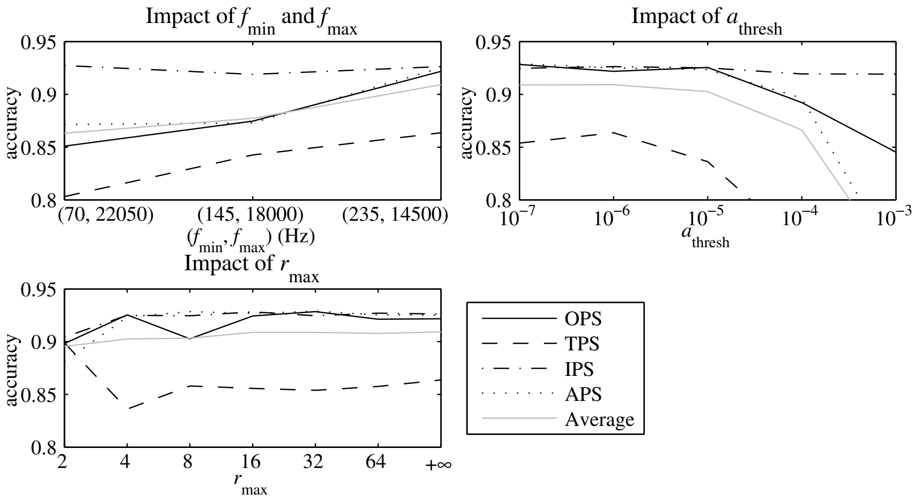


Fig. 2. Accuracy as a function of one of the three parameters (f_{min}, f_{max}), a_{thresh} and r_{max} given the optimal values of the two other parameters, assuming optimal mapping and version of PEMO-Q and PEMO-Q similarity metric

4 Conclusion and Perspectives

We examined the impact of various design parameters over the accuracy of the PEASS metrics. By adopting a consistent training approach together with unconstrained feature selection, we improved the accuracy from 0.738 to 0.868 in a cross-validation setting. By optimizing the parameters of PEMO-Q and the nonlinear mapping, we further increased it to 0.909. These results show that the mapping from the error component saliences to the metrics is crucial, while fine tuning of auditory parameters has smaller impact. The resulting improved metrics have been released as version 2.0 of PEASS and used within SiSEC 2011.

Acknowledgment. This work was supported by the EUREKA Eurostars i3Dmusic project funded by Oseo. We would like to thank Rainer Huber, Volker Hohmann and Valentin Emiya for discussions about this work.

References

1. Dau, T., Kollmeier, B., Kohlrausch, A.: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102(5), 2892–2905 (1997)
2. Dau, T., Püschel, D., Kohlrausch, A.: A quantitative model of the "effective" signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. 99(6), 3615–3622 (1996)
3. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. 19(7), 2046–2057 (2011)
4. Fox, B., Pardo, B.: Towards a model of perceived quality of blind audio source separation. In: Proc. Int. Conf. on Multimedia Expo (ICME), pp. 1898–1901 (2007)
5. Haykin, S.: Neural Networks. Prentice Hall (1999)
6. Hohmann, V.: Frequency analysis and synthesis using a gammatone filterbank. Acta Acustica 88(3), 433–442 (2002)
7. Huber, R.: Objective assessment of audio quality using an auditory processing model. Ph.D. thesis, University of Oldenburg (December 2003)
8. Huber, R., Kollmeier, B.: PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio Speech Lang. Process. 14(6), 1902–1911 (2006)
9. ITU: ITU-R Recommendation BS.1387-1: Method for objective measurements of perceived audio quality (2001)
10. ITU: ITU-R Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems (2003)
11. Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., Duong, N.Q.K.: The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. Signal Processing (to appear)

Musical Audio Source Separation Based on User-Selected F0 Track[*]

Jean-Louis Durrieu[**] and Jean-Philippe Thiran

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS5)
Switzerland
firstname.lastname@epfl.ch,
jean-louis.durrieu@epfl.ch

Abstract. A system for user-guided audio source separation is presented in this article. Following previous works on time-frequency music representations, the proposed User Interface allows the user to select the desired audio source, by means of the assumed fundamental frequency (F0) track of that source. The system then automatically refines the selected F0 tracks, estimates and separates the corresponding source from the mixture. The interface was tested and the separation results compare positively to the results of a fully automatic system, showing that the F0 track selection improves the separation performance.

Keywords: User-guided Audio Source Separation, Graphical User Interface, Non-negative Matrix Factorization.

1 Introduction

Most audio signals are mixtures of different sources, such as a speaker, an instrument, or noise. Applications such as speech enhancement or musical remixing require the identification and the extraction of one such source from the others.

While many existing musical source separation algorithms aim at blindly separating all the different instruments, the aim of the proposed system is to separate the source defined by the user. Let $\{x_t\}_{t=1\dots T}$ be a single-channel mixture signal of duration T. Let $\{v_t\}_t$ and $\{m_{r,t}\}_t$ respectively be the mono signals of the source of interest, usually a singing voice, and of the R remaining sources, *i.e.* the musical accompaniment. These signals are mixed such that:

$$x_t = v_t + \sum_{r=1}^{R} m_{r,t} \qquad (1)$$

The task at hand is to estimate the signal of interest v_t, given user-provided information on the corresponding source. We propose a separation system that

[*] This work was funded by the Swiss CTI agency, project n. 11359.1 PFES-ES, in collaboration with SpeedLingua SA, Geneva, Switzerland.

[**] Corresponding author.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 438–445, 2012.

allows the user to choose the source in an intuitive way, thanks to a representation of the polyphonic pitch content of the audio excerpt. The system was tested by several users on a SiSEC 2011 [10] data set, and the contribution of the users is shown to improve the separation performance compared to the automatic system in [3].

This paper is organized as follows. The relevance of user-guided source separation is first discussed, followed by the presentation of the proposed Graphical User Interface (GUI). The underlying signal model, representation and the algorithm for source separation, mostly derived from previous works from the authors [3], are then briefly stated. The separation guided by the users is thereafter discussed and compared with the automatic separation system. Finally, we conclude with perspectives for the proposed system and concept.

2 User-Guided Source Separation

2.1 Related Works

Audio source separation methods essentially mimic auditory abilities: a human being can focus on the individual instruments of a mixture thanks to their locations, energies, pitch ranges or timbres. With multi-channel signals, such as stereo signals, one can infer spatial information [2], or train models to extract specific sources, even with single-channel signals [1].

The user can be required to provide some meta-information, such as the instrument name in a supervised framework [13], a musical score [5], the time intervals of activity for each instrument [7] or a sung target sound [11]. Musical scores or correct singing are however difficult to acquire, and are often not aligned with the mixture signal.

Expert users can be asked to choose the desired source through its position [14] or selecting components that are played by the desired instrument, thanks to intermediate separation results [15]. In [8], the automatically estimated melody line can be corrected by the user.

2.2 F0-Guided Musical Source Separation

For musical audio excerpts, in particular for vocal sources, many studies have shown the relevance of the fundamental frequency (F0) contours. In [5], the authors use the music score to extract the notes, which helps estimating the actual F0 line of the instrument to remove. In [9], an estimated F0-contour is used to separate the corresponding instrument.

The goal of this work is to study to what extent user input can improve the separation of a specific source. Indeed, some ill-posed issues in the automatic separation problem, using F0 contours, can arise. First, with many interfering sources, it is difficult to automatically decide whether a specific source is present or not. Furthermore, octave and other harmonic-related confusions in the F0 representation can lead to erroneous separations. These errors may easily be corrected by a trained user who uses the context to solve these ambiguities.

3 Graphical User Interface

3.1 Ergonomy Issues

Allowing the user to dynamically choose the desired source requires a representation that clearly displays the possible choices. The waveform would not allow to locate, in time and in frequency, sources that are overlapping in time. Time-frequency representations (TFR), such as the short-term Fourier transform (STFT), are therefore required to visually identify such sources. With a time x-axis and a frequency y-axis, the sinusoids (horizontal lines) or the noises (vertical patterns) corresponding to the desired source easily stand out. Such an approach would however require a significant amount of work, and would not scale well.

Harmonic sources exhibit a characteristic graphical pattern, in the STFT, for each F0: the system in [3] identifies these patterns and provides the energy of the different F0s for each signal frame. From such a representation, the user can select the desired source thanks to its melody line, with little effort.

Furthermore, representing the pitch on the Western musical scale is a visualization that many users can understand. For instance, in [6], Klapuri proposes such a "piano-roll" visualization.

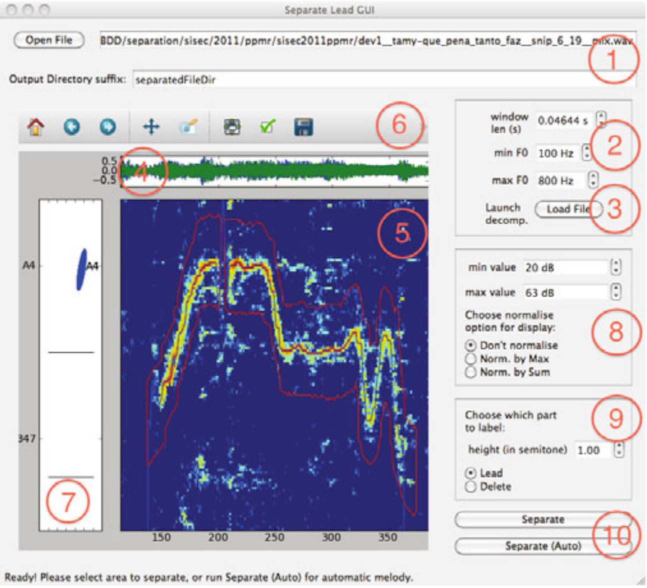
In this article, the mid-level representation introduced in [3] was chosen, because it is easy to configure so as to look like a piano-roll. The method however relies on a fixed dictionary of harmonic spectral shapes, and the proposed system is therefore better suited for the separation of corresponding sources, such as wind instruments, voice or bowed string instruments.

3.2 Practical Solutions

Using Python/NumPy, with the Matplotlib and PyQt4 modules [12], it was possible to design a GUI taking advantage of the representation in [3].

A screen capture of the proposed GUI application is shown on Fig. 1, with the following elements: (1) specify the audio file and the output folder, (2) parameter controls, for the analysis window length, the minimum and maximum candidate F0, (3) a button to "load the file" (computing the decomposition of Sect. 4.1), (4) the waveform of the audio file, (5) the energies for each frame and for each F0 candidate, on which the user can select the melody F0 track (time on x-axis and F0 on y-axis), (6) a toolbar, for zooming and exploring, (7) a representation (musical staff) to indicate the corresponding F0s or notes, (8) normalization choices for the image, (9) buttons toggling between selection ("Lead") and deselection ("Delete"), plus a field to choose the vertical extent of the selection (in semitone), (10) "Separate" and "Separate (Auto)" buttons to launch the separation with or without the user selected track, respectively.

The user can select on (5) a region and thus identify it as a desired F0 range. Once she is finished with her choice, she can start the separation with one of the "Separate" buttons. The underlying mechanisms are further explained in the following section.



BDD/separation/sisec/2011/ppmr/sisec2011ppmr/dev1__tamy-que_pena_tanto_faz__snip_6_19__mix.wav

Output Directory suffix: separatedFileDir

window len (s) 0.04644 s

min F0 100 Hz

max F0 800 Hz

min value 20 dB

max value 63 dB

Choose normalise option for display:

● Don't normalise
○ Norm. by Max
○ Norm. by Sum

Choose which part to label:

height (in semitone) 1.00

● Lead
○ Delete

Ready! Please select area to separate, or run Separate (Auto) for automatic melody.

Fig. 1. GUI for selecting the desired F0 track

4 F0 Representation and Separation Algorithm

The audio signal model presented in [3] is first briefly described. The computation of the F0 representation is then discussed, and at last the user-assisted separation algorithm of the selected source is presented.

4.1 Audio Signal Model

The audio mixture is modelled through its $F \times N$ short-term power spectrum (STPS) matrix $\mathbf{S}$, defined as the power of its STFT $\mathbf{X}$, with F the number of Fourier frequencies and N the number of frames. For simplicity, the model is presented for the single-channel case, but the stereo model of [3] was used for the experiments of this article.

$\mathbf{S}$ is assumed to be the sum of the STPS of the signal of interest $\mathbf{S}^V$ with the residual STPS $\mathbf{S}^M$:

$$\mathbf{S} = \mathbf{S}^V + \mathbf{S}^M \tag{2}$$

$\mathbf{S}^V$ is the element-wise product of a "source" part (F_0) by a "filter" part ($\varPhi$):

$$\mathbf{S}^V = \mathbf{S}^\varPhi \bullet \mathbf{S}^{F_0} \tag{3}$$

All the contributions $\mathbf{S}^\varPhi$, $\mathbf{S}^{F_0}$ and $\mathbf{S}^M$ are further modelled as non-negative matrix products of a spectral shape matrix ($\mathbf{W}^\varPhi$, $\mathbf{W}^{F_0}$ and $\mathbf{W}^M$, with K, U

and R elementary shapes, respectively) by the corresponding amplitude matrix ($\mathbf{H}^{\Phi}$, $\mathbf{H}^{F_0}$ and $\mathbf{H}^{M}$). Finally:

$$\mathbf{S} = \mathbf{W}^{\Phi}\mathbf{H}^{\Phi} \bullet \mathbf{W}^{F_0}\mathbf{H}^{F_0} + \mathbf{W}^{M}\mathbf{H}^{M} \tag{4}$$

In (4), all the parameters of the right hand-side are estimated on the signal, except the matrix $\mathbf{W}^{F_0}$ which is a dictionary of harmonic spectral "comb", parameterized by its F0 frequency. As discussed in [3], a careful choice of the F0s used in that dictionary leads to the desired representation in $\mathbf{H}^{F_0}$: in our case, we chose $\log_2$-spaced F0 values, *i.e.* a scale proportional to the Western musical scale. The number of F0s per semitone is fixed to 16, and the user can choose the extents of the scale, to fit the expected tessitura.

The other parameters are estimated thanks to the Non-negative Matrix Factorization (NMF) algorithm developed in [3]. The resulting matrix $\mathbf{H}^{F_0}$ finally provides the user with an image in which high values correspond to high energies associated with F0 frequencies, as shown on Fig. 1.

4.2 F0 Line Selection and Usage

The user can then, through the GUI of Fig. 1, select the zones containing the F0 values that correspond to the desired melody. A binary mask matrix $\mathcal{H}$, of the same size as $\mathbf{H}^{F_0}$, initialized to 0 everywhere, is updated each time the user draws a curve with the mouse (while holding the left button) over the $\mathbf{H}^{F_0}$ image. All the coefficients along that curve, as well as the coefficients located within a user-defined vertical extent (half a semitone by default) are set to 1. The program superimposes the contour of the selection on the $\mathbf{H}^{F_0}$ image.

Once all the desired tracks have been selected, the user can trigger the separation, given her mask $\mathcal{H}$. Let $\widetilde{\mathbf{H}}^{F_0} = \mathcal{H} \bullet \mathbf{H}^{F_0}$. Assuming the desired source generates smooth melody lines, the melody path is then tracked in $\widetilde{\mathbf{H}}^{F_0}$ with a Viterbi algorithm [4]: the user-defined regions are therefore used to restrict the melody tracking. The user can also refine the chosen regions with a narrower vertical extent, effectively allowing non-smooth melodies if needed.

Finally, the smoothed-out melody line is used to create a refined version of $\widetilde{\mathbf{H}}^{F_0}$, zeroing coefficients lying too far from the melody. The parameters are then re-estimated, using $\widetilde{\mathbf{H}}^{F_0}$ as initial $\mathbf{H}^{F_0}$ matrix. These updated parameters $\{\mathbf{H}^{F_0}, \mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$ are used to compute the separated sources. This second estimation round focuses on voiced patterns, and a third round is done to include more unvoiced elements [3].

4.3 Separating the Selected Source

Wiener filters are used to separate the sources, obtaining the estimates of the STFT $\mathbf{V}$ and $\mathbf{M}$, using [3]:

$$\widehat{\mathbf{V}} = \frac{\mathbf{W}^{\Phi}\mathbf{H}^{\Phi} \bullet \mathbf{W}^{F_0}\mathbf{H}^{F_0}}{\mathbf{W}^{\Phi}\mathbf{H}^{\Phi} \bullet \mathbf{W}^{F_0}\mathbf{H}^{F_0} + \mathbf{W}^{M}\mathbf{H}^{M}} \bullet \mathbf{X} \text{ and } \widehat{\mathbf{M}} = \mathbf{X} - \widehat{\mathbf{V}} \tag{5}$$

The time-domain signals are then retrieved using an inverse STFT (overlap-add procedure).

5 Experiments

5.1 Database and Protocoles

In order to evaluate the usage and the performance of the proposed user-guided source separation system, the development set (5 excerpts) for the SiSEC 2011 "Professionally Produced Music Recordings" task [10] is used.

Three users were asked to try the software. They were all used to handling computer softwares and had some background knowledge in music. The representation and separation principle were explained to each user beforehand. They provided their feedback about the software usage, Sect. 5.2, and the separation scores are discussed in Sect. 5.3. All the systems and users discussed in this section used the same default following parameters: $K = 4$, $U = 577$ (for 16 F0s per semitone, from 100 to 800Hz), $R = 40$, $F = 1025$ (for Fourier tranforms of size 2048, *i.e.* 46.44ms@44.1kHz) and with 25 iterations of the NMF algorithm.

5.2 Usage Feedbacks

The users first tested an early version of the GUI, and their observations were mostly linked with ergonomy issues or missing features (audio feedback, better display). Following their recommendations, we refined the GUI such that the focus was turned to the usability of the F0 representation.

For "easy" songs, with a clearly voiced, sustained vocal track, the F0 representation makes it easy to choose the desired source. However, for near-spoken or weak sources, identifying the vocal tracks was felt as a difficult task: for instance, one user declared not to be able to proceed with two songs for this reason (marked as '-' in Table 1, user #3). In addition, it is interesting to note that other types of sources are also harder to locate (both in time and frequency) than vocals, such as guitar or piano tracks.

5.3 Separation Performance

The Signal-to-Distortion-Ratios (SDRs) for the estimated vocal source for each user (#1, #2 and #3) are reported in Table 1. The results obtained when using the mixture x as the vocals estimation (Mix), when using the fully automatic

Table 1. Source separation results, see text for details

	Mix	Auto		#1		#2		#3		[7]	SiSEC [10]				
Song	-	V	U	V	U	V	U	V	U		S3	S4	S5	S6	S7
dev1_bearlin	-5.3	4.7	4.9	6.1	**6.2**	5.4	5.8	4.7	5.1	-	-	-	3.3	3.2	-
dev1_tamy	0.2	8.6	8.9	10.3	10.1	**10.7**	10.6	8.9	9.2	-	-	-	7.1	8.7	10.3
dev2_another	-3.0	5.1	5.7	5.7	6.2	5.8	**6.6**	-	-	-0.7	-2.9	-2.8	3.4	2.2	-
dev2_fort	-7.2	2.3	2.4	3.2	3.7	3.4	**3.8**	-	-	3.2	-	-5.9	2.5	2.5	-
dev2_ultimate	-7.5	3.2	3.4	4.0	**4.4**	3.8	4.0	3.2	3.5	2.6	-0.9	-10.2	-0.6	1.4	-

system [3] (Auto), those of the algorithm in [7] (from the SiSEC website [10]) and the SiSEC 2011 results for 5 algorithms (S3 to S7) [10] are also given.

The SDRs for the "Auto" and user-guided systems are better than those of the other systems, even the other user-guided system [7]. Furthermore, these examples show that the system is able to take advantage of the user-provided information. Some songs might be more challenging, such as the rap song (dev2_fort), probably because the desired vocal signal is closer to speech than to singing voice. The inadequation of the chosen $\mathbf{H}^{F_0}$ representation for this type of sources was already discussed [3], and the present study shows that even trained users could hardly use it for these signals.

6 Conclusion

A novel user-guided audio source separation system is proposed, allowing the user to easily select a harmonic audio source she desires to separate from a musical audio mixture. The energy of different hypothesized F0 candidates is displayed. Once the user has selected the relevant F0 melody track, the system automatically finds the F0 path maximizing the energy within the regions of interest, estimates the corresponding source and separates it using Wiener filtering and NMF-derived techniques.

The proposed system delegates the source identification to the user, such that there is less ambiguity with the definition of the target source, for the system. The evaluation of the system therefore becomes more relevant. The chosen representation also allows the choice of the source to be straightforward, especially for songs, where the lead singer usually dominates the mixture, providing a fairly readable representation.

The system and GUI could be further improved by adding, for instance, partial separation excerpts allowing the user to listen to what specific chunks of the representation correspond to, before performing the final separation. The user may also want to identify sources from the musical background that are not to be included in the desired source. Such a feature would require to search how to integrate such a prior into the separation stage.

The technique could be used for other applications, such as speech enhancement. The extension to one-speaker signals is straightforward, but many-speakers signals lead to representations that are harder to interprete. Finally, the system could be used as annotation tool: it could assist semi-automatic transcription music signals into musical scores, where an automatic system would infer note boundaries, rhythms, key and time signature from the user inputs.

Acknowledgment. This work was funded by the Swiss CTI agency, project n. 11359.1 PFES-ES, in collaboration with SpeedLingua SA, Lausanne, Switzerland. The authors would also like to thank S. Arberet, F. de Morsier and L. Navarro for their help in testing the GUI.

References

1. Benaroya, L., Bimbot, F., Gribonval, R.: Audio source separation with a single sensor. IEEE Transactions on Audio, Speech and Language Processing 14(1), 191–199 (2006)
2. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non Gaussian signals. IEE Proceedings-F 140(6), 362–370 (1993)
3. Durrieu, J.L., Richard, G., David, B.: A musically motivated representation for pitch estimation and musical source separation. IEEE Journal of Selected Topics on Signal Processing 5(6), 1180–1191 (2011)
4. Durrieu, J.L., Richard, G., David, B., Févotte, C.: Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Transactions on Audio, Speech, and Language Processing 18(3), 564–575 (2010)
5. Han, Y.S., Raphael, C.: Desoloing monaural audio using mixture models. In: Proceedings of the International Conference on Music Information Retrieval, Vienna, Austria, September 23 - 27 (2007)
6. Klapuri, A.: A method for visualizing the pitch content of polyphonic music signals. In: Proceedings of the 10th International Society for Music Information Retrieval Conference, Kobe, Japan, October 26-30, pp. 615–620 (2009)
7. Ozerov, A., Fevotte, C., Blouet, R., Durrieu, J.L.: Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, May 22-27, pp. 257–260 (2011)
8. Pant, S., Rao, V., Rao, P.: A melody detection user interface for polyphonic music. In: Proc. of the National Conference on Communications, Madras, Chennai, India, January 29-31 (2010)
9. Ryynänen, M., Virtanen, T., Paulus, J., Klapuri, A.: Accompaniment separation and karaoke application based on automatic melody transcription. In: IEEE International Conference on Multimedia and Expo., pp. 1417–1420 (2008)
10. SiSEC: Professionally produced music recordings (2011), http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings
11. Smaragdis, P., Mysore., G.: Separation by "humming": User-guided sound extraction from monophonic mixtures. In: Proceedings of IEEE Workshop on Applications Signal Processing to Audio and Acoustics, October 18-21, pp. 69–72 (2009)
12. Tosi, S.: Matplotlib for Python developers. Packt Publishers (2009)
13. Vincent, E.: Musical source separation using time-frequency source priors. IEEE Transactions on Audio, Speech and Language Processing 14(1), 91–98 (2006)
14. Vinyes, M., Bonada, J., Loscos, A.: Demixing commercial music productions via human-assisted time-frequency masking. Convention Paper, The 120th AES Convention, Paris, France, May 20-23 (2006)
15. Wang, B., Plumbley, M.D.: Musical audio stream separation by non-negative matrix factorization. In: Proc. of the DMRN Summer Conference, July 23-24 (2005)

A GMM Sound Source Model for Blind Speech Separation in Under-determined Conditions

Yasuharu Hirasawa, Naoki Yasuraoka, Toru Takahashi,
Tetsuya Ogata, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{hirasawa,yasuraok,tall,ogata,okuno}@kuis.kyoto-u.ac.jp

Abstract. This paper focuses on blind speech separation in under-determined conditions, that is, in the case when there are more sound sources than microphones. We introduce a sound source model based on the Gaussian mixture model (GMM) to represent a speech signal in the time-frequency domain, and derive rules for updating the model parameters using the auxiliary function method. Our GMM sound source model consists of two kinds of Gaussians: sharp ones representing harmonic parts and smooth ones representing nonharmonic parts. Experimental results reveal that our method outperforms the method based on non-negative matrix factorization (NMF) by 0.7dB in the signal-to-distortion ratio (SDR), and by 1.7dB in the signal-to-interference ratio (SIR). This means that our method effectively removes interference coming from other talkers.

Keywords: Blind speech separation, Under-determined condition, GMM sound source model, Auxiliary function method.

1 Introduction

Under-determined blind speech separation is a challenging task in the field of sound separation. Here, the word "blind" means the separation without detailed prior knowledge about mixing model parameters, and "under-determined" refers to the condition in which the number of sounds exceeds that of microphones (Fig. 1). Note that single-channel speech separation is not considered in this paper.

The important characteristic of the human voice is its harmonicity. However, many methods proposed for under-determined source separation, such as the clustering-based method [1], the spatial covariance-based method [2], and the NMF-based source-modeling method [3], do not consider the harmonicity. We believe this is because harmonicity is not easy to handle; estimating parameters for a harmonic model is difficult, especially when there are many speech signals.

In this paper, we propose a GMM sound source model, which can represent the harmonicity in each time frame, and we derive the parameter update rules using an auxiliary function method. Our GMM sound source model also handles nonharmonic parts in the same framework using two kinds of Gaussians: a sharp one for harmonic parts and a smooth one for nonharmonic parts.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 446–453, 2012.

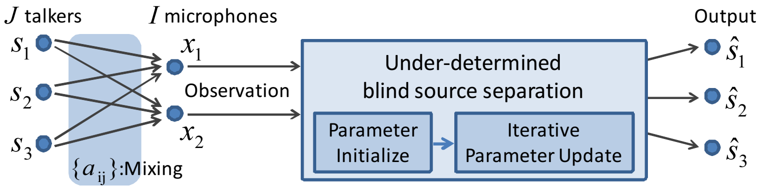


Fig. 1. Our under-determined blind speech separation

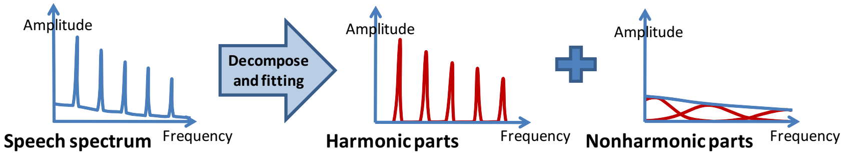


Fig. 2. Proposed GMM sound source model

Since our objective is multiple speech separation, we focus on reducing the interference coming from other talkers. Thus, we accept the existence of distortion because it can be reduced by post-filtering. Experiments reveal that our method achieves the separation with less noise leakage compared to the conventional method using the NMF sound source model.

2 Under-determined Sound Source Separation

2.1 Problem Settings

Table 1 lists the definitions of variables mentioned in this paper. The superscripts ^{H} and ^{N} indicate harmonic and nonharmonic; we do not use ^{H} to express the Hermite transpose. Note that many of the variables are complex-valued because we separate sound mixtures in the time-frequency domain. With these variables, the problem setting for under-determined sound separation is written as follows.

Input	I mixtures of J sound sources: $x_{i,fn}$
Output	Estimated sound source of J sources: $\hat{s}_{j,fn}$
Assumption	Linear time-invariant mixing in the time-frequency domain

2.2 Sound Source Model and Cost Function

We model the sound spectrum of each time frame using two kinds of Gaussians: sharp Gaussians that represent harmonic parts and smooth Gaussians that represent nonharmonic parts. The sharp Gaussians correspond to the sinusoidal representation proposed by McAulay [4]. Figure 2 shows our model visually. Note that although this figure shows the real amplitudes of the spectrum, our model considers the complex amplitudes of the spectrum.

Table 1. Definition of variables

Indices	
i, j, f, n	Index of microphone / talker / frequency bin / time frame
I, J, F, N	Number of microphones / talkers / frequency bins / time frames
m_h, m_n	Index of harmonic overtone / nonharmonic overtone
M_H, M_N	Number of harmonic overtones / nonharmonic overtones
Signals	
$\hat{s}_{j,fn}$	Estimated sound source ($\in \mathbb{C}$)
$x_{i,fn}, \hat{x}_{i,fn}$	Observed / Expected spatial sound image ($\in \mathbb{C}$)
Parameters	
$a_{ij,f}$	Element of mixing matrix ($\in \mathbb{C}$)
$F^H_{0,j,n}$	Fundamental frequency of harmonic Gaussian
p^H_{j,n,m_h}	Complex amplitude of harmonic Gaussian ($\in \mathbb{C}$)
p^N_{j,n,m_n}	Real amplitude of nonharmonic Gaussian
$\phi^N_{j,fn}$	Phase of nonharmonic component ($\in \mathbb{C}$)
Others	
t, T	Symbol to indicate harmonic/nonharmonic ($t \in T = \{H, N\}$)
F^N_0	Base frequency of nonharmonic Gaussian (const.)

More concretely, our sound source model is formalized as follows:

$$\hat{s}_{j,fn} = \textstyle\sum_{m_h} p^H_{j,n,m_h} g^H_{j,fn,m_h} + \sum_{m_n} p^N_{j,n,m_n} g^N_{f,m_n} \phi^N_{j,fn}. \tag{1}$$

Each harmonic Gaussian is transformed from a sinusoidal wave by discrete Fourier transformation with the Gaussian window. Thus time-frequency components in one harmonic Gaussian should have a common phase, hence the peak height of it, p^H_{j,n,m_h}, is complex-valued. On the other hand, the nonharmonic Gaussian does not have such regularity, hence the peak height of it, p^N_{j,n,m_n}, is real-valued, and phase information $\phi^N_{j,fn}$ is added independently.

Harmonic and nonharmonic Gaussians are defined as follows:

$$g^H_{j,fn,m_h} = \exp\left(-\frac{(f - m_h F^H_{0,j,n})^2}{2{\sigma^H}^2}\right), \quad g^N_{f,m_n} = \exp\left(-\frac{(f - m_n F^N_0)^2}{2{\sigma^N}^2}\right). \tag{2}$$

Note that the Gaussians for nonharmonic parts are independent from source number j and time frame n.

From the assumption that the mixing process is time-invariant and linear in the time-frequency domain, we calculate the expected observed signal as follows:

$$\hat{x}_{i,fn} = \textstyle\sum_j a_{ij,f} \hat{s}_{j,fn}. \tag{3}$$

In addition, we employ the following scaling constraints

$$\textstyle\sum_i |a_{ij,f}|^2 = 1, \qquad |\phi^N_{j,fn}| = 1, \tag{4}$$

because there is scaling arbitrariness between $a_{ij,f}$ and p^H_{j,n,m_h}, between $a_{ij,f}$ and p^N_{j,n,m_n}, and between p^N_{j,n,m_n} and $\phi^N_{j,fn}$.

Finally, we define the cost function as the square distance between the real observed signal and the expected observed signal.

$$C = \textstyle\sum_{ifn} \ |x_{i,fn} - \hat{x}_{i,fn}|^2 \qquad (5)$$

Since $\hat{x}_{i,fn}$ contains summation terms which have Gaussians in it, straightforward derivation of parameter update rules is difficult. To overcome this problem, we use the auxiliary function method described in the next subsection.

2.3 Auxiliary Function Method

We use the auxiliary function method [5] to analytically derive update rules for the model parameters. The basic idea of this method is to introduce auxiliary function $C^+(\theta, \psi)$, whose lower bound is the same as that of the original cost function $C(\theta)$, and to derive update rules on the auxiliary function.

More formally, we use the auxiliary function satisfying the followings:

1. $C(\theta) = \min_\psi \ C^+(\theta, \psi)$
2. $\psi_{new} = \text{argmin}_\psi \ C^+(\theta, \psi)$ is analytically solvable
3. $\theta_{new} = \text{argmin}_\theta \ C^+(\theta, \psi)$ is analytically solvable

where ψ is an auxiliary variable. Using these three properties, we update θ using the property 3 following the update of ψ using the property 2, and do the same updates iteratively. These two steps monotonically decrease the value of the original cost because $C(\theta) = C^+(\theta, \psi_{new}) \geq C^+(\theta_{new}, \psi_{new}) \geq C(\theta_{new})$.

2.4 Derivation of Update Rules

We apply the auxiliary function method to Eq. (5) and derive update rules for each parameter. The basic idea of derivation was proposed by Kameoka [6], and the following is its expansion to multi-channel observations. In addition, parameter updates for nonharmonic parts are newly introduced. Please refer the paper [6] for the details about the auxiliary functions used in this paper.

By substituting Eqs. (1) and (3) into Eq. (5), we have

$$C = \textstyle\sum_{ifn} \left| x_{i,fn} - \sum_{jTm_t} \ a_{ij,f} p^T_{j,n,m_t} g^T_{j,fn,m_t} \phi^T_{j,fn} \right|^2, \qquad (6)$$

where $T \in \{H, N\}$ is a variable to select harmonic (H) or nonharmonic (N) parts, and m_t indicates m_h or m_n depending on T. Here, $\phi^H_{j,fn}$ and g^N_{j,fn,m_n} are introduced for simplicity; $\phi^H_{j,fn} = 1$ and $g^N_{j,fn,m_n} = g^N_{f,m_n}$ are satisfied.

Using the first auxiliary function, we have

$$C^+ = \textstyle\sum_{ijfnTm_t} \ {\beta^T_{ij,fn,m_t}}^{-1} \left| \bar{\alpha}^T_{ij,fn,m_t} - a_{ij,f} p^T_{j,n,m_t} g^T_{j,fn,m_t} \phi^T_{j,fn} \right|^2, \qquad (7)$$

where $\bar{\alpha}^T_{ij,fn,m_t} = \alpha^T_{ij,fn,m_t} x_{i,fn} \ \ (\in \mathbb{C})$. Note that $\alpha^T_{ij,fn,m_t} (\in \mathbb{C})$ is an auxiliary variable and β^T_{ij,fn,m_t} is its parameter satisfying the following constraints:

$\sum_{jTm_t} \beta^T_{ij,fn,m_t} = 1, \quad 0 < \beta^T_{ij,fn,m_t} \in \mathbb{R}$. The following equation minimizes the auxiliary function and satisfies the property 1 mentioned in subsection 2.3.

$$\alpha^T_{ij,fn,m_t} = {x_{i,fn}}^{-1}\left\{a_{ij,f}\, p^T_{j,n,m_t}\, g^T_{j,fn,m_t}\, \phi^T_{j,fn} + \beta^T_{ij,fn,m_t}\left(x_{i,fn} - \hat{x}_{i,fn}\right)\right\} \tag{8}$$

Update rules for mixing matrix and amplitudes. We derive parameter update rules using a partial derivation of Eq. (7). Calculating $\partial C^+/\partial a^*_{ij,f} = 0$, $\partial C^+/\partial p^{H*}_{j,n,m_h} = 0$, and $\partial C^+/\partial p^N_{j,n,m_n} = 0$ yields the following update rules.

$$a_{ij,f} = \frac{\sum_{nTm_t} {\beta^T_{ij,fn,m_t}}^{-1} \bar{\alpha}^T_{ij,fn,m_t} p^{T*}_{j,n,m_t} g^T_{j,fn,m_t} \phi^{T*}_{j,fn}}{\sum_{nTm_t} {\beta^T_{ij,fn,m_t}}^{-1} \left|p^T_{j,n,m_t}\right|^2 {g^T_{j,fn,m_t}}^2} \tag{9}$$

$$p^H_{j,n,m_h} = \frac{\sum_{if} {\beta^H_{ij,fn,m_h}}^{-1} \bar{\alpha}^H_{ij,fn,m_h} a^*_{ij,f} g^H_{j,fn,m_h}}{\sum_{if} {\beta^H_{ij,fn,m_h}}^{-1} \left|a_{ij,f}\right|^2 {g^H_{j,fn,m_h}}^2} \tag{10}$$

$$p^N_{j,n,m_n} = \frac{\sum_{if} {\beta^N_{ij,fn,m_n}}^{-1} \Re e\left[\bar{\alpha}^N_{ij,fn,m_n} a^*_{ij,f} g^N_{f,m_n} \phi^{N*}_{j,fn}\right]}{\sum_{if} {\beta^N_{ij,fn,m_n}}^{-1} \left|a_{ij,f}\right|^2 {g^N_{f,m_n}}^2} \tag{11}$$

Here, * indicates a complex conjecture and $\Re e[..]$ is a function to take real part.

Update rule for phase information. When we expand the square-norm term in Eq. (7), only the following term contains $\phi^N_{j,fn}$ from the constraint Eq. (4): $-2\Re e\left[\sum_{im_n} {\beta^N_{ij,fn,m_n}}^{-1} \bar{\alpha}^N_{ij,fn,m_n} a^*_{ij,f} p^N_{j,n,m_n} g^N_{f,m_n} \phi^{N*}_{j,fn}\right]$.

From the constraint Eq. (4), the only thing we can change is the phase of $\phi^N_{j,fn}$. To minimize the above term, we can obtain the following update rule:

$$\phi^N_{j,fn} = phase\left(\sum_{im_n} {\beta^N_{ij,fn,m_n}}^{-1} \bar{\alpha}^N_{ij,fn,m_n} a^*_{ij,f} p^N_{j,n,m_n} g^N_{f,m_n}\right), \tag{12}$$

where $phase(..)$ is the function to return the phase information of the argument.

Update rule for fundamental frequencies. When we expand the square-norm term in Eq. (7), two terms are related to $F^H_{0,j,n}$. However, one term will be independent from $F^H_{0,j,n}$ using the Gaussian integration when the free parameter β^H_{ij,fn,m_h} is defined independently from f. This is because the harmonic Gaussian g^H_{j,fn,m_h} has strong locality, and we can assume $a_{ij,f} g^H_{j,fn,m_h} \approx a_{ij,f_\#} g^H_{j,fn,m_h}$, where $f_\#$ indicates the true frequency.

Thus, the term containing $F^H_{0,j,n}$, which appeared in g^H_{j,fn,m_h}, is only the following one: $-2\sum_{ifm_h} {\beta^H_{ij,fn,m_h}}^{-1} \Re e\left[\bar{\alpha}^H_{ij,fn,m_h} a^*_{ij,f} p^{H*}_{j,n,m_h}\right] g^H_{j,fn,m_h}$. We refer to this as C_F, and using the second auxiliary function, we obtain C^+_F as follows:

$$C^+_F = 2\sum_{ifm_h} {\beta^H_{ij,fn,m_h}}^{-1} \Re e\left[\bar{\alpha}^H_{ij,fn,m_h} a^*_{ij,f} p^{H*}_{j,n,m_h}\right] \times$$
$$e^{-\gamma_{j,fn,m_h}}\left((f - m_h F^H_{0,j,n})^2/(2{\sigma^H}^2) - \gamma_{j,fn,m_h} - 1\right). \tag{13}$$

Note that $\gamma_{j,fn,m_h} (\in \mathbb{R})$ is a new auxiliary variable, and

$$\gamma_{j,fn,m_h} = (f - m_h F^H_{0,j,n})^2 / (2{\sigma^H}^2) \tag{14}$$

minimizes the auxiliary function and satisfies property 1 of auxiliary functions.

Finally, we calculate $\partial C^+_F / \partial F^H_{0,j,n} = 0$ and get the following update rule:

$$F^H_{0,j,n} = \frac{\sum_{ifm_h} {\beta^H_{ij,fn,m_h}}^{-1} \Re e\left[\bar{\alpha}^H_{ij,fn,m_h} a^*_{ij,f} p^{H*}_{j,n,m_h}\right] e^{-\gamma_{j,fn,m_h}} f m_h}{\sum_{ifm_h} {\beta^H_{ij,fn,m_h}}^{-1} \Re e\left[\bar{\alpha}^H_{ij,fn,m_h} a^*_{ij,f} p^{H*}_{j,n,m_h}\right] e^{-\gamma_{j,fn,m_h}} {m_h}^2}. \tag{15}$$

2.5 Parameter Update Ordering

Using the above update rules, we can update parameters in the following order.

1. Update α^T_{ij,fn,m_t} using Eq. (8)
2. Update p^H_{j,n,m_h} and p^N_{j,n,m_n} using Eqs. (10) and (11), respectively
3. Update $\phi^N_{j,fn}$ using Eq. (12)
4. Update $a_{ij,f}$ using Eq. (9) and normalize it to satisfy Eq. (4)
5. Update γ_{j,fn,m_h} using Eq. (14) and update $F^H_{0,j,n}$ using Eq. (15)

This order is just a example. As we mentioned in 2.3, we update the parameters after updating the auxiliary variable used in its update rule. If we comply with this, the other ordering is arbitrary. Experimentally, p^H_{j,n,m_h} and p^N_{j,n,m_n} should be updated frequently because they are sensitive to other parameters.

3 Experiments

To reveal the capability of the proposed method, we separate the sound mixtures given in the community-based Signal Separation Evaluation Campaign (SiSEC) [7]. The testsets used here are `dev1_female3_liverec_130ms_1m` and `dev3_female4_srec_130ms_50cm`, which are the stereo and 3-channel sound mixtures containing 3 and 4 females' simultaneous utterances, respectively. Both are recorded in live condition whose reverberation time is 130ms.

We evaluate the results using the energy ratio criteria: signal-to-distortion ratio (SDR), image-to-spatial-distortion ratio (ISR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [8], as used in SiSEC's evaluation. As we mentioned in the introduction, our goal is to achieve the separation with low noise leakage. Thus, we focus on SIR, the noise reduction measure, as well as SDR, the overall performance measure.

Initial parameters are calculated by the following non-supervised method. First, we roughly estimate $a_{ij,f}$ using the interaural phase difference (IPD). Second, we make time-frequency mask for each source using IPD information, and estimate $F^H_{0,j,n}$ using masked observation signal as follows:

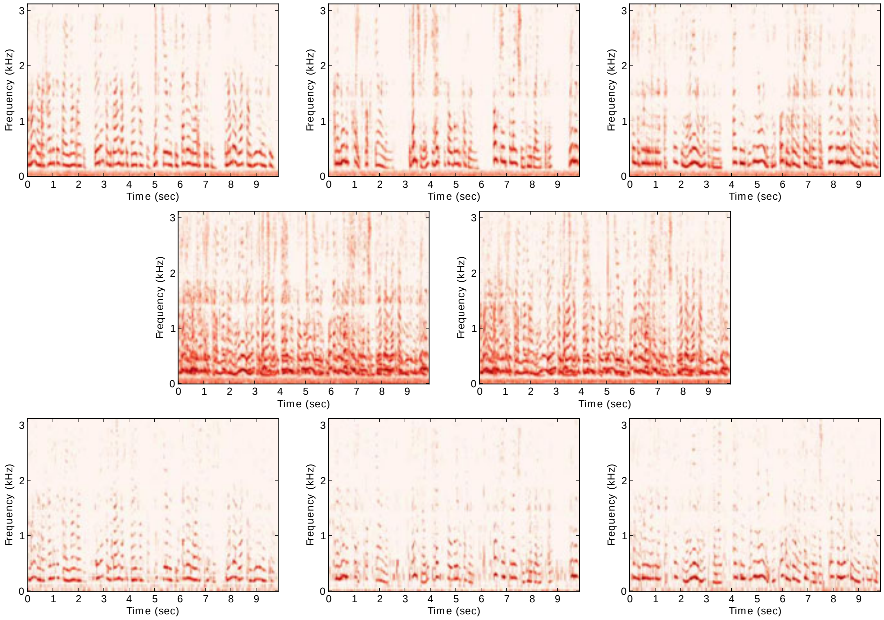


Fig. 3. (Top) Clean speech, (Middle) Sound mixture, (Bottom) Separation result

Table 2. Separation results (dB)

		3 females / 2 mics				4 females / 3 mics			
model for separation	F0	SDR	ISR	SIR	SAR	SDR	ISR	SIR	SAR
NMF sound source model	-	4.3	8.8	6.8	**8.7**	-	-	-	-
GMM sound source model	Given	**6.3**	**11.6**	**11.9**	6.9	5.5	9.9	11.0	6.3
GMM sound source model	Estimated	**5.0**	**9.3**	**8.5**	6.4	4.2	7.9	7.7	5.8

1. Prepare the candidates of F0 for each time frame (5 Hz interval)
2. Update only $F^{H}_{0,j,n}$ and p^{H}_{j,n,m_h} using Eqs. (8), (10), (14), and (15)
3. Choose the F0 that achieves the minimum cost on Eq. (5)

Since this is a very naive method, we plan to improve it in the near future. Initial values of other three parameters, p^{H}_{j,n,m_h}, p^{N}_{j,n,m_n}, and $\phi^{N}_{j,fn}$, are set to be one.

To evaluate the GMM sound source model, we separate the sound mixtures in two conditions; $F^{H}_{0,j,n}$ is initialized by (1) the values estimated by above method and (2) oracle data annotated manually. STFT frame length and shift width are set 1024 points (64 ms) and 256 points (16 ms), respectively. We apply Wiener filtering after the iterations. To compare our separation results with others, We use the method based on NMF sound source model [3]. We use the source code published at its author's website, and modify it in order to use the same initial mixing matrix. We choose the number of basis to realize the best performance. Since its source code is specialized for stereo observation, the second testset, which is 3-channel observation, is separated only by our method.

Figure 3 shows the spectrograms of the clean speech, the sound mixture, and the separation result of our method. Table 2 gives the average separation results, and shows that our method outperforms in the noise reduction measure, SIR, and the overall performance measure, SDR, especially when the F0 is given. This implies that developing more accurate multi-F0 estimation method improves the performance of our separation method.

4 Conclusion and Future Work

In this paper, we aimed to realize the under-determined blind speech separation with less noise leakage. We proposed a GMM sound source model, which consists of two kinds of Gaussians, and derived its parameter update rules using the auxiliary function method. Experimental results show that our method achieves the separation with less interference from other talkers.

The most important task in the future is to develop a robust multi-channel multi-F0 estimation method. Also, we believe that modification of nonharmonic model decreases the noise leakage and increases the overall separation accuracy.

Note that the results of SiSEC 2011 is opened after the notification of this paper, and now there are methods that realizes much higher SDR than ours. We also try to integrate our GMM model to the other methods in the near future.

Acknowledgment. This research was partially supported by Grant-in-Aid for Scientific Research (S), the Global COE Program, and HRI-JP.

References

1. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. on ASLP 19(3), 516–527 (2011)
2. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. IEEE Trans. on ASLP 18(7), 1830–1840 (2010)
3. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. on ASLP 18(3), 550–563 (2010)
4. McAulay, R.J., Quatieri, T.F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. on ASSP 34(4), 744–754 (1986)
5. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems 13, 556–562 (2001)
6. Kameoka, H., Ono, N., Sagayama, S.: Auxiliary function approach to parameter estimation of constrained sinusoidal model for monaural speech separation. In: Proc. of ICASSP 2008, pp. 29–32 (2008)
7. Araki, S., Ozerov, A., Gowreesunker, V., Sawada, H., Theis, F., Nolte, G., Lutter, D., Duong, N.Q.K.: The 2010 Signal Separation Evaluation Campaign (SiSEC2010): Audio Source Separation. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 114–122. Springer, Heidelberg (2010)
8. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. on ASLP 14(4), 1462–1469 (2006)

Model-Driven Speech Enhancement for Multisource Reverberant Environment (Signal Separation Evaluation Campaign (SiSEC) 2011)*

Pejman Mowlaee[1], Rahim Saeidi[2], and Rainer Martin[1]

[1] Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum (RUB)
[2] Centre for Language and Speech Technology, Radboud University Nijmegen
{pejman.mowlaee,Rainer.Martin}@rub.de, rahim.saeidi@let.ru.nl

Abstract. We present a low complexity speech enhancement technique for real-life multi-source environments. Assuming that the speaker identity is known a priori, we present the idea of incorporating speaker model to enhance a target signal corrupted in non-stationary noise in a reverberant scenario. Based on experiments, this helps to improve the limited performance of noise-tracking based speech enhancement methods under unpredictable and non-stationary noise scenarios. Using pretrained speaker model captures a constrained subspace for target speech and is capable to provide enhanced speech estimate by rejecting the non-stationary noise sources. Experimental results on Signal Separation Evaluation Campaign (SiSEC) showed that the proposed approach is successful in canceling the interference signal in the noisy input and providing an enhanced output signal.

Keywords: Model-driven, Speaker model, SiSEC.

1 Introduction

Most of the current speech enhancement methods rely on noise and speech power estimates typically provided by a noise estimation (NE) scheme in a decision-directed manner and a signal-to-noise ratio (SNR) as speech power estimator. These methods assume that the noise signal shows less variations in its second order moment compared to the speech and therefore are limited in performance when the interfering noise signal has a characteristic close to the speech [1–3]. In real-life, the noise signal is time-varying and unpredictable, hence, the stationary assumption is an unrealistic one. Accordingly, the conventional speech enhancement techniques successfully improve those noisy regions where the noise signal

* The work of Pejman Mowlaee was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699. The work of Rahim Saeidi was funded by the European Community's Seventh Framework Programme (FP7 2007-2013) under grant agreement no. 238803.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 454–461, 2012.

is more or less stationary and does not reject the noise burst or non-stationary noise sources. In an unpredictable and non-stationary noise scenario as in [4], stationarity assumption on noise statistics fail. Instead, we need to rely on a strong speech model in order to accurately identify reliable features.

In this paper, we seek to propose a model-driven approach for speech enhancement in reverberant non-stationary noise environment. Incorporating speaker model information into the speech enhancement framework helps to improve the limited performance of the noise-tracking based speech enhancement methods under unpredictable or non-stationary noise scenarios. The experimental results show that the proposed approach compared to other well known speech enhancement methods, provides a reasonable interference rejection capability especially in presence of non-stationary interference signals.

2 Problem Statement

For an acoustic environment, in general, the relationship between the observed time domain samples of noisy data z_n and clean speech x_n is modeled as:

$$z_n = x_n * h_n + d_n^{\text{st}} + d_n^{\text{nst}}, \tag{1}$$

where h_n is the time-domain channel impulse response through which the clean source signal gets distorted and $n \in [0, N-1]$ is the sample index with N the window length. The corrupting noise is comprised of two terms: d_n^{st} that accounts for the stationary part and d_n^{nst} as non-stationary part. Hence, the distorted speech signal experiences two phenomena: h_n accounting for reverberation and $d_n = d_n^{\text{st}} + d_n^{\text{nst}}$ accounting for interference signal which itself is comprised of two parts. As frequency domain symbols, in the rest of the paper, the spectral vector of length K for the unknown clean, the received signal at microphone and the noise signals are denoted by $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{D}$, respectively, with X_k, Z_k and D_k as their kth frequency component.

The problem is formulated as follows; given the noisy speech signal z_n, find the speech estimate which best explains the observed signal according to an optimality criterion. In this paper, we solve the problem under the constraint that the speech estimate are to be selected from a pre-trained speaker codebook used to capture the characteristics of the target source. To train the speaker models we employ the spectral magnitude as the selected feature extracted from the training set data. The speaker model is denoted as $C = \{\boldsymbol{\mu}_i | i \in [1, M]\}$ and $\boldsymbol{\mu}_i \in \mathrm{R}^K$ where M is the model order of the quantizer used, i is an index to refer to the ith codevector denoted by $\boldsymbol{\mu}_i$ composed of K components $\mu_{k,i}$, and k is the frequency bin index with $k \in [0, K-1]$ with K as the number of DFT points. Posing the aforementioned problem in a model-based one, given the observed noisy data, the problem is to find the maximum likelihood (ML) estimate of the speech signal under the constraint that the speech spectrum is a member of the pre-trained codebook.

3 Proposed Method

The proposed method integrates two mechanisms: noise estimation and speaker model. The first mechanism copes with quasi-stationary noise in the background. The second mechanism benefits the good interference rejection capability of model-driven single-channel speech separation systems [5]. For an overview on the large amount of literature about speech separation using codebook see [5,6] and the reference therein. For taking into account the effect of the distortion channel h_n, we train the speaker models on the clean reverberated speech data, *a priori* available for each speaker in the dataset [4]. The reason why we emply the recerberated data instead of the clean ones is because in this way we consider the filter h_n as a part of the speaker models trained per speakers and as a result, the pre-trained codebooks capture a constrained subspace for the spectral shape of target speech (speaker characteristics) as well as learn the average room impulse responses (channel). The codebooks were trained in spectrum amplitude domain. The proposed approach consists of three fundamental steps: (1) estimation of stationary noise spectrum, (2) ML speech estimation with codebook constraint, and (3) signal reconstruction. In the following, each step is explained in detail.

3.1 Noise Spectrum Estimation

To estimate the power spectrum of stationary part ($|D_k^{\mathrm{st}}|^2$), we use the noise-estimation algorithm in [7] proposed for highly non-stationary environments. The periodogram of the noisy data $|Z_k|^2$ is smoothed by recursively updating the first order recursive equation. Based on pilot experiments, in this paper, we set the key parameters in [7] as: $\eta = 0.7$, $\gamma = 0.998$ and $\alpha_d = 0.95$. The estimated noise power spectrum estimate $|\hat{D}_k^{st}|$ is sent to next stage.

3.2 ML Speech Estimation

Based on the noise estimate $|\hat{D}_k^{st}|$ found in previous stage, we produce binary mask $\hat{G}_{k,0}$ as below

$$\hat{G}_{k,0} = \begin{cases} 1\,, & |D_k^{\mathrm{st}}| < |Z_k| \\ 0\,, & \text{Otherwise} \end{cases} . \tag{2}$$

The mask mostly rejects the speech pauses and noise only regions in the observed noisy signal. This is needed to avoid modeling these regions using the codebook inference (presented in the following). The filtered signal, $G_{k,0}|Z_k|$, is then given to pre-trained speaker model to find the maximum likelihood speech estimate $\boldsymbol{\mu}_{i^*}$. We assume that probability density function, $f_{k,i}(x)$, for the kth frequency component of speech follow a normal distribution, i.e. $f_{k,i}(x) = N(x_i; \mu_{k,i}, \sigma_{k,i})$ with $\mu_{k,i}$ and $\sigma_{k,i}$ as the mean and variance for the kth frequency component for the ith state in the speaker model. The ML speech estimate is then obtained by selecting the codevector with the largest log-likelihood $i^* = \arg\max_{\mu_{k,i} \in C} \ln f_{k,i}(|Z_k|)$ given by

$$\boldsymbol{\mu}_{i^*} = \min_{\mu_{k,i} \in C} \sum_{k=0}^{K-1} \left[\frac{(G_{k,0}|Z_k| - \mu_{k,i})^2}{2\sigma_{k,i}^2} + \ln(\sqrt{2\pi}\sigma_{k,i}) \right]. \tag{3}$$

For sake of simplicity, in this work, we assume that the variance terms $\sigma_{k,i}^2$ are equal for all components and dimensions. Training a codebook which only fits the mean vectors $\boldsymbol{\mu}_i$ with $i \in [1, M]$, from the minimization in (3) we obtain the ML speech estimate as $|\hat{\mathbf{X}}^{\mathrm{ML}}| = \boldsymbol{\mu}_{i^*}$.

3.3 Reconstructing the Separated Signals

Given $|\hat{D}_k^{\mathrm{st}}|^2$ as the noise power spectrum estimate (step one), and $|\hat{\mathbf{X}}^{\mathrm{ML}}|$ as an estimate for $h_n * x_n$ (step two), from (1) we find an estimate of d_n^{nst} as

$$\hat{d}_n^{\mathrm{nst}} = z_n - \hat{x}_n^{\mathrm{ML}} - \hat{d}_n^{\mathrm{st}}. \tag{4}$$

To recover the unknown speech and noise signals, we produce the following mask

$$\hat{G}_{k,1} = \begin{cases} \frac{|\hat{X}_k^{\mathrm{ML}}|}{\sqrt{|\hat{Z}_k^{\mathrm{w}}|^2 + |\hat{D}_k^{\mathrm{nst}}|^2}} & , |\hat{X}_k^{\mathrm{ML}}| > |\hat{D}_k^{\mathrm{st}}| \\ G_{\min} & , \text{Otherwise} \end{cases} ,$$

where we define $|\hat{D}_k^{\mathrm{nst}}| = (1 - \tilde{G}_k)|Z_k|$ as non-stationay noise estimation with $\tilde{G}_k = \frac{|\hat{Z_k^w}|}{|Z_k|}$ and $|\hat{Z_k^w}| = \sqrt{|\hat{X}_k^{\mathrm{ML}}|^2 + |\hat{D}_k^{\mathrm{st}}|^2}$. The speech absence gain is set to $20 \log_{10} G_{min} = -25\text{dB}$ as suggested by [8]. Using K-point inverse DFT, the time domain separated speech $\hat{x}_n$is obtained as

$$\hat{x}_n = \mathrm{DFT}^{-1}\{\hat{G}_{k,1} Z_k\}. \tag{5}$$

4 Experiments and Results

System Setup and Dataset. In the proposed method the window size was 32 msec and the frame shift was set to 8 msec with the sampling frequency set to 16 kHz.

Experimental Results on SiSEC. As processing strategy, we process each mixture (isolated sentence) alone. The proposed method was applied to the development data and the test data (24+24 utterances) [1]. The computing station info are as follow: RAM: 4.00 GB, CPU: Intel(R) Core(TM) i5 3.2 GHz. The averaged running time for the algorithm is 1.84 × RT. As our benchmark techniques, we include well-known speech enhancement techniques in order to study the effectiveness of the proposed model-driven idea versus other enhancement techniques relying on noise-trackers. The methods are:

[1] The enhanced wave files can be found at cs.joensuu.fi/pages/saeidi/Sisec2011_wavFiles.tar.gz

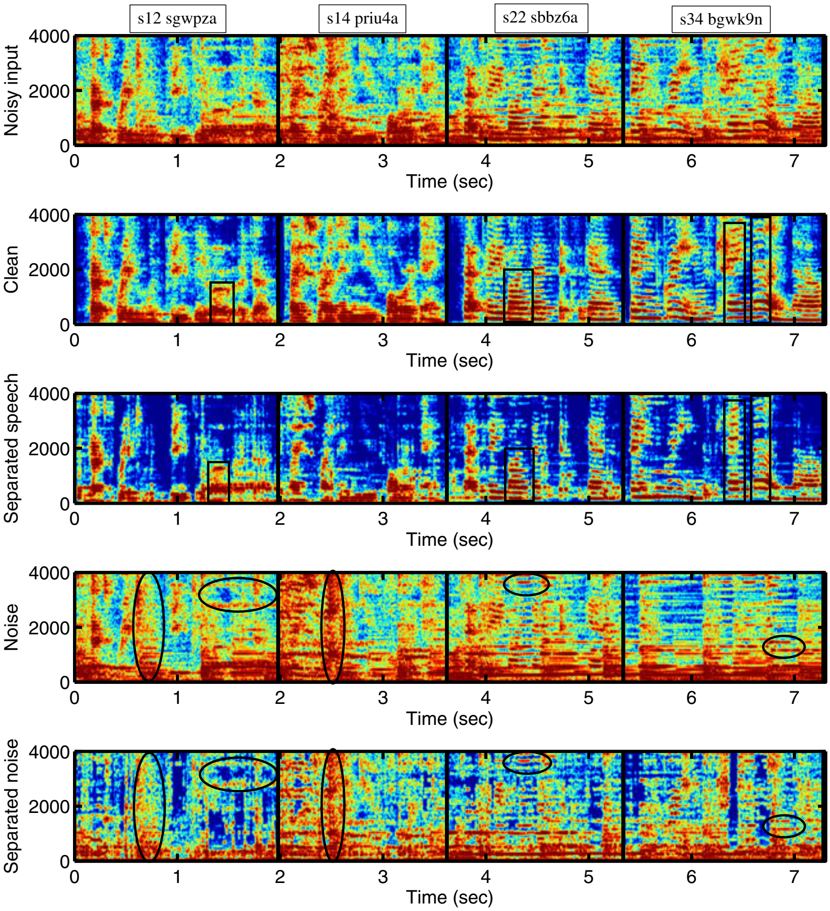


Fig. 1. Showing spectrogram of clean, mixture, separated speech and interference signals using the proposed method. The results are shown for four utterances mixed at 3 dB (transcriptions are shown on the top panel). Absolute improvement compared to noisy mixture data: 5.7 (dB) SDR, 9.6 (dB) SIR, 4.8 (dB) SNR, and 2.9 (dB) SSNR.

- M1 (VAD+LSA): MMSE spectral amplitude estimator (STSA) [1] with a voice activity detector (VAD)-based noise tracker.
- M2: (VAD+MMSE) log-spectral amplitude estimator (LSA) [2] with a VAD-based noise tracker.
- M3 (MMSE+GGD): speech enhancement described in [3].
- M4 (IBM): ideal binary mask [9] known for its maximum SNR performance.

Both M1 and M2 use the decision-directed (DD) approach for a priori SNR estimation [1]. The noise tracker used for both M1 and M2 update its noise estimaete according to a VAD-based decision with a threshold equal to 0.15 was used. In M3 [3] , to estimate the clean speech DFT coefficients, the magnitude-DFT MMSE estimator is used which assumes that speech magnitude-DFT coefficients are generalized Gamma distributed (GGD) with parameters $\gamma = 1$ and $\nu = 0.6$ [10]. For noise tracking it uses the MMSE noise PSD tracker [3].

Improvement in Spectrogram. A useful subjective quality measure is the assessment of spectrograms. Figure 1 is an example to give indications about how the proposed method deals with background noise (as stationary part) and interference signals (as non-stationary part) by integrating noise estimation with codebook constraint. The mixed signal is selected from the development dataset containing speaker 23 corrupted at signal-to-noise ratio of 3 dB. In Figure 1, the top panel shows the clean signal, the second row depicts the noisy input signal, the third row shows the separated speech signal produced by the proposed method and finally the fourth row shows the separated noise signal. From these figures, it is observed that the proposed model-driven approach is capable of rejecting the interference signal while recovering the most part of the target speaker spectrogram. The proposed method also finds lots of noise spectral structure making it as a favorable candidate for noise tracking compared to other well-known methods in speech enhancement literature.

Objective Measurement of Speech Quality. We evaluate the separation performance of the proposed method for the reverberant noisy data consist of multiple sources as described in SiSEC [12]. We report the separation results in terms of the following objective metrics: overall perceptual score (OPS) from [13], signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) from BSS EVAL [11] and finally signal-to-noise ratio (SNR) and segmental signal and segmental SNR. Table 1 shows the separation performance evaluation results averaged over the SiSEC development database. The proposed method consistently improves all objective metrics compared to other methods. It still achieves lower performance compared to ideal binary mask in terms of SNR-based measures confirming the fact that ideal binary mask is known to provide the maximum achievable SNR-based measures [9]. An exception of this, is the OPS results, where the proposed method achieves the highest performance, higher than that obtained by ideal binary mask. In [13], it was shown that the OPS measure outperforms all concurrent measures used for separation performance evaluation.

Table 1. Showing SNR, SSNR, SDR and SIR results over the SiSEC development database. The SDR and SIR results reported in table are calculated using bss_eval_sources.m from BSS EVAL toolbox, well-known for the evaluation of estimated single-channel source signals [11]. The results are reported for the proposed technique versus those obtained by noisy mixture and other benchmark methods.

NE + Method	Metric	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
- + Noisy speech [4]	SNR	-8.2	-3.0	-0.6	2.6	-2.8	6.3	-1.0
	SSNR	-3.8	-4.1	0.8	0.5	-0.3	5.2	-0.3
	SDR	-8.7	-3.6	-2.9	-0.9	5.2	-2.2	-2.2
	SIR	-8.9	-3.6	-2.9	-0.9	5.9	-2.18	-2.1
	OPS	9.4	8.6	25.9	8.6	9.2	18.2	13.3
M1: VAD + LSA [2]	SNR	-6.7	-0.8	1.4	4.3	-1.7	5.4	0.3
	SSNR	-1.7	-1.9	1.8	1.8	1.5	3.7	0.9
	SDR	-7.2	-1.2	1.9	5.1	0.0	9.4	1.3
	SIR	-6.8	-0.5	2.7	7.8	1.1	14.1	3.1
	OPS	19.7	15.7	30.6	28.9	34.4	40.9	28.3
M2: VAD + STSA [1]	SNR	-6.6	-0.8	1.4	4.2	-1.7	5.2	0.3
	SSNR	-1.6	-1.8	1.7	1.8	1.5	3.4	0.8
	SDR	-7.2	-1.3	1.9	5.2	-0.0	9.1	1.3
	SIR	-6.7	-0.5	2.7	8.4	1.3	14.5	3.3
	OPS	20.4	16.2	31.9	29.4	34.4	37.5	28.3
M3: MMSE [3] + GGD [10]	SNR	0.6	0.9	1.2	1.1	0.9	1.2	1.0
	SSNR	-0.8	-0.3	0.1	0.3	0.1	0.8	0.0
	SDR	-6.6	-4.3	-1.1	1.9	-0.8	9.6	-0.2
	SIR	-2.9	-0.6	2.7	3.3	3.1	9.6	2.5
	OPS	21.8	19.3	26.3	27.9	31.7	28.1	25.8
Proposed	SNR	**2.1**	**3.1**	**4.5**	**4.8**	**4.7**	**6.2**	**4.2**
	SSNR	**0.9**	**1.2**	**2.6**	**2.9**	**3.2**	**4.9**	**2.6**
	SDR	**0.2**	**2.4**	**6.0**	**5.7**	**8.1**	**11.3**	**5.6**
	SIR	**2.8**	**5.6**	**8.9**	**9.6**	**15.4**	**16.2**	**9.8**
	OPS	**27.0**	**21.8**	**45.4**	**34.0**	**33.4**	**50.3**	**35.3**
M4: Ideal + IBM [9]	SNR	4.4	5.5	5.1	5.1	3.6	4.3	4.7
	SSNR	3.3	3.4	3.4	3.3	2.5	3.0	3.1
	SDR	6.7	6.9	7.9	6.7	7.5	7.2	7.1
	SIR	21.5	18.7	23.5	19.9	23.4	22.9	21.6
	OPS	16.6	12.4	13.3	14.4	14.0	13.9	14.1

5 Conclusion

We presented a model-driven approach to recover unknown speech signal of a target speaker from a mixtures of multi source reverberant with non-stationary noise environment. The proposed approach makes use of a codebook as pretrained speaker model to reject the interference in the noise corrupted observed signal. The experimental results on SiSEC indicated consistent improvements in terms of spectrograms as well as the objective quality measures by the proposed method. Future work includes the application of the proposed method in binaural speech enhancement as well as in robust automatic speech recognition task.

References

1. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Audio, Speech, and Language Process. 32(6), 1109–1121 (1984)
2. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 33(2), 443–445 (1985)
3. Hendriks, R.C., Heusdens, R., Jensen, J.: MMSE based noise PSD tracking with low complexity. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 4266–4269 (2010)
4. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In: Proc. Interspeech, pp. 1918–1921 (2010)
5. Mowlaee, P.: New Stategies for Single-channel Speech Separation, Ph.D. thesis, Institut for Elektroniske Systemer, Aalborg Universitet (2010)
6. Mowlaee, P., Christensen, M., Jensen, S.: New results on single-channel speech separation using sinusoidal modeling. IEEE Trans. Audio, Speech, and Language Process. 19(5), 1265–1277 (2011)
7. Rangachari, S., Loizou, P.C.: A noise-estimation algorithm for highly non-stationary environments. Speech Communication 48(2), 220–231 (2006)
8. Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Processing 81(11), 2403–2418 (2001)
9. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Speech Separation by Humans and Machines, pp. 181–197. Kluwer (2005)
10. Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J.: Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. IEEE Transactions on Audio, Speech, and Language Processing 15(6), 1741–1752 (2007)
11. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (2006)
12. The third community-based Signal Separation Evaluation Campaign (SiSEC 2011), `http://sisec.wiki.irisa.fr/tiki-index.php`
13. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and objective quality assessment of audio source separation. IEEE Transactions on Audio, Speech, and Language Processing (99), 1 (2011)

Semi-blind Source Separation Based on ICA and Overlapped Speech Detection*

Jiří Málek[1], Zbyněk Koldovský[1,2], and Petr Tichavský[2]

[1] Faculty of Mechatronic and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
jiri.malek@tul.cz,
[2] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic

Abstract. We propose a semi-blind method for separation of stereo recordings of several sources. The method begins with computation of a set of cancellation filters for potential fixed positions of the sources. These filters are computed from one-source-only intervals selected upon cross-talk detection. Each source in some of the fixed positions is canceled by the corresponding filter, by which the other sources are separated. The former source can be then separated by adaptive suppression of the separated sources. To select the appropriate cancellation filter, we use Independent Component Analysis. The performance of the proposed method is verified on real-world SiSEC data with two fixed and/or moving sources.

Keywords: Semi-blind Separation, Audio Source Separation, Cancellation Filter, Independent Component Analysis.

1 Introduction

Separation of multiple audio signals recorded in a natural environment is a discipline comprising several situations. These mainly differ in mutual positions of microphones and sources, room reverberation and variability of the environment. The SiSEC 2012 evaluation campaign[1] defines several tasks. In this paper, we consider the task "Determined convolutive mixtures under dynamic conditions". The goal is here to separate utterances of several speakers where at most two of them speak simultaneously from random fixed positions or moving positions (one source). The scenario is practical as it simulates a meeting situation. Signals recorded by four microphones are available, but we focus on using only two microphones, which are more accessible in practice.

The problem can be solved in a blind way, that is, by using only general assumptions such as the sparsity or independence. The latter assumption enables the use of Independent Component Analysis (ICA) either in the frequency

* This work was supported by Grant Agency of the Czech Republic through the project P103/11/1947.

[1] http://sisec.wiki.irisa.fr/

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 462–469, 2012.

domain or in the time domain. A drawback of blind methods consists in their limited efficiency due to the generality of the conception. The recent effort is therefore to take advantage of blind approaches together with incorporated a priori knowledge. These approaches have the common label *semi-blind*.

The known features of the SiSEC scenario (a priori knowledge) are as follows.

F1 Maximum two sources are active at the same time instant.
F2 At least one of the active sources is located at a fixed position.
F3 There is a finite number of potential fixed positions, and for each such position there exists an interval (even just one second short) in which a source sounds from this position but other sources are silent.
F4 Different sources are mutually independent.

In this paper, we propose a separation method that takes advantage of the above features as much as possible. The method utilizes two basic tools: cancellation filters and the ICA. The use of cancellation filters for separation of audio sources has been already proved to be useful even in difficult environments [2]. The approach is however restricted to sources having fixed known position. In this paper, we go one step further by applying ICA to find the filter assuming that the source is in one of possible (but unknown) positions.

A cancellation filter (CF) is a filter that cancels a targeted signal and passes the other signal through. Its output thus gives, on one hand, a separated (non-target) signal, which, on the other hand, can be suppressed from the original recording by an adaptive filter to separate the targeted signal. The CF is a time-invariant filter, so it cannot cancel a moving source.

In some situations such as a meeting, CFs can be computed for potential positions of sources in advance. Then, when an active speaker is detected at a given position and its speech overlaps with another speaker, the speeches can be separated using the corresponding CF(s). The SiSEC scenario considered here can be seen as one such situation. The CFs can be found based on F3, and the separation is possible thanks to F1 and F2. For easy reference, let the set of the computed CFs be called the cancellation filter-bank (CFB).

The only problem to cope with is the fact that the positions of active sources are not known at a given time. Based on F4, we propose a sophisticated method that uses ICA to separate the signals without knowing their positions. Following the idea of [3] and [4], ICA is applied to a data matrix that is defined using the a priori known CFB. In this sense, the method is "semi-blind". The details are given in Section 4. The following section describes the mixing model and the way to derive a CF. Section 3 describes how the CFB for the SiSEC data was derived. Results of the separation of the SiSEC data are presented in Section 5.

2 Problem Statement

Let s denote a targeted signal whose position is fixed. A stereo mixture of this signal with a noise is, in general, described by

$$\begin{aligned} x_{\mathrm{L}}(n) &= \{h_{\mathrm{L}} * s\}(n) + y_{\mathrm{L}}(n), \\ x_{\mathrm{R}}(n) &= \{h_{\mathrm{R}} * s\}(n) + y_{\mathrm{R}}(n) \end{aligned} \tag{1}$$

where n is the time index, $*$ denotes the convolution, $x_{\mathrm{L}}(n)$ and $x_{\mathrm{R}}(n)$ are, respectively, the signals from the left and right microphone, and $h_{\mathrm{L}}(n)$ and $h_{\mathrm{R}}(n)$ denote the microphone-source impulse responses. The noise signals on respective microphones are denoted by y_{L} and y_{R}. The signals are independent of s and, in our case, they correspond to responses (images) of the other speaker (or may be equal to zero). When the position of the "noise" speaker is fixed, the roles of the target and "noise" are interchangeable.

2.1 Cancellation Filter

To cancel the target s, we can seek a filter g that satisfies

$$\{g * h_{\mathrm{L}}\}(n) = h_{\mathrm{R}}(n), \tag{2}$$

because then the signal

$$\begin{aligned} v(n) &= \{g * x_{\mathrm{L}}\}(n) - x_{\mathrm{R}}(n) \\ &= \{g * h_{\mathrm{L}} * s\}(n) + \{g * y_{\mathrm{L}}\}(n) - \{h_{\mathrm{R}} * s\}(n) - y_{\mathrm{R}}(n) \\ &= \{g * y_{\mathrm{L}}\}(n) - y_{\mathrm{R}}(n) \end{aligned} \tag{3}$$

does not contain any contribution of $s(n)$, while y_{L} and y_{R} are passed through.

The filter g can be found using a noise-free interval $n = N_1, \ldots, N_2$, i.e. when $y_{\mathrm{L}}(n) = y_{\mathrm{R}}(n) = 0$, as a solution to the least square problem

$$g = \arg\min_{g} \sum_{n=N_1}^{N_2} \Big| \{g * x_{\mathrm{L}} - x_{\mathrm{R}}\}(n) \Big|^2. \tag{4}$$

We will call g the cancellation filter, although the true CF is the MISO filter on the right-hand side of (3), comprising of g and $-\delta$ (the unit impulse).

3 Building the CFB

According to F3, it is possible to compute the CF for each potential (fixed) position of a source. Our strategy is therefore to find one-source-only intervals and compute the CF according to (4), for each interval. This can be done manually, that is, in a supervised way, which we take into consideration. On the other hand, an automatic selection may be needed in real-time applications. Therefore, we propose two approaches to find the one-source-only intervals automatically.

The need is to distinguish three possible situations: silence, one speaker active, and two speakers active. The silence is easily detected by thresholding the energy of signals on microphones. It is more challenging to distinguish one speaker talk from a cross-talk.

Our first approach uses single (left) microphone only and is based on the linear predictive coding (LPC) of the observed signal. LPC models the signal as an autoregressive process of a selected order and measures the energy of the residual signal (the prediction error). In the literature, see e.g. [7], it was observed that the prediction error of single speech signal is lower than that of an overlapped speech signal. The first approach therefore does the detection by thresholding the linear prediction error.

The second approach utilizes both microphones and measures the coherence of the signals [5]. The coherence is equal to one when the signal from one microphone is a delayed version of the signal from the other microphone, which ideally happens when the signal comes from a single direction without any reverberation. The reverberation must be taken into account, so the detection is based on thresholding the coherence.

The automatic selection proceeds as follows, examples of selected intervals are shown in Figure 1.

1. The detection criterion is computed throughout available data and smoothed by the moving-average filter (the length is 250 ms).
2. The intervals where the smoothed criterion is lower (higher) than a threshold are selected.
3. For each block of a sufficient length (≥ 1 s), compute the CF according to (4) a store it into the CFB.

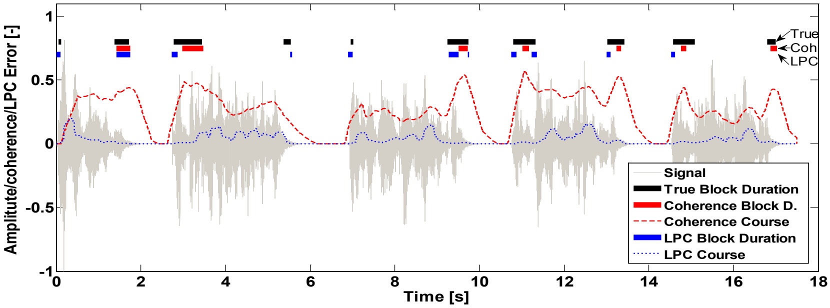


Fig. 1. An example of detected one-source-only blocks by thresholding LPC error and coherence. "True" blocks denote manually selected blocks.

The automatic procedure (but also the manual one) has the potential problem of computing several CFs for the same position. The duplicated CFs can be recognized by using a similarity measure (e.g. the mean square distance). However, there still may be CFs that differ quite much due to estimation errors but correspond to the same position. Fortunately, our method is robust in this respect thanks to the applied ICA, as it is explained in the following section.

4 Source Separation Using ICA and CFB

The SiSEC data can be divided into intervals in which two overlapping sources sound from unknown positions. In this section, we focus on processing one such interval $n = N_1, \ldots, N_2$ and propose a method that separates the signals from the mixtures $x_\mathrm{L}(n)$ and $x_\mathrm{R}(n)$. The features F1-F4 are taken into account, so it is assumed that a CFB containing CF for each potential position of stationary sources is available.

Let g_i, $i = 1, \ldots, P$ denote CFs in the CFB. We define a data matrix as

$$\mathbf{X} = \begin{bmatrix} \{g_1 \star x_\mathrm{L}\}(N_1) & \ldots & \{g_1 \star x_\mathrm{L}\}(N_2) \\ \vdots & \vdots & \vdots \\ \{g_P \star x_\mathrm{L}\}(N_1) & \ldots & \{g_P \star x_\mathrm{L}\}(N_2) \\ x_\mathrm{R}(N_1) & \ldots & x_\mathrm{R}(N_2) \end{bmatrix} \tag{5}$$

and search for its independent components (ICs) by an ICA algorithm[2]. The ICA yields the de-mixing $(P+1) \times (P+1)$ matrix $\mathbf{W}$ and independent components $\mathbf{C} = \mathbf{W}\mathbf{X}$ which are linear combinations of rows of $\mathbf{X}$. It is highly expectable that at least one such combination (independent component) corresponds to the signal in which one source having fixed position is canceled. There are two key reasons for this claim.

1. The output of the kth CF can be expressed by

$$[\underbrace{0, \ldots, 0, 1}_{k}, 0, \ldots, -1] \cdot \mathbf{X} = \{g_k \star x_\mathrm{L}\} - x_\mathrm{R}, \tag{6}$$

   which means that the subspace spanned by rows of $\mathbf{X}$ contains the outputs of all CFs in the CFB. Since one source is in one of the potential positions (although unknown), there exists a linear combination of rows of $\mathbf{X}$ that cancels the source.
2. Such linear combination is an independent signal since it contains the contribution of one source only.

Since the order of the independent components (ICs) is random, the one that corresponds to the signal with canceled source must be found. This problem is easily resolved by finding the largest element (in absolute value) of the last column of $\mathbf{W}$. To explain, the ℓth element of the last column of $\mathbf{W}$ determines how much the last row of $\mathbf{X}$ contributes to the ℓth IC. Since only the last row of $\mathbf{X}$ contains samples of x_R (the other rows contain x_L), its contribution must be significant so that a source in the IC be canceled. Similarly, when there are two stationary sources in the mixture, we select two components corresponding to the two largest elements.

4.1 Separation by Adaptive Post-filtering

Once an independent signal is obtained, it can be considered as a separated one thanks to F1; let it be denoted by $v(n)$. The other source can be obtained by an adaptive Wiener-like filter that suppresses $v(n)$ from x_L and x_R.

[2] An arbitrary ICA algorithm can be used. We utilize the BGSEP algorithm from [6] for its speed and accuracy.

Let $X(k,\ell)$ and $V(k,\ell)$ be the short-time Fourier transform of $x_{\mathrm{L}}(n)$ (or x_{R}) and $v(n)$, respectively, where k is the frequency index and ℓ is the time-frame index. The adaptive filter, which is sometimes called a soft mask or the frequency-domain Wiener filter [8], is defined in the time-frequency domain by

$$W(k,\ell) = \frac{|X(k,\ell)|^2}{|X(k,\ell)|^2 + \tau |V(k,\ell)|^2}. \tag{7}$$

The time-frequency representation of the final output signal is

$$\widehat{S}(k,\ell) = W(k,\ell) X(k,\ell). \tag{8}$$

The free positive parameter τ allows control of the trade-off between the Signal-to-Interference ratio (SIR) and Signal-to-Distortion ratio (SDR) of the output signal.

5 Experiments

The SiSEC datasets “Determined convolutive mixtures under dynamic conditions” were recorded in a room with reverberation time about 700 ms. The sampling rate of signals is 16 kHz. From the four channel recordings in development dataset, we use signals from microphone 2 and 3, whose distance is 2 cm. The distances of the sources from microphones are about 1 m.[3]

The datasets are divided into intervals in which two sources are active. Each interval is processed separately and the separated signals are evaluated using the BSS_EVAL toolbox [9]. We use the criteria SIR, SDR and SAR (Signal-to-Artifact ratio) and SIR improvement (the difference between the SIR of mixed and separated signals). The resulting criteria are averaged over all intervals.

In our experiments, we distinguish the three ways of obtaining the CFB needed for our method (Section 3). *MAN* means the manual selection of one-source-only intervals. The automatic selections are denoted by *LPC* (LPC with the AR order 18) and *COH* (coherences with the length of the FFT window 128 samples and zero overlap).

5.1 Random Sources Activity in Unknown Static Positions

In this situation, two active speakers are located at unknown fixed positions on a semi-circle with radius 1 m. In Setup 1, the competing sources are always located on different angular sides with respect to the center of the array, that is one speaker is in (−90°;0°) while the other one is in (0°;90°). In the Setup 2, the two competing sources can be located in the whole angular space (−90°;90°), but never in the same position. We consider two ways the separated signals

[3] The results for other microphone/source distances achieved on the SiSEC datasets can be found on the SiSEC results web page `http://www.irisa.fr/metiss/SiSEC11/dynamic/main.html`.

could be obtained. They can either be both obtained as ICs (Section 4) in which one source is canceled (denoted by *ica*) or both as the outputs of the adaptive Wiener-like filter (Section 4.1) that suppresses the obtained component from original recordings (denoted by *wf*); the parameter τ in (7) was put equal to 10. The results averaged over both separated sources are summarized in Table 1.

Table 1. Separation of fixed sources with random location

	method	SIR[dB]	SIR impr.[dB]	SDR[dB]	SAR[dB]
Setup 1	MAN (ica)	17.18	15.65	4.02	4.88
	MAN (wf)	12.16	10.62	1.17	3.24
	LPC (ica)	12.76	11.22	2.95	4.56
	LPC (wf)	9.64	8.10	-0.33	2.60
	COH (ica)	14.34	12.80	2.96	4.26
	COH (wf)	10.33	8.80	0.13	2.66
Setup 2	MAN (ica)	14.67	12.79	1.36	2.88
	MAN (wf)	11.42	9.54	0.71	3.33
	LPC (ica)	12.57	10.69	1.61	3.25
	LPC (wf)	8.62	6.74	-0.58	2.89
	COH (ica)	11.99	10.11	0.79	2.89
	COH (wf)	8.20	6.32	-1.15	3.06

The manually selected CFB leads to a better performance in terms of all criteria. The unsupervised approaches give comparable results, which points to their efficiency. The separation is better when signals are taken as the ICs than when they are obtained by the adaptive filter, especially in terms of SDR and SAR. This is explained by the fact that the sources are in fixed positions, so invariant filters (ica), which generate less distortions, are sufficient for the separation.

5.2 A Moving Source

In this scenario, one source is moving within the angular space (0°;90°) and its distance from microphones is varying between 0.5 m and 1.2 m. The position of the second source is fixed within the angular space (−90°;0°) either at one position during the whole dataset (Setup 1) or random position (Setup 2). Here, the moving source can be separated as the IC only, while the stationary source must be separated by the adaptive filter. Table 2 shows the results. In the case of Setup 1 (fixed source at one position), the label *single* denotes the case, when the CFB contains one filter only. This filter is able to supress the fixed source in the whole recording, i.e. the ICA utilization is not necesary.

The performance achieved with the single manually selected filter in Setup 1 confirms the suitability of the CF utilization for this type of separation scenario. In case of automatically constructed CFBs, the performance is lower, because the CFB contain CFs for positions where the moving source appeared for a moment. These CFs cause random confusion of the separated sources and deteriorate the performance.

Table 2. Separation of mixtures of one fixed and one moving source

	method	src. position	SIR[dB]	SIR impr.[dB]	SDR[dB]	SAR[dB]
Setup 1	MAN (single)	moving	13.00	12.62	6.62	8.23
	MAN (wf)	fixed	19.48	16.33	3.74	3.98
	LPC (ica)	moving	7.04	6.66	1.31	4.25
	LPC (wf)	fixed	16.72	13.57	0.17	0.50
	COH (ica)	moving	7.99	7.61	1.33	3.65
	COH (wf)	fixed	16.73	13.57	0.92	1.24
Setup 2	MAN (ica)	moving	10.28	10.26	-1.47	1.81
	MAN (wf)	fixed	14.82	11.29	1.70	2.24
	LPC (ica)	moving	9.10	9.08	1.17	3.73
	LPC (wf)	fixed	15.88	12.35	0.03	0.44
	COH (ica)	moving	10.03	10.01	1.23	3.65
	COH (wf)	fixed	15.29	11.75	-0.60	0.01

6 Conclusion

We presented a solution for the task presented in SISEC evaluation campaign using ICA and cancellation filters. The proposed method can be easily extended to situations where there are more than two sources [10].

References

1. Benesty, J., Makino, S., Chen, J. (eds.): Speech Enhancement, 1st edn. Springer, Heidelberg (2005)
2. Li, J., Sakamoto, S., Hongo, S., Akagi, M., Suzuki, Y.: Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model. In: Proc. of WASPAA 2009, New Paltz, New York, pp. 133–136 (October 2009)
3. Koldovský, Z., Tichavský, P.: Time-Domain Blind Separation of Audio Sources on the basis of a Complete ICA Decomposition of an Observation Space. IEEE Trans. on Speech, Audio and Language Processing 19(2), 406–416 (2011)
4. Koldovský, Z., Tichavský, P., Málek, J.: A Semi-Blind Noise Extraction Using Partially Known Position of the Target Source. Submitted to a Conference (2011)
5. Albouy, B., Deville, Y.: Alternative structures and power spectrum criteria for blind segmentation and separation of convolutive speech mixtures. In: Proc. of ICA 2003, Nara, Japan, April 1-4, pp. 361–366 (2003)
6. Tichavský, P., Yeredor, A.: Fast Approximate Joint Diagonalization Incorporating Weight Matrices. IEEE Trans. on Signal Processing 57(3), 878–891 (2009)
7. Sundaram, N., et al.: Usable speech detection using linear predictive analysis - a model based approach. In: Proc. of ISPACS, Awaji Island, Japan, pp. 231–235 (2003)
8. Koldovský, Z., Nouza, J., Kolorenč, J.: Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA. In: Proc. of Interspeech 2006, Pittsburgh PA, USA, September 17-21, pp. 2578–2581 (2006)
9. Févotte, C., Gribonval, R., Vincent, E.: BSS EVAL toolbox user guide. IRISA, Rennes, France, Tech. Rep. 1706 (2005), http://www.irisa.fr/metiss/bss_eval
10. Rutkowski, T.M., et al.: Identification and tracking of active speaker's position in noisy environments. In: Proc. of IWAENC 2003, Kyoto, Japan, pp. 283–286 (2003)

Nonparametric Modelling of ECG: Applications to Denoising and to Single Sensor Fetal ECG Extraction

Bertrand Rivet, Mohammad Niknazar, and Christian Jutten*

GIPSA-Lab , CNRS UMR-5216, University of Grenoble
Domaine Universitaire, BP 46, 38402 Saint Martin d'Hères cedex, France

Abstract. In this work, we tackle the problem of fetal electrocardiogram (ECG) extraction from a single sensor. The proposed method is based on non-parametric modelling of the ECG signal described thanks to its second order statistics. Each assumed source in the mixture is thus modelled as a second order process thanks to its covariance function. This modelling allows to reconstruct each source by maximizing the related posterior distribution. The proposed method is tested on synthetic data to evaluate its performance behavior to denoise ECG. It is then applied on real data to extract fetal ECG from a single maternal abdominal sensor.

Keywords: non-parametric modelling, source extraction, denoising, fetal ECG extraction.

1 Introduction

Fetal electrocardiogram (f-ECG) extraction from maternal abdominal ECG sensors is an old problem. Since the first works of Cremer [1] who produced a very primitive record of fetal rate activity, this problem is still of interest nowadays since it is a fascinating issue due to the characteristics of the involved signals. Indeed, the f-ECG is definitively less powerful than the mother's ECG (m-ECG), moreover the recorded signals are also contaminated by noise due to electromyogram (EMG) or to power line interference and they are also influenced by the fluctuation of the baseline. Among the several approaches used to tackle the extraction of f-ECG, one can quote methods which require several sensors e.g., adaptive filtering [15], blind source separation [2,16] or quasi-periodic analysis [14].

In this paper, we consider the same issue but assuming that only a single sensor is available. In this case, one can extract f-ECG by singular value decomposition [4] or by nonlinear decomposition such as shrinkage wavelet denoising [7] or nonlinear projections [10]. Moreover, state modelling as Kalman filtering [13] has been applied to overcome the lack of information provided by a single sensor.

[*] Christian Jutten is also with Institut Universitaire de France.

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 470–477, 2012.

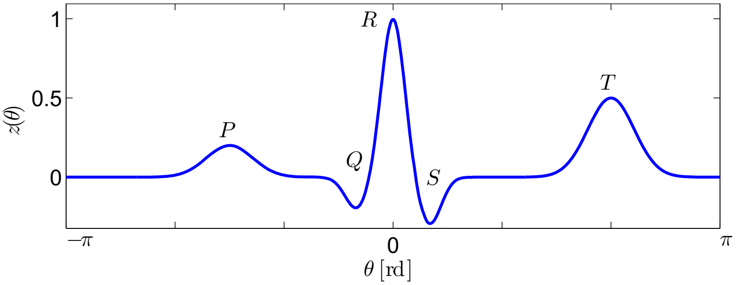


Fig. 1. Modelling (1) of the amplitude of one beat

Among these methods, the latter has been shown to be the most efficient [12]. However, Kalman filtering relies on a very strong assumption: the state equation, which models the dynamical evolution of the unobserved state. As a consequence, Kalman filtering needs reliable prior about the state to perform accurately. To overcome the potential lack of prior information about the system, we propose in this study to model the second order statistics of the signal instead of the signal itself.

This article is organized as follows. Section 2 presents the proposed approach to model a signal thanks to its second order statistics. The proposed algorithm to extract f-ECG is then introduced in Section 3. Numerical experiments and results are given in Section 4 before conclusion and perspectives in Section 5.

2 Nonparametric Modelling of ECG

As already proposed in [13], one can choose a parametric model of ECG: each beat of an ECG signal is a summation of 5 Gaussian functions, each of them modelling the P, Q, R, S and T waves as illustrated in Figure 1:

$$z(\theta) = \sum_{i \in \{P,Q,R,S,T\}} a_i \exp\left(-\frac{(\theta-\theta_i)^2}{2\sigma_i^2}\right), \qquad (1)$$

where a_i, θ_i and σ_i are the amplitude, the position and the width of each wave, respectively. Note that in this model, the beats are defined in phase $\theta \in [-\pi, \pi]$, so that each beat is assumed to have a linear variation of phase with respect to the time, even if each beat has not the same duration. This model can then be used in an extended Kalman filtering to denoise a single ECG or extract f-ECG from a mixture of m-ECG and f-ECG [13,11]. This method is thus a parametric method since the unknown amplitude $z(\theta)$ is explicitly parameterized.

On the other hand, nonparametric methods perform estimation, prediction or denoising without explicitly parameterizing the unknown amplitude $z(\theta)$. For instance, a well known approach is the spline smoothing [5]. If one considers the ECG $z(\theta)$ as a statistical process, it can be fully described at the second order by

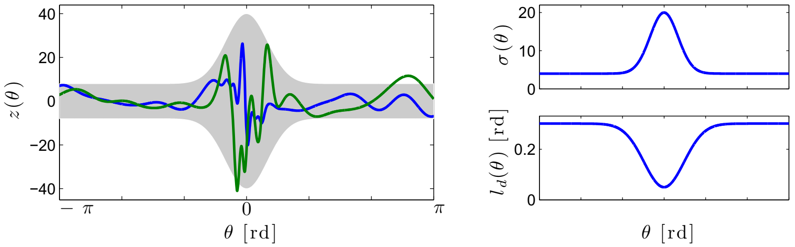


Fig. 2. Two functions drawn at random from a zero-mean GP with covariance function (2). The shaded area represents plus and minus two times the standard deviation for the prior. On the right, the related $\sigma(\theta)$ and $l_d(\theta)$ functions.

its mean function $m(\theta) = \mathbb{E}[z(\theta)]$ and covariance function $k(\theta_1, \theta_2) \overset{\triangle}{=} \mathbb{E}[(z(\theta_1) - m(\theta_1))(z(\theta_2) - m(\theta_2))]$ [8]. Obviously, the ECG signal is almost surely a more complex statistical process than a simple second order one. As a consequence, considering only its second order statistics, it relies among the Gaussian process (GP) framework which is widely used in machine learning e.g., [6,9]. A GP $z(\theta)$ is a distribution over functions denoted as $\mathcal{GP}(m(\theta), k(\theta_1, \theta_2))$. In this case, the statistical latent process $z(\theta)$ is not directly parameterized as in parametric model, but its statistics are it thanks to hyper-parameters. This means that one has to choose a class of semidefinite positive functions $k(\theta_1, \theta_2)$ which describes the expected second order properties of the latent process.

In this study, we propose to use the following non-stationary covariance function

$$k(\theta_1, \theta_2) = \sigma(\theta_1)\sigma(\theta_2)\sqrt{\frac{2l_d(\theta_1)l_d(\theta_2)}{l_d(\theta_1)^2 + l_d(\theta_2)^2}}\, \exp\left(-\frac{\left(\theta_1 - \theta_2\right)^2}{l_d(\theta_1)^2 + l_d(\theta_2)^2}\right), \quad (2)$$

with

$$\sigma(\theta) = a_m + (a_M - a_m)\exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_T^2}\right),$$

$$l_d(\theta) = l_M - (l_M - l_m)\exp\left(-\frac{(\theta - \theta_0)^2}{2\sigma_l^2}\right),$$

where $\sigma(\theta)$ and $l_d(\theta)$ allow to have a time-varying power (between a_m and a_M) and a time-varying length scale correlation (between l_m and l_M), respectively. Indeed, as shown in Fig. 1, an ECG beat can be decomposed into three parts: the P wave, the QRS complex and the T wave. The P and T waves share the same kind of second order statistics: a larger length scale and a lower power than the QRS complex. Fig. 2 shows two functions drawn at random from the zero-mean GP prior with covariance function (2). This figure illustrates the flexibility of such a representation compared to model (1) since with the same prior $\mathcal{GP}\big(0, k(\theta_1, \theta_2)\big)$, it can generate a multitude of different shapes with the same prior.

3 Denoising of ECG and Extraction of Fetal ECG from a Single Sensor

Suppose that the observed values $x(t)$ differ from the ECG, $s(t)$, by additive noise $n(t)$:

$$x(t) = s(t) + n(t), \tag{3}$$

and that this noise is uncorrelated with $s(t)$. The aim of this study is to infer the values of $s(t)$ from $x(t)$, i.e. to denoise or extract the ECG from the observations. Moreover, it is assumed that the ECG signal, $s(t)$, is a succession of beats, each of them following a zero-mean GP defining by (2) and that the additive noise follows a zero-mean GP whose covariance function $k_n(t, t')$ is given by

$$k_n(t, t') = \sigma_n^2 \exp\left(-\frac{(t-t')^2}{2\, l_n^2}\right) + \sigma_w^2 \delta(t - t'), \tag{4}$$

where $\delta(\cdot)$ is the delta Dirac function. In this expression, the first term is useful for instance to model a baseline variation as a stationary process for which the correlation is almost unity between close samples and decreases as their distance increases compared to the length scale l_n. The second term models a white Gaussian noise of power σ_w^2. From (3) and (4), the covariance function of observation $x(t)$ is thus expressed as

$$k_x\big(t, t'\big) = k_s\big(t, t'\big) + k_n\big(t, t'\big), \tag{5}$$

where

$$k_s\big(t, t'\big) = \sum_{n=1}^{N} \sum_{n'=1}^{N} k\big(t - \tau_n, t' - \tau_{n'}\big)$$

and $\{\tau_n\}_{1 \le n \le N}$ is the set of R peak instants that can be estimated easily from the raw signals. From this modelling and assuming that the observed process $x(t)$ has been recorded at times $\{T_m\}_{1 \le m \le M}$, the covariance matrix of this process is thus given by K_x, whose $(i, j)th$ entry is

$$K_x(i, j) = k_x(T_i, T_j). \tag{6}$$

One can then infer on the value $s(t)$ thanks to the maximization of the a posteriori distribution of $s(t)$ given $\mathbf{x} = [x(T_1), \cdots, x(T_M)]^T$ by

$$\hat{s}(t) = \mathbf{k}^T K_x^{-1} \mathbf{x}, \tag{7}$$

where $\mathbf{k} = [k(t, T_1), \cdots, k(t, T_M)]^T$. It is interesting to note that as soon as $\sigma_w^2 \neq 0$, matrix K_x is invertible as the summation of definite positive matrices and a diagonal matrix $\sigma_w^2 I$. This algorithm needs some comments. First of all, the recorded signal, $x(t)$, does not need to be regularly sampled and one can observe from (7) that the value of the latent process, $s(t)$, can be predicted at any time t even for $t \neq T_i, \ \forall i \in \{1, \cdots, T_M\}$. Moreover, the hyper-parameters $\boldsymbol{\theta} = \{a_m, a_M, \sigma_T^2, l_m, l_M, \sigma_l^2, \theta_0, \{\tau_k\}_k, \sigma_n^2, l_n^2, \sigma_w^2\}$ defining $k(\cdot, \cdot)$ and $k_n(\cdot, \cdot)$ need to be estimated. This can be done by maximizing the evidence (or log marginal likelihood) given by

$$\log p\big(\mathbf{x} | \{T_i\}_i, \boldsymbol{\theta}\big) = -\frac{1}{2}\mathbf{x}^T \big(K_s + K_n\big)^{-1}\mathbf{x} - \frac{1}{2}\log\big|K_s + K_n\big| - \frac{M}{2}\log(2\pi). \quad (8)$$

The optimization of the latter equation is obtained thanks to a gradient ascent method, assuming that the initial parameter values are not so far from its actual values.

Fetal ECG extraction from a single abdominal sensor is then a direct extension of the proposed method by modelling the recorded signal $x(t)$ as

$$x(t) = s_m(t) + s_f(t) + n(t), \quad (9)$$

where $s_m(t)$ is the signal related to the mother, $s_f(t)$ is related to the fetus and $n(t)$ is the additive noise. All these signals are modelled by zero-mean GPs with covariance functions $k_m(\cdot, \cdot)$ and $k_f(\cdot, \cdot)$ defined by (2) and $k_n(\cdot, \cdot)$ obtained from (4), respectively. In this case, the estimation of $s_f(t)$ is given by

$$\hat{s}_f(t) = \mathbf{k}_f^T \big(K_m + K_f + K_n\big)^{-1}\mathbf{x}, \quad (10)$$

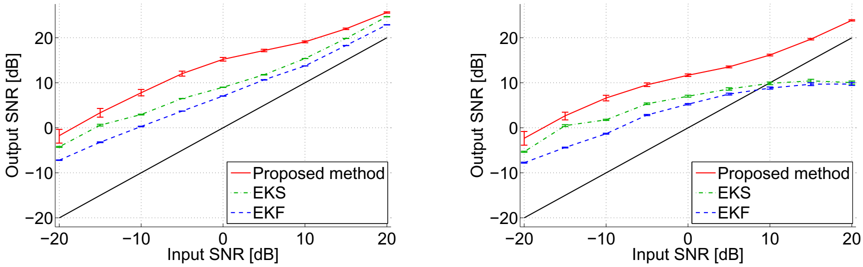
where $\mathbf{k}_f = [k_f(t, T_1), \cdots, k_f(t, T_M)]^T$.

4 Numerical Experiments

In this section we first investigate the performance of the proposed method on synthetic data to denoise ECG (Section 4.1). An illustration of f-ECG extraction is then provided on real data (Section 4.2).

4.1 Synthetic Data: ECG Denoising

The performance of the proposed algorithm to denoise ECG is assessed. In the first experiment, each beat of the ECG signal is generated by model (1). To mimic the variability presented in a real ECG, the waves amplitudes and P-R and R-T intervals are randomly changed (3%) around their average values. The ECG signal is then obtained as the summation of several beats with random global amplitudes and random R-R intervals. To ensure the consistency of the results, the whole procedure has been repeated one thousand times by regenerating all random parameters of the signal and noise samples. In this experiment, 1500 samples are used with 15 heart beats simulated at 100Hz sampled frequency. It is worth noting that the proposed method does not assume that the maxima



(a) Without parameters variability

(b) With 3% parameters variability

Fig. 3. ECG denoising: output SNRs vs. the input SNR without Fig. 3(a) and with Fig. 3(b) parameters variability. In the two figures, the black line corresponds to the same input and output SNRs. In each case, the median are plotted, as well as the first and last quartiles as error bars.

of the R peaks are located at observed samples but can also appear in between samples. The proposed method is compared to the extended Kalman filtering (EKF) and smoothing (EKS) [13]. The state model is chosen equal to (1) (i.e. the same model than the one used to generate data) whose parameters are equal to average values.

Quantitative results are shown in Fig. 3 which compares the output signal-to-noise ratio (SNR) achieved after denoising versus different input SNRs. As one can see (Fig. 3(b)), the proposed method increases the SNR with a gain between 3dB to 18dB. Contrary to extended Kalman filtering, the proposed method always improves the SNR. Indeed, in the case of high input SNR, EKS and EKF deteriorate the SNR: this can be explained by the variability of the simulated ECG as confirmed by Fig. 3(a), since this phenomenon is not observed without variability. Moreover, one can see that the variability decreases the overall performance, but the proposed method keeps the best performance by a smaller decrease than EKS or EKF.

4.2 Real Data: f-ECG Extraction

In this section, we illustrate (Fig. 4) the proposed method to extract f-ECG from a single sensor on the well-known DaISy fetal ECG database [3]. As one can see, the proposed method provides suitable estimations of both maternal and fetal ECG even when mother's and fetus's R peaks are concomitant (e.g., the fourth, seventh and tenth mother's beats). Moreover, a visual inspection of the residual noise $\hat{n}(t) = x(t) - \hat{s}_m(t) - \hat{s}_f(t)$ confirms the validity of the assumed modelling (9). Indeed, this residual noise is effectively composed of a smooth varying baseline (dark curve) related to the first term of covariance function (4) plus a quasi white noise (validated by its covariance function empirical estimation). Moreover, both contributions are decorrelated with the mother's and fetus's ECG signals.

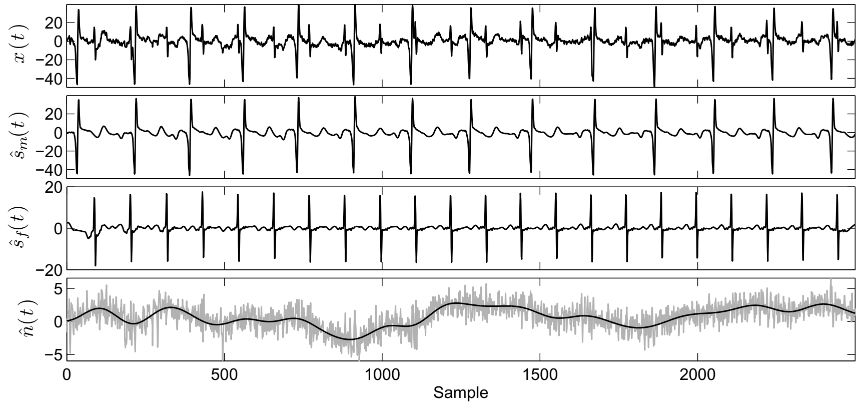


Fig. 4. Fetal ECG extraction. Top to bottom: recorded signal $x(t)$, estimated mother's ECG $\hat{s}_m(t)$, estimated fetal's ECG $\hat{s}_f(t)$ and residual noise $\hat{n}(t)$ (light gray curve) with estimated baseline (dark curve), respectively.

5 Conclusions and Perspectives

In this paper, a non-parametric model of ECG signals is derived. By considering them as second order processes, which are fully defined by their mean and covariance functions, one can model a large class of signals with few hyper-parameters. From this modelling, denoising or extraction methods are directly obtained as the maximization of the posterior distribution. Numerical experiments show that the proposed method outperforms an extended Kalman filtering especially in presence of slightly random state parameters. Indeed, Gaussian processes realize a tradeoff between the suitable description of the signal by its second order statistics and its intrinsic variabilities. Finally, the main advantage of the proposed method is its flexibility and it provides a mix between purely data based methods as principal component analysis and parametric model based methods as Kalman filtering.

Future work will deal with a computationally efficient implementation of hyper-parameters estimation of the proposed method as well as an online implementation.

References

1. Cremer, M.: Über die direkte Ableitung der Aktionsströme des menschlichen Herzens vom Ösophagus und über das Elektrokardiogramm des fötus. München. med. Wchnschr. 53, 811 (1906)
2. De Lathauwer, L., Callaerts, D., De Moor, B., Vandewalle, J.: Fetal electrocardiogram extraction by source subspace separation. In: Proc. IEEE Workshop on HOS, Girona, Spain, June 12–14, pp. 134–138 (1995)

3. De Moor, B., De Gersem, P., De Schutter, B., Favoreel, W.: Daisy: A database for identification of systems. Journal A 38(3), 4–5 (1997)
4. Kanjilal, P.P., Palit, S., Saha, G.: Fetal ecg extraction from single-channel maternal ecg using singular value decomposition. IEEE Transactions on Biomedical Engineering 44(1), 51–59 (1997)
5. Kimeldorf, G.S., Wahba, G.: A correspondence between bayesian estimation on stochastic processes and smoothing by splines. Annal of Mathematical Statistics 41(2), 495–502 (1970)
6. MacKay, D.J.C.: Introduction to Gaussian processes. Neural Networks and Machine Learning 168, 1–32 (1998)
7. Mochimaru, F., Fujimoto, Y., Ishikawa, Y.: Detecting the fetal electrocardiogram by wavelet theory-based methods. Progress in Biomedical Research 7(3), 185–193 (2002)
8. Papoulis, A.: Probability, Random Variables, and Stochastic Processes, 3rd edn. McGraw-Hill (1991)
9. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
10. Richter, M., Schreiber, T., Kaplan, D.T.: Fetal ecg extraction with nonlinear state-space projections. IEEE Transactions on Biomedical Engineering 45(1), 133–137 (1998)
11. Sameni, R.: Extraction of Fetal Cardiac Signals from an Array of Maternal Abdominal Recordings. PhD thesis, Institut Polytechnique de Grenoble, Grenoble-INP (2008)
12. Sameni, R., Clifford, G.D.: A review of fetal ECG signal processing; issues and promising directions. The Open Pacing, Electrophysiology & Therapy Journal (TOPETJ) 3, 4–20 (2010)
13. Sameni, R., Shamsollahi, M.B., Jutten, C., Clifford, G.D.: A nonlinear bayesian filtering framework for ECG denoising. IEEE Transactions on Biomedical Engineering 54(12), 2172–2185 (2007)
14. Tsalaile, T., Sameni, R., Sanei, S., Jutten, C., Chambers, J.: Sequential blind source extraction for quasi-periodic signals with time-varying period. IEEE Transactions on Biomedical Engineering 56(3), 646–655 (2009)
15. Widrow, B., Glover Jr., J.R., McCool, J.M., Kaunitz, J., Williams, C.S., Hearn, R.H., Zeidler, J.R., Dong Jr., E., Goodlin, R.C.: Adaptive noise cancelling: Principles and applications. Proceedings of the IEEE 63(12), 1692–1716 (1975)
16. Zarzoso, V., Nandi, A.K.: Noninvasive fetal electrocardiogram extraction: blind source separation versus adaptative noise cancellation. IEEE Transactions on Biomedical Engineering 48(1), 12–18 (2001)

Nesterov's Iterations for NMF-Based Supervised Classification of Texture Patterns

Rafal Zdunek[1] and Zhaoshui He[2]

[1] Institute of Telecommunications, Teleinformatics and Acoustics,
Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
[2] Faculty of Automation, Guangdong University of Technology, Guangzhou, China
rafal.zdunek@pwr.wroc.pl, zhshhe@gdut.edu.cn

Abstract. Nonnegative Matrix Factorization (NMF) is an efficient tool for a supervised classification of various objects such as text documents, gene expressions, spectrograms, facial images, and texture patterns. In this paper, we consider the projected Nesterov's method for estimating nonnegative factors in NMF, especially for classification of texture patterns. This method belongs to a class of gradient (first-order) methods but its convergence rate is determined by $O(1/k^2)$. The classification experiments for the selected images taken from the UIUC database demonstrate a high efficiency of the discussed approach.

1 Introduction

Nonnegative Matrix Factorization (NMF) [1] is a relevant tool for extracting low-dimensional, nonnegative, sparse, and parts-based feature vectors that are particularly useful for classification of various objects. Qin *et al* [2] successfully applied NMF to unsupervised classification of texture patterns. Their approach combines several strategies from pattern analysis, such as a local invariant affine region detection, scale-invariant feature transform, model reduction and feature extraction, and finally multi-label nearest-neighbor classification. NMF was used for extracting low-dimensional nonnegative encoding vectors from a set of scale, affine and rotation invariant key-points that intimately characterize the images to be classified. Unfortunately, the multiplicative algorithms used in their approach for estimating the nonnegative factors in NMF are slowly-convergent and do not guarantee convergence to a local minimizer.

Another NMF-based approach to texture pattern classification is to use the NMF algorithm that was proposed by Sandler and Lindenbaum [3] for minimizing the Earth Mover's Distance (EMD). The EMD is more suitable for measuring similarity between images with local deformations. However, this approach involves solving a huge system of linear equations with the Linear Programming (LP) technique, which is a computationally very expensive.

In this paper, we discuss the Nesterov's method [4] that belongs to a class of gradient methods but it has $O(1/k^2)$ iteration complexity for the functions $f(\cdot) \in C_L^{1,1}$ (continuously differentiable with a Lipschitz continuous gradient).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 478–485, 2012.

This condition is intimately satisfied for the Euclidean distance or the alpha- and beta-divergences [5]. Since a computational complexity of the Nesterov's method is very low (roughly the same as for the Landweber iterations), we applied this method for updating both factors of NMF in training as well as testing stages.

The Nesterov's method has already found several relevant applications in signal and image processing [6,7]. Guan *et al.* [8] efficiently applied this method to NMF in the context of text document clustering. Following this way, we propose this method to other NMF-application, however, we discuss another version of the Nesterov's method and our alternating optimization algorithm is somehow different than given in [8].

The paper is organized in the following way. The next section discusses the tools used for preprocessing the images to be classified. Section 3 is concerned with the NMF algorithm. The Nesterov's iterations are briefly discussed in Section 4. The experiments for texture image classification are presented in Section 5. Finally, the conclusions are given in Section 6.

2 Image Preprocessing

The process of supervised classification of texture patterns consists of a few stages. The preprocessing aims at extracting some local features that uniquely identify the analyzed images.

The first step of preprocessing is to identify local regions in a given image that are similar by geometrical transformations such as scaling, rotation, and shearing. This task can be achieved with the Harris-Laplace invariant detector [10] that extracts blobs of homogeneous intensity. Thus, this stage of the preprocessing provides a number of local regions (subimages) for each training or testing image.

The number of local regions should be large for each image to maximize the amount of global information on local patterns in each image. This approach considerably enlarges both sets of training and testing images, leading to a substantial increase in a computational complexity, and to a strong redundancy. To tackle the redundancy problem, we propose a simple approach that selects a certain, predefined number of the least correlated images from each analyzed set. This approach assures the images in both sets are considerably diversified but their redundancy is strongly decreased.

The least correlated local regions are then processed by the Scale-Invariant Feature Transform (SIFT) descriptor to get rotation invariant descriptors for each image. SIFT descriptor was proposed by D. Lowe in 1999 [11] to detect local image features that are invariant to scale, orientation, affine distortion, and partially to illumination changes. The SIFT extracts a certain number of keypoints from the analyzed image, and provides rotation and scale invariant descriptors for each keypoint. In our approach, we create SIFT descriptors using the Matlab software taken from the Lowe's homepage[1]. It provides a 128-dimension

[1] http://www.cs.ubc.ca/∼lowe/keypoints/

vector of a descriptor for each detected key-point, which gives a $128 \times K$ matrix of descriptors for each local region extracted from the affine detector, where K is the number of key-points. Assuming we have P training images, and we extract R_p local regions for the p-th training image, thus the total number of training vectors for processing with NMF is $T = \sum_{p=1}^{P} \sum_{r=1}^{R_p} K_r$, where K_r is the number of key-points for the r-th local region. The input matrix for NMF is $\boldsymbol{Y} \in \mathbb{R}^{128 \times T}$.

3 NMF Algorithm

The aim of NMF is to find such lower-rank nonnegative matrices $\boldsymbol{A} = [a_{ij}] \in \mathbb{R}^{I \times J}$ and $\boldsymbol{X} = [x_{jt}] \in \mathbb{R}^{J \times T}$ that $\boldsymbol{Y} = [y_{it}] \cong \boldsymbol{A}\boldsymbol{X} \in \mathbb{R}^{I \times T}$, given the matrix $\boldsymbol{Y}$, the lower rank J, and possibly *a priori* knowledge on the matrices $\boldsymbol{A}$ and $\boldsymbol{X}$. Assuming each column vector of $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ represents a rotation invariant descriptor (a datum point in $\mathbb{R}^I$) of a given key-point, and J is *a priori* known number of clusters (usually it is equal to the number of classes), we can interpret the column vectors of $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J]$ as the feature vectors or centroids (indicating the directions of central points of clusters in $\mathbb{R}^I$) and the columns vectors in $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ are the low-dimensional nonnegative encoding vectors that contain coefficients of a convex combination of the feature vectors, and have discriminant nature. Each $\boldsymbol{x}_t$ for $t = 1, \ldots, T$ corresponds to the t-th keypoint.

To estimate the matrices $\boldsymbol{A}$ and $\boldsymbol{X}$, we assume the Euclidean function:

$$D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}) = \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}||_F^2, \tag{1}$$

which is then alternatingly minimized by the Algorithm 1.

The matrices $\boldsymbol{A}$ and $\boldsymbol{X}$ are updated in Algorithm 1 with the function `NestIter` which executes the Nesterov's iterations that are not fixed but changing with the alternating steps. The rule given in step 4 of Algorithm 1 aims at adapting the nature of alternating steps, starting from projected gradient updates and exploring local minima more and more when the alternating steps proceed. This is a similar rule as proposed in [12], however, here the regularization is achieved by truncated iterations instead of the damping parameter.

After estimating the nonnegative factors $\boldsymbol{A}$ and $\boldsymbol{X}$, the classification can be readily performed in the J-dimensional space of the encoding vectors in $\boldsymbol{X}$ with some multi-label nearest-neighbor classification. Note that for each testing image we have a couple of key-points. We assumed the following strategy: all the key-points from the testing image are projected onto the column space of the matrix $\boldsymbol{A}$ estimated in the training process. Then for each testing key-point in the "reduced" space, the corresponding training key-point in the same "reduced" space is found using the 1-NN rule with a given metrics. The tests are carried out for the Euclidean distance metrics. Finally, we get a set of class-labels for a given test image. The most frequently occurring label indicates the right class.

Algorithm 1. MN-NMF Algorithm

```
Input  : $\boldsymbol{Y} \in \mathbb{R}^{I \times T}$, $J$ - lower rank
Output: Factors $\boldsymbol{A}$ and $\boldsymbol{X}$
Initialize (randomly) $\boldsymbol{A}$ and $\boldsymbol{X}$, $s = 0$;
repeat
    $s \leftarrow s + 1$;
    $k_{inner} = \min\{10, s\}$ ;                                                   // Inner iterations
    $\boldsymbol{X} \leftarrow \texttt{NestIter}(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{X}, k_{inner})$ ;                                  // Update for $\boldsymbol{X}$
    $d_j^{(X)} = \sum_{t=1}^{T} x_{jt}$, ;                                  // $l_1$ norms of rows in $\boldsymbol{X}$
    $\boldsymbol{X} \leftarrow \text{diag}\left\{(d_j^{(X)})^{-1}\right\}\boldsymbol{X}, \quad \boldsymbol{A} \leftarrow \boldsymbol{A}\text{diag}\left\{d_j^{(X)}\right\}$, ;           // Scaling
    $\bar{\boldsymbol{A}} \leftarrow \texttt{NestIter}(\boldsymbol{X}^T, \boldsymbol{Y}^T, \boldsymbol{A}^T, k_{inner})$ ;                          // Update for $\boldsymbol{A}$
    $\boldsymbol{A} = \bar{\boldsymbol{A}}^T$;
    $d_j^{(A)} = \sum_{i=1}^{I} a_{ij}$, ;                             // $l_1$ norms of columns in $\boldsymbol{A}$
    $\boldsymbol{X} \leftarrow \text{diag}\left\{d_j^{(A)}\right\}\boldsymbol{X}, \quad \boldsymbol{A} \leftarrow \boldsymbol{A}\text{diag}\left\{(d_j^{(A)})^{-1}\right\}$, ;           // Scaling
until Stop criterion is satisfied ;
```

4 Nesterov's Iterations

The Nesterov's method [4] solves the problem of unconstrained minimization of the convex function $f(\cdot)$. Its convergence rate is determined by $O(1/k^2)$ if $f(\cdot) \in C_L^{1,1}$. It is easy to notice that the cost function (1) belongs to the class $C_L^{1,1}$ with respect to either $\boldsymbol{A}$ or $\boldsymbol{X}$. Thus:

$$||\nabla_X D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}) - \nabla_X D(\boldsymbol{Y}||\boldsymbol{A}\bar{\boldsymbol{X}})||_F \leq L_X ||\boldsymbol{X} - \bar{\boldsymbol{X}}||_F, \tag{2}$$

where $\{\boldsymbol{X}, \bar{\boldsymbol{X}}\} \in \mathbb{R}^{J \times T}$, and $L_X = ||\boldsymbol{A}^T\boldsymbol{A}||_2$ is the Lipschitz constant. From (2), we have $D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}) \leq F_X(\boldsymbol{X}, \bar{\boldsymbol{X}})$, where

$$F_X(\boldsymbol{X}, \bar{\boldsymbol{X}}) = D(\boldsymbol{Y}||\boldsymbol{A}\bar{\boldsymbol{X}}) + \left\langle \nabla_X D(\boldsymbol{Y}||\boldsymbol{A}\bar{\boldsymbol{X}}), \boldsymbol{X} - \bar{\boldsymbol{X}} \right\rangle + \frac{\tilde{L}_X}{2}||\boldsymbol{X} - \bar{\boldsymbol{X}}||_F^2$$

is the majorization function with $L_X \geq \tilde{L}_X$. The similar expressions can be written for $\boldsymbol{A}$. Assuming the proximal-gradient approach, one obtains:

$$\boldsymbol{X} = \text{prox}_h(\bar{\boldsymbol{X}}) = \arg\min_{\boldsymbol{X}} \left(h(\boldsymbol{X}) + F_X(\boldsymbol{X}, \bar{\boldsymbol{X}})\right), \tag{3}$$

where $\text{prox}_h(\bar{\boldsymbol{X}})$ is the proximal mapping of the convex function $h(\boldsymbol{X})$. For

$$h(x_{jt}) = \begin{cases} 0 & \text{if } x_{jt} \in \Omega, \\ \infty & \text{else} \end{cases} \qquad \text{where} \qquad \Omega = \{\xi : \xi \geq 0\}, \tag{4}$$

the updates for $\boldsymbol{X}$ are as follows:

$$\boldsymbol{X} = \left[\bar{\boldsymbol{X}} - \tilde{L}_X^{-1}\boldsymbol{G}_X(\bar{\boldsymbol{X}})\right]_+, \tag{5}$$

where $\boldsymbol{G}_X(\bar{\boldsymbol{X}}) = \nabla_X D(\boldsymbol{Y}||\boldsymbol{A}\bar{\boldsymbol{X}}) = \boldsymbol{A}^T(\boldsymbol{A}\bar{\boldsymbol{X}} - \boldsymbol{Y})$, and $[\xi]_+ = \max\{0, \xi\}$.

There are several strategies for determining the approximation point $\bar{\boldsymbol{X}}$. If $\bar{\boldsymbol{X}} = \boldsymbol{X}^{(k-1)}$ for the k-th iterative step, the update (5) can be considered as the projected Landweber iterations. In the Nesterov's method, $\bar{\boldsymbol{X}}$ is computed as extrapolated directions from the previous updates, that is:

$$\bar{\boldsymbol{X}} = \boldsymbol{X}^{(k-1)} + \beta^{(k)}(\boldsymbol{X}^{(k-1)} - \boldsymbol{X}^{(k-2)}). \tag{6}$$

According to [4], an optimal rate of convergence can be achieved if the factor $\beta^{(k)}$ in the k-th iteration is expressed by $\beta^{(k)} = \frac{\gamma^{(k-1)}-1}{\gamma^{(k)}}$, where $\gamma^{(k)}$ solves the quadratic equation $(\gamma^{(k)})^2 - \gamma^{(k)} - (\gamma^{(k-1)})^2 = 0$.

The final function $\texttt{NestIter}(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{X}, k_{inner})$ used in the steps 5 and 8 of Algorithm 1 is given by Algorithm 2.

Algorithm 2. NestIter

```
Input  : $\boldsymbol{A} \in \mathbb{R}^{I \times J}$, $\boldsymbol{Y} \in \mathbb{R}^{I \times T}$, $\boldsymbol{X} \in \mathbb{R}^{J \times T}$, $k_{inner}$ - number of inner iterations
Output: $\boldsymbol{X}$ - estimated factors
Initialize $\boldsymbol{X}^{(0)} = \boldsymbol{Z}^{(0)} = \boldsymbol{X}$, $L_X = ||\boldsymbol{A}^T\boldsymbol{A}||_2$, $\gamma^{(0)} = 1$;
for $k = 1, 2, \ldots, k_{inner}$ do
    $\boldsymbol{G}_X^{(k)} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Z}^{(k-1)} - \boldsymbol{Y})$ ;                // Gradient at $\boldsymbol{Z}^{(k)}$
    $\boldsymbol{X}^{(k)} = \left[\boldsymbol{Z}^{(k-1)} - L_X^{-1}\boldsymbol{G}_X^{(k)}\right]_+$ ;          // Projected updates
    $\gamma^{(k)} = \frac{1+\sqrt{4(\gamma^{(k-1)})^2-1}}{2}$,   $\beta^{(k)} = \frac{\gamma^{(k-1)}-1}{\gamma^{(k)}}$;
    $\boldsymbol{Z}^{(k)} = \boldsymbol{X}^{(k)} + \beta^{(k)}(\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k-1)})$ ;         // Search direction
```

Theorem 1. *Let the sequences* $\{\boldsymbol{X}^{(k)}\}$ *and* $\{\boldsymbol{Z}^{(k)}\}$ *be generated by Algorithm 2 for* $k \geq 1$, *and* $\boldsymbol{X}^{(*)}$ *be the limit point achieved by Algorithm 2, then*

$$D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}^{(k)}) - D(\boldsymbol{Y}||\boldsymbol{A}\boldsymbol{X}^{(*)}) \leq \frac{2L_X||\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(*)}||_F^2}{(k+2)^2}.$$

The proof of Theorem 1 is given in [6].

5 Classification Results

The experiments are carried out for the texture images taken from the UIUC database[2]. We selected 18 categories whose sample images are shown in Fig. 1. Each class consists of 16 images with inhomogeneous texture patterns and significant nonrigid deformations. To create the training set 14 images are randomly selected, and the remainder (2 images) forms the testing set. For each training or testing image 30 local regions are extracted with the Harris-Laplace affine

[2] http://perso.telecom-paristech.fr/∼xia/texture.html

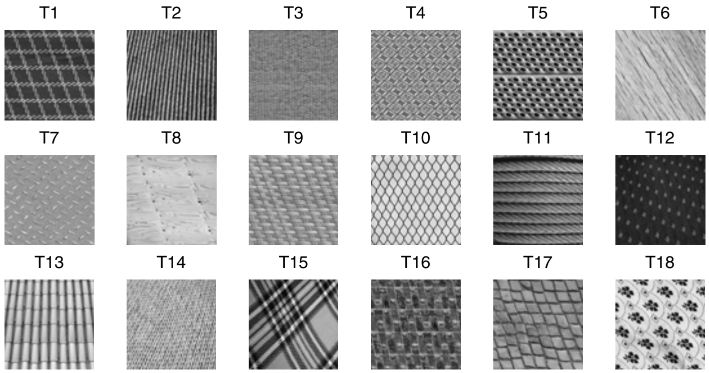
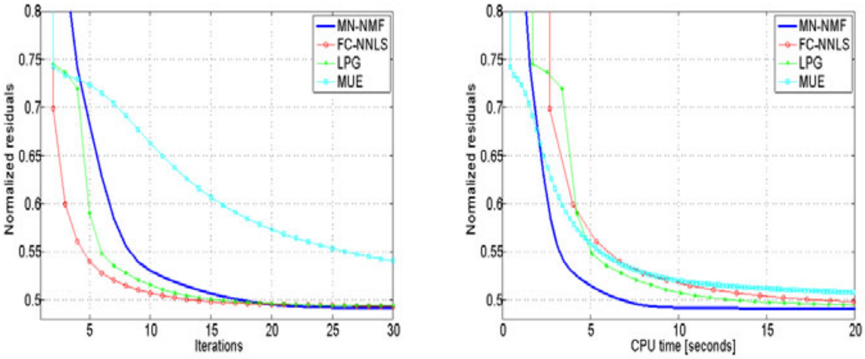


Fig. 1. Examples of texture images from each class



(a) (b)

Fig. 2. Normalized residuals: $||\boldsymbol{Y} - \boldsymbol{AX}||_F / ||\boldsymbol{Y}||_F$ versus: (a) iterations; (b) CPU time

detector, and then we select 20 the most uncorrelated sub-images. The number of key-points is adaptively selected by the SIFT descriptor, and this value ranges from few to a few dozen.

We tested the following NMF algorithms: MUE, LPG, FC-NNLS, and MN-NMF. The MUE stands for the standard Lee-Seung algorithm for minimizing the Euclidean distance [1]. The LPG[3] denotes the projected gradient NMF proposed by C. Lin [13]. The FC-NNLS [14] is a modified version of the standard NNLS algorithm that was originally proposed by Lawson and Hanson [15]. Kim and Park [16] applied the FC-NNLS to NMF-based processing of gene expression microarrays. This algorithm has been also extended and efficiently applied to supervised classification of texture patterns in [9]. The MN-NMF refers to the algorithm discussed here.

[3] http://www.csie.ntu.edu.tw/∼cjlin/nmf/

Each tested NMF algorithm is initiated randomly and runs for 50 iterations with $J = 18$. The initial regularization parameter η_0 in the FC-NNLS [9] was set to $\eta_0 = 10^{-4}$.

The normalized residuals $||\boldsymbol{Y} - \boldsymbol{AX}||_F / ||\boldsymbol{Y}||_F$ versus alternating steps are shown in Fig. 2. The averaged recognition rates and the elapsed time over 100 Monte Carlo (MC) runs are presented in Table 1.

Table 1. Statistics for testing NMF algorithms with 100 MC runs. Each algorithm and in each MC run performs 50 alternating steps.

	MUE	LPG	FC-NNLS	MN-NMF
Time [seconds]	10.7	42.5	67.1	19.6
Mean recognition rate	96.4	98.5	98.6	98.6
Std.	2.7	1.9	1.8	1.8

6 Conclusions

Our tests (Table 1) demonstrated that the FC-NNLS, LPG and MN-NMF algorithms give nearly the same accuracy of the classification results, measured in terms of the mean recognition rate and the standard deviation. However, according to Table 1 the MN-NMF algorithm after 50 alternating steps is more than twice faster than the LPG, and more than 3 times faster than the FC-NNLS. Fig. 2(a) shows that the FC-NNLS and LPG are initially faster than the MN-NMF but a single iteration of both the FC-NNLS and LPG involves more computational effort than the MN-NMF, hence we observe the faster convergence of the MN-NMF versus the elapsed time in Fig. 2(b).

The computational complexity of the MN-NMF algorithm for updating $\boldsymbol{X}$ is $O(IJ^2 + IJT) + k\left(O(J^2T) + 2O(JT)\right)$, where k is the number of inner iterations. Note that the MUE has the computational complexity $O(IJ^2 + IJT) + O(J^2T) + 2O(JT)$ per one update of $\boldsymbol{X}$. In our example, $I = 128$, $J = 18$, and $T = 25200$ (roughly estimated). Thus, the MN-NMF and MUE perform 50 alternating steps with 7.033 and 3.359 Giga floating-point operations, respectively. This justifies the elapsed time in Table 1. The computational complexity of the remaining algorithms is rather difficult to estimate since it depends on the dataset. The number of inner iterations in the LPG can variate in each alternating step (governed by the Armijo rule). Motivated by the step 4 in Algorithm 1, we set the maximum number of inner iterations in the LPG to 10.

Summing up, our experiments demonstrate that there is some computational potential in the Nesterov's method and replacing multiplicative algorithms in NMF with the optimal gradient method might be beneficial for texture pattern classification.

Acknowledgment. This work was partially supported by the habilitation grant N N515 603139 (2010-2012) from the Ministry of Science and Higher Education, Poland.

References

[1] Lee, D.D., Seung, H.S.: Learning of the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)

[2] Qin, L., Zheng, Q., Jiang, S., Huang, Q., Gao, W.: Unsupervised texture classification: Automatically discover and classify texture patterns. Image and Vision Computing 26(5), 647–656 (2008)

[3] Sandler, R., Lindenbaum, M.: Nonnegative matrix factorization with earth mover's distance metric. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1873–1880. IEEE Computer Society, Los Alamitos (2009)

[4] Nesterov, Y.: A method of solving a convex programming problem with convergence rate $o(1/k^2)$. Soviet Mathematics Doklady 27(2), 372–376 (1983)

[5] Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. Wiley and Sons (2009)

[6] Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1), 183–202 (2009)

[7] Zhou, T., Tao, D., Wu, X.: NESVM: A fast gradient method for support vector machines. In: ICDM, pp. 679–688 (2010)

[8] Guan, N., Tao, D., Luo, Z., Yuan, B.: NeNMF: An optimal gradient method for solving non-negative matrix factorization and its variants. Technical report, Mendeley database (2010)

[9] Zdunek, R.: Supervised classification of texture patterns with nonnegative matrix factorization. In: The 2011 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV 2011), vol. II, pp. 544–550. CSREA Press, Las Vegas (2011); WORLDCOMP 2011

[10] Lindeberg, T., Garding, J.: Shape-adapted smoothing in estimation of 3-d depth cues from affine distortions of local 2-d brightness structure. Image and Vision Computing 15, 415–434 (1997)

[11] Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, pp. 1150–1157 (1999)

[12] Zdunek, R., Phan, A., Cichocki, A.: Damped Newton iterations for nonnegative matrix factorization. Australian Journal of Intelligent Information Processing Systems 12(1), 16–22 (2010)

[13] Lin, C.J.: Projected gradient methods for non-negative matrix factorization. Neural Computation 19(10), 2756–2779 (2007)

[14] Benthem, M.H.V., Keenan, M.R.: Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. Journal of Chemometrics 18, 44–450 (2004)

[15] Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. Prentice-Hall, Englewood Cliffs (1974)

[16] Kim, H., Park, H.: Non-negative matrix factorization based on alternating nonnegativity constrained least squares and active set method. SIAM Journal in Matrix Analysis and Applications 30(2), 713–730 (2008)

Detection of Aliasing in Image Sequences Using Nonlinear Factor Analysis

Scott C. Douglas

Department of Electrical Engineering
Bobby B. Lyle School of Engineering
Southern Methodist University
Dallas, Texas 75275 USA
douglas@lyle.smu.edu

Abstract. In computational imaging, reconstructing a single high-resolution scene from multiple low-resolution aliased images is most efficient if done only over those regions where significant aliasing occurs. This paper presents a framework for detecting pixel locations exhibiting the most-prominent effects of aliasing in a sequence of subpixel-shifted images. The process employs nonlinear factor analysis of the image sequence, in which the latent variables are the relative position offsets for each image in the sequence, followed by outlier detection on the error residuals from the joint estimation process. Numerical examples illustrate the capabilities of the methodology.

Keywords: aliasing, computational imaging, image reconstruction, nonlinear factor analysis, signal detection.

1 Introduction

Computational imaging refers to imaging system designs for producing high-quality visual images through numerical procedures on the collected data. In these systems, the image content is often split across the measurements in non-conventional ways, and a final image is obtained only after mathematical reconstruction is performed. One well-known problem in computational imaging is the reconstruction of high-quality images from multiple low-resolution aliased and shifted versions of these images. In this context, aliased information is desirable, as it provides additional information beyond the Nyquist limit of the sensor array to enhance spatial detail using multiple observations [1].

Most algorithms for multi-frame image reconstruction under aliasing assume that all regions of the scene are aliased and apply advanced reconstruction processes everywhere in the scene. In practice, large portions of the scene may not be aliased, and such an approach is wasteful of both computational and imaging resources. The PANOPTES architecture attempts to achieve a matching of computational imaging resources to scene complexity through careful management of both imaging resources and computational effort [2]. In this context, a key challenge is the following: How does one determine which portions of the scene

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 486–493, 2012.

suffer from excessive aliasing? This paper considers the problem of identifying aliased regions within image sequences as a nonlinear factor analysis problem, in which image offsets are the latent variables. The error residuals from this estimation process are used to detect those regions that exhibit significant aliasing. Application of the procedure to both synthetic and real-world image sequences shows the efficacy of the method.

2 Nonlinear Factor Analysis for Modeling Image Offsets

Assume that a digital image $g[m,n]$ is generated by ideal sampling of a image field $f(x,y)$ via the relation

$$g[m,n] = f(x,y)|_{x=mT,y=nT}\,, \tag{1}$$

where T is the spatial sampling period. The image field models all effects due to blurring by the optical imaging system and sensor averaging, and sensor noise is neglected. We collect N such images $g_i[m,n]$ that differ from each other due to spatial offsets $\{\Delta_x[i], \Delta_y[i]\}$, $1 \leq i \leq N$ as

$$g_i[m,n] = f(x-\Delta_x[i], y-\Delta_y[i])|_{x=mT,y=nT}\,. \tag{2}$$

The goal is to model this image set assuming that the N images are suitably "smooth," such that aliasing is neglected. Those regions that violate this model will suffer from significant aliasing artifacts.

The smoothness model used for the image set is similar to that used in determining optical flow [3] and is given by the two-dimensional Taylor series

$$\begin{aligned} g_i[m,n] = {} & g[m,n] + p_2[m,n]\Delta_x[i] + p_3[m,n]\Delta_y[i] \\ & + p_4[m,n](\Delta_x[i])^2 + p_5[m,n](\Delta_y[i])^2 + p_6[m,n]\Delta_x[i]\Delta_y[i] \\ & + \{\text{higher order terms}\}\,. \end{aligned} \tag{3}$$

where $p_j[m,n]$, $2 \leq j \leq 6$ are spatial derivative terms to be described. In what follows, we assume that the higher order terms in the expansion can be neglected for image pixels with no aliasing. Indexing the data in a 1-D raster format, we can express the qth image sample of the ith image as $g_i[q] = \mathbf{p}^T[q]\mathbf{d}[i]$, where $p_1[q] = g[q]$, $1 \leq q \leq P$, P is the number of pixels in each aliased image, and

$$d[i] = [1 \quad \Delta_x[i] \quad \Delta_y[i] \quad (\Delta_x[i])^2 \quad (\Delta_y[i])^2 \quad (\Delta_x[i]\Delta_y[i])]^T. \tag{4}$$

Collected all image pixels across all scenes, we form the imaging model

$$\mathbf{G} = \mathbf{PD}. \tag{5}$$

The dimensions of $\mathbf{G}$, $\mathbf{P}$, and $\mathbf{D}$ are $(P \times N)$, $(P \times 6)$ and $(6 \times N)$, respectively, where P is the number of pixels in each aliased image. Eqn. (5) is in the form of a nonlinear factor model employing a second-order bivariate Taylor series [4]. Solutions to this problem are reasonable if (a) the number of pixels P in each aliased image much larger than the number of collected images N, and (b) the number of aliased images collected is significantly greater than the dimension of the nonlinear factor model. In practice, we desire $N \gg 6$, although there are only $2N$ unknowns in the definition of $\mathbf{D}$.

To jointly estimate $\mathbf{P}$ and $\mathbf{D}$ from $\mathbf{G}$, we use the standard approach of alternating least-squares. Starting with a chosen initial $\mathbf{P}^{(0)}$, we iterate the following:

$$\mathbf{D}^{(k)} = (\mathbf{P}^{(k-1)T}\mathbf{P}^{(k-1)})^{-1}\mathbf{P}^{(k-1)T}\mathbf{G} \tag{6}$$

$$\mathbf{P}^{(k)} = \mathbf{G}\mathbf{D}^{(k)T}(\mathbf{D}^{(k)}\mathbf{D}^{(k)T})^{-1} \tag{7}$$

In the above, we only need to solve for the second and third row of $\mathbf{D}^{(k)}$ at each iteration, as the first row of $\mathbf{D}^{(k)}$ is forced to be unity, and the last three rows of $\mathbf{D}^{(k)}$ are quadratic forms of the second and third rows. To initialize the entries of $\mathbf{P}^{(0)}$, we generate a prior $p_1^{(0)}[m,n] \approx g[m,n]$ to be described shortly, and then form estimates of the spatial derivatives of this prior over the entire image, where we use the linear convolutional models

$$p_q^{(0)}[m,n] = \sum_{i,j} h_q[i,j]p_1^{(0)}[m-i,n-j], \quad 2 \leq q \leq 6, \quad \text{where} \tag{8}$$

- $p_2^{(0)}[m,n]$ is the horizontal difference of $p_1^{(0)}[m,n]$, where $\mathbf{H}_2 = [-1 \quad 1]$.
- $p_3^{(0)}[m,n]$ is the vertical difference of $p_1^{(0)}[m,n]$, where $\mathbf{H}_3 = [-1 \quad 1]^T$.
- $p_4^{(0)}[m,n]$ is the twice-horizontal-difference of $p_1^{(0)}[m,n]$, where $\mathbf{H}_4 = [1 \quad -2 \quad 1]$.
- $p_5^{(0)}[m,n]$ is the twice-vertical-difference of $p_1^{(0)}[m,n]$, where $\mathbf{H}_5 = [1 \quad -2 \quad 1]^T$.
- $p_6^{(0)}[m,n]$ is the local Laplacian of $p_1^{(0)}[m,n]$, where $\mathbf{H}_6 = \mathbf{H}_5 \cdot \mathbf{H}_4$.

Other convolutional kernels involving difference-of-Gaussians and Laplacian-of-Gaussians were also tried but resulted in no significant performance differences.

As for the initial prior $p_1^{(0)}[m,n]$, we do not have access to an undistorted version of the downsampled image, only its aliased versions in the columns of $\mathbf{G}$. Without any prior information, a reasonable selection is the average of the measured images at each pixel position, or

$$p_1[m,n] = \frac{1}{N}\sum_{i=1}^{N} g_i[m,n]. \tag{9}$$

Other estimators are also possible and are the subject of further investigation.

Table 1 provides the entire algorithm in a convenient listing. Convergence of the iterative portion is fast and typically takes only a few iterations. The error matrix $\mathbf{E}$ shown at convergence is equivalent to

$$\mathbf{E} = \mathbf{G} - \mathbf{P}^{(f)}\mathbf{D}^{(f)}. \tag{10}$$

Remark: The complexity of the iterative portion of the algorithm is much-reduced when one considers the fact that $\mathbf{P}^{(k)}$ need not be formed at each iteration k. The algorithm manipulates matrices with a maximum dimension of N once $\mathbf{D}^{(1)}$ is found. Since $P \gg N$ in practice, the most-significant computations are the calculation of $\mathbf{D}^{(1)}$ and $\mathbf{R_G}$ at initialization and $\mathbf{E}$ at convergence.

Table 1. Nonlinear Factor Analysis for Detecting Aliased Regions

```
Initialize. Compute $\mathbf{P}^{(0)}$, $\mathbf{R_G} = \mathbf{G}^T\mathbf{G}$, and the second and third rows of
    $\mathbf{D}^{(1)} = (\mathbf{P}^{(0)T}\mathbf{P}^{(0)})^{-1}(\mathbf{P}^{(0)T}\mathbf{G})$.
Iteration: For $k = 1, 2, \ldots$, do until convergence ("iteration f")
    1. Let $d_{pq}^{(k)}$ be the $(p, q)$th element of $\mathbf{D}^{(k)}$. Then, compute the first, fourth,
       fifth, and sixth rows of $\mathbf{D}^{(k)}$ as follows: for all $1 \leq q \leq N$,
       $d_{1q}^{(k)} = 1, \quad d_{4q}^{(k)} = \left(d_{2q}^{(k)}\right)^2 \quad d_{5q}^{(k)} = \left(d_{3q}^{(k)}\right)^2 \quad d_{6q}^{(k)} = d_{2q}^{(k)}d_{3q}^{(k)}$
    2. Compute the second and third row of
       $\mathbf{D}^{(k+1)} = (\mathbf{D}^{(k)}\mathbf{D}^{(k)T})(\mathbf{D}^{(k)}\mathbf{R_G}\mathbf{D}^{(k)T})^{-1}\mathbf{D}^{(k)}\mathbf{R_G}$.
Finalize: Set the first, fourth, fifth, and sixth rows of $\mathbf{D}^{(\mathrm{f})}$ as in Step 1, and
    $\mathbf{E} = \mathbf{G} - \mathbf{G}\mathbf{D}^{(\mathrm{f})T}(\mathbf{D}^{(\mathrm{f})}\mathbf{D}^{(\mathrm{f})T})^{-1}\mathbf{D}^{(\mathrm{f})}$.
```

3 Detection of Aliased Regions

Regions that are not well-modeled by the model in (5) are not spatially-smooth, and these positions likely correspond to areas with significant aliasing. Clearly, such regions can be identified by those rows in **E** where the errors are large. The key concept is determining the detection mechanism.

We have explored numerous examples involving sequences of aliased and non-aliased imagery under subpixel translation. When aliasing is present, the errors in the nonlinear factor analysis model are "heavy-tailed" and appear as outliers in the error images. Fig. 1 shows results of three particular synthetic numerical examples showing the logs of the histograms of the magnitudes of the elements of **E** for images with only smooth features, for images with only aliased features, and for images with both smooth and aliased features. When the underlying features are smooth, the percentage of errors decreases to zero as $|e|$ is increased. When only aliasing is present, the error distribution is considerably flatter, indicating the presence of strong outliers. When there is a combination of aliased and non-aliased features, the outliers due to aliasing are also clearly present. Thus, we can detect the locations of aliased regions by calculating the histogram of the absolute values of the elements of **E**, determining a threshold from the distribution where the log of the p.d.f. experiences an upwards inflection, and using this threshold to detect those pixels where the errors are large. In the plots of Fig. 1, the chosen threshold is $\tau = 0.015$.

In practice, the detection of aliased regions is jointly performed over all of the error images, as we expect these regions to exhibit large errors for several image frames. In the examples that follow, we use the following detection criterion: a pixel is identified to be part of an aliased region when $|e_i[m, n]|$ exceeds τ for at least 10% of the images in the sequence. Other criteria are possible and are under investigation.

4 Numerical Examples

The first image sequence tested using the described procedure is synthetically-generated to show its capabilities under complete sequence knowledge. The

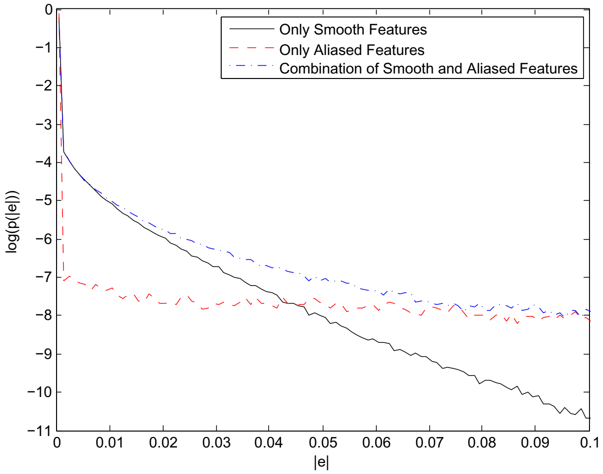


Fig. 1. The distributions of $|e_i[m,n]|$ for three different synthetic image sequences

upper-left image in Fig. 3 shows the high-resolution image, which consists of a set of 80 binary squares that are of size 3×3 to size 16×16 in high-resolution pixel space, along with the same 80 squares that have been filtered by a Gaussian kernel with $\sigma = 5$. This (600×600) pixel image is offset by 7 different subpixel shifts in both x and y directions, and the resulting images are downsampled via pixel averaging to produce 49 different (120×120) aliased images. The upper-right image in Fig. 3 shows one image of this sequence. The proposed algorithm is applied to this data. The middle-left image in Fig. 3 is the sum of the absolute errors of the estimation procedure across all 49 images, inverted for clarity. The aliased portions of the image clearly have higher error magnitudes. The middle-right image in Fig. 3 shows those pixel positions for which $|e_i[m,n]| > \tau$ for at least 4 images in the sequence. As can be seen, the regions where aliasing occurs are clearly detected, whereas the smoothed regions are largely not detected.

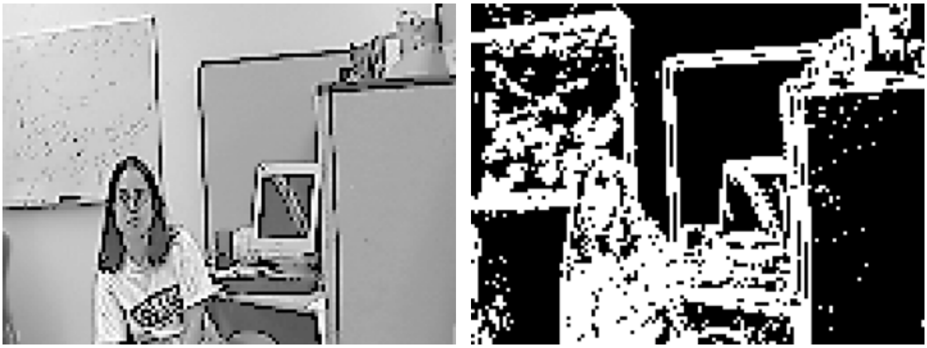


Fig. 2. *Real-world image sequence* (Left) One image from the first thirty frames of the aliased Emily sequence [5]. (Right) Aliased regions as bright pixels, detected as described in the text.

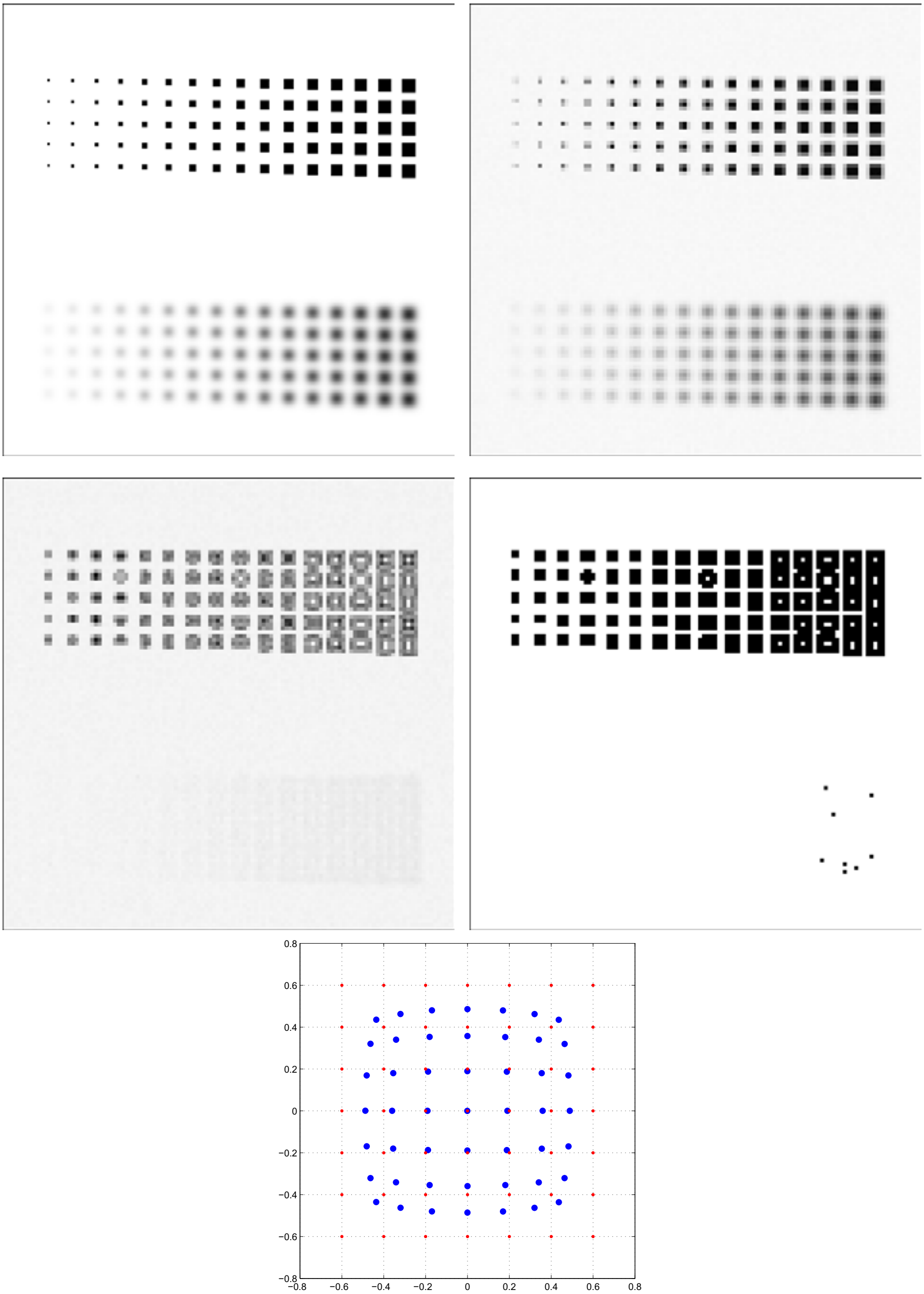


Fig. 3. *Synthetic image sequence example* (Upper Left) Original image. (Upper Right) One of the 49 downsampled images. (Middle Left) Sum of absolute errors across all 49 images after nonlinear factor analysis, shown in inverted grayscale. (Middle Right) Aliased regions as dark pixels, detected as described in the text. (Bottom) Estimated offsets [large blue dots] and actual offsets [small red dots].

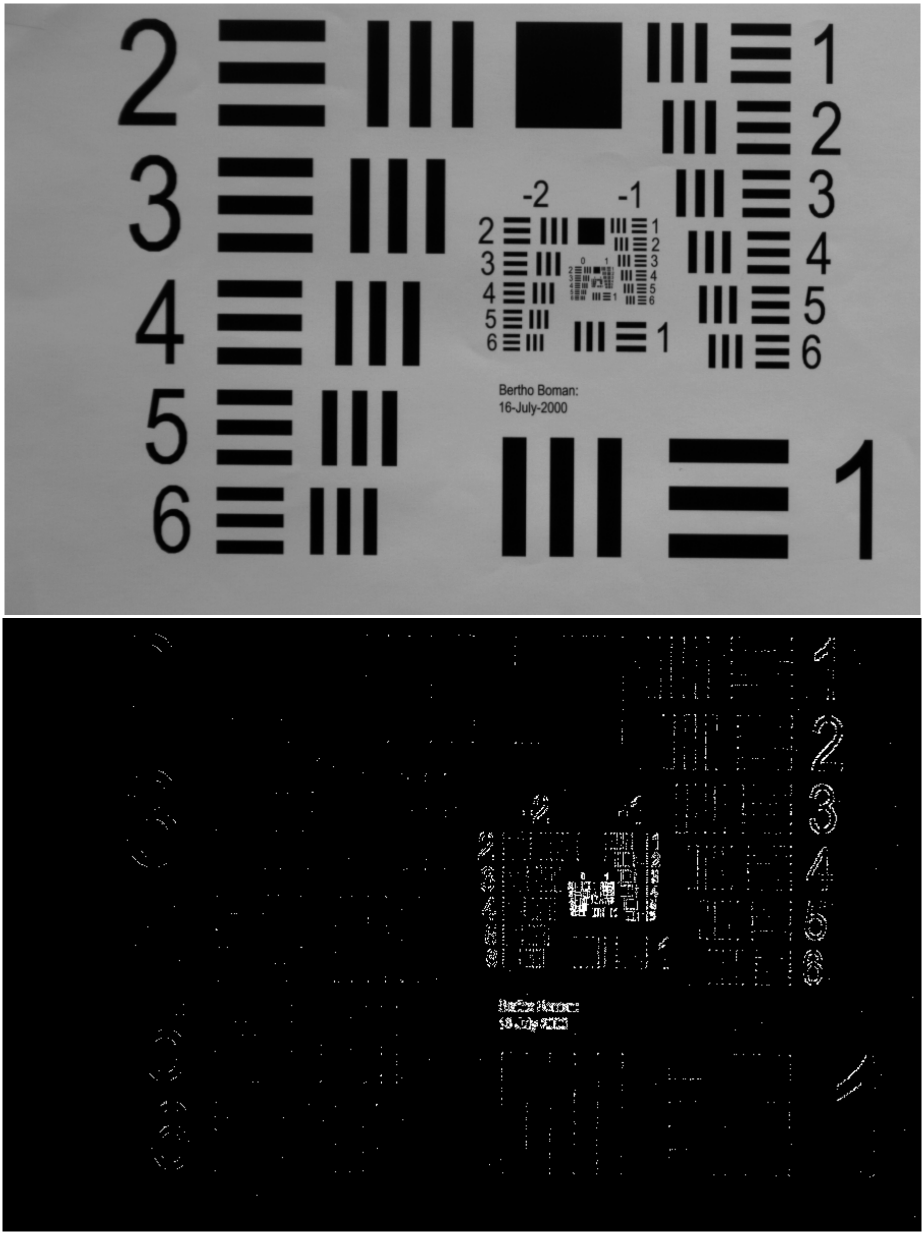


Fig. 4. *Real-world image sequence* (Top) One image from the 25-frame aliased Air Force target sequence. (Bottom) Aliased regions as bright pixels, detected as described in the text.

The bottom plot in Fig. 3 shows the estimated offsets determined by the nonlinear factor analysis procedure in blue as well as the true offsets for the images in red. The general distribution and gridding of the estimated offsets is similar to the actual offsets, but the estimates are clearly biased. We believe the issue is the accuracy of the estimated prior $p_1^{(0)}[m, n]$ in the presence of significant aliasing in the synthetic example. The estimated positions depend on the accuracy of this prior, but the algorithm's ability to detect aliased regions is not significantly hampered by these differences.

We now explore the behavior of the procedure on real-world imagery. The top image of Fig. 4 is one of a 25 image sequence collected as part of a precision imaging collection system at SMU. Subpixel shifts in the image capture were obtained by computer-controlled precision micrometer adjustment of the camera position for a fixed Air Force resolution target. These 25 images were modeled using the nonlinear factor analysis procedure, and the absolute values of the error residuals were used to detect aliased regions in which $\tau = 0.0185$ was selected. The bottom image of Fig. 4 shows those pixels whose error magnitudes exceeded the chosen threshold for two or more images. The pixels identified by the process occur along edges of large objects that exhibit ±45-degree slopes where the spatial sampling rate is lower, as well as in the finer resolution portions that are known to be aliased for this data set. Only a small portion of the image region – 0.949% of the (601×901) pixel area – is identified in this process, suggesting that much of the image area does not require resolution enhancement.

Fig. 2 shows the application of the procedure on the Emily sequence available from [5]. The first thirty frames of this sequence exhibit largely translational motion. Shown on the left is one image from this 30-frame sequence. Shown on the right are the aliased pixel regions identified by the procedure in this paper, in which $\tau = 0.0062$ was selected. As can be seen, the portions of the image selected represent areas where spatial detail can be improved, including regions of the white board, the seated individual, and borders of larger objects.

Acknowledgements. The author would like to thank Indranil Sinharoy and Vikrant Bhakta for supplying the Air Force target image sequence, and the PANOPTES group for fruitful discussions.

References

1. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and robust multi-frame super-resolution. IEEE Trans. Image Processing 13, 1327–1344 (2004)
2. Bhakta, V.R., Somayaji, M., Douglas, S.C., Christensen, M.P.: Experimentally validated computational imaging with adaptive multiaperture folded architecture. Applied Optics 49(10), B51–B58 (2010)
3. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intell. 17, 185–203 (1981)
4. Maxwell, A.E.: Contour analysis. Educ. Psychol. Measmt. 17, 347–360 (1957)
5. Data available from the Multi-Dimensional Signal Processing (MDSP) research lab at the University of California at Santa Cruz, `http://users.soe.ucsc.edu/~milanfar/software/sr-datasets.html`

Geometrical Method Using Simplicial Cones for Overdetermined Nonnegative Blind Source Separation: Application to Real PET Images

Wendyam S.B. Ouedraogo[1,2,3], Antoine Souloumiac[1],
Meriem Jaidane[2], and Christian Jutten[3]

[1] CEA, LIST, Laboratoire d'Outils pour l'Analyse de Données,
Gif-sur-Yvette, F-91191, France
[2] Unité Signaux et Systèmes, École Nationale d'Ingénieurs de Tunis,
BP 37, 1002 Tunis, Tunisie
[3] GIPSA-lab, UMR 5216 CNRS, Université de Grenoble, 961 rue de la Houille
Blanche BP 46 F-38402 Grenoble Cedex, France
wendyam-serge-boris.ouedraogo@cea.fr

Abstract. This paper presents a geometrical method for solving the overdetermined Nonnegative Blind Source Separation (N-BSS) problem. Considering each column of the mixed data as a point in the data space, we develop a Simplicial Cone Shrinking Algorithm for Unmixing Nonnegative Sources (SCSA-UNS). The proposed method estimates the mixing matrix and the sources by fitting a simplicial cone to the scatter plot of the mixed data. It requires weak assumption on the sources distribution, in particular the independence of the different sources is not necessary. Simulations on synthetic data show that SCSA-UNS outperforms other existing geometrical methods in noiseless case. Experiment on real Dynamic Positon Emission Tomography (PET) images illustrates the efficiency of the proposed method.

Keywords: Blind Source Separation, Nonnegativity, Simplicial Cone, Minimum Volume, Facet.

1 Introduction

We deal with the problem of Nonnegative Blind Source Separation (N-BSS) in noiseless, linear intantaneous mixture case. The mixture model is given by:

$$\underset{m\times p}{X} = \underset{m\times n}{A}\underset{n\times p}{S} \tag{1}$$

where m is the number of observations, n is the number of sources and p is the number of samples. X, A and S are respectively the given observations matrix, the unknown mixing matrix and the hidden nonnegative sources matrix. N-BSS consists on retrieving S and A given only X. Possible directions for solving problem (1) are the geometrical approaches. These methods are very natural and intuitive and require weak assumption on the sources distribution. The first geometrical method

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 494–501, 2012.

was introduced by Puntonet et al. [1] for separating two sources having bounded probability density functions. The mixing matrix is retrieved by finding the slopes of the parallelogram containing the scatter plot of mixed data. Babaie-Zadeh et al. [2] propose another geometrical method applicable to more than two sources based on clustering the observations points and fitting a line (hyper-plane) to each cluster to recover the mixing matrix. By assuming the (very strong) condition ***local dominance of the sources*** (i.e. for every source there is at least one instant where the underlined source is active and all the others are not), several authors propose estimating the mixing matrix by looking for the vertices of the convex hull of the scatter plot of mixed data [3][4]. These methods can unfortunately be very slow and demanding very large size samples, specially for large scale problem due to the convex hull computing. Noting that when the sources are nonnegative, the scatter plot of mixed data is contained in the simplicial cone generated by the mixing matrix, other geometrical methods were proposed for solving problem (1) by looking for the Minimum Volume (MV) simplicial cone containing the mixed data [5][6]. MV like methods do not require local dominance of sources. But the simplicial cone generated by the mixing matrix must be well recognizable from the scatter plot of mixed data. This weaker condition implies that there should be at least $n-1$ mixed data points on, or close to, each ***facet*** of this cone. It's also necessary to specify that beside the nonnegativity, some MV like methods developed for hyperspectral data processing [5][6] require ***full additivity of the sources*** (i.e. the sum on every column of the sources matrix is equal to one).

This paper presents a MV like method for solving overdetermined N-BSS problem called Simplicial Cone Shrinking Algorithm for Unmixing Nonnegative Sources (SCSA-UNS). This work establishes an extension of [9] in which we only consider the determined case with nonnegative mixing matrix. Section 2 reviews the geometrical view of the N-BSS problem and derives the main idea of MV like methods. In section 3, we describe the proposed SCSA-UNS method. Section 4 presents simulation results on synthetic data and real Dynamic Positon Emission Tomography (PET) data and comparisons with other MV like methods. Finally section 5 presents the conclusions and future works.

2 Geometrical View of N-BSS Problem

Let's review the geometrical view of the N-BSS problem we described in [9]. For a given matrix $W = [w_1, w_2, \cdots, w_n]$, we define the ***Simplicial Cone*** $Span^+(W)$ generated by W by :

$$Span^+(W) = \left\{ z \,\middle|\, z = Wy \text{ with } y \in \mathbb{R}^n_+ \right\} \tag{2}$$

We also define the ***positive orthant*** $\mathbb{R}^n_+$ as being the simplicial cone generated by the identity matrix I_n: $\mathbb{R}^n_+ = Span^+(I_n)$

By considering each column x_i of X $(1 \leq i \leq p)$ as a point in the n dimension data space, it comes that when the sources are nonnegative, the mixed data form a cloud of points contained in the simplicial cone generated by the mixing matrix.

$$\{x_i \,|\, x_i \in X, 1 \leq i \leq p\} \subseteq Span^+(A) \tag{3}$$

One can thus imagine to estimate A (up to the classical (positive here) scaling and permutation BSS indetermination's) by finding a simplicial cone containing all the mixed data. But without any additional condition there is an infinite number of such cones. So for recovering the mixing matrix, we require the scatter plot of the mixed data to ***fill enough*** $Span^+(A)$ and we look for the Minimum Volume (i.e. the smallest) simplicial cone containing the mixed data [7]. Filling enough means that $Span^+(A)$ must be well recognizable from $\{x_i, x_i \in X, 1 \leq i \leq p\}$ (i.e. there should be at least $n-1$ mixed data points on, or close to, each ***facet*** of $Span^+(A)$ [8]). This intuitive and natural condition will be defined more formally in a future work.

3 Geometrical Method Using Simplicial Cones for Overdetermined Nonnegative Blind Source Separation

3.1 Determined Case : Simplicial Cone Shrinking Algorithm for Unmixing Nonnegative Sources (SCSA-UNS)

We first restrict to determined case with full column rank nonnegative mixing matrix (i.e. $m = n$ and $A \geq 0$). This case is considered in [9]. SCSA-UNS aims at estimating A by finding the Minimum Volume simplicial cone containing all the mixed data. In this objective, we propose a criterion for measuring the volume of a given simplicial cone and an algorithm to minimize this criterion.

Proposed Criterion : We define $V(W)$, the volume of a simplicial cone $Span^+(W)$ generated by a given square matrix $W = [w_1, w_2, \cdots, w_n]$, where w_i is the i-th column of W, by :

$$V(W) = \frac{|\det(W)|}{\|w_1\| \times \|w_2\| \times \cdots \times \|w_n\|} \tag{4}$$

$V(W)$ strictly represents the "aperture" of the simplicial cone $Span^+(W)$, it is positive and upper bounded by 1 (Hadamard's Inequality).
The task of estimating the mixing matrix can then be reduced to solving the following optimization problem:

$$W^* = \underset{W \geq 0,\ W^{-1}X \geq 0}{\arg\min} \; V(W) \tag{5}$$

Proposed Algorithm : We define the R_k^l like matrices by :

$$R_k^l = \begin{pmatrix} 1 & 0 & \cdots & 0 & r_{1k}^l & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & r_{2k}^l & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & r_{k-1k}^l & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & r_{k+1k}^l & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & r_{nk}^l & 0 & \cdots & 1 \end{pmatrix} \text{ where } r_{ik}^l \geq 0,\ \forall\ 1 \leq i \leq n, i \neq k \tag{6}$$

Proposition 1. *For a given nonnegative matrix W, the volume $V(W)$ of $Span^+(W)$ decreases when W is multiplied to the right by R^l_k, $\forall\ l$ and $\forall\ 1 \leq k \leq n$.*

Proof. For fixed l, fixed k $(1 \leq k \leq n)$ and a given nonnegative square matrix $W = [w_1, w_2, \cdots, w_n]$, let $U = WR^l_k$ $(U = [u_1, u_2, \cdots, u_n])$. The task is to demonstrate that $V(U) \leq V(W)$:
1. $|\det(U)| = \left|\det(WR^l_k)\right| = |\det(W)| \left|\det(R^l_k)\right| = |\det(W)|$ because $\det(R^l_k) = 1$
2. $\forall\ j \neq k, u_j = w_j \Longrightarrow \|u_j\| = \|w_j\|$
3. For $j = k$, $\forall\ 1 \leq i \leq n$, $u_{ik} = w_{ik} + \sum_{q=1,q\neq k}^{n} w_{iq} r^l_{qk} \geq w_{ik}$ because w_{iq} and r^l_{qk} are all nonnegative. Therefore $(u_{ik})^2 \geq (w_{ik})^2$ which leads to $\|u_k\| \geq \|w_k\|$
According to the definition of the volume of a simplicial cone given by (4), 1., 2. and 3. allow us to conclude that $V(U) \leq V(W)$

For solving problem (5), SCSA-UNS starts from an initial simplicial cone containing all the mixed data (typically, the positive orthant $Span^+(I_n)$) and iteratively decreases its volume by performing several sweeps of n right-multiplications of the matrix which generated this initial cone ($W_0 = I_n$) by the R^l_k matrices $(1 \leq k \leq n)$. At each iteration, the matrix R^l_k is computed so to keep the mixed data inside the new simplicial cone. Details for computing the R^l_k are given in Appendix 1. Let $W \geq 0$ be the matrix which generated the current simplicial cone (W is the current estimation of the mixing matrix and $Y = W^{-1}X \geq 0$ is the current estimation of the sources). The algorithm stops when one cannot decrease anymore $V(W)$ without creating negative values in Y. This often corresponds to the convergence of W to the true mixing matrix A. However, it may happen that $V(W)$ does not decrease anymore during the iterations while W has not converged yet to A. This freezing[1] situation arises when there is at least one zero value on each row of Y. To avoid this problem, we suggest applying, after each sweep l, an orthogonal linear transformation Q_l to Y (and Q_l^T to W) to delete the zeros values of Y without increasing $V(W)$. The details of computing the unfreezing matrices Q_l are given in Appendix 2.

The whole estimated mixing matrix and sources are given by:

$$A = W_0 \prod_l \left[\left(\prod_{k=1}^{n} R^l_k \right) Q_l^T \right] \quad \text{and } S = A^{-1}X \tag{7}$$

3.2 Proposed Method for Overdetermined Case

When the number of observations is greater than the number of sources ($m > n$), we propose to first perform a dimension reduction of the mixed data and afterward run the SCSA-UNS algorithm on the reduced mixed data.

Dimension Reduction : The dimension reduction is performed by the classical Principle Component Analysis (PCA). We compute the Singular Value Decomposition (SVD) of the mixed data and we keep only the n largest singular values and the corresponding singular vectors.

[1] See Appendix 2 for more details.

$$\underset{m\times p}{X} \approx \underset{m\times n}{E}\underset{n\times n}{F}\underset{n\times p}{G^T} \tag{8}$$

The SCSA-UNS algorithm should be executed on G^T but G^T is not necessarily nonnegative, so its scatter plot is not necessarily contained in the positive orthant. One must then find an initial simplicial cone containing all the dimension reduced mixed data G^T, since $Span^+(I_n)$ is not necessarily suited any more.

Widening Procedure for Initialization : If all the observations are not nonnegative, we have developed a procedure for finding a convenient initial simplicial cone containing all the mixed for the SCSA-UNS algorithm. This procedure is based on widening $Span^+(I_n)$ by multiplying I_n by D_k^l matrices, with a structure similar to R_k^l, but with negative entries in order to increase the volume of $Span^+(I_n)$ up to enclose all the mixed data. The details of computation of the D_k^l are not given here due to lack of space. This widening procedure is used on G^T to compute W_1 and Y_1 so that:

$$\underset{n\times p}{G^T} = \underset{n\times n}{W_1}\underset{n\times p}{Y_1} \quad \text{with} \quad Y_1 \geq 0 \tag{9}$$

Estimating the Mixing Matrix and the Sources : The SCSA-UNS algorithm is finally executed on Y_1 to give $\underset{n\times p}{Y_1} = \underset{n\times n}{W_2}\underset{n\times p}{Y_2}$.
The whole mixing matrix and sources are estimated by:

$$\underset{m\times n}{A} \approx \underset{m\times n}{E}\underset{n\times n}{F}\underset{n\times n}{W_1}\underset{n\times n}{W_2} \text{ and } \underset{n\times p}{S} \approx \underset{n\times p}{Y_2} \tag{10}$$

4 Simulations and Discussions

Case 1: Synthetic Data

The proposed method is first evaluated on synthetic data and compared with two other MV methods: MVSA [5] and MVES [6]. To make sure that the mixed data *fill enough* the simplicial cone generated by the mixing matrix, the nonnegative sources have been generated using "random sparse uniform distribution generator" with 64% of non-zero elements in the sources matrix. We consider two cases: full additive sources and non full additive sources. The mixing matrix has Gaussian random entries. We set $m = 20$, $n = 5$ and $p = 10000$. Comparison criteria are the CPU time to converge T and the separation error E_{sep} defined by (11). The smaller E_{sep}, the better the separation.

$$E_{sep} = \frac{1}{n(n-1)}[\sum_i(\sum_j \frac{|(W^{-1}A)_{ij}|}{\max_k |(W^{-1}A)_{ik}|} - 1) + \sum_j(\sum_i \frac{|(W^{-1}A)_{ij}|}{\max_k |(W^{-1}A)_{kj}|} - 1)] \tag{11}$$

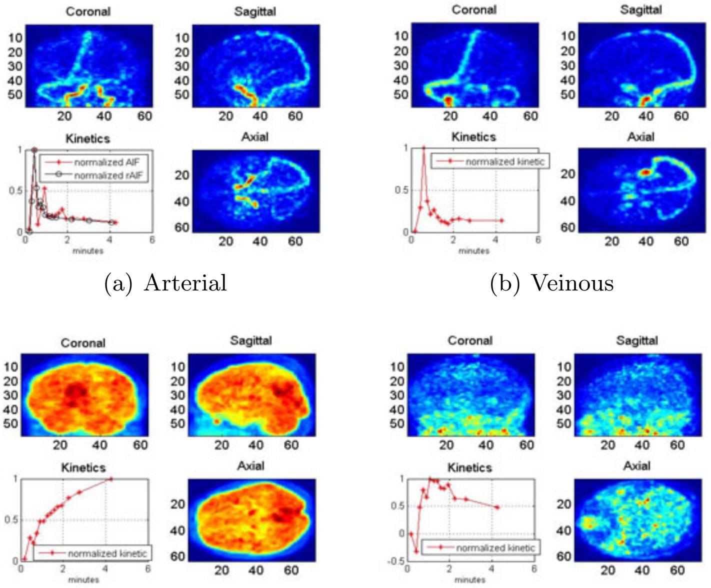
Table 1 records the average performance indices for 50 independent Monte-Carlo runs. One may note that when the sources are full additive, the three methods perform a good separation but SCSA-UNS is faster than MVSA and MVES. However when the sources are not full additive, SCSA-UNS still performs a perfect separation while MVSA and MVES do not.

Table 1. Average performance indices

	Index	SCSA-UNS	MVSA	MVES
Full additive sources	E_{sep}	$3,56.10^{-10}$	$1,23.10^{-12}$	$2,67.10^{-8}$
	$T(s)$	0.69	2, 21	8, 37
Non full additive sources	E_{sep}	$1,54.10^{-9}$	7, 89	6, 01
	$T(s)$	0, 83	4, 76	5, 59

Case 2: Real Dynamic Positon Emission Tomography (PET) Images Simulations have also been performed on real Dynamic Positon Emission Tomography (PET) data to study the pharmacokinetics of the [18F]-FDG (FluoroDeoxyGlucose) tracer on human brain. The main objective is to estimate the arterial pharmacokinetic also called Arterial Input Function (AIF) using only the dynamic TEP images with no arterial blood sampling (rAIF) which is too invasive for routine clinic use. The rAIF is only considered here as the reference AIF to assess the proposed estimator accuracy. As a matter of fact, an accurate estimation of the AIF allows a quantitative measurement which is indispensable for an efficient treatment evaluation in oncology. We have 19 human brain PET images recorded during $33mn$. Each 3D PET image is reshaped to form one row of the observations matrix X. The number of observations is $m = 19$ and the number of samples is $p = 266742$. We set the number of sources to $n = 4$.

Fig 1 shows the pharmacokinetics compartments estimated by the SCSA-UNS algorithm. Every subfigure represents the normalized kinetics (estimated mixing matrix) over the first four minutes (lower left) and the corresponding spatial



(a) Arterial

(b) Veinous

(c) Tissue

(d) Unidentified

Fig. 1. Estimated pharmacokinetics compartments

distributions (estimated sources) according to the three views coronal (upper left), sagittal (upper right) and axial (lower right). Three of the estimated compartments are identified to be the Arterial compartment (Fig 1.a), the Veinous one (Fig 1.b) and the Tissue one (Fig 1.c). Fig 1.a (lower left) shows that the normalized estimated AIF correctly approximates the normalized rAIF obtained by blood sampling, which was the main objective. However, one may note negative values on the kinetic of the unidentified compartment (lower left of Fig 1.d) which we attribute to measure noise.

5 Conclusions an Future Works

In this paper, we present a geometrical method for separating nonnegative sources. In overdetermined case, the proposed method first reduces the dimension of mixed data by performing the classical PCA and afterward runs the SCSA-UNS algorithm on the reduced data. The SCSA-UNS algorithm, estimates the mixing matrix by fitting a minimum volume simplicial cone to the scatter plot of observations. Unlike other geometrical methods that require local dominance or full additivity of the sources, SCSA-UNS only requires the mixed data to fill enough the simplicial cone generated by the mixing matrix. Simulations on synthetic data have showned that the proposed method always performs a good separation in noiseless case and runs faster than other MV methods. The proposed method also gave very promising results on real Dynamic PET images. As future works, we will investigate how to avoid the freezing situations without having to compute unfreezing matrices and we will consider the noisy case.

References

1. Puntonet, C., Mansour, A., Jutten, C.: A Geometrical Algorithm for Blind Source Separation. In: Gretsi 1995, Juan-les-Pins, pp. 273–276 (September 1995)
2. Babaie-Zadeh, M., Mansour, A., Jutten, C., Marvasti, F.: A Geometric Approach for Separating Several Speech Signals. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 798–806. Springer, Heidelberg (2004)
3. Lazar, C., Nuzillard, D., Nowé, A.: A New Geometrical BSS Approach for Non Negative Sources. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 530–537. Springer, Heidelberg (2010)
4. Chan, T.H., Ma, W.K., Chi, C.Y., Wang, Y.: A Convex Analysis Framework for Blind Separation of Non-Negative Sources. IEEE Transactions on Signal Processing 56, 5120–5134 (2008)
5. Li, J., Bioucas-Dias, J.M.: Minimum Volume Simplex Analysis: A Fast Algorithm to Unmix Hypersectral Data. In: IEEE International Symposium on Geoscience and Remote Sensing, vol. 3, pp. 250–253 (2008)
6. Chan, T.H., Chi, C.Y., Huang, Y.M., Ma, W.K.: A Convex Analysis-Based Minimum-Volume Enclosing Simplex Algorithm for Hyperspectral Unmixing. IEEE Transactions on Signal Processing 57, 4418–4432 (2009)
7. Craig, M.D.: Minimum-Volume Transforms for Remotely Sensed Data. IEEE Transactions on Geoscience and Remote Sensing 32, 542–552 (1994)

8. Henry, R.C.: History and fundamentals of multivariate air quality receptor models. Chemometrics and Intelligent Laboratory Systems 37, 37–42 (1997)
9. Ouedraogo, W.S.B., Souloumiac, A., Jaidane, M., Jutten, C.: Simplicial Cone Shrinking Algorithm for Unmixing Nonnegative Sources. In: Accepted to ICASSP 2012 (2012)

Appendix 1: Computing R_k^l

$W \geq 0$ and $Y = W^{-1}X \geq 0$ are the current estimations of A and S at iteration $k-1$ of sweep l, respectively. At iteration k, we look for a matrix R_k^l, so that $U = WR_k^l$ verify the following three conditions:
i) $U \geq 0$
ii) $V(U) \leq V(W)$
iii) $Z = U^{-1}X \geq 0$ $\left(Z = U^{-1}X = (R_k^l)^{-1}W^{-1}X = (R_k^l)^{-1}Y\right)$
The first two conditions are automatically satisfied because W and R_k^l are nonnegative and due to Proposition 1. From the definition of R_k^l given by (6), one may demonstrate that $[Z]_{ij} = [Y]_{ij} - r_{ik}^l[Y]_{kj}$, $\forall\ 1 \leq i \leq n$, $\forall\ 1 \leq j \leq p$. For fixed i, $[Z]_{ij} \geq 0 \Leftrightarrow r_{ik}^l \leq \frac{[Y]_{ij}}{[Y]_{kj}}$, $\forall\ 1 \leq j \leq p$.

A convenient R_k^l matrix can then be computed by taking:

$$r_{kk}^l = 1 \text{ and } r_{ik}^l = \min_{1\leq j\leq p} \frac{[Y]_{ij}}{[Y]_{kj}}, [Y]_{kj} \neq 0, \text{ for } i \neq k \tag{12}$$

Appendix 2: Computing the Unfreezing Matrix Q_l

Before giving details of computation of the unfreezing matrix Q_l, lets explain how arises the freezing situation. Given $W \geq 0$ and $Y = W^{-1}X \geq 0$ the current estimated mixing matrix sources respectively and $U = WR_k^l$:
$V(U) = V(W) \Leftrightarrow R_k^l = I_n \iff r_{ik}^l = \delta_{ik}$. For $i \neq k$ and according to (12), $r_{ik}^l = 0 \iff \exists\ 1 \leq j \leq p$ so $[Y]_{ij} = 0$ and $[Y]_{kj} \neq 0$.

The freezing arises if, at sweep l, the algorithm finds $R_k^l = I_n$, $\forall\ 1 \leq k \leq n$. This situation happens when there are at least one zero value on each row of the current estimated sources matrix (i.e. when $\forall\ 1 \leq i \leq n, \exists\ 1 \leq j \leq p, [Y]_{ij} = 0$). To avoid this problem, we suggest applying an orthogonal linear transformation Q_l to Y (and Q_l^T to W) to delete the zeros values of Y without increasing $V(W)$. We then introduce the unfreezing matrix Q_l so that $X = WQ_l^TQ_lY = HT$. The current estimated mixing matrix and sources become $H = WQ_l^T$ and $T = Q_lY$. We look for a matrix Q_l so that $Q_l^TQ_l = I_n$ and $T > 0$.

For findind such a Q_l matrix, we define the criterion J by:

$$J(Q) = \sum_{i=1}^{n}\sum_{j=1}^{p} 1_{T_{ij}} \text{ where } 1_{T_{ij}} = \begin{cases} 1 \text{ if } T_{ij} = 0 \\ 0 \text{ elsewhere} \end{cases} \tag{13}$$

The unfreezing matrix Q_l can be computed by solving the following equation:

$$Q_l = \underset{Q^TQ=I_n}{\arg\min}\ J(Q) \tag{14}$$

We developed a regularized gradient algorithm for solving the optimization problem (14) [9] but this method is not described here due to lack of space.

Multi-domain Feature of Event-Related Potential Extracted by Nonnegative Tensor Factorization: 5 vs. 14 Electrodes EEG Data

Fengyu Cong[1], Anh Huy Phan[2], Piia Astikainen[3], Qibin Zhao[2],
Jari K. Hietanen[4], Tapani Ristaniemi[1], and Andrzej Cichocki[2]

[1] Department of Mathematical Information Technology, University of Jyväskylä, Finland
[2]Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan
[3]Department of Psychology, University of Jyväskylä, Finland
[4]Human Information Processing Laboratory, School of Social Science and Humanities, University of Tampere, Finland
{fengyu.cong,piia.astikainen,tapani.ristaniemi}@jyu.fi,
{phan,qbzhao,cia}@brain.riken.jp, {jari.hietanen}@uta.fi

Abstract. As nonnegative tensor factorization (NTF) is particularly useful for the problem of underdetermined linear transform model, we performed NTF on the EEG data recorded from 14 electrodes to extract the multi-domain feature of N170 which is a visual event-related potential (ERP), as well as 5 typical electrodes in occipital-temporal sites for N170 and in frontal-central sites for vertex positive potential (VPP) which is the counterpart of N170, respectively. We found that the multi-domain feature of N170 from 5 electrodes was very similar to that from 14 electrodes and more discriminative for different groups of participants than that of VPP from 5 electrodes. Hence, we conclude that when the data of typical electrodes for an ERP are decomposed by NTF, the estimated multi-domain feature of this ERP keeps identical to its counterpart extracted from the data of all electrodes used in one ERP experiment.

Keywords: Event-related potential, feature extraction, multi-domain feature, N170, nonnegative tensor factorization.

1 Introduction

Event-related potentials (ERPs) have become a very useful method to reveal, for example, the specific perceptual and cognitive processes [11]. To achieve this goal, it is necessary to represent the information carried by data of an ERP with a feature or features for analysis. Generally, the peak amplitude of an ERP measured from its waveform in the time domain has become a mostly used feature to symbolize the ERP for statistical analysis [11], [12]. Furthermore, an ERP can also be represented by features in the frequency domain and in the time-frequency domain for analysis [9], [15]. Combined with the source localization method, these measurements can be applied to formulate the topography of an ERP in the spatial domain [2]. Indeed, the above mentioned features are very conventional to analyze ERPs. With the

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 502–510, 2012.

development of advanced signal processing technologies, some new features of ERPs can be formulated, for example, the multi-domain feature of an ERP [6], [7] extracted by nonnegative tensor factorization (NTF) [5]. In contrast to an ERP's conventional features which exploit the ERP's information in one or more domains sequentially, the multi-domain feature of the ERP can reveal the properties of the ERP in the time, frequency, and spatial domains simultaneously [4], [5], [6], [7]. Hence, this new feature may be less affected by the heterogeneousness of datasets [7].

Generally, when an ERP is statistically analyzed, the EEG data at the typical electrodes for the ERP are often used. For example, regarding a visual N170, the data at P7 and P8 are mostly analyzed [14] and for the auditory mismatch negativity (MMN), the data at Fz is frequently studied [13]. In our previous report to extract the multi-domain feature of MMN by NTF from the time-frequency representation of EEG, we used all the data collected at frontal, central, parietal and mastoid sites [7]. Since NTF is particularly useful for the problem of the underdetermined linear transformation model where the number of sensors is smaller than that of sources, it is possible to apply NTF for data collected at one scalp area (the model of such data is underdetermined since the number of electrodes is smaller than that of brain sources). Hence, it can be very interesting to examine whether NTF can extract the desired multi-domain feature of an ERP not from data collected at sites distributed along the whole scalp, but just from data recorded at a typical or restricted area of the scalp. This is very significant in EEG data collection when the target of research is not the source localization, but the more conventional analysis of ERPs.

In this study, we performed NTF on the multi-way representation of ERPs elicited by pictures of human faces in adult participants with and without depressive symptoms. We expected to obtain the identical multi-domain features of N170 from the data of 14 electrodes and the data of five typical electrodes for N170.

2 Method

2.1 Data Description

Twenty two healthy adults (control group, denoted as CONT hereinafter, 18 females, age range 30-58 years, mean 46.1 years) and 29 adults with depressive symptoms (depressive symptom group, denoted as DEPR hereinafter, 24 females, age range 29-61 years, mean 49.1 years) participated in the experiment. Pictures of neutral facial expressions served as a repeated standard stimulus (probability = 0.8), and pictures of happy and fearful expressions (probability = 0.1 for each) as rarely presented deviant stimuli. At least two standards were presented between randomly presented consecutive deviants. The stimulus duration was 200 ms, and the stimulus onset asynchrony was 700 ms. Altogether, a total of 1600 stimuli presented. During the recordings, the participants were seated in a chair, and were instructed to pay no attention to the visual stimuli but instead attended to a radio play presented via loud speakers. Enhanced face sensitive N170 responses for the emotional faces have shown to be elicited also in this type of an oddball paradigm [1].

Brain Vision Recorder software (Brain Products GmbH, Munich, Germany) was used to record the EEG with 14 electrodes at Fz, F3, F4, Cz, C3, C4, Pz, P3, P4, P7, P8, Oz, O1 and O2 according to the international 10-20 system. An average reference was used. Data were on-line digitally filtered from 0.1 to 100 Hz, and the down sampling frequency was 1000 Hz. Then, the obtained data were offline processed with Brain Vision Analyzer software and MATLAB (Version R2010b, The Mathworks, Inc., Natick, MA). EEG data were segmented into ERP responses of 200 ms pre-stimulus period and 500 ms after the stimulus onset, and the baseline was corrected based on the average amplitude of the 200-ms pre-stimulus period. Segments with signal amplitudes beyond the range between -100 and 100 μV in any recording channel, were rejected from further analysis. The number of kept trials for the averaging was about 100 in average. The recordings of the artifact-free single trials were averaged at each channel for each subject. For the present study, the data from the happy deviants were chosen for the analysis as the processing of positive information is known to be especially impaired in the depressed individual [16]. Therefore, comparison of the brain activity between the happy and the fearful expressions, or between the rare emotional (happy and fearful) stimuli and the frequently presented (neutral) standard stimuli are out of the scope of this study.

2.2 Nonnegative Tensor Factorization for Multi-domain Feature Extraction

In the form of tensor products, the NTF model [5] can also be written as

$$\underline{\mathrm{Y}} \approx \underline{\mathrm{I}} \times_1 \mathrm{U}^{(1)} \times_2 \mathrm{U}^{(2)} \cdots \times_{\mathrm{N}} \mathrm{U}^{(\mathrm{N})} = \hat{\underline{\mathrm{Y}}} \, , \tag{1}$$

where $\hat{\underline{\mathrm{Y}}}$ is an approximation of the N-order tensor $\underline{\mathrm{Y}} \in \Re_+^{\mathrm{I}_1 \times \mathrm{I}_2 \times \cdots \times \mathrm{I}_\mathrm{N}}$, and $\underline{\mathrm{I}}$ is an identity tensor [5], $\mathrm{U}^{(\mathrm{n})} = \left[\mathrm{u}_1^{(\mathrm{n})}, \mathrm{u}_2^{(\mathrm{n})}, \cdots, \mathrm{u}_\mathrm{J}^{(\mathrm{n})}\right] \in \Re_+^{\mathrm{I}_\mathrm{n} \times \mathrm{J}}$ is the nonnegative matrix, $\mathrm{n} = 1,2,\cdots,\mathrm{N}$, and $\left\|\mathrm{u}_\mathrm{j}^{(\mathrm{n})}\right\|_2 = 1$, for $\mathrm{n} = 1,2,\cdots,\mathrm{N}-1$, $\mathrm{j} = 1,2,\cdots,\mathrm{J}$. Each factor $\mathrm{U}^{(\mathrm{n})}$ explains the data tensor along a corresponding mode. Most algorithms for NTF are to minimize a squared Euclidean distance as the following cost function [5]

$$\mathrm{D}\left(\underline{\mathrm{Y}} | \hat{\underline{\mathrm{Y}}}\right) = \frac{1}{2} \left\|\underline{\mathrm{Y}} - \underline{\mathrm{I}} \times_1 \mathrm{U}^{(1)} \times_2 \mathrm{U}^{(2)} \cdots \times_\mathrm{N} \mathrm{U}^{(\mathrm{N})}\right\|_\mathrm{F}^2. \tag{2}$$

In this study, we applied the hierarchical alternating least squares (HALS) algorithm [5] whose simplified version for NMF has been proved to be superior to the multiplicative algorithms [8]. The HALS is related to the column-wise version of the ALS algorithm for 3-D data [3]. The HALS algorithm sequentially updates components $\mathrm{u}_\mathrm{j}^{(\mathrm{n})}$ by a simple update rule

$$\begin{aligned} \mathrm{u}_\mathrm{j}^{(\mathrm{n})} \leftarrow \underline{\mathrm{Y}} \, \bar{\times}_1 \, \mathrm{u}_\mathrm{j}^{(1)} \, \bar{\times}_2 \, \mathrm{u}_\mathrm{j}^{(2)} \cdots \bar{\times}_{\mathrm{n}-1} \, \mathrm{u}_\mathrm{j}^{(\mathrm{n}-1)} \, \bar{\times}_{\mathrm{n}+1} \, \mathrm{u}_\mathrm{j}^{(\mathrm{n}+1)} \cdots \bar{\times}_\mathrm{N} \, \mathrm{u}_\mathrm{j}^{(\mathrm{N})} \\ -\mathrm{U}_{-\mathrm{j}}^{(\mathrm{n})} \left(\circledast \, \mathrm{U}_{-\mathrm{j}}^{(\mathrm{k})^\mathrm{T}} \mathrm{u}_\mathrm{j}^\mathrm{k}\right) \end{aligned} \tag{3}$$

where, '$\circledast$' denotes the Hadarmard product, $k \neq n$, and $\underline{\mathrm{Y}} \, \bar{\times}_n \, \mathrm{u}_j^{(n)}$ represents the n-mode product between tensor and vector [5]. The factor except the last one will be normalized to be unit vectors during iterations $\mathrm{u}_j^{(n)} \leftarrow \mathrm{u}_j^{(n)} / \left\|\mathrm{u}_j^{(n)}\right\|_2$, $n = 1,2,\cdots,N-1$. It should be noted that this study does not tend to propose an NTF algorithm. Therefore, any NTF algorithm can work for the data.

In detail, regarding the study N170 with NTF, we formulated a fourth-order tensor $\underline{\mathrm{Y}}$ including modes of the frequency by time by channel by subject. The number of frequency bins ($\mathrm{I_f}$), timestamps ($\mathrm{I_t}$), channels ($\mathrm{I_c}$), and subjects ($\mathrm{I_s}$) compose the dimensions of the tensor $\underline{\mathrm{Y}}$. Decomposition of $\underline{\mathrm{Y}}$ results in four matrices:

$$\underline{\mathrm{Y}} \approx \underline{\mathrm{I}} \times_1 \mathrm{U}^{(1)} \times_2 \mathrm{U}^{(2)} \times_3 \mathrm{U}^{(3)} \times_4 \mathrm{U}^{(4)} = \underline{\mathrm{I}} \times_1 \mathrm{U}^{(\mathrm{f})} \times_2 \mathrm{U}^{(\mathrm{t})} \times_3 \mathrm{U}^{(\mathrm{c})} \times_4 \mathrm{F}, \quad (4)$$

where, the last factor is the feature matrix ($\mathrm{I_s} \times \mathrm{J}$) consisting of J extracted multi-domain features of brain responses in the N170 experiment onto the subspaces spanned by the spectral (i.e., $\mathrm{U}^{(\mathrm{f})}(\mathrm{I_f} \times \mathrm{J})$), temporal (i.e., $\mathrm{U}^{(\mathrm{t})}(\mathrm{I_t} \times \mathrm{J}\,)$) and spatial (i.e., $\mathrm{U}^{(\mathrm{c})}(\mathrm{I_c} \times \mathrm{J})$) factors. This is that each subject $\mathrm{i}\,(\mathrm{i} = 1,2,\cdots,\mathrm{I_s})$ is characterized by the $\mathrm{i^{th}}$ row of $\mathrm{U}^{(\mathrm{n})}$ ($\mathrm{n} = 1,2,3,4$) in this study. Furthermore, for one feature, i.e., one component among J components in the feature factor matrix, the values of different participants, i.e., the data at the same column of feature factor matrix $\mathrm{U}^{(4)}$, are comparable since they are extracted under the identical subspaces; but due to the variance ambiguity of NTF [5], the variances of the different features/components in any factor matrix are not comparable. Moreover, the extracted J multi-domain features should be associated with different sources of brain activities. Then, it is necessary to determine which multi-domain feature corresponds to the desired ERP.

Regarding the multi-domain feature of N170, firstly, the temporal components in the temporal factor matrix extracted by NTF have different peak latencies and the desired one for N170 may look like the waveform with a sole peak whose latency should be around 170 ms. Secondly, when the subjects in the fourth-order tensor as denoted by the tensor $\underline{\mathbf{Y}}$ include two groups, the spatial pattern extracted by NTF can be the difference topography between the two groups of participants because NTF also decomposes the multi-way representation of data in the spatial dimension. We will show in the next section that N170 has different peak amplitudes at P7 for the two groups. Thus, in this study, we assume the desired spatial component reveals the difference topography around P7 for N170. Finally, the desired spectral structure of an ERP elicited by the passive oddball paradigm may possess its largest energy between 1 and 5 Hz [7]. These are the criteria to choose the desired multi-domain feature of N170 from all the extracted multi-domain features. Furthermore, in our experiment, the vertical positive potential (VPP) and N170 probably correspond to the identical brain processes [10]. Hence, the multi-domain feature of VPP was also extracted here. The difference in the topography of VPP between two groups of participants would probably appear at the right hemisphere in this study. For detail of the multi-domain feature selection for an ERP, please refer to our previous report [7].

2.3 Data Processing and Analysis

In this study, NFT was performed on all subjects' data consisting of the time-frequency representation of the ordinary averaged traces at all 14 electrodes, as well as at five electrodes including P7, P8, O1, Oz and O2 which are typical electrode sites to analyze N170 [14], and at five electrodes including Fz, F3, F4, C3 and C4 which are typical sites to analyze VPP [10]. In order to obtain the time-frequency representation (TFR) of ERPs, the complex Morlet wavelet transformation [17] was performed on the averaged trace at each channel. For the Morlet, the half wavelet length was set to be six for the optimal resolutions of the frequency and the time [17]; the frequency range was set from 1 to 10 Hz, and 91 frequency bins were uniformly distributed within this frequency range. Next, the fourth-order tensor with the dimensions of frequency (91 bins) by time (700 samples) by channel (14 or 5) by subject (51) was formulated in terms of TFR of all subjects at chosen channels. And then, from the formed tensor, multi-domain features respectively were extracted by 10 NTF models with numbers of components ranging from 15 till 24 based on the experience learned from our previous report [7]. Subsequently, in each model, the desired multi-domain feature of N170 was selected according to properties of N170 in the time, frequency, and spatial domains as mentioned above. So did for VPP. After the desired feature component which was a vector including values for all subjects was chosen in one NTF model, it was normalized according to its L-2 norm. Finally, we obtained multi-domain features of N170 and VPP with the data of 3 multi-domain features of ERPs by 51 participants by 10 models.

After features of N170 were ready, statistical tests were performed to examine the difference of N170 between two groups with the Bonferroni correction and with 0.05 as the level of significance. For peak amplitudes of N170 in the time domain measured from ordinary averaged traces (i.e., 'raw data'), a General Linear Model (GLM) multivariate procedure for a 4×2 design was applied using the channel (P7, P8, O1 and O2) as the independent variable and the group (CONT and DEPR) as the fixed factor. Regarding the multi-domain feature of N170 extracted by NTF, a GLM multivariate procedure was implemented. The GLM multivariate procedure for an 10 × 2 design was made using the NTF-model as the dependent variable and the group (CONT and DEPR) as the fixed factor.

3 Results

In this section we compare discriminability of various features of N170 between two groups of participants, as well as the coherence between the multi-domain feature of an ERP extracted from data of 14 electrodes and that from data of five electrodes.

Fig. 1 demonstrates the grand averaged waveforms of ERPs. The significant difference between two groups only appeared at P7 (F(1,49) = 5.185, p = 0.027). For illustration on how the multi-way data can be decomposed by a multi-way analysis method, Fig. 2 shows the demo for the common components factors in different domains extracted by NTF from the data of 14 electrodes when 20 components were extracted in each mode of Eq. (4). In this model as the third component in the temporal, spectral and spatial components matrices matched the properties of N170, the third feature was chosen as the desired multi-domain feature of N170 which is the one for model-20 in Fig.3. Fig.3 presents the desired multi-domain features of N170 extracted through 10 NTF models from the data of 14 electrodes for demonstration.

As illustrated in Table-1, the difference between two groups of participants was better revealed by the multi-domain features of N170 no matter they were extracted from the data of 14 electrodes or 5 typical electrodes, and the multi-domain feature of N170 outperformed that of VPP in discriminating the two groups based on the degree of significance of difference. Moreover, Table-2 tells that the multi-domain features of N170 between the data of 14 electrodes and the data of 5 typical electrodes were more highly correlated (correlations in Table-2 were all significant) than any other two pairs, which means they reflect absolutely similar information. Furthermore, these indicate that although VPP and N170 possess the identical latency and conform to identical brain processes [10], the multi-domain feature of N170 extracted by NTF from its typical electrodes better represented the brain processes than that of VPP from its typical electrodes to categorize different groups of participants.

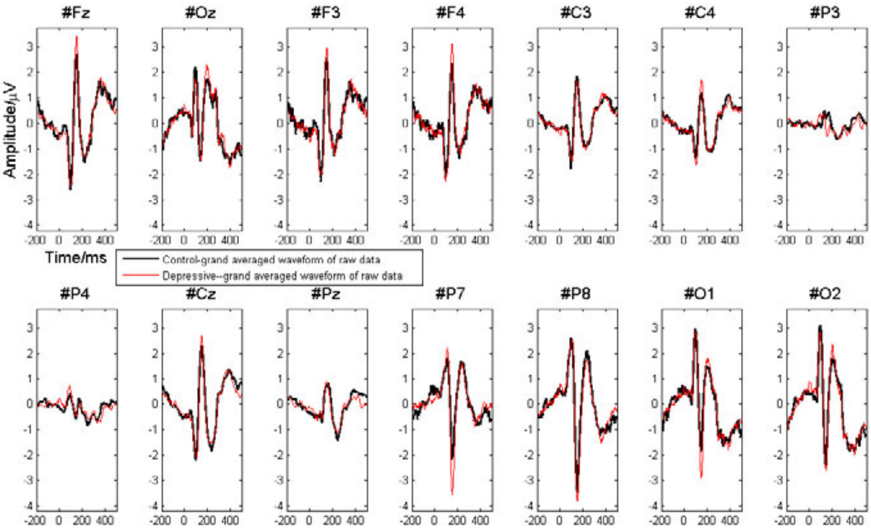
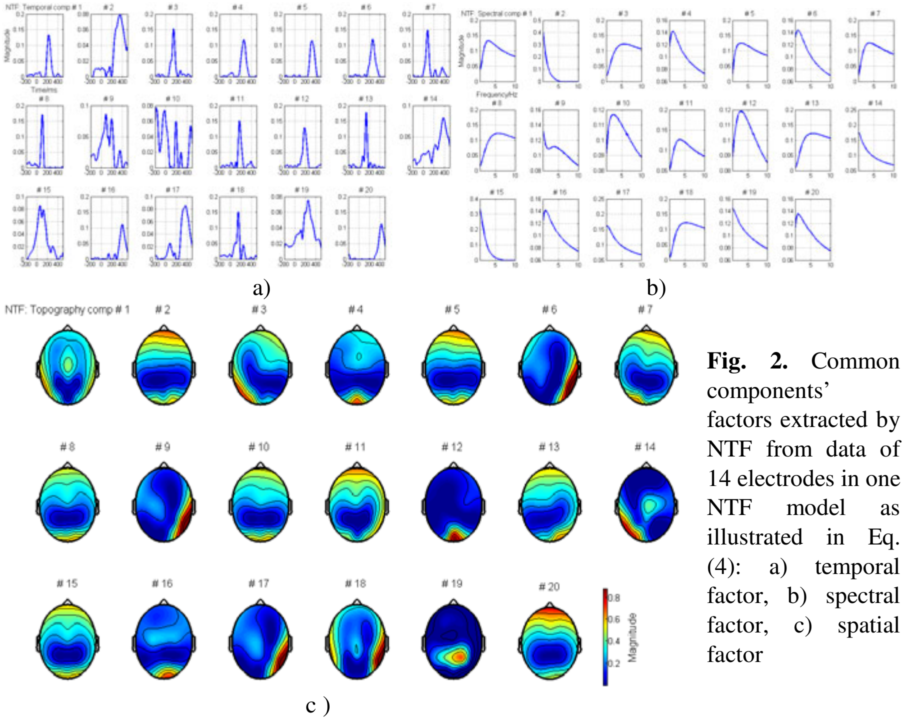


Fig. 1. Grand averaged waveforms of ERPs



a)

b)

c)

Fig. 2. Common components' factors extracted by NTF from data of 14 electrodes in one NTF model as illustrated in Eq. (4): a) temporal factor, b) spectral factor, c) spatial factor

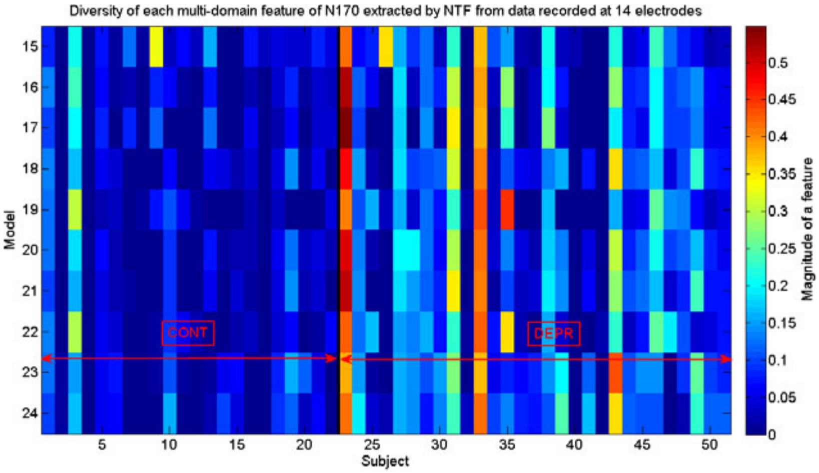


Fig. 3. Multi-domain features of N170 extracted by 10 NTF models from data of 14 electrodes

Table 1. Statistical tests of extracted features

NTF Model	N170-14 electrodes		N170 -five electrodes		VPP - five electrodes	
	F(1,49)	p	F(1,49)	p	F(1,49)	p
15	2.666	0.109	9.275	0.004	5.515	0.023
16	7.444	0.009	8.611	0.005	5.464	0.024
17	5.255	0.026	8.034	0.007	4.334	0.043
18	10.397	0.002	8.951	0.004	5.955	0.018
19	6.493	0.014	8.623	0.005	2.470	0.122
20	10.391	0.002	8.142	0.006	3.878	0.055
21	10.258	0.002	8.256	0.006	2.252	0.140
22	8.162	0.006	9.474	0.003	5.507	0.023
23	9.727	0.003	8.707	0.005	2.620	0.112
24	11.271	0.002	8.685	0.005	3.079	0.086

Table 2. Correlation coefficient of features

NTF Model	N170-14 electrodes vs. -5 electrodes	N170-14 electrodes vs. VPP-5 electrodes	N170-5 electrodes vs. VPP-5 electrodes
15	0,614	0,652	0,653
16	0,897	0,751	0,597
17	0,803	0,831	0,592
18	0,941	0,649	0,59
19	0,756	0,863	0,573
20	0,919	0,679	0,515
21	0,858	0,494	0,434
22	0,722	0,738	0,607
23	0,903	0,523	0,543
24	0,854	0,478	0,671

4 Conclusions

Though NTF from data of fewer electrodes which are typical to analyze an ERP, the extracted multi-domain feature of the ERP may be as identical as that from the data of much more electrodes distributed all over the scalp surface. Furthermore, in one ERP experiment, different components with different polarities in different scalp sites may have the same latency and reveal identical brain activities, such as, VPP in frontal-central sites and N170 in occipital-temporal sites [10], the multi-domain feature of prime component may better represent the brain activities than other components do.

Acknowledgments. Cong F. thanks the sponsorship from Research and Innovation Office of University of Jyväskylä for the international mobility grants (2009, 2010). The study was supported by Academy of Finland (research grant for Professor Raimo Lappalainen). The authors thank Professor Lappalainen for his help in recruiting the depressed participants for the study.

References

1. Astikainen, P., Hietanen, J.K.: Event-Related Potentials to Task-Irrelevant Changes in Facial Expressions. Behav. Brain Funct. 5, 30 (2009)
2. Blankertz, B., Lemm, S., Treder, M., et al.: Single-Trial Analysis and Classification of ERP Components - A Tutorial. Neuroimage 56, 814–825 (2011)
3. Bro, R.: Multi-way analysis in the food industry - models, algorithms, and applications. PhD thesis, University of Amsterdam, Holland (1998)
4. Cichocki, A., Washizawa, Y., Rutkowski, T.M., et al.: Noninvasive BCIs: Multiway Signal-Processing Array Decompositions. Computer 41, 34–42 (2008)
5. Cichocki, A., Zdunek, R., Phan, A.H., et al.: Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis. John Wiley (2009)
6. Cong, F., Phan, A.H., Lyytinen, H., Ristaniemi, T., Cichocki, A.: Classifying Healthy Children and Children with Attention Deficit through Features Derived from Sparse and Nonnegative Tensor Factorization Using Event-Related Potential. In: Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., Vincent, E. (eds.) LVA/ICA 2010. LNCS, vol. 6365, pp. 620–628. Springer, Heidelberg (2010)
7. Cong, F., Phan, A.H., Zhao, Q., et al.: Multi-Domain Feature of Mismatch Negativtiy Extracted by Nonnegative Tensor Factorization to Discriminate Reading Disabled Children and Children with Attention Deficit. B5 (27pages), Series B. Scientific Computing Reports of Department of Mathematical Information Technology, University of Jyväskylä (2011)
8. Gillis, N., Glineur, F.: Accelerated Multiplicative Updates and Hierarchical ALS Algorithms for Nonnegative Matrix Factorization. CORE Discussion Paper, 30 (2011)
9. Görsev, G.Y., Basar, E.: Sensory Evoked and Event Related Oscillations in Alzheimer's Disease: A Short Review. Cogn. Neurodyn. 4, 263–274 (2010)
10. Joyce, C., Rossion, B.: The Face-Sensitive N170 and VPP Components Manifest the Same Brain Processes: The Effect of Reference Electrode Site. Clin. Neurophysiol. 116, 2613–2631 (2005)
11. Luck, S.J.: An introduction to the event-related potential technique. The MIT Press (2005)

12. Näätänen, R.: Attention and brain functions. Lawrence Erlbaum Associates, Hillsdale (1992)
13. Näätänen, R., Pakarinen, S., Rinne, T., et al.: The Mismatch Negativity (MMN): Towards the Optimal Paradigm. Clin. Neurophysiol. 115, 140–144 (2004)
14. Rossion, B., Jacques, C.: Does Physical Interstimulus Variance Account for Early Electrophysiological Face Sensitive Responses in the Human Brain? Ten Lessons on the N170. Neuroimage 39, 1959–1979 (2008)
15. Stefanics, G., Haden, G., Huotilainen, M., et al.: Auditory Temporal Grouping in Newborn Infants. Psychophysiology 44, 697–702 (2007)
16. Surguladze, S.A., Young, A.W., Senior, C., et al.: Recognition Accuracy and Response Bias to Happy and Sad Facial Expressions in Patients with Major Depression. Neuropsychology 18, 212–218 (2004)
17. Tallon-Baudry, C., Bertrand, O., Delpuech, C., et al.: Stimulus Specificity of Phase-Locked and Non-Phase-Locked 40 Hz Visual Responses in Human. J. Neurosci. 16, 4240–4249 (1996)

The Use of Linear Feature Projection for Precipitation Classification Using Measurements from Commercial Microwave Links

Dani Cherkassky, Jonatan Ostrometzky, and Hagit Messer

School of Electrical Engineering,
Tel Aviv University, Israel
{dani.cherkassky,yonstero}@gmail.com,
messer@eng.tau.ac.il

Abstract. High frequency electromagnetic waves are highly influenced by atmospheric conditions, namely wireless microwave links with carrier frequency of tens of GHz can be used for precipitation monitoring. In the scope of this paper we present a novel detection/classification system capable of detecting wet periods, with the ability to classify the precipitation type as rain or sleet, given an attenuation signal from spatially distributed wireless commercial microwave links. Fade (attenuation) dynamics was selected as a *discriminating feature* providing the data for classification. Linear Feature Extraction method is formulated; thereafter, the efficiency is evaluated based on real data. The detection/classification system is based on the Fisher's *linear discriminant* and *likelihood ratio test*. Its performance is demonstrated using actual Received Signal Level measurements from a cellular backhaul network in the northern part of Israel. In particular, the use of the raw data as well as its derivatives to achieve better classification performance is suggested.

Keywords: Environmental monitoring, Received Signal Level (RSL) measurements, feature extraction, rain sleet events classification/detection, fade dynamics.

1 Introduction

Microwave communication links are used in the backhaul network of cellular systems, making them widespread in most countries. Since the carrier frequency of those links is typically above 10GHz, the wave propagation is highly influenced by precipitation. The impairment caused by precipitation has been extensively studied, whereas the main objective is designing a reliable communication network. However, those 'impairments' can be used for monitoring precipitation, as first suggested by Messer *et al* [1]. Following these finding, a number of metrological applications using microwave Received Signal Level (RSL) recording were explored [2]. Dual-frequency links find applications in calibration of weather radar [3], correction of X-band radar rainfall estimates [4] and identification of melting snow [5]. The use of microwave links to study evaporation was explored by Leijnse *et al.* in 6]. A method

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 511–519, 2012.

for estimation of two-dimensional rainfall intensity field based on multiple path-integrated RSLs was suggested in [7]. A method for the detection of RSL attenuation in a single link caused by sleet events was proposed in [8].

In this paper we deal, for the first time, with the challenge of classifying precipitation (rain, sleet and snow) given RSL measurements. RSL signals from 3 commercial microwave links were recorded in the area of Ortal Mountain in the northern part of Israel. Following the basic observation of [8] that sleet/snow and rain events could be distinguished by observing the dynamics (magnitude, duration and slope) of attenuation; optimal feature vector for classification of rain, sleet and snow events were selected by maximizing the Fisher's Linear criterion, and by applying Linear Mapping (LM). Finally, the classification of the events was done using the Likelihood Ration Test (LRT).

The paper is organized as follows: Section 2 describes the classification method and the resulting algorithm. Section 3 provides experimental results, and in Section 4 discussion and conclusions are provided.

2 Method

2.1 Setup

A backhaul system composed of fixed terrestrial line-of-sight radio microwave links, employed for transmission purposes by an Israeli cellular operator named Cellcom (http://www.cellcom.com/), was used to demonstrate the classification method. In particular, three microwave links in the northern part of Israel were used, as described in Table 1. Microwave Links Information. Cellcom system was designed to secure reliable communication and not to precipitate monitoring, and the RSL records were pre-processed accordingly, introducing two main challenges once used for precipitation monitoring: i) The RSL signal measured at the base station with quantization level of 1dB; ii) Only a maximum RSL (MRSL) and minimum RSL (mRSL) values within every 15 (non-overlapping) minutes were transmitted and recorded at the control center. The RSL records at a control center were used as input for the classification system described in Fig.1.

Table 1. Microwave Links Information

Link #	*Microwave Link Name*	*Frequency (GHz)*	*Length (km)*	*Site 1 Height(m)*	*Site 2 Height(m)*
1	HAR ODEM - ORTAL	19.3	12.8	1080	898
2	ORTAL- HAR ODEM	19.3	12.8	898	1080
3	KATZRIN - ORTAL	18.36	11.9	375	898

As seen in Table 1, the three links had a common base station in Ortal, whereas a Parsivellaser based disdrometer [9] was installed as a control device by our research group. This disdrometer measures the size and velocity of the particles passing through the laser beam and classifies them into rain/snow/drizzle/hail [9], to name a few. The disdrometer output signal (precipitate type and intensity, measured every 10 sec) was used as a "ground truth" (baseline) to validate the performance of the classification system depicted in Fig.1.

2.2 Signal Processing

The proposed signal processing system for the classification of precipitates, based on Pattern Recognition (PR), is schematically illustrated in Fig.1. A preliminary task (preprocessing) in PR problems is to select suitable features allowing to distinct between the classes. Following [8], the dynamic properties of the fade (RSL) are used as a distinctive feature. Moreover, it is essential to select the best possible class indicators (features) from the data obtained by the sensors, prior to classifying the data. A common method for the selection of indicators (features) is Feature Extraction (FE). In the following section, a general linear method for FE is described [15], aiming to select an optimal feature for classification. In the last part of this section, Likelihood Ratio Test for comparison of two hypotheses will be presented.

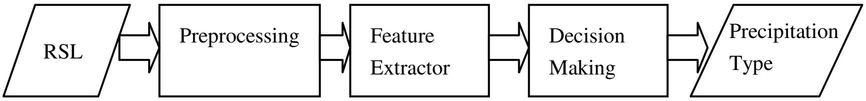


Fig. 1. Precipitates Classification System - Data Flowchart

2.2.1 Fade Dynamics

The dynamics of the attenuation caused by specific precipitation can be characterized by a *fade magnitude*, *fade duration* and *fade slope* [10,11]. *Fade duration* is simply the time interval during which attenuation exceeds a certain threshold value. *Fade slope* indicates the rate of change in the attenuation. The *fade slope* attenuation caused by rain or other meteorological events is very important for designing fade mitigation techniques, since it determines the required tracking speed of fade mitigation techniques. The ITU-R Model [12] defines the fade slope ξ at a certain point in time by filter data as:

$$\xi(t) = \frac{A\left(t+\frac{1}{2}\Delta t\right) - A\left(t-\frac{1}{2}\Delta t\right)}{\Delta t} \tag{1}$$

Where A is the attenuation in dB, and Δt is time interval length inwhich the fade slope is calculated, in seconds. The following model for *Conditional Probability Distribution* (*CPD*) of the fade slope ξ for a given attenuation A was suggested by Van de Kamp [13].The model was developed using measurements collected in16 months:

$$p(\xi \mid A) = \frac{2}{\pi\sigma_\xi \cdot \left(1+\left(\xi/\sigma_\xi\right)^2\right)^2} \quad \left(dB/s\right)^{-1} \tag{2}$$

Whereas σ_ξ is the standard deviation of the conditional fade slope at a given attenuation level. The standard deviation is a function of: attenuation level, properties of the microwave link, climate and type of precipitate - drop diameter, and a fraction of melting water the drop contains [12, 13]. An experimental study [8] demonstrated that the *CPD - p(ζ|A)* is different in casesofattenuation caused by rain compared with attenuation caused by sleet. Fig.2 demonstrates qualitatively the difference between the fade slope *CPD* function caused by rain vs. the functioncaused by sleet, according to [8].

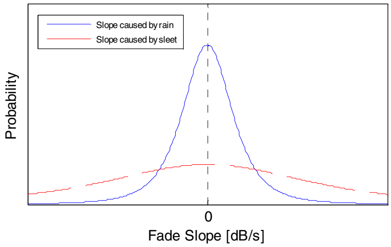


Fig. 2. Qualitativeshapeof the Conditional Probability Distribution Functions of fade slope

2.2.2 Linear Discriminant

The goal of discriminant analysis can be summarized as finding a function returning values,allowing a good discrimination between different classes of the input data. More formally, one is looking for a function $f: X \to R^D$, whereas $f(\mathbf{x})$ and $f(\mathbf{z})$ are similar whenever $\mathbf{x}$ and $\mathbf{z}$ are similar, and different otherwise. Similarity is usually measured by class membership and Euclidean distance. In a uniquecase of linear discriminant analysis, one is seeking for the linear function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}, \quad \mathbf{w} \in R^{N \times D}$. The most renowned linear discriminant was presented by Fisher [14]. Fisher's idea was to look for a direction $\mathbf{w}$, whereas the distance between class measuresis as large as possible, while achieving thesmallest possible variance for every class.Meaning, given a training set $X = \{(\mathbf{x}_1, y_1), .., (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \ \forall i = 1, \ldots, l$ are the input vectors and $y_i \ \forall i = 1, \ldots, l$ are the class labels, Fisher's linear discriminant is given by vector $\mathbf{w}$ maximizing [14,15]:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{3}$$

Where:

$$S_B \triangleq (\mathbf{m}_1 - \mathbf{m}_2) \cdot (\mathbf{m}_1 - \mathbf{m}_2)^T, S_W \triangleq \sum_i \sum_{\mathbf{x} \in X_i} (\mathbf{x} - \mathbf{m}_i) \cdot (\mathbf{x} - \mathbf{m}_i)^T \tag{4}$$

are the *between class* and *within class scattering matrices*, respectively, $\mathbf{m}_i$ is the sample average of class i, defined by $\mathbf{m}_i \triangleq \frac{1}{l_i} \sum_{j=1}^{l_i} \mathbf{x}_j^i$, l_i is the number of samples in class i, $\mathbf{x}^i = \{(\mathbf{x}, y) \in X \mid y = i\}$, and i is the class index. In order to find the optimal $\mathbf{w}$, one should differentiate (3) with respect to $\mathbf{w}$ and set the result to zero:

$$\left(\mathbf{w}^T S_B \mathbf{w}\right) S_W \mathbf{w} - \left(\mathbf{w}^T S_W \mathbf{w}\right) S_B \mathbf{w} = 0 \Leftrightarrow S_B \mathbf{w} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \cdot S_W \mathbf{w} \qquad (5)$$
$$S_B \mathbf{w} = \lambda \cdot S_W \mathbf{w}$$

Two important properties of Fisher’s discriminant[16,17] are: i) it has a global solution (maximum), although not necessarily unique, ii) this global maximum of (3) can be found by solving a generalized eigenvalue problem. It is well known that $\mathbf{w}$ maximizing (3) is the leading eigenvector of the eigenvalue problem (5). A linear model is relatively robust to noise and most likely will not overfit; on the other hand, the performance is naturallylimited by the linearity assumption.

2.2.3 Likelihood Ratio Test

In the scope of this work, our main objective is to classify the various physical phenomena responsible for the measured attenuations on a microwave link. In the previous sections we described analgorithm for selecting optimal features (or projection of the RSL signal to optimal space for classification); following such projection, we will have to assign each feature to one of the classes. In other words, a decision to which class each feature belongs must be made. A fundamental approach for pattern classification induced by *Bayesian Decision Theory* is *a Likelihood Ration Test*, which iswidely and extensivelydiscussed in the literature, e.g., [18]. At a glance, in a scenario of decision between two possible classes ω_1, ω_2 with known a priori probabilities $P(\omega_1), P(\omega_2)$, respectively, a sample $\mathbf{x}$ will be associated with ω_1 or ω_2 according to:

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{p(\omega_1)}{p(\omega_2)} \qquad (6)$$

Where λ_{ml} is the loss associated with determining ω_m where the true class is ω_l ; therefore, it is reasonable to assume that no loss is caused by correct classification, and to set $\lambda_{11} = \lambda_{22} = 0$.

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12}}{\lambda_{21}} \cdot \frac{p(\omega_1)}{p(\omega_2)} \qquad (7)$$

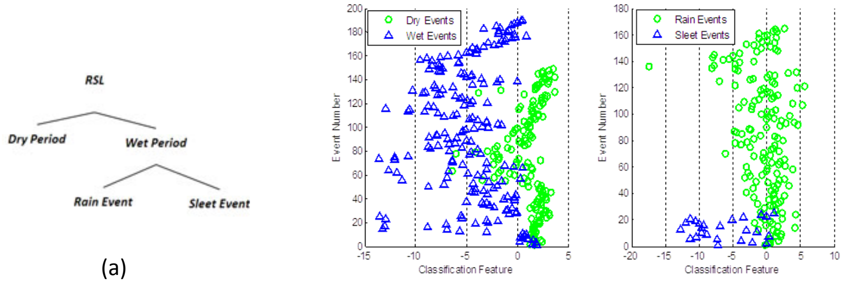
The selection of the ratio $\lambda_{12}/\lambda_{21}$ can increase the probability of *detection*, while increasing the probability of *false alarm* and vice versa. For the *symmetrical* case (zero-one loss) where $\lambda_{12}/\lambda_{21} = 1$ and equal priors $p(\omega_1) = p(\omega_2)$, we obtain the decision rule that minimizes the average probability of error in theclassification process.

3 Experiment and Results

The preprocessing unit converts the sequence of l RSL measurements from the 3 links (mRSL and MRSL from each) into a *12 by (l-1)* matrix *X,* with the first 6 columns presenting the raw data (mRSL and MRSL for each link), and the next 6 columns presenting an approximation to the derivative of the measurementsin order to provide information on the dynamics of the RSL. Matrix X is given by:

$$X \triangleq \begin{bmatrix} mRSL_1[0] & MRSL_1[0] & \dots & mRSL_1[0]-mRSL_1[1] & MRSL_1[0]-MRSL_1[1] & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ mRSL_1[l-2] & MRSL_1[l-2] & \dots & mRSL_1[l-2]-mRSL_1[l-1] & MRSL_1[l-2]-MRSL_1[l-1] & \dots \end{bmatrix} \tag{8}$$

The classification system was designed a decision making tree, as illustrated in 0. By applying a tree classification method we need only a single feature at each decision node, since under an optimal projection only *c-1* features are needed to distinguish between *c* classes [15]. In Fig.3, scatter plots of two different features are demonstrated: on the left hand side, a scatter diagram of a feature for classification between wet and dry events, f_{dw}, is depicted; while on the right hand side, a scatter plot of a feature distinguishing between rain and sleet events, f_{rs}, is shown. The featuresf_{dw},f_{rs} were calculated as a linear combination of thematrix *X*together withthe projecting vectors$\mathbf{w}_{\text{dw}}$,$\mathbf{w}_{\text{rs}}$, respectively. Theprojecting vectors were calculated using the LDA method (3-5), based on a training set*(X,***y***)*, where y_i is a class label $y_i \in \{dry, rain, sleet\}$, as measured by a disdrometer.



(a)

(b)

Fig. 3. (a) Classification Tree. (b) Scatter plots of an optimal feature obtained by applying LDA. On the left hand side, a scatter diagram of a feature for classification between wet and dry events; on the right hand side, a scatter plot of a feature distinguishing between rain and sleet events.

Fig.4 presents the performance of the proposed classification system, whereas data from the 3 day storm in mid-December, 2010, was used. It describes the probability of detection as a function of probability of false alarm for different decision rules. The ratio of $\lambda_{12}/\lambda_{21}$ was varied from 0 to infinity and for each probability of detection value and probability of false alarm value was evaluated. As a reference, the classifier of Fig. 1 has been applied on the raw RSL data only (without the derivative); subsequently, the X matrix contains only the first 6 columns. As expected in this case, where the different dynamics of rain/sleet events are less emphasized, the performance of the classifier is worse (Fig. 4b). The classification between dry and wet events, however, is less influenced once the derivative is not used.

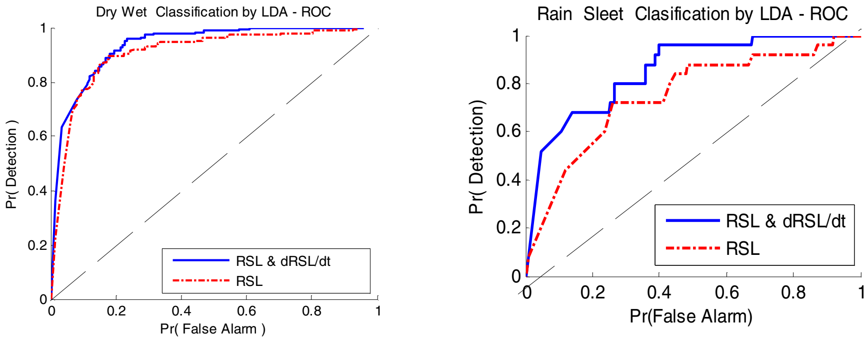


Fig. 4. ROCs of the classification system for various preprocessing methods (RSL only vs. using RSL with its derivative dRSL/dt.) and LDA feature extractor. On the left hand side, "Dry" / "Wet" classification ROC. On right hand side, "Rain" / "Sleet" classification ROC.

4 Discussion and Summary

In the scope of this paper, a novel detection/classification system capable of detecting wet periods, and of classifying the type of precipitationas rain or sleet, based on a measured attenuation signal obtained from spatially distributed wireless microwave links, was presented. The performance of the proposed classification system is summarized in Table 2.

Table 2. Classification system with LDA feature extraction – performace summary

	"Dry", "Wet" classification	*"Rain", "Sleet" classification*
Definition of "*False Alarm*"	*"dry"* classified as *"wet"*	*"rain"* classified as *"sleet"*
Definition of "*Detection*"	*"wet"* classified as *"wet"*	*"sleet"* classified as *"sleet"*
dRSL/dt contribution to classication performance	Low	High
Pr(Error)@ Pr(detection)	12%@ 83%	9% @ 52% , 25% @ 80%

This work presents, for the first time, the applicability of RSL signal measured on a commercial wireless microwaves network to monitor the type of precipitation causingthe attenuation. As part of this work, it was shown that the dynamic behavior of the Received Signal Level as expressed by the derivative of the attenuation signal is a sufficient discriminator for the classification of rain and sleet events. Following this work, it wouldbe interesting to go one step further to estimate the amount/intensity of the precipitation, which is relatively straightforward in case of pure rain and/or pure snow [19]. However, once sleet is involved,a solid theory relating to how this can be done could not be found. In searchfor a better classification performance, one may also consider a non-linear classification system as the physical system that is defiantly non-linear (quantization, minimum/maximum signals).

Acknowledgements. This research was partially supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 173/08).

References

1. Messer, H., Zinevich, A., Alpert, P.: Environmental monitoring by wireless communication networks. Science 312, 713 (2006)
2. Leijnse, H., Uijlenhoet, R., Stricker, J.: Rainfall measurement using radio links from cellular communication networks. Water Resource Res. 43(3) (2007)
3. Rahimi, A., Holt, A., Upton, G.: Attenuation calibration of an X-band weather radar using a microwave link. J. Atmos. Ocean. Technol. 23(3), 295–405 (2006)
4. Krämer, S., Verworn, H., Redder, A.: Microwave links A precipitation measurement method filling the gap between rain gauge and radar data? In: Proc. 6th Int. Workshop Precip. Urban Areas (2003)
5. Upton, G., Cummings, R., Holt, A.: Identification of melting snow using data from dual-frequency microwave links. Microw. Antennas Propag. 1(2), 282–288 (2007)
6. Leijnse, H., Uijlenhoet, R., Stricker, J.: Hydrometeorological application of a microwave link: 1. Evaporation, Water Resource Res. 43 (2007)
7. Goldshtein, O., Messer, H., Zinevich, A.: Rain rate estimation using measurements from commercial telecommunications links. IEEE Trans. Signal Proc. 57(4), 1616–1625 (2009)
8. Heder, B., Bertok, A.: Detection of sleet attenuation in data series measured on microwave links. Info Communication Journal LXIV(III), 2–8 (2009)
9. `http://www.ott.com/web/ott_de.nsf/id/pa_parsivel2_e.html`
10. Tjelta, T., Braten, L.E., Bacon, D.: Predicting the attenuation distribution on line-of-sight radio links due to melting snow. In: Proc. ClimDiff, Cleveland, U.S.A. (2005)
11. Dao, H., Rafiqul, M.D., Al-Khateeb, K.: Fade Dynamics review of Microwave Signals on Earth-Space Paths at Ku-Band. In: Proceeding ICCCE 2008 (May 2008)
12. ITU-R P.1623, Prediction method of fade dynamics on Earth space paths, ITU, Geneva, Switzerland (2003)
13. Castanet, L., de Kamp, M.V.: Modeling the Dynamic properties of the propagation channel. In: 5th Management Committee Meeting of the COST 280 Action (May 2003)
14. Fisher, R.: The statistical utilization of multiple measurements. Ann. Eugen. 8, 376–386 (1938)
15. Fukanaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, San Diego (1990)

16. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.-R.: Fisher discriminant analysis with kernels. In: Hu, Y.-H., Larsen, J., Wilson, E., Douglas, S. (eds.) Neural Networks for Signal Processing IX, pp. 41–48. IEEE, Piscataway (1999)
17. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)
18. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley-Interscience, New York (2001)
19. Frey, T.: The Effects of the Atmosphere and Weather on the Performance of a mm-Wave Communication Link. Applied Microwave & Wireless Magazine 11(2) (February 1999)

Bayesian Inference of Latent Causes in Gene Regulatory Dynamics

Sabine Hug[1,2] and Fabian J. Theis[1,2]

[1] Institute of Bioinformatics and Systems Biology,
Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany
[2] Institute for Mathematical Sciences, TU München, 85747 Garching, Germany
{sabine.hug,fabian.theis}@helmholtz-muenchen.de

Abstract. In the study of gene regulatory networks, more and more quantitative data becomes available. However, few of the players in such networks are observed, others are latent. Focusing on the inference of multiple such latent causes, we arrive at a blind source separation problem. Under the assumptions of independent sources and Gaussian noise, this condenses to a Bayesian independent component analysis problem with a natural dynamic structure. We here present a method for the inference in networks with linear dynamics, with a straightforward extension to the nonlinear case. The proposed method uses a maximum a posteriori estimate of the latent causes, with additional prior information guaranteeing independence. We illustrate the feasibility of our method on a toy example and compare the results with standard approaches.

Keywords: Independent component analysis, Bayesian inference, latent causes.

1 Introduction

In the field of bioinformatics, one prominent task is the study and inference of gene regulatory networks (GRNs). Even though more and more quantitative data becomes available, we still don't observe all players in these networks. The time courses of the unobserved players, also called latent causes, should then be inferred from the data through their influence on the observed players. For this, we use an ordinary differential equation model for the network with linear dynamics which can easily be extended to include also nonlinear dynamics. If we then assume that the latent causes are independent of each other and that the noise in our system is Gaussian, we can infer multiple latent causes by solving a Bayesian independent component analysis (ICA) problem. We here present a straightforward approach, where the maximum a posteriori (MAP) estimate for the latent causes is computed by a gradient descent algorithm in both linear and nonlinear models. Our method allows inference of all parameters involved in the system and naturally incorporates prior knowledge. When provided with suitable prior knowledge, it is able to break the sign indeterminacy present in usual ICA

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 520–527, 2012.

solutions. As example we study a small model, the so-called repressilator [1,3]. We introduce latent causes in the form of time courses of gene activity influencing the feedback loop. Our model can become nonlinear by the use of a sigmoidal activation function. We then show that Bayesian inference is able to retrieve the latent causes in the linear case. In the following, boldface capital/lowercase letters stand for matrices/vectors, non-bold capital/lowercase letters for their entries or scalars. We now first present some background on Bayesian inference in Section 2, then details on the method and the model in Section 3 and finally the results for the example in Section 4.

2 Bayesian Inference

When attempting blind source separation (BSS) of time course data, one difficulty is noisy data, due to measurement errors, biological variability etc. For this reason, we chose a Bayesian inference approach to BSS since it naturally incorporates the noise model for the data, here we will focus on Gaussian noise. We base our approach on Bayes' Theorem for the log of the posterior probability

$$\log p(\boldsymbol{\theta}|\mathbf{X}) \propto \log p(\mathbf{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of the model and $\mathbf{X}$ is the measured time course data, derived from the observed players. Thus the log posterior is proportional to the log likelihood plus the log prior. In our case, the parameters are the unobserved time courses of the latent causes and the mixing matrix of how they interact to produce the data $\mathbf{X}$, plus further parameters such as the strength of the noise.

In comparison to more involved methods such as mean field ICA [4], we directly determine an optimal set of parameters by computing the MAP estimate. We do so by minimization of the negative log posterior by gradient descent. This is convenient since with the assumption of normally distributed noise, the log posterior can be derived analytically such that we can compute explicit update formulas for the gradient descent. We will refer to our method as EM-MAP. We will now further describe the model and then provide details on the Bayesian inference.

3 Model Setup and Method Description

3.1 Data Description

Gene regulatory networks can be modeled by using a recent approach to gene expression data analysis [2] which uses generalized continuous time recurrent neural networks (CTRNN) as abstract dynamical models of regulatory systems, leading to ODEs of the form

$$\dot{g}_i(\tau) = \lambda_i \left(-g_i(\tau) + \sum_l \underbrace{W^{(o)}_{i,l}}_{\text{known}} \varphi_l(g_l(\tau)) + \sum_j \underbrace{W^{(u)}_{i,j}}_{\text{unknown}} \varphi_j(h_j(\tau)) \right) . \tag{2}$$

The different $\mathbf{g}_i$ in equation (2) represent the time courses of the measured genes, i.e. the observed players, while the $\mathbf{h}_j$ represent the time courses of the latent causes, all varying with time τ. Interactions are incorporated via the possibly linear or nonlinear activation function φ, where an example for a nonlinear activation could be $\varphi(x) = (1 + \mathrm{e}^{-ax})^{-1}$. In the linear case, which we will use to compare our method to standard approaches, we take $\varphi(x) = x$. The indexes $i = 1, ..., n_i$ and $j = 1, ..., n_j$ indicate the n_i observed genes and the n_j latent causes, respectively. We assume here that n_j is known, otherwise we face the more difficult task of model selection.

The parameter λ_i in equation (2) denotes the degradation rate. The matrix $\mathbf{W}$ is the interaction matrix, containing the interactions between all involved players, whether observed or latent.

$$\mathbf{W} = \begin{array}{cc} \begin{array}{cc} \mathbf{g} & \mathbf{h} \end{array} & \\ \begin{pmatrix} \mathbf{W}^{(o)} & \mathbf{W}^{(u)} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} & \begin{array}{c} \mathbf{g} \\ \mathbf{h} \end{array} \end{array} \tag{3}$$

We assume that the interactions between the observed players, i.e the block $\mathbf{W}^{(o)}$ in equation (3), are known and the block of $\mathbf{W}$ belonging to the interactions on the latent causes is equal to zero, see equation (3). The matrix $\mathbf{W}^{(u)}$ in equation (3) corresponds to the interactions of the latent causes on the observed players, we will call it the mixing matrix from now on. It corresponds to the mixing matrix commonly found in standard BSS problems. Furthermore, a positive mixing matrix entry means an activating interaction, while a negative entry stands for inhibition.

We now resort equation (2) according to observed and unobserved players and obtain the mixing model

$$x_i\left(\tau\right) = \lambda_i \sum_j W_{i,j}^{(u)} \varphi_j\left(h_j\left(\tau\right)\right) , \tag{4}$$

where the **data** $x_i\left(\tau\right)$ can be computed from the observed time courses $g_i\left(\tau\right)$ according to:

$$x_i\left(\tau\right) = \dot{g}_i\left(\tau\right) + \lambda_i g_i\left(\tau\right) - \lambda_i \sum_l W_{i,l}^{(o)} \varphi_l\left(g_l\left(\tau\right)\right) \tag{5}$$

The derivative $\dot{g}_i\left(\tau\right)$ in eq. (5) can be estimated from $g_i\left(\tau\right)$ using e.g. splines. In order to be able to solve the system numerically, we abandon the notion of continuous time τ and make a transition to $t = 1, ..., n_t$ discrete time points, yielding discrete versions of equations (4) and (5), cf. also equation (6).

3.2 Likelihood Setup

By assuming that our data is subject to time-independent, signal-specific Gaussian noise

$$x_{i,t} = \lambda_i \sum_{j=1}^{n_j} W_{i,j}^{(u)} \varphi_j\left(h_{j,t}\right) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right) \text{ for } i = 1, ..., n_i, t = 1, ..., n_t \tag{6}$$

we are able to set up the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$. Now we can describe in detail what the parameters $\boldsymbol{\theta}$ of the system are: first the time courses of the latent causes $\mathbf{H}$, then the entries of the mixing matrix corresponding to the influence of the latent causes on the observed players $\mathbf{W}^{(u)}$ and furthermore the additional parameters of the mixing model, namely λ_i, σ_i, $i = 1, ..., n_i$ and possibly the parameters of the activation function a_i, $i = 1, ..., n_i$. For easier computation, we rearrange all parameters lexicographically into a vector, i.e. $\boldsymbol{\theta}$. An advantage of our method in comparison with non-Bayesian methods like FastICA lies in the fact that by our approach, we naturally get an estimate for all parameters of the system, i.e. also for the strength of the noise and not only for the two matrices $\mathbf{W}^{(u)}$ and $\mathbf{H}$.

We take the natural logarithm of the likelihood and get

$$F(\boldsymbol{\theta}) = \log p(\mathbf{X}|\boldsymbol{\theta}) = \log \prod_{t=1}^{n_t} \prod_{i=1}^{n_i} \mathcal{N}_{x_{i,t}} \left(\lambda_i \sum_{j=1}^{n_j} W_{i,j}^{(u)} \varphi_j \left(h_{j,t} \right), \sigma_i^2 \right) = \tag{7}$$

$$= -n_t \sum_{i=1}^{n_i} \log \left(\sqrt{2\pi} \sigma_i \right) - \sum_{t=1}^{n_t} \sum_{i=1}^{n_i} \frac{1}{2\sigma_i^2} \left(x_{i,t} - \lambda_i \sum_{j=1}^{n_j} W_{i,j}^{(u)} \varphi_j \left(h_{j,t} \right) \right)^2 .$$

This yields a derivative (provided all φ_j are differentiable) of

$$\frac{\partial F}{\partial h_{j,t}} = -\varphi_j' \left(h_{j,t} \right) \sum_{i=1}^{n_i} \frac{\lambda_i W_{i,j}^{(u)}}{\sigma_i^2} \left(x_{i,t} - \lambda_i \sum_{b=1}^{n_j} W_{i,b}^{(u)} \varphi_b \left(h_{b,t} \right) \right) \tag{8}$$

and

$$\frac{\partial F}{\partial W_{i,j}^{(u)}} = -\frac{\lambda_i}{\sigma_i^2} \sum_{t=1}^{n_t} \varphi_j \left(h_{j,t} \right) \left(x_{i,t} - \lambda_i \sum_{b=1}^{n_j} W_{i,b}^{(u)} \varphi_b \left(h_{b,t} \right) \right). \tag{9}$$

The derivatives with respect to σ_i, λ_i and a_i are similar in form and thus omitted for the sake of brevity.

3.3 Prior Distributions

Since the likelihood is just a consequence of the assumed error model, we need to use the prior distribution of the latent causes to enforce their independence. This is equivalent to a prior that factorizes over the causes, such as

$$p(\mathbf{H}) = \prod_{j=1}^{n_j} p \left(\mathbf{h}_j \right), \quad p \left(\mathbf{h}_j \right) = \prod_{t=1}^{n_t} p \left(h_{j,t} \right). \tag{10}$$

To give the prior a shape which is well suited to represent a latent cause which switches on and off again, we choose a mixture of two Gaussians, which is a commonly used approach [8], to yield

$$p \left(h_{j,t} \right) = \sum_{k=1}^{2} \alpha_{j,k} \, \mathcal{N} \left(h_{j,t} | \phi_{j,t,k}, \beta_{j,k}^2 \right) \tag{11}$$

where

$$\phi_{j,t,k} = \nu_{j,k} \, \mathcal{N}\left(t|\xi_{j,k}, \delta_{j,k}\right) \tag{12}$$

The prior modes $\xi_{j,k}$ should be chosen according to the largest change in the data or any prior knowledge on the on/off time of the latent cause. For all other parameters, we choose uniform priors. A prior of this form yields a derivative of

$$\begin{aligned}\frac{\partial p\left(h_{j,t}\right)}{\partial h_{j,t}} =& \sum_{k=1}^{2} \alpha_{j,k} \frac{1}{\sqrt{2\pi\beta_{j,k}^{2}}} \cdot \\ & \cdot \exp\left(-\frac{1}{2\beta_{j,k}^{2}}\left(h_{j,t}-\phi_{j,t,k}\right)^{2}\right) \cdot \left(-\frac{1}{\beta_{j,k}^{2}}\left(h_{j,t}-\phi_{j,t,k}\right)\right)\end{aligned} \tag{13}$$

Together with eq. (8) and eq. (9), equation (13) yields the gradient necessary for a standard gradient descent. In the actual implementation, we use adaptive step sizes and an EM-like update scheme (hence the name EM-MAP), where the parameters of the mixing matrix and all the other parameters are updated alternately. To exclude the possibility of the optimization getting stuck in a local optimum, we run it 100 times with randomly drawn initial values until convergence. We ensure uniqueness of the decomposition by first whitening the data $\mathbf{X}$ and furthermore whitening the current matrix $\mathbf{H}$ after each step.

4 Results

4.1 Setup of the Toy Example

We demonstrate the potential of our method by applying it to a toy model. We chose the so-called *repressilator* [3], a GRN of three observed genes, where each gene inhibits the next gene in the loop. Tying into that loop are two additional players which are unobserved and thus latent causes [1]. The aim is to reconstruct the latent causes from their influence on the observed gene activities. We do so by inferring the entries of the mixing matrix $\mathbf{W}^{(u)}$ for the latent causes which corresponds to finding out which of the observed genes are or are not targeted by which of the latent causes. Indeed, we will focus on a reconstruction of the mixing matrix which is as good as possible, since the entries describe the type of interaction and may thus yield testable predictions. The complete setup of the toy data is shown in Figure (1), the corresponding mixing matrix $\mathbf{W}^{(u)}$ is shown in equation (14).

$$\mathbf{W}^{(u)} = \begin{pmatrix} -2 & 1.2 \\ -1 & 0 \\ 0 & 0 \end{pmatrix} \tag{14}$$

To obtain the data from our toy model, we solved the differential equations numerically on the time interval $\tau \in [0, 50]$. By inserting the solution into the right hand side of the differential equation, we obtain the derivatives and can then compute the data $\mathbf{X}$ from the discrete version of equation (5), based on $n_t = 20$ time points.

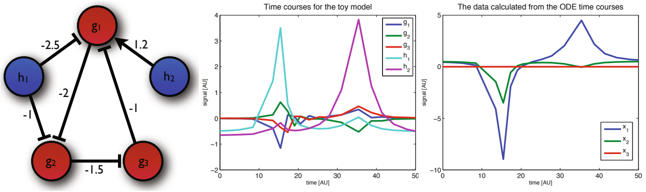


Fig. 1. The toydata: on the left a schematic of the setup with the observed players $\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3$ in red and the latent causes $\mathbf{h}_1, \mathbf{h}_2$ in blue, in the middle the time courses of all five players and on the right the three signals computed from the players according to the discrete version of equation (5)

4.2 Results for the Toy Example

For the evaluation of EM-MAP, we focused on the linear case with $\varphi(x) = x$, since we can then compare our results to existing ICA methods. For this reason, we also fixed the noise strengths σ_i and set $\lambda_i = 1$. We obtained results for our toy model for four scenarios with different noise strengths, first with no noise, then low (SNR $\approx$ 20 dB), medium (SNR $\approx$ 10 dB) and high noise intensity (SNR $\approx$ 5 dB). For all four settings, the gradient descent of EM-MAP was run a 100 times to yield the results depicted in Figure (2). We see from Figure (2) that the runs corresponding to good posterior values, i.e. near the beforehand known maximum, also correspond to good fits with the true entries of the mixing matrix. About 10 % of runs converge to a value near the true solution. On the left in Figure (2), we see the results for the noise free case, where the match is very good. Notice especially that practically all runs correctly reconstruct the fact that $\mathbf{g}_3$ is not targeted by the latent causes, i.e. that the entries 3 and 6 are zero. In the case for low noise intensity, we see that the quality of the reconstruction is still very good, while it deteriorates for medium and high noise levels, depicted in the two plots on the right. Note that now entries 3 and 6 of the mixing matrix are not zero anymore in many runs. Also, we tested several variants of the prior with different widths and locations of the modes and found that results stay about the same, as long as prior modes are not close together.

4.3 Comparison of Our Method with Standard Approaches

We also compared our results with existing methods, namely with FastICA [6] and mean field ICA [4], correcting for permutations and sign changes in the mixing matrix for these two methods. Due to the influence of the prior, which prefers positive time courses to negative ones, our method does not suffer from the sign indeterminacy. For the comparison, we randomly chose ten additional settings of the repressilator with two latent causes and ran both EM-MAP and FastICA 100 times. The settings varied in the values in both $\mathbf{W}^{(u)}$ and in $\mathbf{W}^{(o)}$, however we kept the structure of the three known players inhibiting each other in a loop. Note that we use only $n_t = 20$ time points, since data for GRNs is

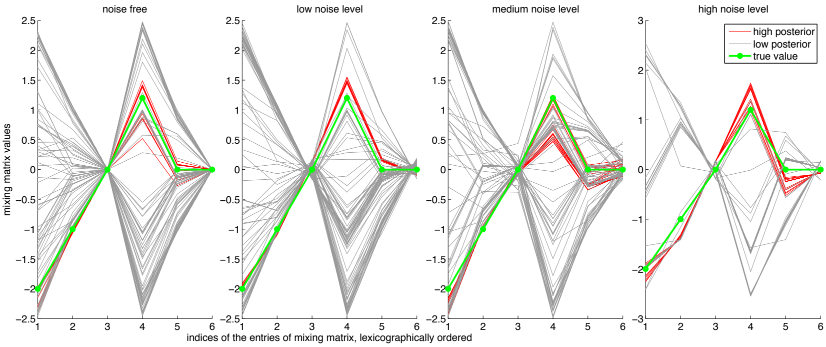


Fig. 2. The six entries of the mixing matrix, shown in a parallel coordinates plot. Each line corresponds to one found mixing matrix, grey lines for solutions with low posterior values, red lines for solutions with high posterior value, for all four noise levels, respectively. The green line represents the true matrix.

typically sparse. To evaluate our method, we picked the gradient descent run with the best posterior value. We then looked at the absolute differences from the true mixing matrix for all three methods, see Table (1). We find that icaMF, the mean field ICA method, does not perform very well on our toy example, while our results are comparable to the results for FastICA. Indeed FastICA performs better than EM-MAP in the settings with no noise and low noise intensity, for medium noise intensity, both methods are equally good. EM-MAP clearly outperforms FastICA for high noise intensity, a situation we will typically have in GRNs combined with only few time points. Furthermore, FastICA did not converge for the first 100 initial values for one of the settings, while convergence for EM-MAP stays constant over all noise levels. This suggests a high robustness to noise of our method, which is especially important for the study of GRNs such that we expect our method to perform better than FastICA on real data. All methods also work with only 10 time points, however the quality of the solution deteriorates.

Table 1. The absolute differences from the true mixing matrix, for the four noise levels and the three methods: our method (EM-MAP), icaMF and FastICA, mean and standard error computed from the 11 settings of the repressilator

Data	**EM-MAP**	**icaMF**	**FastICA**
noise free	0.5287 ± 0.060	3.0458 ± 0.29	0.2970 ± 0.03
low noise intensity	0.6166 ± 0.049	3.0526 ± 0.36	0.4566 ± 0.032
medium noise intensity	0.8098 ± 0.078	3.0964 ± 0.28	0.7794 ± 0.058
high noise intensity	1.2969 ± 0.15	3.7274 ± 0.34	2.2310 ± 0.21

5 Conclusions and Outlook

In this contribution, we propose a direct method to compute the MAP estimate for Bayesian ICA we call EM-MAP. We demonstrated its ability to correctly reconstruct the mixing matrix on a toy example. Furthermore, we showed that it is robust to noise in the data. We expect it to be that also when challenged with real data. It is easily possible to extend our method to nonlinear settings by using a nonlinear activation function. We conclude that Bayesian inference provides a powerful framework for BSS which can use existing prior knowledge on the latent causes. Also, in a next step, it would be desirable to extend the inference by using MCMC sampling instead of a MAP estimate in order to provide whole distributions for the unknown parameters. Another possible extension is the use of a mixture of Gaussian processes as a prior distribution which would then capture the time dependence between adjacent data points in the signals. Future work could also include model selection through thermodynamic integration [7]. Since we will probably only very rarely have the case in GRNs that all players are observed, we expect methods for reconstructing latent causes and especially EM-MAP to be of particular relevance in that field.

Acknowledgements. This work was supported by the Federal Ministry of Education and Research (BMBF) in its MedSys initiative (project “SysMBO”), and the European Union within the ERC grant “LatentCauses.” The authors are grateful to Florian Blöchl for his initial work on the project and helpful discussions, to Katrin Illner and Ivan Kondofersky for critical comments on the manuscript and to Daniel Schmidl for helpful discussions.

References

1. Blöchl, F., Theis, F.J.: Estimating Hidden Influences in Metabolic and Gene Regulatory Networks. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 387–394. Springer, Heidelberg (2009)
2. Busch, H., Camacho-Trullio, D., Rogon, Z., Breuhahn, K., Angel, P., Eils, R., Szabowski, A.: Gene network dynamics controlling keratinocyte migration. Molecular Systems Biology 4 (2008)
3. Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators. Nature 4(6767), 335–338 (2000)
4. Højen-Sørensen, P.A., Winther, O., Hansen, L.K.: Mean-field approaches to independent component analysis. Neural Computation 14, 889–918 (2002)
5. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York (2001)
6. Hyvärinen, A.: Fast and robust fixedpoint algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)
7. Lartillot, N., Philippe, H.: Computing Bayes factors using thermodynamic integration. Systematic Biology 55, 195–207 (2006)
8. Marin, J., Mengersen, K., Robert, C.P.: Bayesian modelling and inference on mixtures of distributions. Bayesian Thinking (2005)

Bayesian Fuzzy Clustering of Colored Graphs

Fabian J. Theis

Institute for Bioinformatics and Systems Biology
Helmholtz Zentrum München, Germany
fabian.theis@helmholtz-muenchen.de
http://cmb.helmholtz-muenchen.de

Abstract. With the increasing availability of interaction data stemming form fields as diverse as systems biology, telecommunication or social sciences, the task of mining and understanding the underlying graph structures becomes more and more important. Here we focus on data with different types of nodes; we subsume this meta information in the color of a node. An important first step is the unsupervised clustering of nodes into communities, which are of the same color and highly connected within but sparsely connected to the rest of the graph. Recently we have proposed a fuzzy extension of this clustering concept, which allows a node to have membership in multiple clusters. The resulting gradient descent algorithm shared many similarities with the multiplicative update rules from nonnegative matrix factorization. Two issues left open were the determination of the number of clusters of each color, as well as the non-defined integration of additional prior information. In this contribution we resolve these issues by reinterpreting the factorization in a Bayesian framework, which allows the ready inclusion of priors. We integrate automatic relevance determination to automatically estimate group sizes. We derive a maximum-a-posteriori estimator, and illustrate the feasibility of the approach on a toy as well as a protein-complex hypergraph, where the resulting fuzzy clusters show significant enrichment of distinct gene ontology categories.

1 Introduction

We are studying the question of clustering a k-colored graph into multiple overlapping ('fuzzy') clusters within each color. Here, a *k-colored weighted graph* denotes a (positively) weighted graph $G = (V, E)$ together with a partition of the vertices V into k disjoint sets V_i. G is called partite if no two vertices in the same subset are adjacent, i.e. edges are only allowed between different subsets. In the following we denote the different sets as *colors*. Let $n_i := |V_i|$ be the number of vertices in partition i. We represent the graph as a set of $n_i \times n_j$ weight matrices $\mathbf{A}^{(ij)}$ with nonnegative entries for $1 \leq i < j \leq k$. So node r of color i is linked with weight $\mathbf{A}^{(ij)}_{rl}$ to node l of color j. The graph may be directed if $\mathbf{A}^{(ij)}_{rl} \neq \mathbf{A}^{(ij)}_{lr}$.

Let m_i denote the number of clusters of V_i. We say that a non-negative $n_i \times m_i$-matrix $\mathbf{C}^{(i)}$ is a *fuzzy clustering* of V_i, if it is right-stochastic i.e. $\sum_s c^{(i)}_{rs} = 1$ for

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 528–535, 2012.

all k. In other words $c_{rs}^{(i)}$ quantifies the contribution of the r-th vertex of V_i to the V_i-cluster s. A *fuzzy clustering of* G is then defined as the approximation of G by a smaller 'backbone network' H, defined on fuzzy clusters of same-colored vertices of G. So we search for a k-colored graph H with $m_i \times m_j$ weight matrices $\mathbf{B}^{(ij)}$ and $n_i \times m_i$ fuzzy clustering matrices $\mathbf{C}^{(i)}$ such that the connectivity explained by H is as close as possible to G after clustering, i.e. such that

$$\mathbf{A}^{(ij)} \approx \mathbf{C}^{(i)}\mathbf{B}^{(ij)}(\mathbf{C}^{(j)})^\top =: \hat{\mathbf{A}}^{(ij)}. \tag{1}$$

In the following we will measure approximation quality by some matrix norm or divergence.

We have previously addressed this question [2,3] as extension of a discrete graph clustering method [6]. Our main contribution was a novel efficient minimization procedure, mimicking the multiplicative update rules employed in algorithms for non-negative matrix factorization [4]. However just as in k-means, the user had to provide the algorithm with the desired number of clusters for each color. Moreover, it was unclear how to include additional prior information such as already known interactions or clusterings in the method.

Here we propose a Bayesian extension of our previously proposed algorithm. The number of clusters will be determined using automatic relevance determination (ARD), which only needs a single hyper parameter to determine a degree of coarse graining simultaneously and comparably across all graph colors. ARD has been initially proposed by [7], with later successful applications in PCA [1] and more recently NMF [11] and community detection [9].

In the following, we will first derive the algorithm and then present an NMF-type multiplicative update algorithm, solving a maximum-a-posteriori optimization. We finish with a toy and an example from Bioinformatics.

2 Method

2.1 Bayesian Fuzzy Clustering Model

We want to approximate each adjacency matrix by its clustered backbone according to equation (1). A Bayesian interpretation of the previously used least-squares cost function [2,3] implies a probabilistic noise model resulting in the likelihood $p(\mathbf{A}^{(ij)}|\hat{\mathbf{A}}^{(ij)}) = \mathcal{N}(\mathbf{A}^{(ij)}|\hat{\mathbf{A}}^{(ij)}, \sigma)$, with normal distribution $\mathcal{N}$ and some fixed noise variance σ. Instead we now choose the parameter-free noise model of a Poisson prior $p(a_{rs}^{(ij)}|\hat{a}_{rs}^{(ij)}) = \mathcal{P}(a_{rs}^{(ij)}|\hat{a}_{rs}^{(ij)})$ with $\mathcal{P}(x|\lambda) = e^{-\lambda}\lambda^x/\Gamma(x+1)$ as motivated e.g. in [8,11]. The corresponding negative log-likelihood can be rewritten as

$$-\log p(\mathbf{A}^{(ij)}|\hat{\mathbf{A}}^{(ij)}) = \sum_{rs} \hat{a}_{rs}^{(ij)} - a_{rs}^{(ij)} \log \hat{a}_{rs}^{(ij)} + \log \Gamma(a_{rs}^{(ij)} + 1),$$

which essentially measures the Kullback-Leibler divergence between matrices $\mathbf{A}^{(ij)}$ and $\hat{\mathbf{A}}^{(ij)}$, interpreted as i.i.d. samples of two random variables. Since the last term is independent of $\hat{\mathbf{A}}$, maximizing the above negative log-likelihood is equivalent to minimizing the KL-divergence.

Using the shrinkage approach from automatic relevance determination [7], successfully employed in Bayesian PCA [1] and NMF [11], we now constrain each of the clustering matrices $\mathbf{C}^{(i)}$ along the rows according to

$$p(c_{rk}^{(i)}|\beta_k^{(i)}) = \mathcal{HN}\left(c_{rk}^{(i)}\middle| 0, \beta_k^{(i)-1}\right)$$

with

$$\mathcal{HN}(x|0,\beta^{-1}) = \begin{cases} \sqrt{\frac{2}{\pi}}\,\beta^{1/2}\exp\left(-\frac{1}{2}\beta x^2\right) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

being the half-normal distribution cut off at 0 with precision (of the corresponding full normal distribution) β. This implies the negative log-priors

$$-\log p(\mathbf{C}^{(i)}|\boldsymbol{\beta}^{(i)}) = \frac{1}{2}\sum_k \left(\sum_{r=1}^{m_i} \beta_k^{(i)} c_{rk}^{(i)2}\right) - m_i \log \beta_k^{(i)} + \text{const}\,.$$

Here m_i is the (maximal) number of allowed dimensions in the reduction, which may be as high as n_i, but in practice should be lower.

The key point now lies in tying the prior for the middle factor $\mathbf{B}^{(ij)}$ to the same hyper-parameters $\boldsymbol{\beta}$:

$$p(b_{kl}^{(ij)}|\beta_k^{(i)}, \beta_l^{(j)}) = \mathcal{HN}\left(b_{kl}^{(ij)}\middle| 0, \beta_k^{(i)-1} + \beta_l^{(j)-1}\right)$$

We need to include these terms in the fuzzy clustering from left (by $\mathbf{C}^{(i)}$) and from right (by $\mathbf{C}^{(j)}$), so we add up the two variances. Note that instead of normalizing the clustering matrices as before to remove the scaling indeterminacies, we now include the scale parameter β, which does not overconstrain the problem. This results in the following negative log-prior

$$\begin{aligned} -\log p(\mathbf{B}^{(ij)}|\boldsymbol{\beta}^{(i)}, \boldsymbol{\beta}^{(j)}) &= \\ &= \frac{1}{2}\sum_{kl} \frac{\beta_k^{(i)}\beta_l^{(j)}}{\beta_k^{(i)} + \beta_l^{(j)}} b_{kl}^{(ij)2} - \log \beta_k^{(i)} - \log \beta_l^{(j)} + \log(\beta_k^{(i)} + \beta_l^{(j)}) + \text{const} \\ &= \frac{1}{2}\left(\sum_{kl} \frac{\beta_k^{(i)}\beta_l^{(j)}}{\beta_k^{(i)} + \beta_l^{(j)}} b_{kl}^{(ij)2} + \log(\beta_k^{(i)} + \beta_l^{(j)})\right) \\ &\quad - \frac{1}{2}m_j \sum_k \log \beta_k^{(i)} - \frac{1}{2}m_i \sum_l \log \beta_l^{(j)} + \text{const}\,. \end{aligned}$$

Finally, we specify each precision $\beta_k^{(i)}$ by a Gamma distribution, conjugate to the half-normal density, with hyper-parameters shape g and rate h chosen to be independent of k and i:

$$p(\beta_k^{(i)}|g,h) = \frac{h^g}{\Gamma(g)}\beta_k^{(i)g-1}\exp(-\beta_k^{(i)}h)$$

This leads to the negative log-prior

$$-\log p(\boldsymbol{\beta}^{(i)}|g,h) = \sum_k h\beta_k^{(i)} - (g-1)\log\beta_k^{(i)} + \text{const}\,.$$

The resulting overall posterior of the model is

$$p(\mathbf{B},\mathbf{C},\boldsymbol{\beta}|\mathbf{A}) \sim p(\mathbf{A}|\mathbf{B},\mathbf{C})p(\mathbf{B}|\boldsymbol{\beta})p(\mathbf{C}|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

according to Bayes theorem, so altogether after some simplification we get the total negative log posterior of the model as

$$\begin{aligned} f_{\text{MAP}}(\mathbf{B},\mathbf{C},\boldsymbol{\beta}) = &\sum_{ij}\sum_{rs}\left(\sum_{kl} c_{rk}^{(i)} b_{kl}^{(ij)} c_{sl}^{(j)} - a_{rs}^{(ij)} \log \sum_{kl} c_{rk}^{(i)} b_{kl}^{(ij)} c_{sl}^{(j)}\right) \\ &+ \sum_i\sum_k \left(h + \frac{1}{2}\sum_r c_{rk}^{(i)2}\right)\beta_k^{(i)} \\ &+ \frac{1}{2}\sum_{ij}\sum_{kl} b_{kl}^{(ij)2}\frac{\beta_k^{(i)}\beta_l^{(j)}}{\beta_k^{(i)}+\beta_l^{(j)}} + \frac{1}{2}\sum_{ij}\sum_{kl}\log(\beta_k^{(i)}+\beta_l^{(j)}) \\ &- \sum_i\sum_k\left(\frac{1}{2}m_i + M + (g-1)\right)\log\beta_k^{(i)} + \text{const}, \end{aligned}$$

with $M = \sum_j m_j$ being the number of all latent dimensions. Here the regulating influence of $\boldsymbol{\beta}$ can be seen: for larger $\beta_k^{(i)}$, the last (negative) term will decrease i.e. become more negative, whereas the second to fourth (positive) terms will all increase, thus in effect forcing some $\beta_k^{(i)}$ to stay small.

2.2 Algorithm

We will minimize the negative log posterior f_{MAP} using a multiplicative update rule, based on local gradient descent, generalizing the multiplicative updates from NMF [4,5]. For this we first determine partial derivatives of f_{MAP}. If we are to minimize f by alternating gradient descent, we then simply start from an initial guess of $\mathbf{B}^{(ij)}, \mathbf{C}^{(i)}, \boldsymbol{\beta}^{(i)}$ and alternate between updates of $\mathbf{B}^{(ij)}$, $\mathbf{C}^{(i)}$ and $\boldsymbol{\beta}^{(i)}$ for all i, j:

$$\begin{aligned} b_{kl}^{(ij)} &\leftarrow b_{kl}^{(ij)} - \eta_{kl}^{(ij)}\frac{\partial f}{\partial b_{kl}^{(ij)}} \\ c_{rk}^{(i)} &\leftarrow c_{rk}^{(i)} - \eta_{rk}^{(i)}\frac{\partial f}{\partial c_{rk}^{(i)}} \\ \beta_k^{(i)} &\leftarrow \beta_k^{(i)} - \eta_k^{(i)}\frac{\partial f}{\partial \beta_k^{(i)}} \end{aligned}$$

These update rules have two disadvantages: for one, choice of update rate η (possibly different for $\mathbf{B}$, $\mathbf{C}$, $\boldsymbol{\beta}$ and i, j) is unclear; in particular, for too small η convergence may take too long or may not be achieved at all, whereas for too large η we may easily overshoot the minimum. Moreover, the resulting matrices may become negative. Hence we follow the Lee and Seung's idea for NMF [4] and rewrite this into multiplicative update rules. Let us choose update rates

$$\eta_{kl}^{(ij)} := \frac{b_{kl}^{(ij)}}{b_{kl}^{(ij)} \frac{\beta_k^{(i)} \beta_l^{(j)}}{\beta_k^{(i)} + \beta_l^{(j)}} + \sum_{rs} c_{rk}^{(i)} c_{sl}^{(j)}}$$

Plugging this into the gradient descent equations, this results in the desired multiplicative update rules

$$\begin{aligned} b_{kl}^{(ij)} &\leftarrow b_{kl}^{(ij)} - \eta_{kl}^{(ij)} \left(b_{kl}^{(ij)} \frac{\beta_k^{(i)} \beta_l^{(j)}}{\beta_k^{(i)} + \beta_l^{(j)}} + \sum_r c_{rk}^{(i)} \sum_s c_{sl}^{(j)} \right) + \eta_{kl}^{(ij)} \sum_{rs} \frac{a_{rs}^{(ij)} c_{rk}^{(i)} c_{sl}^{(j)}}{\sum_{uv} c_{ru}^{(i)} b_{uv}^{(ij)} c_{sv}^{(j)}} \\ &= \frac{b_{kl}^{(ij)}}{b_{kl}^{(ij)} \frac{\beta_k^{(i)} \beta_l^{(j)}}{\beta_k^{(i)} + \beta_l^{(j)}} + \sum_r c_{rk}^{(i)} \sum_s c_{sl}^{(j)}} \sum_{rs} \frac{a_{rs}^{(ij)} c_{rk}^{(i)} c_{sl}^{(j)}}{\sum_{uv} c_{ru}^{(i)} b_{uv}^{(ij)} c_{sv}^{(j)}} \end{aligned}$$

$$c_{rk}^{(i)} \leftarrow \frac{c_{rk}^{(i)}}{c_{rk}^{(i)} \beta_k^{(i)} + \sum_{sjl} \left(b_{kl}^{(ij)} + b_{lk}^{(ji)} \right) c_{sl}^{(j)}} \sum_{sjl} \frac{\left(a_{rs}^{(ij)} b_{kl}^{(ij)} + a_{sr}^{(ji)} b_{lk}^{(ji)} \right) c_{sl}^{(j)}}{\sum_{uv} c_{ru}^{(i)} b_{uv}^{(ij)} c_{sv}^{(j)}}$$

$$\beta_k^{(i)} \leftarrow \frac{\frac{1}{2} m_i + M + (g-1)}{h + \frac{1}{2} \sum_r c_{rk}^{(i)2} + \sum_j \sum_l \left(\frac{\left(b_{kl}^{(ij)2} + b_{lk}^{(ji)2} \right)}{2(\beta_k^{(i)} / \beta_l^{(j)} + 1)^2} + \frac{1}{\beta_k^{(i)} + \beta_l^{(j)}} \right)}$$

Note that we can rewrite the update rules in the commonly used matrix form; however in our present setting they turn out to be more complicated than the above component-wise rules, so we omit them for brevity.

In practice we set the shape hyper-parameter in the Gamma prior on β to $g = 2$, and only include the rate hyper-parameter h as tunable scale, which can be increased to get more clusters overall. An abort criterion is defined via a minimally required rate of increase of the cost function.

3 Results

3.1 Algorithm Performance on a 2-Partite Toy Example

As starting example, we choose a 2-partite toy example already proposed in [6] and further studied in [2]. 6 nodes of color 1 ('upper', see figure 3.1a) connect 4 nodes of color 2 ('lower'), whereby the first two upper connect only the first two lower, same for the last two, and the center two upper nodes connect all lower ones.

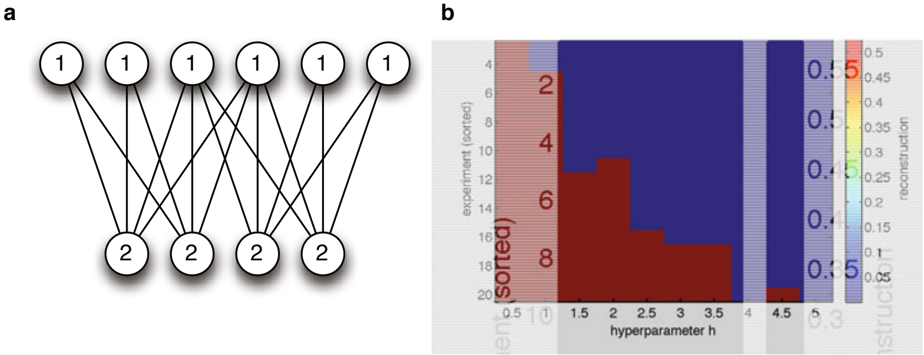


Fig. 1. Fuzzy clustering on a toy example. (a) The 2-partite example from [6] is chosen, with $(n_1, n_2) = (6, 4)$. (b) Result of structure recovery for different hyper-parameters h over 20 runs.

Depending on the hyper-parameter h, the fuzzy clustering finds different number of clusters; this also depends on initialization due to local convergence of the search algorithm. However for sufficiently larger h, the algorithm much more robustly determines the correct answer, see figure 3.1b, where we measure reconstruction by the norm of $s_1\mathbf{A}^{(ij)} - s_2\mathbf{C}^{(i)}\mathbf{B}^{(ij)}(\mathbf{C}^{(j)})^\top$ with scalings s_1 and s_2 chosen as the respective factor's inverse norm in order to account for such indeterminacies, occurring due to use of Kullback-Leibler distance. We observe that for sufficiently high hyper-parameter h, the correct reconstruction and indeed the correct number of clusters is always found. This is not the case for lower h. Hence the algorithm needs to be initialized with sufficiently large h, which however only needs to be chose across all colors. Actual choice does not matter as much as the cluster choice in previous work [2].

3.2 Fuzzy Clusters of a Protein-Complex Hypergraph

We finish by briefly illustrating the applicability of our method to heterogeneous biological data; here we choose an example network from bioinformatics consisting of proteins and protein complexes. A protein complex is defined as a group of two or more associated polypeptide chains. Proteins in such a complex are linked by non-covalent protein?protein interactions. Protein complexes are a key building block of many biological processes. Essentially, they form the molecular machinery that perform a vast array of biological functions in a cell.

We ask the question how proteins are organized in their functional protein complexes. For this we constructed a hypergraph i.e. a two-partite graph with nodes protein and protein complex, and links if a protein is involved in a complex. The hypergraph was extracted from the CORUM data base [10], which represents a non-redundant catalogue of experimentally verified mammalian protein complexes manually annotated at MIPS. Proteins were mapped onto their corresponding gene names in order to compile such a hypergraph across multiple

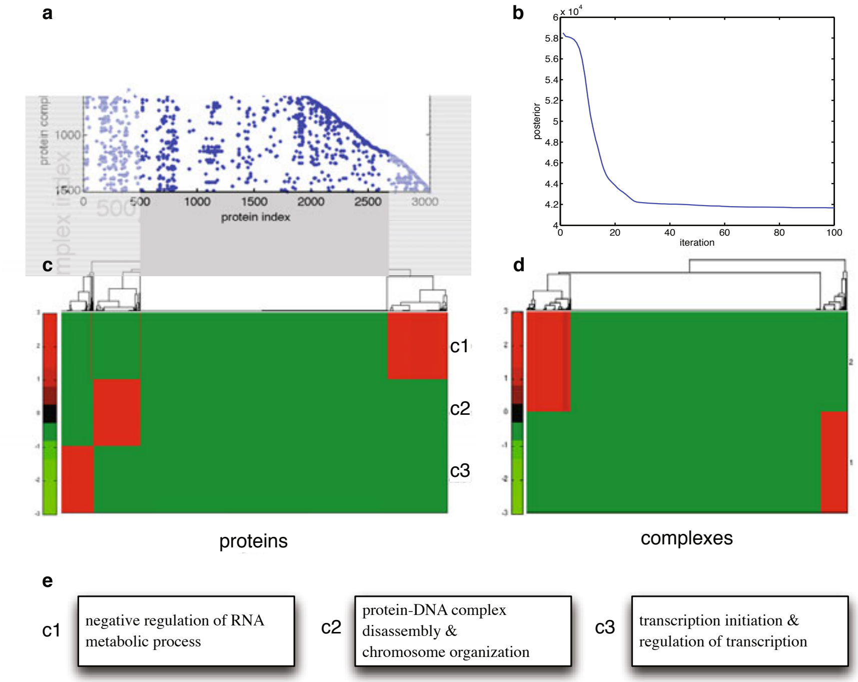


Fig. 2. Fuzzy clustering of a protein-complex-hypergraph. (a) The bipartite unweighted adjacency matrix of proteins and protein complexes. (b) Algorithm convergence (posterior) over 100 iterations. The resulting fuzzy protein clusters (c) and complex clusters (d) are mostly non-overlapping. A gene ontology enrichment ($p < 10^{-9}$) of each protein cluster (e) shows clearly separated biological processes per cluster.

species (mammals only). The resulting bipartite graph $G = ((V_1, V_2), E)$ consisted of $|V_1| = 3041$ proteins and $|V_2| = 1515$ complexes, and was linked with $|E| = 8255$ edges, see figure 3.2(a).

Bayesian fuzzy clustering was applied with sufficiently high h and 10 as maximal number of clusters for both proteins and complexes. The algorithm was run for 100 iterations, and resulted in 3 protein and 2 complex clusters, see figure 3.2(b-d). In order to test for relevance of each of the three protein clusters, we tested whether 'similar' proteins are grouped together. This we determine by gene ontology enrichment, which measures whcih biological processes are overrepresented in which cluster (Figure 3.2e). We find that each of the clusters represents processes essential for cell survival and progression. Most notably, the three major processes have a dynamic nature. On protein level, the coordination of proteins is accomplished by proteins transiently forming complexes to carry out their tasks. Depending on the e.g. signaling pathway activated, proteins may build many different complexes each fulfilling the majority of cellular processes. In summary, the proposed fuzzy clustering method was able to identify important biological processes such as translation as protein clusters.

4 Conclusions

We have extended a previously proposed fuzzy colored graph clustering method to a Bayesian setting, thus being able to determine adequate cluster sizes by appropriate prior choice. The resulting NMF-type update rules allow efficient search for the maximum-a-posteriori solution. In the future, we will study realistic use-cases in more detail and if necessary choose more informative priors for better clusterings.

Acknowledgements. This work was partially supported by the Helmholtz Alliance on Systems Biology (project 'CoReNe') and the European Research Council (starting grant 'LatentCauses'). The author thanks Nikola Müller for support in the interpretation of the clusters.

References

1. Bishop, C.: Bayesian pca. In: Proc. NIPS 1999 (1999)
2. Blöchl, F., Hartsperger, M., Stümpflen, V., Theis, F.: Uncovering the structure of heterogeneus biological data: fuzzy graph partitioning in the k-partite setting. In: Proc. GCB 2010 (2010)
3. Hartsperger, M., Blöchl, F., Stümpflen, V., Theis, F.: Structuring heterogeneous biological information using fuzzy clustering of k-partite graphs. BMC Bioinformatics 11(522) (2010)
4. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 40, 788–791 (1999)
5. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. In: Proc. NIPS 2000, vol. 13, pp. 556–562. MIT Press (2001)
6. Long, B., Wu, X., Zhang, Z., Yu, P.: Unsupervised learning on k-partite graphs. In: Proc. SIGKDD 2006, pp. 317–326 (2006)
7. MacKay, D.: Probable networks and plausible predictions – a review of practical bayesian models for supervised neural networks. Network: Computation in Neural Systems 6(3), 469–505 (1995)
8. Neher, R., Mitkovski, M., Kirchhoff, F., Neher, E., Theis, F., Zeug, A.: Blind source separation techniques for the decomposition of multiply labeled fluorescence images. Biophysical Journal 96(9), 3791–3800 (2009)
9. Psorakis, I., Roberts, S., Sheldon, B.: Efficient bayesian community detection using non-negative matrix factorisation (2010) (submitted)
10. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stumpflen, V., Mewes, H.W.: Corum: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 36(Database issue), D646–D650 (2008)
11. Tan, V., Févotte, C.: Automatic relevance determination in nonnegative matrix factorization. In: SPARS 2009 - Signal Processing with Adaptive Sparse Structured Representations, pp. 1–19 (2009)

Author Index