

**Yannick Deville
Sharon Gannot
Russell Mason
Mark D. Plumbley
Dominic Ward (Eds.)**

LNCS 10891

Latent Variable Analysis and Signal Separation

**14th International Conference, LVA/ICA 2018
Guildford, UK, July 2–5, 2018
Proceedings**

 **Springer**

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, Lancaster, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Zurich, Switzerland

John C. Mitchell

Stanford University, Stanford, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

C. Pandu Rangan

Indian Institute of Technology Madras, Chennai, India

Bernhard Steffen

TU Dortmund University, Dortmund, Germany

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany


More information about this series at <http://www.springer.com/series/7407>

Yannick Deville · Sharon Gannot
Russell Mason · Mark D. Plumbley
Dominic Ward (Eds.)


Latent Variable Analysis and Signal Separation


14th International Conference, LVA/ICA 2018
Guildford, UK, July 2–5, 2018
Proceedings


Editors

Yannick Deville 
Paul Sabatier University
Toulouse
France

Sharon Gannot 
Bar-Ilan University
Ramat Gan
Israel

Russell Mason 
University of Surrey
Guildford
UK

Mark D. Plumbley 
University of Surrey
Guildford
UK

Dominic Ward 
University of Surrey
Guildford
UK

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-93763-2 ISBN 978-3-319-93764-9 (eBook)
<https://doi.org/10.1007/978-3-319-93764-9>

Library of Congress Control Number: 2018946632

LNCS Sublibrary: SL1 – Theoretical Computer Science and General Issues

© Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume gathers the full articles presented at the 14th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA 2018, which was held at the University of Surrey, Guildford, UK, during July 2–5, 2018. The conference was organized by the Centre for Vision, Speech and Signal Processing (CVSSP) and by the Institute of Sound Recording (IoSR).

Since its inception in 1999, under the name “Independent Component Analysis and Blind Source Separation,” ICA, the series of LVA/ICA conferences (held approximately every 18 months) has attracted hundreds of researchers and practitioners. The conference has continuously broadened its horizons and scope of applications. The LVA/ICA research topics encompass a wide range of general mixtures of latent variable models but also theories and tools drawn from a great variety of disciplines such as signal processing, applied statistics, machine learning, linear and multilinear algebra, numerical analysis, optimization, etc. These research areas are of interest in numerous application fields ranging from audio, image, telecommunications, biochemistry, and quantum physics to biomedical engineering or observation sciences, to cite a few. Thus, this conference offers very exciting interdisciplinary interactions between researchers in various fields and domains. It also constitutes a multidisciplinary discussion forum for scientists and engineers where they can gain access to a broad understanding of the state of the research in the field, keep up to date with active research areas, discover or address the main theoretical challenges, and also face real-world problems and share experiences.

This volume of Springer’s *Lecture Notes in Computer Science* (LNCS) continues the tradition, which began in ICA 2004 (held in Granada, Spain), of publishing the conference proceedings in this form. We thank the editorial board of LNCS for their ongoing commitment and confidence in our conference.

For this 14th issue of the LVA/ICA international conference, 61 full papers were submitted to both regular and special sessions. Each submission of a regular full paper was peer reviewed by at least three members of our Technical Program Committee (TPC) or by competent additional reviewers assigned by the TPC members. From these 61 submitted full papers, 52 were accepted, 28 as oral papers and 24 as poster presentations. The conference program included two special sessions: “Structured Tensor Decompositions and Applications,” proposed by Laurent Albera (Université de Rennes, France), Taylan Cemgil (Bogazici University, Turkey), and Umut Şimşekli (Télécom ParisTech, France); and “Advances in Phase Retrieval and Applications,” proposed by Antoine Deleforge (Inria, Nancy, Grand-Est, France) and Angélique Dremeau (ENSTA Bretagne, France). Regular topics included: ICA methods, matrix and tensor factorizations, nonlinear mixtures, audio data and methods, deep learning and data-driven methods, sparsity-related methods, biomedical data methods, and applications of LVA and ICA.

The Organizing Committee was pleased to invite three leading experts for keynote addresses:

- Tuomas Virtanen (Tampere University of Technology, Finland)
- Orly Alter (University of Utah, USA)
- Danilo Mandic (Imperial College London, UK) Joint work with Andrzej Cichocki (Skolkovo Institute of Science and Technology, Skoltech, Moscow, Russia)

The Organizing Committee also decided to precede the conference by a one-day Summer School including plenary lectures given by:

- Evrim Acar (Simula Research Laboratory, Oslo, Norway)
- Richard Turner (University of Cambridge, UK)
- Russell Mason, Ryan Chungun Kim, Dominic Ward (University of Surrey, Guildford, UK)

The LVA/ICA conference was followed by a special one-day workshop organized on “Audio Applications” with the support of the UK Engineering and Physical Sciences Research Council (EPSRC) through the projects “Musical Audio Repurposing Using Source Separation” (EP/L027119/2), S3A “Future Spatial Audio in the Home” (EP/L000539/1), and “Making Sense of Sounds” (EP/N014111/1), and the “Audio Commons Initiative” funded by the European Commission Horizon 2020 grant 688382. The conference also provided a forum for the seventh community-based Signal Separation Evaluation Campaign (SiSEC 2018), organized by Antoine Liutkus (Inria, Montpellier, Languedoc-Roussillon, France). SiSEC 2018 successfully continued the series of evaluation campaigns initiated at ICA 2007, in London.

The success of the LVA/ICA 2018 conference was the result of the hard work of many people whom we wish to warmly thank here. We wish to thank all the authors, keynote speakers, and tutorial lecturers, as well as all the members of the TPC, without whom this high-quality edition of LVA/ICA 2018 would not exist. We also wish to express our gratitude to the members of the International LVA/ICA Steering Committee for their support to the conference, to the SiSEC 2018 organizers, and finally to the local Organizing Committee and Events Team, for their hard work behind the scenes to ensure that the conference ran smoothly and seamlessly.

May 2018

Yannick Deville
Sharon Gannot
Russell Mason
Mark D. Plumbley
Dominic Ward

Organization

General Chairs

Mark D. Plumbley University of Surrey, UK
Russell Mason University of Surrey, UK

Program Chairs

Sharon Gannot Bar-Ilan University, Israel
Yannick Deville University of Toulouse, France

Area Chairs

Laurent Albera University of Rennes, France
Massoud Babaie-Zadeh Sharif University of Technology, Iran
Nancy Bertin University of Rennes, Inria, CNRS, IRISA, France
Jose Bioucas-Dias University of Lisbon, Portugal
Ali Taylan Cemgil Bogazici University, Turkey
Antoine Deleforge Inria, Rennes, France
Yannick Deville University of Toulouse, France
Sharon Gannot Bar-Ilan University, Israel
Dorien Herremans Singapore University of Technology and Design,
Singapore
Antoine Liutkus Inria and LIRMM, University of Montpellier, France
Mark D. Plumbley University of Surrey, UK
Petr Tichavský The Czech Academy of Sciences, Czech Republic

Special Session Organizers

Laurent Albera University of Rennes, France
Ali Taylan Cemgil Bogazici University, Turkey
Antoine Deleforge Inria, Rennes, France
Angélique Drémeau ENSTA Bretagne, France
Antoine Liutkus Inria and LIRMM, University of Montpellier, France
Umut Şimşekli Télécom ParisTech, France

SiSEC Chair

Antoine Liutkus Inria and LIRMM, University of Montpellier, France

Local Organization

Helen Cooper	University of Surrey, UK
Dominic Ward	University of Surrey, UK
Hagen Wierstorf	University of Surrey, UK
Emad Grais	University of Surrey, UK
Philip Coleman	University of Surrey, UK

International Liaison

Nobutaka Ono	Tokyo Metropolitan University, Japan
Leonardo Duarte	University of Campinas, Brazil
Moussa Karoui	Agence Spatiale Algérienne, Algeria

International Steering Committee

Tülay Adali	University of Maryland Baltimore County, USA
Andrzej Cichocki	Riken Brain Science Institute, Japan
Lieven De Lathauwer	K.U. Leuven, Belgium
Rémi Gribonval	Irisa, Rennes, France
Christian Jutten	Grenoble-Alpes University, France
Shoji Makino	University of Tsukuba, Japan
Nobutaka Ono	Tokyo Metropolitan University, Japan
Mark D. Plumbley	University of Surrey, UK
Paris Smaragdis	University of Illinois at Urbana-Champaign, USA
Petr Tichavský	The Czech Academy of Sciences, Czech Republic
Emmanuel Vincent	Inria, France
Arie Yeredor	Tel Aviv University, Israel

Program Committee

Alessandro Perelli	University of Edinburgh, UK
Ken O'Hanlon	Queen Mary University of London, UK
Radu Balan	University of Maryland, USA
Nancy Bertin	University of Rennes, Inria, CNRS, IRISA, France
Esa Ollila	Aalto University, Finland
Lisandro Lovisolo	UERJ, Brazil
Matthieu Kowalski	University of Paris-Sud, France
Salman Asif	University of California, Riverside, California
Gilles Roussel	ULCO, France
Karin Schnass	University of Innsbruck, Austria
Hendrik Kayser	University of Oldenburg, Germany
Shuyang Ling	New York University, USA
Thomas Martinetz	University of Lübeck, Germany
Vincent Vigneron	Université d'Evry, Université Paris-Saclay, France
Hao Shen	fortiss GmbH, Germany

Rodrigo Cabral Farias	I3S, UCA, CNRS, France
Ivan Dokmanić	University of Illinois at Urbana-Champaign, USA
Anh-Huy Phan	Brain Science Institute, RIKEN, Japan
Nicolas Gillis	University of Mons, Belgium
Stephen Becker	University of Colorado Boulder, USA
Andreas Tillmann	RWTH Aachen, Germany
Stephane Chretien	National Physical Laboratory, UK
Philippe Loubaton	University of Paris-Est Marne-la-Vallée, France
George Karystinos	Technical University of Crete, Greece
Eleftherios Kofidis	University of Piraeus, Greece
Stefan Kunis	Universität Osnabrück, Germany
Lieven De Lathauwer	KU Leuven, Belgium
Christian Rohlfing	RWTH Aachen University, Germany
Sergio Cruces	University of Seville, Spain
Szymon Drgas	Poznan University of Technology, Poland
Roland Badeau	Télécom ParisTech, France
Chun-Guang Li	Beijing University of Posts and Telecommunications, China
Noboru Murata	Waseda University, Japan
Nobutaka Ono	Tokyo Metropolitan University, Japan
Jean-Francois Cardoso	CNRS, France
Marc Castella	Telecom SudParis, France
Thomas Hueber	CNRS/GIPSA-lab, France
Antoine Deleforge	Inria, Rennes, France
Cecile Chenot	University of Edinburgh, UK
Sebastian Miron	Université de Lorraine, France
Matthieu Puigt	LISIC, ULCO, France
Francois Malgouyres	Université Paul Sabatier, France
Abdeldjalil Aissa El Bey	IMT Atlantique, France
Markus Haltmeier	University of Innsbruck, Austria
Flavio Teixeira	University of Innsbruck, Austria
Jiri Malek	Technical University of Liberec, Czech Republic
Yannick Deville	University of Toulouse, France
Vincent Duval	Inria, France
Tim Conrad	Freie Universität Berlin, Germany
Saeid Sanei	Nottingham Trent University, UK
Bertrand Rivet	GIPSA-Lab, Grenoble-Alpes University, France
Björn Schuller	Imperial College London, UK
Konstantin Usevich	CNRS and Université de Lorraine, France
Ricardo Suyama	UFABC, Brazil
Sharon Gannot	Bar-Ilan University, Israel
Thomas Blumensath	University of Southampton, UK
Ahmad Nimr	TU-Dresden Vodafone Chair Mobile Communications Systems, Germany
Jörn Anemüller	University of Oldenburg, Germany
Yoshitatsu Matsuda	University of Tokyo, Japan

Leonardo Tomazeli Duarte	University of Campinas (UNICAMP), Brazil
Ante Jukić	University of Oldenburg, Germany
Paris Smaragdis	University of Illinois at Urbana-Champaign, USA
Ankit Parekh	Icahn School of Medicine at Mount Sinai, USA
Jérémy Cohen	FNRS, UMONS, Belgium
Yong Xu	University of Surrey, UK
Herzet Cédric	Inria, Rennes, France
Martin Kleinstauber	TU Munich, Germany
Jitong Chen	Baidu Silicon Valley AI Lab
Waheed Bajwa	Rutgers University, USA
Zbynek Koldovsky	Technical University of Liberec, Czech Republic
Pavel Rajmic	Brno University of Technology, Czech Republic
Olivier Michel	GIPSA-Lab, France
Petr Tichavský	The Czech Academy of Sciences, Czech Republic
Estefania Cano Cerón	Fraunhofer IDMT, Germany
Haardt Martin	Ilmenau University of Technology, Germany
Chandra Sekhar Seelamantula	Indian Institute of Science, Bangalore, India
João Romano	UNICAMP, Brazil
Alexey Ozerov	Technicolor, France
Keith Dillon	Formulens, LLC, USA
Philippe Dreesen	Vrije Universiteit Brussel, Belgium
Mark D. Plumbley	University of Surrey, UK
Ngoc Duong	Technicolor, France
Hicham Ghennioui	Sidi Mohamed Ben Abdellah University, Morocco
Timo Gerkmann	Universität Hamburg, Germany
Guillaume Tochon	Graduate School of Computer Science and Advanced Technologies, France
Hossein Rabbani	Medical Image and Signal Processing Research Center, Iran
Alexandre Gramfort	Inria, France
Mariya Ishteva	Vrije Universiteit Brussel, Belgium
Bjoern Menze	Technical University of Munich, Germany
Yannis Kopsinis	University of Athens, Greece
Pierre Comon	CNRS, Univ Grenoble Alpes, France
Shinji Watanabe	Johns Hopkins University, USA
David Westwick	University of Calgary, Canada
Laurent Jacques	ISPGROUP, ICTEAM/ELEN, UCLouvain, Belgium
Dana Lahat	Gipsa-Lab, France
Bruno Torresani	Aix-Marseille Université, France
Aline Roumy	Inria, Rennes, France
Russell Mason	University of Surrey, UK
Ali Taylan Cemgil	Bogazici University, Turkey
Giacomo Boracchi	Politecnico di Milano, Italy
Lucas Drumetz	IMT Atlantique, France

Amin Jalali	University of Wisconsin-Madison, USA
Gilles Puy	Technicolor, France
Philip Coleman	University of Surrey, UK
Samaneh Kouchaki	University of Oxford, UK
Ronen Talmon	Technion, Israel Institute of Technology, Israel
Christian Jutten	Grenoble-Alpes University, France
Lior Weizman	Technion, Israel Institute of Technology, Israel
Wenwu Wang	University of Surrey, UK
Florent Sureau	CEA Saclay, France
Jonathon Chambers	University of Leicester, UK
Minje Kim	Indiana University, USA
Denis Gustavo Fantinato	Federal University of ABC, Brazil
Louis Chevallier	Technicolor, France
Niko Lietzén	Aalto University School of Science, Finland
Mário Figueiredo	Instituto Superior Técnico, Portugal
David Brie	Université de Lorraine, France
Jare Tanner	University of Oxford, UK
Laurent Albera	University of Rennes, France
Jerome Bobin	CEA Saclay, France
Antoine Liutkus	Inria and LIRMM, University of Montpellier, France
Ondřej Tichý	Institute of Information Theory and Automation, Czech Republic
Gilles Delmaire	LISIC ULCO, France
Zafar Rafii	Gracenote, USA
Jonathan Le Roux	Mitsubishi Electric Research Labs, USA
Sandrine Anthoine	CNRS, France
Boris Mailhe	Siemens Healthineers, USA
Peter Balazs	Austrian Academy of Sciences, Austria
Stefania Petra	Heidelberg University, Germany
Thomas Coavess	KU Leuven, Belgium
Charles Cavalcante	Federal University of Ceará, Brazil
Pauliina Ilmonen	Aalto University School of Science, Finland
Weiss Pierre	CNRS, France
Romis Attux	University of Campinas (UNICAMP), Brazil
Mhammed Lahbabi	FST, Morocco
Vaclav Smidl	Institute of Information Theory and Automation, Czech Republic
Mehrdad Yaghoobi	University of Edinburgh, UK
Gerard Roma	University of Huddersfield, UK
Yoann Altmann	Heriot-Watt University, UK
Angélique Drémeau	ENSTA Bretagne, France
Kenji Nose Filho	Federal University of ABC, Brazil
Dirk Lorenz	Braunschweig University of Technology, Germany
Petros Boufounos	Mitsubishi Electric Research Labs, USA
Stanislaw Gorlow	Dolby Sweden, Sweden
Eric Tramel	Owkin, Inc., France

Jocelyn Chanussot	GIPSA-Lab, Grenoble-Alpes University, France
Souleyman Sahnoun	Situ8ed SA, France
Nadège Thirion-Moreau	LSIS, UMR CNRS 7296, France
Tülay Adalı	University of Maryland Baltimore County, USA
Marius Miron	Universitat Pompeu Fabra, Spain
Kyong Jin	EPFL, Switzerland
Emmanuel Vincent	Inria, France
Valentin Emiya	Aix-Marseille University, France
Ivica Kopriva	Rudjer Boskovich Institute, Croatia

Contents

Structured Tensor Decompositions and Applications

Robust Multilinear Decomposition of Low Rank Tensors	3
<i>Xu Han, Laurent Albera, Amar Kachenoura, Huazhong Shu, and Lotfi Senhadji</i>	
Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition	13
<i>Mathieu Fontaine, Fabian-Robert Stöter, Antoine Liutkus, Umut Şimşekli, Romain Serizel, and Roland Badeau</i>	
Sum Conditioned Poisson Factorization	24
<i>Gökhan Çapan, Semih Akbayrak, Taha Yusuf Ceritli, and Ali Taylan Cemgil</i>	
Curve Registered Coupled Low Rank Factorization	36
<i>Jeremy Emile Cohen, Rodrigo Cabral Farias, and Bertrand Rivet</i>	
Source Analysis and Selection Using Block Term Decomposition in Atrial Fibrillation	46
<i>Pedro Marinho R. de Oliveira and Vicente Zarzoso</i>	
Some Issues in Computing the CP Decomposition of NonNegative Tensors . . .	57
<i>Mohamad Jouni, Mauro Dalla Mura, and Pierre Comon</i>	

Matrix and Tensor Factorizations

A Grassmannian Minimum Enclosing Ball Approach for Common Subspace Extraction	69
<i>Emilie Renard, Kyle A. Gallivan, and P.-A. Absil</i>	
Decoupling Multivariate Functions Using Second-Order Information and Tensors	79
<i>Philippe Dreesen, Jeroen De Geeter, and Mariya Ishteva</i>	
Nonnegative PARAFAC2: A Flexible Coupling Approach	89
<i>Jeremy E. Cohen and Rasmus Bro</i>	
Applications of Polynomial Common Factor Computation in Signal Processing	99
<i>Ivan Markovsky, Antonio Fazzi, and Nicola Guglielmi</i>	

Joint Nonnegative Matrix Factorization for Underdetermined Blind Source Separation in Nonlinear Mixtures	107
<i>Ivica Kopriva</i>	
Image Completion with Nonnegative Matrix Factorization Under Separability Assumption	116
<i>Tomasz Sadowski and Rafał Zdunek</i>	
Feature Selection in Weakly Coherent Matrices	127
<i>Stéphane Chrétien and Olivier Ho</i>	
Variable Projection Applied to Block Term Decomposition of Higher-Order Tensors	139
<i>Guillaume Olikier, P.-A. Absil, and Lieven De Lathauwer</i>	
ICA Methods	
Accelerating Likelihood Optimization for ICA on Real Signals	151
<i>Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort</i>	
Orthogonally-Constrained Extraction of Independent Non-Gaussian Component from Non-Gaussian Background Without ICA	161
<i>Zbyněk Koldovský, Petr Tichavský, and Nobutaka Ono</i>	
A New Link Between Joint Blind Source Separation Using Second Order Statistics and the Canonical Polyadic Decomposition	171
<i>Dana Lahat and Christian Jutten</i>	
Nonlinear Mixtures	
A Blind Source Separation Method Based on Output Nonlinear Correlation for Bilinear Mixtures.	183
<i>Andréa Guerrero, Yannick Deville, and Shahram Hosseini</i>	
Using Taylor Series Expansions and Second-Order Statistics for Blind Source Separation in Post-Nonlinear Mixtures.	193
<i>Denis G. Fantinato, Leonardo T. Duarte, Yannick Deville, Christian Jutten, Romis Attux, and Aline Neves</i>	
New Classes of Blind Quantum Source Separation and Process Tomography Methods Based on Spin Component Measurements Along Two Directions	204
<i>Yannick Deville and Alain Deville</i>	

Audio Data and Methods

Blind Signal Separation by Synchronized Joint Diagonalization	217
<i>Hiroshi Sawada</i>	
Exploiting Structures of Temporal Causality for Robust Speaker Localization in Reverberant Environments	228
<i>Christopher Schymura, Peng Guo, Yanir Maymon, Boaz Rafaely, and Dorothea Kolossa</i>	
Relative Transfer Function Estimation from Speech Keywords	238
<i>Ryan M. Corey and Andrew C. Singer</i>	
On the Number of Signals in Multivariate Time Series	248
<i>Markus Matilainen, Klaus Nordhausen, and Joni Virta</i>	
A Generative Model for Natural Sounds Based on Latent Force Modelling.	259
<i>William J. Wilkinson, Joshua D. Reiss, and Dan Stowell</i>	
Independent Vector Analysis Exploiting Pre-learned Banks of Relative Transfer Functions for Assumed Target's Positions	270
<i>Jaroslav Čmejla, Tomáš Kounovský, Jiří Málek, and Zbyněk Koldovský</i>	
Does k Matter? k-NN Hubness Analysis for Kernel Additive Modelling Vocal Separation.	280
<i>Delia Fano Yela, Dan Stowell, and Mark Sandler</i>	

Signal Separation Evaluation Campaign (SiSEC 2018)

The 2018 Signal Separation Evaluation Campaign.	293
<i>Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito</i>	
Improving Single-Network Single-Channel Separation of Musical Audio with Convolutional Layers	306
<i>Gerard Roma, Owen Green, and Pierre Alexandre Tremblay</i>	

Deep Learning and Data-driven Methods

Training Strategies for Deep Latent Models and Applications to Speech Presence Probability Estimation	319
<i>Shlomo E. Chazan, Sharon Gannot, and Jacob Goldberger</i>	
Jointly Detecting and Separating Singing Voice: A Multi-Task Approach.	329
<i>Daniel Stoller, Sebastian Ewert, and Simon Dixon</i>	

Multi-Resolution Fully Convolutional Neural Networks for Monaural Audio Source Separation	340
<i>Emad M. Grais, Hagen Wierstorf, Dominic Ward, and Mark D. Plumbley</i>	
Long-Term SNR Estimation Using Noise Residuals and a Two-Stage Deep-Learning Framework	351
<i>Xuan Dong and Donald S. Williamson</i>	
Improving Reverberant Speech Separation with Binaural Cues Using Temporal Context and Convolutional Neural Networks.	361
<i>Alfredo Zermeni, Qiuqiang Kong, Yong Xu, Mark D. Plumbley, and Wenwu Wang</i>	
Generating Talking Face Landmarks from Speech	372
<i>Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan</i>	
Advances in Phase Retrieval and Applications	
An Approximate Message Passing Approach for DOA Estimation in Phase Noisy Environments	385
<i>Guillaume Beaumont, Ronan Fablet, and Angélique Drémeau</i>	
An Expectation-Maximization Approach to Tuning Generalized Vector Approximate Message Passing	395
<i>Christopher A. Metzler, Philip Schniter, and Richard G. Baraniuk</i>	
A Study on the Benefits of Phase-Aware Speech Enhancement in Challenging Noise Scenarios.	407
<i>Martin Krawczyk-Becker and Timo Gerkmann</i>	
Phase Reconstruction for Time-Frequency Inpainting.	417
<i>A. Marina Krémé, Valentin Emiya, and Caroline Chaux</i>	
Sparsity-Related Methods	
Revisiting Synthesis Model in Sparse Audio Declipper	429
<i>Pavel Záváška, Pavel Rajmic, Zdeněk Průša, and Vítězslav Veselý</i>	
Consistent Dictionary Learning for Signal Declipping	446
<i>Lucas Rencker, Francis Bach, Wenwu Wang, and Mark D. Plumbley</i>	
Learning Fast Dictionaries for Sparse Representations Using Low-Rank Tensor Decompositions	456
<i>Cássio F. Dantas, Jérémy E. Cohen, and Rémi Gribonval</i>	

Truncated Variational Sampling for ‘Black Box’ Optimization of Generative Models	467
<i>Jörg Lücke, Zhenwen Dai, and Georgios Exarchakis</i>	
Using Hankel Structured Low-Rank Approximation for Sparse Signal Recovery	479
<i>Ivan Markovsky and Pier Luigi Dragotti</i>	
Probabilistic Sparse Non-negative Matrix Factorization	488
<i>Jesper Løve Hinrich and Morten Mørup</i>	
Biomedical Data and Methods	
Application of Independent Component Analysis to Tumor Transcriptomes Reveals Specific and Reproducible Immune-Related Signals.	501
<i>Urszula Czerwinska, Laura Cantini, Ulyzbek Kairov, Emmanuel Barillot, and Andrei Zinovyev</i>	
Probit Latent Variables Estimation for a Gaussian Process Classifier: Application to the Detection of High-Voltage Spindles	514
<i>Rémi Souriau, Vincent Vigneron, Jean Lerbet, and Hsin Chen</i>	
Spatial Filtering of EEG Signals to Identify Periodic Brain Activity Patterns	524
<i>Dounia Mulders, Cyril de Bodt, Nicolas Lejeune, André Mouraux, and Michel Verleysen</i>	
Static and Dynamic Modeling of Absence Epileptic Seizures Using Depth Recordings	534
<i>Saeed Akhavan, Ronald Phlypo, Hamid Soltanian-Zadeh, Mahmoud Kamarei, and Christian Jutten</i>	
Applications of LVA/ICA	
Multichannel Audio Source Separation Exploiting NMF-Based Generic Source Spectral Model in Gaussian Modeling Framework	547
<i>Thanh Thi Hien Duong, Ngoc Q. K. Duong, Cong-Phuong Nguyen, and Quoc-Cuong Nguyen</i>	
Perceptual Evaluation of Blind Source Separation in Object-Based Audio Production	558
<i>Philip Coleman, Qingju Liu, Jon Francombe, and Philip J. B. Jackson</i>	
Muticriteria Decision Making Based on Independent Component Analysis: A Preliminary Investigation Considering the TOPSIS Approach	568
<i>Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, and João Marcos Travassos Romano</i>	
Author Index	579

Structured Tensor Decompositions and Applications



Robust Multilinear Decomposition of Low Rank Tensors

Xu Han^{1,2,4}, Laurent Albera^{1,2,4}(✉), Amar Kachenoura^{1,2,4}, Huazhong Shu^{3,4},
and Lotfi Senhadji^{1,2,4}

¹ LTSI, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France

² INSERM, U 1099, 35402 Rennes Cedex, France

³ LIST, Southeast University, 2 Sipailou, Nanjing 210096, China

⁴ Centre de Recherche en Information Biomédicale Sino-Français, Rennes, France

hanxu.list@gmail.com, shu.list@seu.edu.cn,

{laurent.albera, amar.kachenoura, lotfi.senhadji}@univ-rennes1.fr

Abstract. Although several methods are available to compute the multilinear rank of multi-way arrays, they are not sufficiently robust with respect to the noise. In this paper, we propose a novel Multilinear Tensor Decomposition (MTD) method, namely R-MTD (Robust MTD), which is able to identify the multilinear rank even in the presence of noise. The latter is based on sparsity and group sparsity of the core tensor imposed by means of the l_1 norm and the mixed-norm, respectively. After several iterations of R-MTD, the underlying core tensor is sufficiently well estimated, which allows for an efficient use of the minimum description length approach and an accurate estimation of the multilinear rank. Computer results show the good behavior of R-MTD.

Keywords: Multilinear tensor rank · MTD · Low rank
Sparse · Mixed-norm · l_1 norm

1 Introduction

The rank estimation problem has been considered for several decades. The noisy matrix case was firstly considered. Methods based on the computation of singular values and the use of either Akaike's information criterion [1], the Bayesian Information Criterion (BIC) [2] or the Minimum Description Length (MDL) principle [3] were proposed. The singular values of a low-rank noiseless matrix can be sorted in decreasing order (see the red diamonds in Fig. 1). Then the estimated rank R^{est} can be computed by searching the breaking point of the singular value curve, which minimizes the MDL criterion [3]:

$$R^{est} = \arg \min_r - 2 \log \left\{ \frac{\prod_{i=r+1}^I \lambda_i^{1/(I-r)}}{\frac{1}{I-r} \sum_{i=r+1}^I \lambda_i} \right\}^{J(I-r)} + r(2I - r) \log(J) \quad (1)$$

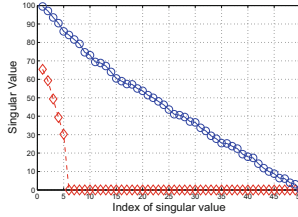


Fig. 1. Singular values of a matrix (red diamonds) and those of its noisy version for an SNR value of 0 dB (blue circles). (Color figure online)

where λ_i is the i -th highest singular value and where $(I \times J)$ is the size of the considered matrix. Unfortunately, when the matrix is strongly noisy the MDL approach has difficulty finding the breaking point (see the blue circles in Fig. 1).

The rank estimation problem was also considered for arrays with more than two entries, commonly named tensors. Contrarily to the matrix case, there are several definitions of tensor rank. In the following, we will consider the definition related to the Multilinear Tensor Decomposition (MTD) model [4]:

$$\mathcal{F} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = \sum_{i=1}^{R_1} \sum_{j=1}^{R_2} \sum_{k=1}^{R_3} \mathcal{G}_{i,j,k} \mathbf{A}_{:,i} \circ \mathbf{B}_{:,j} \circ \mathbf{C}_{:,k}, \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{I_1 \times R_1}$, $\mathbf{B} \in \mathbb{R}^{I_2 \times R_2}$ and $\mathbf{C} \in \mathbb{R}^{I_3 \times R_3}$ are the loading matrices and where $\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ is the core tensor of a third order tensor \mathcal{F} . Note that $\mathbf{X}_{:,j}$ denotes the j -th column of \mathbf{X} and \circ denotes the vector outer product operator. The multilinear rank of \mathcal{F} is given by the minimal values of R_1 , R_2 and R_3 for which the equality (2) holds. The classical Higher-Order Singular Value Decomposition (HOSVD) [5] is a special MTD, which is a direct extension of the matrix SVD to tensors. While the multilinear rank can be derived from the rank of the unfolding matrices of the considered tensor, the latter matrix ranks are difficult to estimate as explained previously. Thus authors proposed to minimize the nuclear norm of each unfolding matrix in order to find the minimal MTD [8] and consequently the multilinear rank. More recently, the SCORE algorithm [7], based on the modified eigenvalues of the considered noisy tensor, was proposed too. Unfortunately, such methods are not sufficiently robust with respect to the presence of noise.

In this paper, we propose a novel MTD method, namely R-MTD (Robust MTD), which is able to identify the multilinear rank even in the presence of noise. The latter is based on sparsity and group sparsity of the core tensor imposed by means of the l_1 norm and the mixed-norm, respectively. Note that the mixed-norm was proved to be a convex envelope of rank and used to provide robust canonical polyadic and block term decomposition algorithms [9, 10]. After several iterations of R-MTD, the underlying core tensor is sufficiently well estimated, which allows for an efficient use of the MDL approach and an accurate estimation of the multilinear rank. Computer results show the good behavior of R-MTD.

2 Notions and Preliminaries

A scalar is denoted by an italic letter, e.g., x and I . A vector is denoted by a bold lowercase letter, e.g., $\mathbf{x} \in \mathbb{R}^I$ and a matrix is represented by a bold capital letter, e.g., $\mathbf{X} \in \mathbb{R}^{I \times J}$ and specially, \mathbf{I} is the identity matrix. The vectorization of \mathbf{X} is denoted by $\text{vec}(\mathbf{X}) \in \mathbb{R}^{IJ \times 1}$. \mathbf{X}^\dagger is the pseudo inverse of matrix \mathbf{X} . The nuclear norm of a matrix $\mathbf{X} \in \mathbb{R}^{I \times J}$ with rank r is equal to the sum of its singular values, i.e., $\|\mathbf{X}\|_* = \sum_{i=1}^r \sigma_i$, and $\|\mathbf{X}\|$ is the largest singular value. The symbol $\text{rank}(\cdot)$ denotes the rank operator. The l_0 norm, l_1 norm and Frobenius norm of $\mathbf{X} \in \mathbb{R}^{I \times J}$ are defined by $\|\mathbf{X}\|_0 = \sum_{i=1}^I \sum_{j=1}^J \mathbf{X}_{i,j} / \mathbf{X}_{i,j}$ with $\mathbf{X}_{i,j} \neq 0$, $\|\mathbf{X}\|_1 = \sum_{i=1}^I \sum_{j=1}^J |\mathbf{X}_{i,j}|$ and $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^I \sum_{j=1}^J \mathbf{X}_{i,j}^2}$, respectively. The mixed-norm pair of $\mathbf{X} \in \mathbb{R}^{I \times J}$ is given by $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^I \sqrt{\sum_{j=1}^J \mathbf{X}_{i,j}^2} = \text{Tr}[\mathbf{X}^\top \Phi \mathbf{X}]$ and $\|\mathbf{X}\|_{1,2} = \sum_{j=1}^J \sqrt{\sum_{i=1}^I \mathbf{X}_{i,j}^2} = \text{Tr}[\mathbf{X} \Psi \mathbf{X}^\top]$, where $\text{Tr}[\cdot]$ is the trace operator, where Φ is a diagonal matrix with $\Phi_{i,i} = 1/\sqrt{\sum_{j=1}^J \mathbf{X}_{i,j}^2}$ denoting the (i, i) -th component of Φ and where Ψ is a diagonal matrix with $\Psi_{j,j} = 1/\sqrt{\sum_{i=1}^I \mathbf{X}_{i,j}^2}$ standing for the (j, j) -th component of Ψ . Note that this trace compact definition is convenient for a convex optimization if Φ and Ψ are fixed as weight matrices. A high-order tensor is denoted by a bold calligraphic letter, e.g., $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$. The symbol \otimes denotes the Kronecker product operator. The n -th mode unfolding matrix of the tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is denoted by $\mathbf{X}^{(n)} \in \mathbb{R}^{I_n \times (I_{n+1} I_{n+2} \dots I_N I_1 I_2 \dots I_{n-1})}$. The sub-tensor $\mathcal{X}_{i_n=k}$ is obtained by fixing the n -th index to k . The scalar product of two tensors $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ and $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ is defined by $\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1=1}^{I_1} \dots \sum_{i_N=1}^{I_N} \mathcal{X}_{i_1, \dots, i_N} \mathcal{Y}_{i_1, \dots, i_N}$. The Frobenius norm of tensor \mathcal{X} is defined by $\|\mathcal{X}\|_F = \sqrt{\sum_{i_1, \dots, i_N} \mathcal{X}_{i_1, \dots, i_N}^2}$.

3 The R-MTD Method

3.1 Preliminary Tools

First let's recall the definition of a convex envelope which will be used in the following analysis.

Definition 1 (Lower convex envelope). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real-valued function. The convex envelope of f , represented by \mathcal{C}_f , is the convex pointwise largest function $\mathcal{C}_f : \mathbb{R}^n \rightarrow \mathbb{R}$ which is pointwise less than f . In other words, we have $\mathcal{C}_f = \sup\{g : \mathbb{R}^n \rightarrow \mathbb{R} \mid g \text{ is convex and } g(x) \leq f(x) \text{ for any } x \in \mathbb{R}^n\}$.*

Now let's study the lower convex envelopes of rank. The nuclear norm of \mathbf{X} is one of them in the norm ball, i.e., $\|\mathbf{X}\| \leq 1$, as shown in [12, Theorem 1] and since the low rank constraint involves a NP-hard optimization problem it is better to minimize the nuclear norm. On the other hand, an interesting relationship between the nuclear norm and the mixed-norm was established in [11, Proposition 1] for a thin matrix case, i.e., $\mathbf{X} \in \mathbb{R}^{m \times n}$ with $m > n$. We extended this result by means of the following Theorem 1 and Corollary 1.

Theorem 1. *Given any matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ and its orthonormal subspace decomposition denoted by $\mathbf{X} = \mathbf{D}\boldsymbol{\alpha}$ and $\mathbf{X} = \boldsymbol{\theta}\mathbf{Z}$, where $\mathbf{D} \in \mathbb{R}^{m \times m}$ and $\mathbf{Z} \in \mathbb{R}^{n \times n}$ are orthonormal matrices, with $\boldsymbol{\alpha} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\theta} \in \mathbb{R}^{m \times n}$, the mixed-norms of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are larger than or equal to the nuclear norm of \mathbf{X} , i.e. $\|\boldsymbol{\alpha}\|_{2,1} \geq \|\mathbf{X}\|_*$ and $\|\boldsymbol{\theta}\|_{1,2} \geq \|\mathbf{X}\|_*$.*

Then we can easily derive the following corollary by fixing \mathbf{D} and \mathbf{Z} equal to the identity matrix in Theorem 1.

Corollary 1. *We have $\|\mathbf{X}\|_{2,1} \geq \|\mathbf{X}\|_*$ and $\|\mathbf{X}\|_{1,2} \geq \|\mathbf{X}\|_*$.*

In fact, for a matrix with full column rank or full row rank, the nuclear norm is not the tightest convex envelope of rank and the mixed-norm is more qualified to be a lower convex envelop of rank compared with the nuclear norm based on the Definition 1 and the explanation is given in the following theorem.

Theorem 2. *Given any matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ with linearly independent rows or columns, we have:*

1. *If \mathbf{X} has linearly independent rows, which means that $\text{rank}(\mathbf{X}) = m$ and $m \leq n$, then we get $\text{rank}(\mathbf{X}) \geq \frac{\|\mathbf{X}\|_{2,1}}{\|\mathbf{X}\|} \geq \frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|}$.*
2. *If \mathbf{X} has linearly independent columns, which means that $\text{rank}(\mathbf{X}) = n$ and $m \geq n$, then we get $\text{rank}(\mathbf{X}) \geq \frac{\|\mathbf{X}\|_{1,2}}{\|\mathbf{X}\|} \geq \frac{\|\mathbf{X}\|_*}{\|\mathbf{X}\|}$.*

Consequently, we will use the mixed-norm in the following subsection in order to impose the low rank property. Besides, we will also minimize the l_1 norm of the over-estimated core tensor in order to vanish the residual nonzero elements for which the absolute values are close to zero.

3.2 Towards the R-MTD Algorithm

In order to guarantee a sufficient group sparsity of each unfolding matrix along the different dimensions, and each unfolding matrix consist with linear independent rows, the proposed method is implemented under the following assumption: $R_i \ll \min\{I_i, \prod_{k=1, k \neq i}^3 I_k\}$ and $R_i \leq \prod_{k=1, k \neq i}^3 R_k$, $i = 1, 2, 3$.

R-MTD for Rank Estimation: The mixed-norm $\|\widehat{\mathbf{G}}^{(i)}\|_{2,1}$ is considered as a lower convex envelope of rank in the objective function and the other mixed-norm $\|\widehat{\mathbf{G}}^{(i)}\|_{1,2}$ is also adopted as a convex upper bound of the nuclear norm for a more robust estimation. The objective function is presented below for a third order multilinear tensor as an example but it is not hard to generalize it to higher orders:

$$\begin{aligned}
 & \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}, \widehat{\mathcal{G}}} \sum_{i=1}^3 \lambda_i \left(\|\widehat{\mathbf{G}}^{(i)}\|_{2,1} + \|\widehat{\mathbf{G}}^{(i)}\|_{1,2} \right) + \|\widehat{\mathcal{G}}\|_1 \\
 & \text{s.t. } \mathcal{T} = \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}} \times_2 \widehat{\mathbf{B}} \times_3 \widehat{\mathbf{C}} + \mathcal{N}, \widehat{\mathcal{G}} \in \mathbb{R}^{\widehat{R}_1 \times \widehat{R}_2 \times \widehat{R}_3}, \\
 & \widehat{\mathbf{A}} \in \mathbb{R}^{I_1 \times \widehat{R}_1}, \widehat{\mathbf{B}} \in \mathbb{R}^{I_2 \times \widehat{R}_2}, \widehat{\mathbf{C}} \in \mathbb{R}^{I_3 \times \widehat{R}_3}, \quad (3)
 \end{aligned}$$

where $\widehat{\mathcal{G}}$ is the over-estimated core tensor, where $\widehat{\mathbf{G}}^{(i)}$ are the unfolding matrices along the different dimensions and where $\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}$ are the corresponding over-estimated loading matrices. Here, we should note that \widehat{R}_i should be larger than R_i . The minimization problem (3) is solved by means of the Alternating Directions Method of Multiplier (ADMM) [14]:

$$\begin{aligned} \min_{\widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}, \widehat{\mathcal{G}}} \quad & \sum_{i=1}^3 \lambda_i \left(\|\widehat{\mathbf{G}}^{(i)}\|_{2,1} + \|\widehat{\mathbf{G}}^{(i)}\|_{1,2} \right) + \|\widehat{\mathcal{P}}\|_1 \\ & + \frac{u}{2} \|\mathcal{T} - \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}} \times_2 \widehat{\mathbf{B}} \times_3 \widehat{\mathbf{C}}\|_F^2 \\ \text{s.t.} \quad & \widehat{\mathcal{G}} - \widehat{\mathcal{P}} = \mathbf{0}. \end{aligned} \quad (4)$$

The augmented Lagrangian function \mathcal{L} of the variables $\widehat{\mathcal{G}}, \widehat{\mathbf{A}}, \widehat{\mathbf{B}}, \widehat{\mathbf{C}}, \widehat{\mathcal{P}}, \widehat{\mathcal{Y}}, \beta, \Phi^{(i)}$ and $\Psi^{(i)}$ can be written as follows using the mixed-norm definition given in Sect. 2:

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^3 \lambda_i & \left[\text{Tr}(\widehat{\mathbf{G}}^{(i)\top} \Phi^{(i)} \widehat{\mathbf{G}}^{(i)}) + \text{Tr}(\widehat{\mathbf{G}}^{(i)} \Psi^{(i)} \widehat{\mathbf{G}}^{(i)\top}) \right] \\ & + \|\widehat{\mathcal{P}}\|_1 + \frac{u}{2} \|\mathcal{T} - \widehat{\mathcal{G}} \times_1 \widehat{\mathbf{A}} \times_2 \widehat{\mathbf{B}} \times_3 \widehat{\mathbf{C}}\|_F^2 \\ & + \langle \widehat{\mathcal{G}} - \widehat{\mathcal{P}}, \widehat{\mathcal{Y}} \rangle + \frac{\beta}{2} \|\widehat{\mathcal{G}} - \widehat{\mathcal{P}}\|_F^2, \end{aligned} \quad (5)$$

where the parameters λ_i, u and β are penalty coefficients and where $\widehat{\mathcal{Y}}$ is the Lagrangian multiplier tensor. Now let's derive the update rule of each variable. The mode-1 unfolding matrix of the core tensor $\widehat{\mathcal{G}}$ is computed from the following gradient equation:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \widehat{\mathbf{G}}^{(1)}} = 2\lambda_1 \Phi^{(1)} \widehat{\mathbf{G}}^{(1)} + 2\lambda_1 \widehat{\mathbf{G}}^{(1)} \Psi^{(1)} + \beta \widehat{\mathbf{G}}^{(1)} + u \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}} \widehat{\mathbf{G}}^{(1)} (\widehat{\mathbf{B}} \otimes \widehat{\mathbf{C}})^\top (\widehat{\mathbf{B}} \otimes \widehat{\mathbf{C}}) \\ + \widehat{\mathcal{Y}}^{(1)} - \beta \widehat{\mathcal{P}}^{(1)} - u \widehat{\mathbf{A}}^\top \mathbf{T}^{(1)} (\widehat{\mathbf{B}} \otimes \widehat{\mathbf{C}}) = 0. \end{aligned} \quad (6)$$

The solution of (6) can be calculated by Encapsulating Sum in [15], the $(k+1)$ -th iteration of $\widehat{\mathbf{G}}^{(1)}$ is given by:

$$\begin{aligned} \text{vec}(\widehat{\mathbf{G}}_{k+1}^{(1)}) = \left\{ \mathbf{I} \otimes (2\lambda_1 \Phi_k^{(1)}) + (2\lambda_1 \Psi_k^{(1)} + \beta_k \mathbf{I}) \otimes \mathbf{I} + [(\widehat{\mathbf{B}}_k \otimes \widehat{\mathbf{C}}_k)^\top (\widehat{\mathbf{B}}_k \otimes \widehat{\mathbf{C}}_k) \right. \\ \left. \otimes (u \widehat{\mathbf{A}}_k^\top \widehat{\mathbf{A}}_k) \right\}^{-1} \text{vec} \left[\beta_k \widehat{\mathcal{P}}_k^{(1)} - \widehat{\mathcal{Y}}_k^{(1)} + u \widehat{\mathbf{A}}_k^\top \mathbf{T}^{(1)} (\widehat{\mathbf{B}}_k \otimes \widehat{\mathbf{C}}_k) \right]. \end{aligned} \quad (7)$$

Consecutively, $\widehat{\mathbf{G}}^{(2)}$ and $\widehat{\mathbf{G}}^{(3)}$ at the $(k+1)$ -th iteration can be computed as follows:

$$\begin{aligned} \text{vec}(\widehat{\mathbf{G}}_{k+1}^{(2)}) = \left\{ \mathbf{I} \otimes (2\lambda_2 \Phi_k^{(2)}) + (2\lambda_2 \Psi_k^{(2)} + \beta_k \mathbf{I}) \otimes \mathbf{I} + [(\widehat{\mathbf{C}}_k \otimes \widehat{\mathbf{A}}_k)^\top (\widehat{\mathbf{C}}_k \otimes \widehat{\mathbf{A}}_k) \right. \\ \left. \otimes (u \widehat{\mathbf{B}}_k^\top \widehat{\mathbf{B}}_k) \right\}^{-1} \text{vec} \left[\beta_k \widehat{\mathcal{P}}_k^{(2)} - \widehat{\mathcal{Y}}_k^{(2)} + u \widehat{\mathbf{B}}_k^\top \mathbf{T}^{(2)} (\widehat{\mathbf{C}}_k \otimes \widehat{\mathbf{A}}_k) \right], \end{aligned} \quad (8)$$

$$\text{vec}(\widehat{\mathbf{G}}_{k+1}^{(3)}) = \left\{ \mathbf{I} \otimes (2\lambda_i \boldsymbol{\Phi}_k^{(3)}) + (2\lambda_i \boldsymbol{\Psi}_k^{(3)} + \beta_k \mathbf{I}) \otimes \mathbf{I} + [(\widehat{\mathbf{A}}_k \otimes \widehat{\mathbf{B}}_k)^\top (\widehat{\mathbf{A}}_k \otimes \widehat{\mathbf{B}}_k)] \right. \\ \left. \otimes (u \widehat{\mathbf{C}}_k^\top \widehat{\mathbf{C}}_k) \right\}^{-1} \text{vec} \left[\beta_k \widehat{\mathbf{P}}_k^{(3)} - \widehat{\mathbf{Y}}_k^{(3)} + u \widehat{\mathbf{C}}_k^\top \mathbf{T}^{(3)} (\widehat{\mathbf{A}}_k \otimes \widehat{\mathbf{B}}_k) \right]. \quad (9)$$

Note that, the weighting matrices $\boldsymbol{\Phi}_k^{(1)}$ and $\boldsymbol{\Psi}_k^{(1)}$ are derived from $\widehat{\mathbf{G}}_k^{(1)}$ as the definition in Sect. 2. For the computation of $\boldsymbol{\Phi}_k^{(2)}$ and $\boldsymbol{\Psi}_k^{(2)}$, they are calculated using the 2-th mode unfolding matrix of the updated core tensor $\widehat{\mathbf{G}}_{k+1}^{(1)}$. Identically, $\boldsymbol{\Phi}_k^{(3)}$ and $\boldsymbol{\Psi}_k^{(3)}$ are given using the updated core tensor $\widehat{\mathbf{G}}_{k+1}^{(2)}$. Then, updating factor matrices by using tensor $\widehat{\mathcal{G}}_{k+1}$, i.e., $\widehat{\mathbf{G}}_{k+1}^{(3)}$. The sub-problem function about factor matrix $\widehat{\mathbf{A}}$ in $(k+1)$ -th iteration is given as following:

$$\min_{\widehat{\mathbf{A}}_{k+1}} \left\| \mathbf{T}^{(1)} - \widehat{\mathbf{A}}_{k+1} \widehat{\mathbf{G}}_{k+1}^{(1)} (\widehat{\mathbf{B}}_k \otimes \widehat{\mathbf{C}}_k)^\top \right\|_F^2. \quad (10)$$

The updated $\widehat{\mathbf{A}}_{k+1}$ matrix at the $(k+1)$ -th iteration is given by:

$$\widehat{\mathbf{A}}_{k+1} = \mathbf{T}^{(1)} \left[\widehat{\mathbf{G}}_{k+1}^{(1)} (\widehat{\mathbf{B}}_k \otimes \widehat{\mathbf{C}}_k)^\top \right]^\dagger. \quad (11)$$

Using $\widehat{\mathbf{A}}_{k+1}$, $\widehat{\mathbf{B}}_{k+1}$ at the $(k+1)$ -th iteration can be computed as follows:

$$\widehat{\mathbf{B}}_{k+1} = \mathbf{T}^{(2)} \left[\widehat{\mathbf{G}}_{k+1}^{(2)} (\widehat{\mathbf{C}}_k \otimes \widehat{\mathbf{A}}_{k+1})^\top \right]^\dagger. \quad (12)$$

Similarly, $\widehat{\mathbf{C}}_{k+1}$ at the $(k+1)$ -th iteration is given by:

$$\widehat{\mathbf{C}}_{k+1} = \mathbf{T}^{(3)} \left[\widehat{\mathbf{G}}_{k+1}^{(3)} (\widehat{\mathbf{A}}_{k+1} \otimes \widehat{\mathbf{B}}_{k+1})^\top \right]^\dagger. \quad (13)$$

Regarding variable $\widehat{\mathcal{P}}$, we have to minimize:

$$\min_{\widehat{\mathcal{P}}} \left\| \widehat{\mathcal{P}} \right\|_1 + \left\langle \widehat{\mathcal{G}} - \widehat{\mathcal{P}}, \widehat{\mathcal{Y}} \right\rangle + \frac{\beta}{2} \left\| \widehat{\mathcal{G}} - \widehat{\mathcal{P}} \right\|_F^2. \quad (14)$$

An equivalent problem of (14) is given by:

$$\min_{\widehat{\mathcal{P}}} \left\| \widehat{\mathcal{P}} \right\|_1 + \frac{\beta}{2} \left\| \widehat{\mathcal{G}} - \widehat{\mathcal{P}} + \frac{1}{\beta} \widehat{\mathcal{Y}} \right\|_F^2. \quad (15)$$

The solution of $\widehat{\mathcal{P}}$ in problem (15) has been solved in [13] as follows:

$$\widehat{\mathcal{P}}_{k+1} = \mathbb{S}_{\frac{1}{\beta_k}} \left(\widehat{\mathcal{G}}_{k+1} + \frac{1}{\beta_k} \widehat{\mathcal{Y}}_k \right), \quad (16)$$

where $\mathbb{S}_\tau(x) = \text{sign}(x)(|x| - \tau, 0)$, $x \in \mathbb{R}$ is the soft-thresholding operator and it is used elementwise. The multiplier tensor $\widehat{\mathcal{Y}}$ in $(k+1)$ -th iteration is calculated based on the following update rule:

$$\widehat{\mathcal{Y}}_{k+1} = \widehat{\mathcal{Y}}_k + \beta_k (\widehat{\mathcal{G}}_{k+1} - \widehat{\mathcal{P}}_{k+1}). \quad (17)$$

The parameter β in $(k + 1)$ -th iteration is updated as:

$$\beta_{k+1} = \rho\beta_k, \rho > 1. \quad (18)$$

Finally, the proposed algorithm R-MTD stops when the convergence criterion is reached, i.e.,

$$\frac{\text{error}_{k+1} - \text{error}_k}{\text{error}_k} < \textit{tolerance} \quad (19)$$

with $\text{error}_{k+1} = \|\mathcal{T} - \widehat{\mathcal{G}}_{k+1} \times_1 \widehat{\mathbf{A}}_{k+1} \times_2 \widehat{\mathbf{B}}_{k+1} \times_3 \widehat{\mathbf{C}}_{k+1}\|_F$ and $\text{error}_k = \|\mathcal{T} - \widehat{\mathcal{G}}_k \times_1 \widehat{\mathbf{A}}_k \times_2 \widehat{\mathbf{B}}_k \times_3 \widehat{\mathbf{C}}_k\|_F$. Eventually, the values R_i^{est} are estimated using the MDL criterion (2) on the singular values derived from the SVD of the unfolding matrices $\widehat{\mathbf{G}}_{k+1}^{(i)}$, $i = 1, 2, 3$.

R-MTD for Denoising: Once the multilinear rank is estimated, we can use it to replace the over-estimated rank in (3) and the penalty function for denoising is rewritten as follows:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathcal{G}} \quad & \sum_{i=1}^3 \lambda_i \left(\|\mathbf{G}^{(i)}\|_{2,1} + \|\mathbf{G}^{(i)}\|_{1,2} \right) + \gamma \|\mathcal{G}\|_1 \\ \text{s.t.} \quad & \mathcal{T} = \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} + \mathcal{N}, \mathcal{G} \in \mathbb{R}^{R_1^{est} \times R_2^{est} \times R_3^{est}}, \\ & \mathbf{A} \in \mathbb{R}^{I_1 \times R_1^{est}}, \mathbf{B} \in \mathbb{R}^{I_2 \times R_2^{est}}, \mathbf{C} \in \mathbb{R}^{I_3 \times R_3^{est}}. \end{aligned} \quad (20)$$

Here, the weight parameters of the constraints are set to zero, i.e., $\lambda_i = \gamma = 0$, $i = 1, 2, 3$. With these zero weight parameters, we can repeat the same procedure as described previously.

4 Experiment Results

In this section, the noisy simulated data is utilized to test the performance of R-MTD on rank estimation and signal denoising. A comparative study of our method with the recent and efficient SCORE algorithm [7] is also provided. For the proposed method, the selection of parameter λ_i , $i = 1, 2, 3$ should be larger when the size of underlying core tensor is smaller than the over-estimated core tensor size. Typically, these coefficients are set as $\lambda_i = 5$, $i = 1, 2, 3$ and the other parameters μ , β and ρ are fixed as 0.025, 0.3 and 20, respectively. The selected parameters of SCORE method are chosen as the paper [7] advised. All results reported in this section are averaged over 100 Monte Carlo (MC) realizations.

Experimental Setting. The size of \mathcal{F} is fixed with $(I_1 = I_2 = I_3 = 100)$ and we consider different sizes of the core tensor \mathcal{G} : $(R_1 = R_2 = R_3 = 3)$, $(R_1 = R_2 = R_3 = 4)$ and $(R_1 = R_2 = R_3 = 5)$. All the elements of factor matrices and the core tensor follow a Gaussian distribution. To implement our method, the over-estimated core size is set to $\widehat{R}_1 = \widehat{R}_2 = \widehat{R}_3 = 10$. The noise

tensor $\mathcal{E} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ also follows Gaussian distribution with different Signal-to-Noise Ratio (SNR). Finally, the noisy tensor \mathcal{T} is obtained as follows:

$$\mathcal{T} = \mathcal{F} + \sigma \frac{\|\mathcal{F}\|_F}{\|\mathcal{E}\|_F} \mathcal{E},$$

where the parameter σ controls the SNR, i.e. $\text{SNR} = -20\log_{10}(\sigma)$.

Rank Estimation Performance. To evaluate the performance of all methods on rank estimation, we propose two criteria. The first one named Accuracy Rate (AR) measures the good estimate rate of all ranks, i.e. $R_1 = R_1^{est}$, $R_2 = R_2^{est}$ and $R_3 = R_3^{est}$, and is defined as the following:

$$\text{AR} : \text{Times of } R_i = R_i^{est}, i = 1, 2, 3.$$

The second criterion, called Average Rank Estimation Error (AREE), measures the deviation between the estimated ranks and exact ranks. It is given by:

$$\text{AREE} : \frac{1}{100} \sum_{\text{times}=1}^{100} \left(\sum_{i=1}^3 |R_i - R_i^{est}| \right).$$

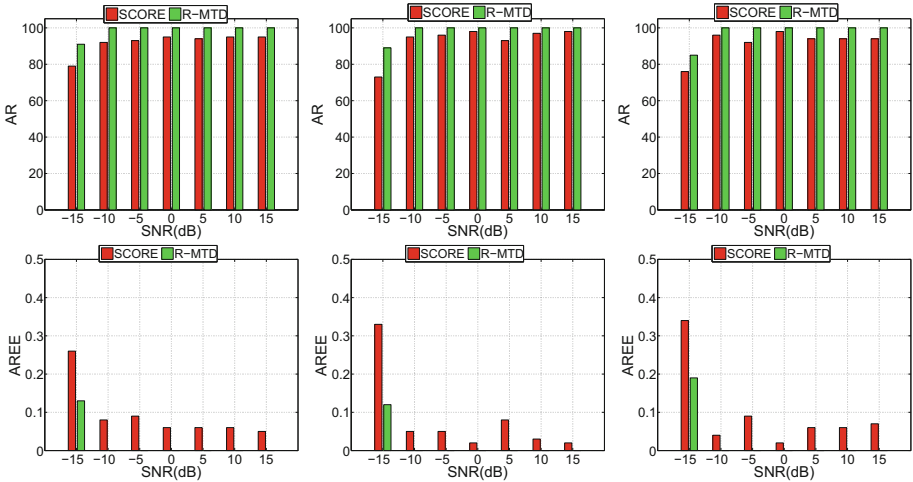


Fig. 2. The Accuracy Rate (AR) and the Average Rank Estimation Error (AREE) criteria for different sizes of the core tensor as a function of SNR.

Figure 2, displays the obtained results of the rank estimation for different SNR and different sizes of the core tensor. We can see that the AR criterion obtained for R-MTD and SCORE is increasing as the SNR increases (first line of Fig. 2). We also observed that, for all cases of the size of the core tensor, the robustness to noise of the R-MTD outperforms the SCORE whatever the SNR is. The second

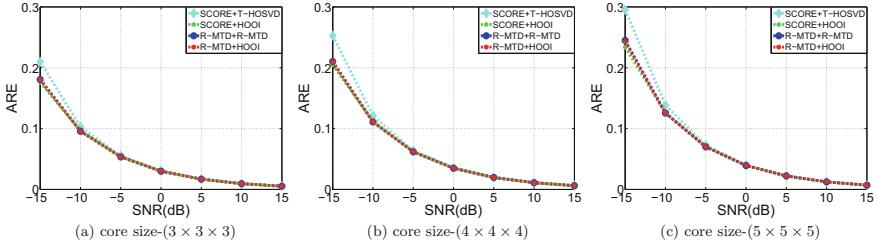


Fig. 3. The Average Relative Error (ARE) criterion for different sizes of the core tensor as a function of SNR.

line of Fig. 2 shows the AREE criterion. Clearly, the rank estimation errors of R-MTD are always smaller than those of the SCORE method for all scenarios and SNR.

Signal Denoising Efficiency. In this section, we consider the effectiveness and robustness of the proposed method on signal denoising. To do so, four combination cases of rank estimation methods and tensor decomposition ones are considered: R-MTD+R-MTD, R-MTD+HOOI [6], SCORE+T-HOSVD [5] and SCORE+HOOI. The reconstruction error is defined as the Average Relative Error (ARE), given by:

$$\text{ARE} : \frac{1}{100} \sum_{\text{times}=1}^{100} \frac{\|\mathcal{F}^{est} - \mathcal{F}\|_F}{\|\mathcal{F}\|_F}.$$

Figure 3 gives the effectiveness of the four compared methods. For SNR higher or equal to -5 dB, all methods seem to have similar performances. Regarding the lower SNR, we can see that the performances of the SCORE+HOOI, R-MTD+R-MTD and R-MTD+HOOI are better than the SCORE+T-HOSVD. This can be explained by the fact that the T-HOSVD method only extract the first several components corresponding to the estimated rank without any denoising process, it cannot gives excellent denoising result for lower SNR cases. Besides, although AR results of the R-MTD is more effective than the SCORE, the denoising efficiency of SCORE+HOOI is almost the same as the efficiency of R-MTD+R-MTD and R-MTD+HOOI because of the smaller difference of AREE between the SCORE and the R-MTD.

5 Conclusion

We have proposed a new MTD method, named R-MTD, that permits to identify the multilinear rank of noisy data. To do so, a sparsity and group sparsity of the core tensor are imposed by means of the l_1 norm and the mixed-norm, respectively. More precisely, we first over-estimate the core tensor, then an optimal core tensor is estimated from noisy multilinear tensor after several iterations of

the R-MTD algorithm. We also assessed the importance of a good estimation of the rank of the core tensor to signal denoising. Different experiments conducted using simulated data demonstrated the effectiveness of the R-MTD algorithm both for rank estimation and for data denoising.

References

1. Wax, M., Kailath, T.: Detection of signals by information theoretic criteria. *IEEE Trans. Acoust. Speech Signal Process.* **33**(2), 387–392 (1985)
2. Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
3. Rissanen, J.: Modeling by shortest data description. *Automatica* **14**(5), 465–471 (1978)
4. Tucker, L.R.: Implications of factor analysis of three-way matrices for measurement of change. In: *Problems in Measuring Change*, pp. 122–137 (1963)
5. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**(4), 1253–1278 (2000)
6. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank- (R_1, R_2, \dots, R_n) approximation of higher-order tensors. *SIAM J. Matrix Anal.* **21**(4), 1324–1342 (2000)
7. Yokota, T., Lee, N., Cichocki, A.: Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria. *IEEE Trans. Signal Process.* **65**(5), 1196–1206 (2017)
8. Xie, Q., Zhao, Q., Meng, D., Xu, Z.: Kronecker-basis-representation based tensor sparsity and its applications to tensor recovery. *IEEE Trans. Pattern Anal. Mach. Intell.* (2017)
9. Han, X., Albera, L., Kachenoura, A., Senhadji, L., Shu, H.: Low rank canonical polyadic decomposition of tensors based on group sparsity. In: *IEEE 25th European Signal Processing Conference (EUSIPCO)*, pp. 668–672 (2017)
10. Han, X., Albera, L., Kachenoura, A., Senhadji, L., Shu, H.: Block term decomposition with rank estimation using group sparsity. In: *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)* (2017)
11. Shu, X., Porikli, F., Ahuja, N.: Robust orthonormal subspace learning: efficient recovery of corrupted low-rank matrices. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3874–3881 (2014)
12. Fazel, M.: Matrix rank minimization with applications. Ph.D. dissertation, Ph.D. thesis, Stanford University (2002)
13. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90**(432), 1200–1224 (1995)
14. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends[®] Mach. Learn.* **3**(1), 1–122 (2011)
15. Petersen, K.B., Pedersen, M.S., et al.: *The Matrix Cookbook*. Technical University of Denmark, pp. 7–15 (2008)



Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition

Mathieu Fontaine¹(✉), Fabian-Robert Stöter²(✉), Antoine Liutkus²(✉),
Umut Şimşekli³(✉), Romain Serizel¹(✉), and Roland Badeau³(✉)

¹ Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
{mathieu.fontaine,romain.serizel}@inria.fr

² Inria and LIRMM, Montpellier, France

{fabian-robert.stoter,antoine.liutkus}@inria.fr

³ LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
{umut.simsekli,roland.badeau}@telecom-paristech.fr

Abstract. This paper introduces a new method for multichannel speech enhancement based on a versatile modeling of the residual noise spectrogram. Such a model has already been presented before in the single channel case where the noise component is assumed to follow an alpha-stable distribution for each time-frequency bin, whereas the speech spectrogram, supposed to be more regular, is modeled as Gaussian. In this paper, we describe a multichannel extension of this model, as well as a Monte Carlo Expectation - Maximisation algorithm for parameter estimation. In particular, a multichannel extension of the Itakura-Saito nonnegative matrix factorization is exploited to estimate the spectral parameters for speech, and a Metropolis-Hastings algorithm is proposed to estimate the noise contribution. We evaluate the proposed method in a challenging multichannel denoising application and compare it to other state-of-the-art algorithms.

1 Introduction

In many contexts, speech denoising is studied and applied in order to obtain, among other things, a comfortable listening or broadcast of a talk [2], by exploiting the observed noisy signal, obtained by several microphones. From an audio source separation perspective, this denoising is achieved through a probabilistic model, where the observed signal is divided into two latent sources: a noise component and a target source.

Both speech and noise components are usually considered in the *time frequency* (TF) domain where all TF-bins are supposed to be independent and follow a Gaussian law [5, 13]. A common approach to speech enhancement is the spectral subtraction method [6]. Its principle is to estimate an a priori *signal to noise ratio* (SNR) in order to infer a *short-time spectral amplitude* (STSA) estimator of the noise which will be subtracted from the STSA of the observations. Another popular trend is to decompose the *power spectral densities* (PSD) of sources into a product of two matrices. The *non-negative matrix factorization*

(NMF) model assumes that the PSDs admit low-rank structures and it performs well in denoising.

To the best of our knowledge, NMF models for multichannel speech enhancement have been proposed only in a Gaussian probabilistic context, whereas a non-Gaussian approach could offer a more flexible model for noise and speech. Moreover, a good initialization in a Gaussian NMF model is crucial to avoid staying stuck in a local minimum [3]. Many studies in the single-channel case have shown a better robustness to initialization when the signal is modeled in the TF domain with as heavy tail distribution [19,22].

Among this type of distributions, α -stable distributions preserve interesting properties satisfied by Gaussian laws, and they can model distributions ranging from light tails as in the *Gaussian case* to heavy tails as in the *Cauchy case*. Indeed, α -stable distributions are the only ones which admit a central limit theorem and stability by summation [16]. Various studies have been carried out on audio modeling using alpha-stable processes [12,19]. Especially in the TF domain, a generalization of wide-sense stationary (WSS) processes [13] has been established in the α -stable case [12] and applied to noise reduction [8]. More precisely, in [20] it was considered that the target source is Gaussian and the residual noise is α -stable, in order to get a greater flexibility on noise modeling.

This paper introduces a generalization of [20] to the multichannel case. The goal is to develop a Gaussian NMF model for speech that assumes a low-rank structure for speech covariances [5], while the noise part is taken as an α -stable process. Parameters are estimated through a combination of the multichannel extension of Itakura Saito NMF (IS-NMF) [17] for speech and a Markov Chain Monte Carlo (MCMC) strategy for estimating the noise part. The proposed method is evaluated for multichannel denoising, and compared to other state-of-the-art algorithms.

2 Probabilistic and Filtering Models

2.1 Mixture Model

Let $\mathbf{x} \in \mathbb{C}^{F \times T \times K}$ be the observed data in the short-time Fourier transform (STFT) domain where F, T and K denote the number of frequency bands, time frames and microphones, respectively. The observation \mathbf{x} will be assumed to be the sum of two latent audio sources represented by two tensors: the first one is written $\mathbf{y} \in \mathbb{C}^{F \times T \times K}$ and accounts for the *speech signal*. The second one is written $\mathbf{r} \in \mathbb{C}^{F \times T \times K}$ and called the *residual component*. We have:

$$\mathbf{x}_{ft} = \mathbf{y}_{ft} + \mathbf{r}_{ft}, \quad (1)$$

where each term belongs to \mathbb{C}^K . The main goal in this paper is to estimate the tensors \mathbf{y} and \mathbf{r} knowing \mathbf{x} , by using a probabilistic model described below.

2.2 Source Model

At short time scales, the speech signal may be assumed stationary and does not feature strong impulsiveness. This motivates modeling it as a locally stationary Gaussian process [13]. Furthermore, we also assume that the different channels for \mathbf{y}_{ft} are correlated, accounting for the *spatial* characteristics of the signal. Consequently, we assume that each \mathbf{y}_{ft} is an isotropic complex Gaussian vector¹ of mean $\mathbf{0}$ and covariance matrix $\mathbf{C}_{ft}^y \triangleq \mathbf{R}_f v_{f,t}$, where the *spatial covariance matrix* $\mathbf{R}_f \in \mathbb{C}^{K \times K}$ encodes the time-invariant correlations of speech in the different channels and $v_{f,t}$ is the PSD of the speech signal [5]. To exploit the redundancy of speech, we further decompose $v_{f,t}$ through NMF and obtain:

$$\forall f, t \quad \mathbf{y}_{ft} \sim \mathcal{N}_c \left(\mathbf{y}_{ft}; \mathbf{0}, \mathbf{R}_f v_{f,t} \triangleq \mathbf{R}_f \sum_{l=1}^L w_{fl} h_{lt} \right). \quad (2)$$

where \triangleq means “equals by definition” and $\mathbf{W} \in \mathbb{R}_+^{F \times L}$, $\mathbf{H} \in \mathbb{R}_+^{L \times T}$ are matrices which respectively contain all non-negative scalars w_{fl} and h_{lt} . While \mathbf{W} is understood as L spectral bases, \mathbf{H} stands for their activations over time. To make notations simpler, let $\Theta \triangleq \{\mathbf{W}, \mathbf{H}, \mathbf{R}\}$ be the parameters to be estimated with $\mathbf{R} \triangleq \{\mathbf{R}_f\}_f$. Note that the decomposition of $v_{f,t}$ is not unique: it is defined up to multiplicative constant.

In contrast to the speech signal, the model of the residual component should allow for outliers and impulsiveness. To do so, the residual part is modeled by an heavy-tailed distribution in the time domain. Recent works proposed a stationary model called α -harmonizable process with $\alpha \in (0, 2]$ in the single-channel case. It is shown in [12, 16] that such a model is equivalent to assuming that the signal at every time-frequency bin f, t follows a complex isotropic symmetric α -stable distribution. With the aim of extending the previous model to a multichannel one, we take all \mathbf{r}_{ft} as distributed with respect to an *elliptically contoured multivariate stable distribution* (ECMS, denoted $\mathcal{E}\alpha S$) and independent of one another. These distributions, which are a particular case of α -stable distributions, have the particularity of requiring only two parameters [11, 16]:

1. A *characteristic exponent* $\alpha \in (0, 2]$: the smaller α , the heavier the tails of the distribution.
2. A positive definite Hermitian *scatter matrix* in $\mathbb{C}^{K \times K}$.

In this article, the scatter matrices for all \mathbf{r}_{ft} are taken equal to $\sigma_f \mathbf{I}_K$, where $\mathbf{I}_K \in \mathbb{R}^{K \times K}$ is the identity matrix and $\sigma_f > 0$ is a positive scalar that does not depend on time. We have:

$$\forall f, t \quad \mathbf{r}_{ft} \sim \mathcal{E}\alpha S^K (\sigma_f \mathbf{I}_K). \quad (3)$$

¹ The probability density function (PDF) of an isotropic complex Gaussian vector is $\mathcal{N}_C(\mathbf{x}|\mu, \mathbf{C}) = \frac{1}{\pi^K \det \mathbf{C}} \exp(-(\mathbf{x} - \mu)^* \mathbf{C}^{-1} (\mathbf{x} - \mu))$.

2.3 Filtering Model

As mentioned in Subsect. 2.1, we aim to reconstruct the sources \mathbf{y} and \mathbf{r} from the observed data \mathbf{x} . From a signal processing point of view, when parameters $\boldsymbol{\sigma}$, \mathbf{W} , \mathbf{H} , \mathbf{R} are known, one would like to compute the Minimum Mean Squared Error (MMSE) estimates of both sources. In our probabilistic context, these can be expressed as the posteriori expectations $\mathbb{E}(\mathbf{y}_{ft}|\mathbf{x}_{ft}, \boldsymbol{\Theta}, \boldsymbol{\sigma})$.

To compute such estimates, a property specific to ECMS distributions can be exploited to represent \mathbf{r} as a complex normal distribution \mathcal{N}_c of dimension K , whose variance is randomly multiplied by a positive random *impulse variable* ϕ_{ft} distributed as $\mathcal{P}_{\frac{\alpha}{2}}S\left(2\cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha}\right)$, where $\mathcal{P}_{\frac{\alpha}{2}}S$ is the *positive $\alpha/2$ -stable distribution* (see [19] for more details):

$$\forall f, t \quad \mathbf{r}_{ft}|\phi_{ft} \sim \mathcal{N}_c(\mathbf{r}_{ft}; 0, \phi_{ft}\sigma_f\mathbf{I}_K), \quad (4)$$

If we assume for now that $\boldsymbol{\Phi} \triangleq \{\phi_{ft}\}_{f,t}$ are known in (4), we get the distribution of the mixture as:

$$\forall f, t \quad \mathbf{x}_{ft}|\phi_{ft} \sim \mathcal{N}_c(\mathbf{x}_{ft}; 0, \mathbf{C}_{ft}^{\mathbf{x}|\phi}), \quad (5)$$

where $\mathbf{C}_{ft}^{\mathbf{x}|\phi} \triangleq \mathbf{R}_f \sum_{l=1}^L w_{fl}h_{lt} + \phi_{ft}\sigma_f\mathbf{I}_K$. This in turns allows to build a multichannel Wiener filter as [2]:

$$\mathbb{E}(\mathbf{y}_{ft}|\mathbf{x}_{ft}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{\sigma}) = \mathbf{C}_{ft}^{\mathbf{y}} \left(\mathbf{C}_{ft}^{\mathbf{x}|\phi} \right)^{-1} \mathbf{x}_{ft}, \quad (6)$$

with \cdot^{-1} standing for matrix inversion.

Now, the strategy we adopt here is to marginalize this expression over $\boldsymbol{\Phi} | x$, to get:

$$\hat{\mathbf{y}}_{ft} = \mathbb{E}_{\boldsymbol{\Phi}|x} [\mathbb{E}[\mathbf{y}_{ft}|\mathbf{x}_{ft}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \boldsymbol{\sigma}]] = \mathbf{G}_{ft}\mathbf{x}_{ft},$$

where

$$\mathbf{G}_{ft} \triangleq \mathbf{C}_{ft}^{\mathbf{y}} \boldsymbol{\Xi}_{ft} \quad (7)$$

is the marginal Wiener filter, and $\boldsymbol{\Xi}_{ft} \triangleq \mathbb{E}_{\boldsymbol{\Phi}|x} \left[\left(\mathbf{C}_{ft}^{\mathbf{x}|\phi} \right)^{-1} \right]$ is the average inverse mixture covariance matrix. We will explain how to compute $\boldsymbol{\Xi}$ later in Sect. 3.3.

3 Parameter Estimation

3.1 Expectation-Maximization (EM) Algorithm

Assuming that the observations \mathbf{x} and the impulse variable ϕ are known, we first aim to estimate the parameters $\boldsymbol{\Theta}$. We choose a maximum likelihood estimator in order to get the most likely source NMF parameters \mathbf{W} , \mathbf{H} :

$$(\mathbf{W}^*, \mathbf{H}^*, \mathbf{R}^*) = \arg \max_{\mathbf{W}, \mathbf{H}, \mathbf{R}} \log \mathbb{P}(\mathbf{x}, \boldsymbol{\Phi} | \boldsymbol{\Theta}, \boldsymbol{\sigma}), \quad (8)$$

where Φ is a latent variable and $\log \mathbb{P}(\mathbf{x}, \Phi | \Theta, \sigma)$ is the log-likelihood. As in [20], we propose an EM algorithm. This method aims to minimize an upper-bound of $\mathcal{L}_n(\mathbf{W}, \mathbf{H}, \mathbf{R}) = -\log \mathbb{P}(\mathbf{x}, \Phi | \Theta, \sigma)$. This approach is summarized in the following two steps:

$$\text{E-Step: } \quad \mathcal{Q}_n(\mathbf{W}, \mathbf{H}, \mathbf{R}) = -\mathbb{E}_{\Phi | \mathbf{x}, \mathbf{W}^{(n-1)}, \mathbf{H}^{(n-1)}} [\mathcal{L}_n(\mathbf{W}, \mathbf{H}, \mathbf{R})], \quad (9)$$

$$\text{M-Step: } \quad (\mathbf{W}^{(n)}, \mathbf{H}^{(n)}, \mathbf{R}^{(n)}) = \arg \max_{\mathbf{W}, \mathbf{H}, \mathbf{R}} \mathcal{Q}_n(\mathbf{W}, \mathbf{H}, \mathbf{R}). \quad (10)$$

E-Step: We first introduce a positive function that upper-bounds the negative log-likelihood $\mathcal{L}_n(\mathbf{W}, \mathbf{H}, \mathbf{R})$, which is equal to [17]:

$$\mathcal{L}_n(\mathbf{W}, \mathbf{H}, \mathbf{R}) = \sum_{f,t} \left[\text{tr} \left(\tilde{\mathbf{X}}_{ft} \left(\mathbf{C}_{ft}^{x|\phi} \right)^{-1} \right) + \log \det \mathbf{C}_{ft}^{x|\phi} \right] \quad (11)$$

where $\tilde{\mathbf{X}}_{ft} \triangleq \mathbf{x}_{ft} \mathbf{x}_{ft}^*$ and $*$ stands for the Hermitian transposition. A positive auxiliary function $\mathcal{L}_n^+(\mathbf{W}, \mathbf{H}, \mathbf{R}, \mathbf{U}, \mathbf{V}) = \sum_{f,t} \left[\sum_l \frac{\text{tr}(\tilde{\mathbf{X}}_{ft} \mathbf{U}_{lft} (\mathbf{C}_{lft}^{x|\phi})^{-1} \mathbf{U}_{lft})}{w_{fl} h_{lt}} + \frac{\text{tr}(\tilde{\mathbf{X}}_{ft} \mathbf{U}_{lft}^2)}{\sigma_f \phi_{ft}} + \log \det \mathbf{V}_{ft} + \frac{\det \mathbf{C}_{ft}^{x|\phi} - \det \mathbf{V}_{ft}}{\det \mathbf{V}_{ft}} \right]$ which satisfies:

$$\mathcal{L}_n^+(\mathbf{W}, \mathbf{H}, \mathbf{R}, \mathbf{U}, \mathbf{V}) \geq \mathcal{L}_n(\mathbf{W}, \mathbf{H}, \mathbf{R}) \quad (12)$$

is introduced in [17]. Using (12) and the definition of \mathcal{Q}_n in (9), we obtain:

$$\mathbb{E}_{\Phi | \mathbf{x}} \mathcal{L}_n(\cdot) \leq \mathbb{E}_{\Phi | \mathbf{x}} \mathcal{L}_n^+(\cdot) \triangleq \mathcal{Q}_n^+(\cdot) \quad (13)$$

with:

$$\begin{aligned} \mathcal{Q}_n^+(\mathbf{W}, \mathbf{H}, \mathbf{R}, \mathbf{U}, \mathbf{V}) &= \sum_{f,t} \left[\sum_l \frac{\mathbb{E}_{\Phi | \mathbf{x}} \left(\text{tr} \left[\tilde{\mathbf{X}}_{ft} \mathbf{U}_{lft} (\mathbf{C}_{lft}^{x|\phi})^{-1} \mathbf{U}_{lft} \right] \right)}{w_{fl} h_{lt}} \right] \\ &+ \mathbb{E}_{\Phi | \mathbf{x}} \left(\text{tr} \left[\tilde{\mathbf{X}}_{ft} \mathbf{U}_{lft}^2 \right] \right) \sigma_f^{-1} \phi_{ft}^{-1} + \mathbb{E}_{\Phi | \mathbf{x}} \left(\log \det \mathbf{V}_{ft} + \det \left(\mathbf{V}_{ft}^{-1} \mathbf{C}_{lft}^{x|\phi} \right) - 1 \right) \end{aligned} \quad (14)$$

The form in (14) admits partial derivatives that will be useful as part of a multiplicative update [7] in the M-Step.

M-Step: Solving the M-Step in (10) is equivalent to zeroing the partial derivatives $\frac{\partial \mathcal{Q}_n^+}{\partial w_{fl}}$ and $\frac{\partial \mathcal{Q}_n^+}{\partial h_{lt}}$ and to set \mathbf{U}, \mathbf{V} such that the equality in (13) is verified. A multiplicative update approach yields:

$$w_{fl} \leftarrow w_{fl} \sqrt{\frac{\sum_t h_{lt} \text{tr}(\mathbf{R}_f \mathbf{P}_{ft})}{\sum_t h_{lt} \text{tr}(\mathbf{R}_f \mathbf{\Xi}_{ft})}}; \quad h_{lt} \leftarrow h_{lt} \sqrt{\frac{\sum_f w_{fl} \text{tr}(\mathbf{R}_f \mathbf{P}_{ft})}{\sum_f w_{fl} \text{tr}(\mathbf{R}_f \mathbf{\Xi}_{ft})}} \quad (15)$$

where the quantity $\Xi_{ft} = \mathbb{E}_{\Phi|x} \left[\left(C_{ft}^{x|\varphi_i} \right)^{-1} \right]$ has been used above in (7) and $P_{ft} = \mathbb{E}_{\Phi|x} \left[\left(C_{ft}^{x|\varphi_i} \right)^{-1} \tilde{X}_{ft} \left(C_{ft}^{x|\varphi_i} \right)^{-1} \right]$. We will explain how to compute these expectations in Subsect. 3.3.

3.2 Estimation of Spatial Covariance Matrices and Noise Gains σ

We update the spatial covariance matrix \mathbf{R} in the M-step as in [5], further using the trick proposed in [14] to use a weighted update, resulting in:

$$\mathbf{R}_f \leftarrow \left(\sum_t v_{ft} \right)^{-1} \times \sum_t \left(C_{ft}^{yy^*|x} \right), \quad (16)$$

where: $C_{ft}^{yy^*|x} \triangleq \mathbf{G}_{ft} \tilde{X}_{ft} \mathbf{G}_{ft} + C_{ft}^y - \mathbf{G}_{ft} C_{ft}^y$ is the total posterior variance for the speech source.

Concerning the estimation of the noise gain σ in (3), we exploit a result in [4] that if $z \sim \mathcal{E}\alpha S(\sigma)$, then $\mathbb{E}[\|z\|^p]^{\frac{\alpha}{p}} \propto \sigma$, for $p < \alpha$, with \propto standing for proportionality. The strategy we adopt is to apply this estimation only once at the beginning of the algorithm to the mixture itself, by taking a robust estimation like the median \mathbb{M} instead of the average, to account for the fact that not all TF bins pertain to the noise, but that a small portion also pertain to speech. We thus pick $p = \alpha/2$ and take:

$$\sigma_f \leftarrow \mathbb{M} \left(\left\| \sum_t \mathbf{x}(f, t) \right\|^{\alpha/2} \right)^2. \quad (17)$$

3.3 Expectation Estimation Using Metropolis-Hastings Algorithm

We still have to calculate the expectations Ξ_{ft} and P_{ft} . Unfortunately, they cannot be calculated analytically. To address this issue, we set up a Markov Chain Monte Carlo (MCMC) algorithm in order to approximate the expectations for each iteration. We are focusing on the Metropolis-Hastings algorithm through an empirical estimation of Ξ_{ft} and P_{ft} as follows:

$$\bar{\Xi}_{ft} \simeq \frac{1}{I} \sum_{i=1}^I \left(C_{ft}^{x|\varphi_i} \right)^{-1}; \quad \bar{P}_{ft} \simeq \frac{1}{I} \sum_{i=1}^I \left(\left(C_{ft}^{x|\varphi_i} \right)^{-1} \tilde{X}_{ft} \left(C_{ft}^{x|\varphi_i} \right)^{-1} \right) \quad (18)$$

with $\left(C_{ft}^{x|\varphi_i} \right)^{-1} = [\sum_l (\mathbf{R}_{fl} w_{fl} h_{lt}) + \varphi_{ft, i} \sigma_f \mathbf{I}_k]^{-1}$ and $\varphi_{ft, i}$ are sampled as follows:

First Step (Sampling Process): Generate a sampling via the prior distribution $\varphi'_{ft} \sim \mathcal{P}'_{\frac{\alpha}{2}} S \left(2 \cos \left(\frac{\pi\alpha}{4} \right)^{2/\alpha} \right)$.

Second Step (Acceptance):

- Draw $u \sim \mathcal{U}([0, 1])$ where \mathcal{U} denotes the uniform distribution.
- Compute the following acceptance probability:

$$\text{acc}(\varphi_{ft} \rightarrow \varphi'_{ft}) = \min \left(1, \frac{\mathcal{N}_c(\mathbf{x}_{ft}; 0, \varphi'_{ft} \sigma_f \mathbf{I}_K + \mathbf{C}_{ft}^y)}{\mathcal{N}_c(\mathbf{x}_{ft}; 0, \varphi_{ft} \sigma_f \mathbf{I}_K + \mathbf{C}_{ft}^y)} \right)$$

- Test the acceptance:
 - if $u < \text{acc}(\varphi_{ft, i-1} \rightarrow \varphi'_{ft})$, then $\varphi_{ft, i} = \varphi'_{ft}$ (acceptance)
 - otherwise, $\varphi_{ft, i} = \varphi_{ft, i-1}$ (rejection)

4 Single-Channel Speech Signal Reconstruction

Let $\hat{\mathbf{y}}$ be the multichannel signal obtained after Wiener filtering (7). In the context of speech enhancement, the desired speech is rather a single-channel signal, that we write $\hat{\mathbf{s}} \in \mathbb{C}^{F \times T}$. In this study, we take $\hat{\mathbf{s}}$ as a linear combination of $\hat{\mathbf{y}}$ with a time-invariant *beamformer* $\mathbf{B}_f \in \mathbb{C}^K$ [21]:

$$\hat{\mathbf{s}}_{ft} \triangleq \mathbf{B}_f^* \hat{\mathbf{y}}_{ft},$$

where \cdot^* denotes the Hermitian transposition. There are many ways to devise the beamformer \mathbf{B}_f . In this study, we choose to maximize the energy of $\mathbf{B}_f^* \mathbf{y}_{ft} \mid \mathbf{x}$:

$$\begin{aligned} \frac{1}{T} \sum_t \mathbb{E} \left(|\mathbf{B}_f^* \mathbf{y}_{ft}|^2 \mid \mathbf{x}_{ft} \right) &= \mathbf{B}_f^* \mathbb{E}(\mathbf{y}_{ft} \mathbf{y}_{ft}^* \mid \mathbf{x}) \mathbf{B}_f. \\ &= \mathbf{B}_f^* \frac{1}{T} \sum_t \left(\mathbf{C}_{ft}^{y y^* \mid \mathbf{x}} \right) \mathbf{B}_f. \end{aligned}$$

This is solved by taking \mathbf{B}_f as the eigenvector associated to the largest eigenvalue of the Hermitian matrix $\frac{1}{T} \sum_t \left(\mathbf{C}_{ft}^{y y^* \mid \mathbf{x}} \right)$ [5].

5 Evaluation

We investigate both the quality of speech enhancement and the audio source separation performance. Our proposed approach will be compared to two baseline methods:

- ARC** The proposed Alpha Residual component. We take $N = 10$ iterations for the EM and pick $\alpha = 1.9$.
- MWF** The classic multi-channel Wiener filter [5] which assumes Gaussianity for both noise and speech.
- GEVD** The generalized eigenvalue decomposition [18] is based on a low-rank approximation of the autocorrelation matrix of the speech signal.

5.1 Experimental Setup

The corpus for evaluation is made up of mono speech excerpts from Librispeech [15] and three different environmental noises taken from Aurora [10]: babble noise, restaurant and train. A groundtruth voice activity detection (VAD) is used on all three methods.

Mixtures were obtained for two 15 cm separated microphones, with the Roomsimove simulator with room dimensions of $5 \times 4 \times 3$ meters and RT60=0 ms and 500 ms. The sources are taken 1 m from the microphones, with different SNR values of $-5, 0, 5, 10$ dB and an angular distance of 30° or 90° . This results in 48 experiments.

5.2 Performance Measures

For the evaluation, two scores will be measured: the first one is a speech intelligibility weighted spectral distortion (SIW-SD) measure and the second one is a speech intelligibility-weighted SNR (SIW-SNR) [9].

The SIW-SD measure is defined as:

$$\text{SIW} - \text{SD} = \sum_i I_i \text{SD}_i \quad (19)$$

where I_i is the band importance function [1] and SD_i the average SD (in dB) in the i -th one third octave band,

$$\text{SD}_i = \frac{1}{(2^{1/6} - 2^{-1/6})f_i^c} \int_{2^{-1/6}f_i^c}^{2^{1/6}f_i^c} |10 \log_{10} G^y(f)| df \quad (20)$$

with center frequencies f_i^c and $G^y(f)$ is given by:

$$G^y(f) = \frac{P_{\mathbf{y}}(f)}{P_{\hat{\mathbf{y}}}(f)} \quad (21)$$

where $P_{\mathbf{y}}(f)$ and $P_{\hat{\mathbf{y}}}(f)$ are the power, for the frequency f , of the speech component of the input signal \mathbf{y} and the speech component output signal $\hat{\mathbf{y}}$, respectively.

The SIW-SNR [9] is used here to compute the *SIW-SNR improvement* which is defined as

$$\Delta \text{SNR}_{\text{intellig}} = \sum_i I_i (\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{in}}) \quad (22)$$

where $\text{SNR}_{i,\text{out}}$ and $\text{SNR}_{i,\text{in}}$ represent the output SNR of the noise reduction filter and the SNR of the signal in the first microphone in the i^{th} band, respectively.

5.3 Results

Results are displayed on Fig. 1 and present the SIW-SNR and SIW-SD scores averaged over noise types and spatial scenarios, against the input SNR.

We first investigate the impact of reverberation. While we see that ARC is outperformed by other methods in the anechoic case, we see it is much less sensitive to reverberation and becomes competitive compared to the other algorithms in terms of SIW-SD at low input SNR.

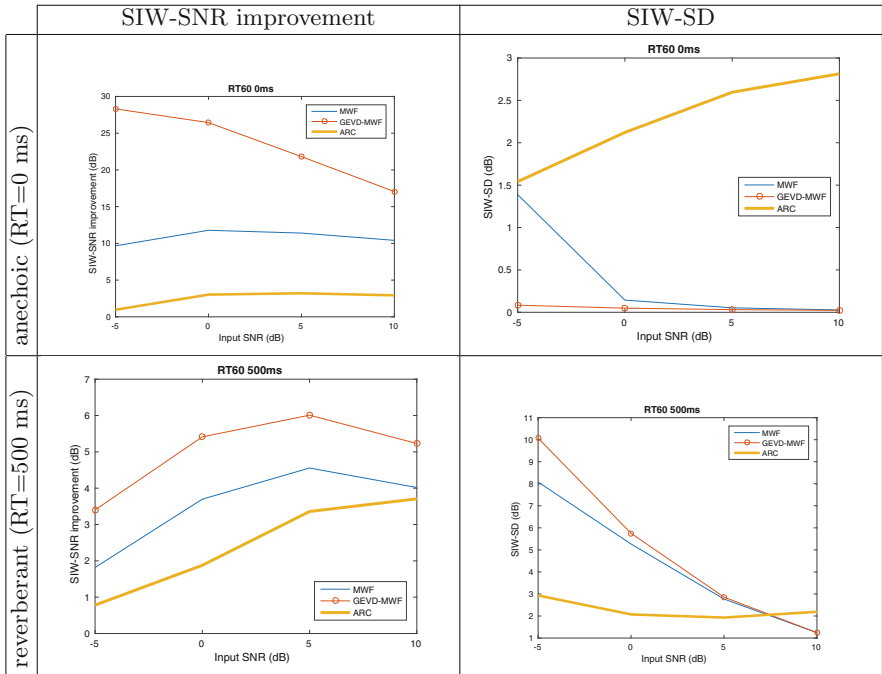


Fig. 1. SIW (left, higher is better) and SNR & SIW-SD (right, lower is better) for: (top) an anechoic scenario and (bottom) a reverberant room.

6 Conclusion

We proposed a new method ARC for denoising that is more robust to reverberation than competing approaches, although less effective in the anechoic case. It is based on modeling the speech signal as a Gaussian process and noise as an α -stable sub-Gaussian process. Interestingly, that approach can be combined with existing methods, which could be an interesting avenue for future work.

Acknowledgments. This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01), EDiSon3D (ANR-13-CORD-0008-01), FBIMATRIX (ANR-16-CE23-0014) funded by ANR, the French State agency for research.

References

1. ANSI: S3. 5–1997, Methods for the calculation of the speech intelligibility index. New York: American National Standards Institute 19, 90–119 (1997)
2. Van den Bogaert, T., Doclo, S., Wouters, J., Moonen, M.: Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *J. Acoust. Soc. Am.* **125**(1), 360–371 (2009)
3. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognit.* **41**(4), 1350–1362 (2008)
4. Cambanis, S., Keener, R., Simons, G.: On α -symmetric multivariate distributions. *J. Multivar. Anal.* **13**(2), 213–233 (1983)
5. Duong, N., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1830–1840 (2010)
6. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984)
7. Févotte, C., Idier, J.: Algorithms for nonnegative matrix factorization with the β -divergence. *Neural Comput.* **23**(9), 2421–2456 (2011)
8. Fontaine, M., Liutkus, A., Girin, L., Badeau, R.: Parameterized Wiener filtering for single-channel denoising. In: *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2017)
9. Greenberg, J., Peterson, P., Zurek, P.: Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *J. Acoust. Soc. Am.* **94**(5), 3009–3010 (1993)
10. Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)* (2000)
11. Leglaive, S., Simsekli, U., Liutkus, A., Badeau, R., Richard, G.: Alpha-stable multichannel audio source separation. In: *42nd International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017)
12. Liutkus, A., Badeau, R.: Generalized Wiener filtering with fractional power spectrograms. In: *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 266–270. IEEE (2015)
13. Liutkus, A., Badeau, R., Richard, G.: Gaussian processes for underdetermined source separation. *IEEE Trans. Signal Process.* **59**(7), 3155–3167 (2011)
14. Nugraha, A.A., Liutkus, A., Vincent, E.: Multichannel music separation with deep neural networks. In: *24th European Signal Processing Conference (EUSIPCO)* 2016. pp. 1748–1752. IEEE (2016)
15. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an ASR corpus based on public domain audio books. In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210. IEEE (2015)
16. Samoradnitsky, G., Taqqu, M.: *Stable non-Gaussian random processes: stochastic models with infinite variance*, vol. 1. CRC Press, Boca Raton (1994)
17. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization. In: *37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–264. IEEE (2012)

18. Serizel, R., Moonen, M., Van Dijk, B., Wouters, J.: Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(4), 785–799 (2014)
19. Şimşekli, U., Liutkus, A., Cemgil, A.: Alpha-stable matrix factorization. *IEEE Signal Process. Lett.* **22**(12), 2289–2293 (2015)
20. Şimşekli, U., et al.: Alpha-stable low-rank plus residual decomposition for speech enhancement. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2018)
21. Van Veen, B.D., Buckley, K.M.: Beamforming: a versatile approach to spatial filtering. *IEEE assp magazine* **5**(2), 4–24 (1988)
22. Yoshii, K., Itoyama, K., Goto, M.: Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 51–55. IEEE (2016)



Sum Conditioned Poisson Factorization

Gökhan Çapan¹, Semih Akbayrak^{1(✉)}, Taha Yusuf Ceritli^{2,3},
and Ali Taylan Cemgil¹

¹ Computer Engineering Department, Boğaziçi University, Istanbul, Turkey

{gokhan.capan, semih.akbayrak, taylan.cemgil}@boun.edu.tr

² School of Informatics, University of Edinburgh, Edinburgh, UK

t.y.ceritli@sms.ed.ac.uk

³ Alan Turing Institute, London, UK

Abstract. We develop an extension to Poisson factorization, to model Multinomial data using a moment parametrization. Our construction is an alternative to the canonical construction of generalized linear models. This is achieved by defining K component Poisson Factorization models and constraining the sum of observation tensors across components. A family of fully conjugate tensor decomposition models for binary, ordinal or multinomial data is devised as a result, which can be used as a generic building block in hierarchical models for arrays of such data. We give parameter estimation and approximate inference procedures based on Expectation Maximization and variational inference. The flexibility of the resulting model on binary and ordinal matrix factorizations is illustrated. Empirical evaluation is performed for movie recommendation on ordinal ratings matrix, and for knowledge graph completion on binary tensors. The model is tested for both prediction and producing ranked lists.

1 Introduction

Approximate matrix and tensor factorization methods have found diverse applications all across in machine learning, information retrieval, bioinformatics and signal processing [5]. For many applications, they provide a good balance between accurate, interpretable models and efficient, scalable and parallelizable algorithms. This balance, supported with a rich background theory has arguably contributed to the popularity of approximate matrix decompositions.

The approximate matrix factorization problem is typically cast as an optimization problem: for a given matrix X find matrices W^*, H^* such that

$$(W^*, H^*) = \arg \min_{W, H} D(X || WH) \quad (1)$$

where the error function D is a divergence. Note that this formulation is equivalent to finding a nearby matrix \hat{X} that can be written as an exact matrix product $\hat{X} = WH$.

T. Y. Ceritli—Work was done at Boğaziçi University.

Error functions for matrices are almost always chosen as elementwise, where $D(X||\hat{X}) = \sum_{ij} d(X(i, j)||\hat{X}(i, j))$, with the most popular choice being the square Euclidean divergence $d(x||\hat{x}) = (x - \hat{x})^2/2$. When matrix element $X(i, j)$ and the corresponding approximation $\hat{X}(i, j)$ are positive, the information divergence with $d(x||\hat{x}) = x \log(x/\hat{x}) - x + \hat{x}$ is also often used. When the matrices X and \hat{X} are suitably normalized, the information divergence is also known as the Kullback-Leibler (KL) divergence.

Approximate matrix factorization can also be cast as a statistical estimation problem. Such a probabilistic perspective is attractive as it opens up the possibility of solving more challenging tasks such as active learning or model selection, via an hierarchical Bayesian treatment. In the sequel, we will detail on two closely related approaches for the construction of probabilistic matrix factorization models as generative models [9, 18].

The first approach is defining the observation density for each matrix element $X(i, j)$ as an exponential family:

$$X(i, j) \sim p(\cdot; \theta(i, j)) = \exp(\psi(X(i, j))\theta(i, j) - A(\theta(i, j))) \quad (2)$$

Here ψ is known as the sufficient statistics and A is the log-partition function [1, 21]. We will refer to this model as the canonical parametrization. The exponential family form arises naturally as the maximum entropy (Gibbs) distribution when the data distribution is solely described in terms of fixed dimensional sufficient statistics ψ . Many models, such as exponential family PCA [6], Logistic Matrix Factorization [20], Ordinal Matrix Factorization [10, 17], or various binary tensor factorization models such as RESCAL [15] can be viewed as a factorization of the canonical parameters.

In the second approach, one directly parametrizes the moment parameters. Here, each matrix entry $X(i, j)$ is modelled with a random variable having a density of the form

$$X(i, j) \sim p(\cdot; \mu(i, j)) = \frac{1}{Z} \exp(-D(X(i, j)||\mu(i, j))) \quad (3)$$

where $\mu(i, j)$ is an expectation parameter. One key aspect is that the normalizing constant Z does not depend on the expectation μ . Many popular matrix decomposition methods, such as Probabilistic PCA [14], Factor analysis, Nonnegative Matrix factorization with KL cost (KL-NMF) also known as Poisson Factorization [3, 4, 7, 8] Gaussian Matrix Factorization [18] can be viewed as a low-rank decompositions of moment parameters. For example, Poisson factorization closely related to NMF with KL cost assumes $X(i, j) \sim \mathcal{PO}(\sum_k W(i, k)H(k, j))$ where $\mathcal{PO}(\mu)$ denotes the Poisson distribution with intensity μ [3].

In the absence of any parameter tying or regularization, the moment and canonical forms would be equivalent in the sense that there exists a bijective mapping g between the two parametrizations as $g^{-1}(\theta(i, j)) = \mu(i, j)$ known as the *canonical link*. In fact, when D is a Bregman divergence, there is a one-to-one correspondance between exponential family distributions and divergences [2]. However, as matrix factorization achieves parameter tying via a low rank decomposition, the choice of the parametrization can sometime have a dramatic effect

on the prediction power as well as the interpretability of the resulting model. In other words, it makes a difference when $\mu(i, j)$ or $\theta(i, j)$ is set to be equal to $\sum_k W(i, k)H(k, j)$. The difference is sometimes explained informally as additive versus multiplicative combination of the inferred latent representations $W(i, :)$ and $H(:, j)$. For example, the canonical form alternative to Poisson factorization is $X(i, j) \sim \mathcal{PO}(\exp(\sum_k W(i, k)H(k, j)))$, where the canonical parameter $\theta = \log \mu$ is factorized, is far less often used, possibly because the parameters would lack sparsity and are not easily interpretable.

This observation brings us to the motivation of the present paper: to develop a simple method to extend Poisson factorization, hence KL-NMF, to Multinomial models using a moment parametrization. The resulting approach enjoys all the attractive properties of Poisson factorization such as conjugacy and can be very easily implemented. The model can also be used as a building block in hierarchical models for arrays of binary, ordinal or categorical data, as the multinomial family of distributions include Bernoulli, Binomial and categorical distributions. The derivation is simple and hinges on two well known properties of Poisson distributions [11]: (i) the sum $n = \sum_{i=1}^K x_i$ of a collection of independent Poisson random variables x_i with intensities μ_i is also Poisson with the intensity $\mu = \sum_i \mu_i$; and (ii) conditioned on n , the joint posterior $p(x_1, \dots, x_K | n)$ is multinomial with the i 'th cell probability given as μ_i / μ . A similar construction for source separation but for conditionally Gaussian models has been described in [19].

2 Model

The Sum Conditioned Poisson Factorization (SCPF) model defines K component Poisson Factorization (PF) models and constrains the sum of observation matrices across components. More precisely, the generative model for SCPF is:

$$\begin{aligned} w_{k,i,r} &\sim \mathcal{G}(a^w, b^w/a^w) & h_{k,r,j} &\sim \mathcal{G}(a^h, b^h/a^h) & s_{k,i,j,r} &\sim \mathcal{PO}(w_{k,i,r}h_{k,r,j}) \\ x_{k,i,j} &= \sum_{r=1}^{R_k} s_{k,i,j,r} & n_{i,j} &= \sum_{k=1}^K x_{k,i,j} \end{aligned} \quad (4)$$

where $\mathcal{G}(a, b/a)$ denotes the Gamma distribution with shape a and mean b .

Using the superposition property, each matrix entry $x_{k,i,j}$ is a-priori Poisson with intensity $\sum_{r=1}^{R_k} w_{k,i,r}h_{k,r,j}$ where R_k is the model order, that is the rank of the latent intensity matrix. The model is completed by constraining the total sum across k components to $n_{i,j}$ that we will refer as the cardinality matrix. In practice, we will assume that the cardinality matrix $n_{i,j}$ is always known and is equal to the cardinality of the discrete variable and K is the number of categories in an one-hot encoding schema. Conditioning on $n_{i,j}$ couples the random variables $x_{:,i,j}$ across the k index to have jointly a Multinomial distribution. As a specific example, for modelling a binary matrix, we choose $K = 2$ and let $n_{i,j} = 1$ to be a matrix of ones. As such, the full conditional $p(x_{1:2,i,j} | n_{i,j}, w_{:,i,:}, h_{:,i,:})$ is a Bernoulli in a one-hot encoding $[1, 0]$ or $[0, 1]$. For a categorical distribution, we

choose K as the number of categories and let $n_{i,j} = 1$. For a binomial random variable with a range of $\{0, \dots, n\}$, we let $n_{i,j} = n$ and $K = 2$. A schematic representation with ‘cubic’ plots is shown in Fig. 1.

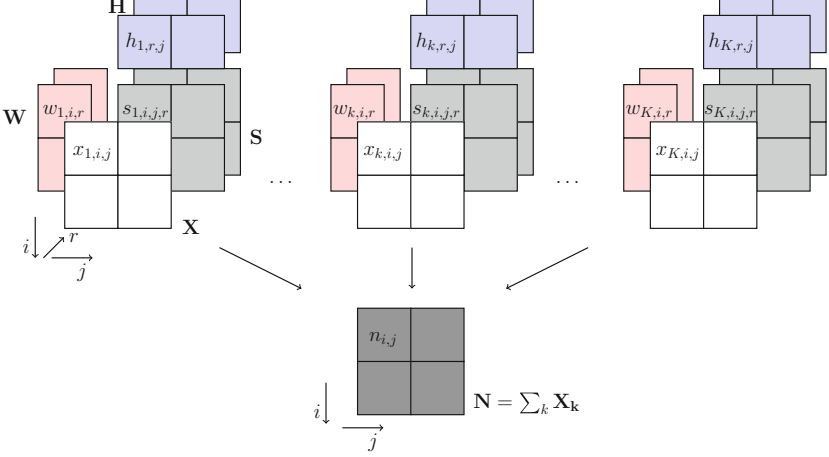


Fig. 1. A schematic description of SCPF model.

The SCPF model is particularly suitable for modeling bounded ordinal data such as movie ratings by a Binomial distribution. Suppose we are given an incomplete matrix with entries $Y(i, j) \in \{0, \dots, n\}$ with observed (missing) entries denoted by a mask matrix $\tilde{m}_{i,j} = 1$ ($=0$). We would define a SCPF model with $K = 2$ and $n_{i,j} = n$ for all variables and clamp $x_{1,i,j} = Y(i, j)$. For the special case of $K = 2$, we can infer that $x_{2,i,j} = n - x_{1,i,j}$. Here, the matrices $x_{1,:}$ and $x_{2,:}$ correspond to ‘like’ and ‘dislike’ scores respectively. If there were no missing entries in the original matrix Y , the corresponding SCPF model would reduce to two conditionally independent PF models. However, when there are missing entries, the predictive distribution $p(x_{1,i,j} | n_{i,j}, W, H)$ is Binomial with index parameter n and the probability $\sum_{r=1}^{R_1} w_{1,i,r} h_{1,r,j} / (\sum_{r=1}^{R_1} w_{1,i,r} h_{1,r,j} + \sum_{r=1}^{R_2} w_{2,i,r} h_{2,r,j})$. In a sense, the model balances ‘like’ and ‘dislike’ features to generate predictions.

For the general case $n > 1$ and $K > 2$, we introduce for each component $k = 1 \dots K$ a 0–1 mask matrix $m_{k,i,j}$, that is 1 or 0 if $x_{k,i,j}$ is observed or missing. Since the cardinality matrix $n_{i,j}$ is always fixed, the unobserved $x_{k,i,j}$ for a fixed i, j pair are still coupled. To simplify the notation, we define a residual matrix \tilde{N} where $\tilde{n}_{i,j} = n_{i,j} - \sum_{k=1}^K x_{k,i,j} m_{k,i,j}$. \tilde{N} provides the fraction of data that needs to be explained by the unobserved component matrices X_k .

3 Learning and Inference

In this section we present an Expectation Maximization (EM) approach for parameter estimation in a SCPF model and compare it to KL-NMF and PF. We also present full Bayesian inference with Variational inference.

For notational clarity, the procedures given in this section would, without loss of generality, assume 3-way component and cardinality tensors and 3 factor matrices per each component. Also for brevity, we will omit symmetrical update rules.

The generative model becomes:

$$\begin{aligned}
 w_{k,i,r} &\sim \mathcal{G}(a^w, b^w/a^w) & h_{k,j,r} &\sim \mathcal{G}(a^h, b^h/a^h) & g_{k,l,r} &\sim \mathcal{G}(a^g, b^g/a^g) \\
 s_{k,i,j,l,r} &\sim \mathcal{PO}(w_{k,i,r}h_{k,j,r}g_{k,l,r}) & x_{k,i,j,l} &= \sum_{r=1}^{R_k} s_{k,i,j,l,r} & n_{i,j,l} &= \sum_{k=1}^K x_{k,i,j,l}
 \end{aligned} \tag{5}$$

3.1 Expectation Maximization

Given observed entries of component tensors X , the mask M and cardinality tensors N , our goal is to infer the factors W , H and G such that the likelihood is maximized.

The derivations are straightforward and follow closely [3]. The SCPF log-likelihood is:

$$\begin{aligned}
 \log p(N, X|W, H, G) &= \log \left[\sum_S p(N|X)p(X|S)p(S|W, H, G) \right] \\
 &= \sum_{k,i,j,l} m_{k,i,j,l} \log(\mathcal{PO}(x_{k,i,j,l}; \sum_r w_{k,i,r}h_{k,j,r}g_{k,l,r})) \\
 &+ \sum_{i,j,l} \log(\mathcal{PO}(\tilde{n}_{i,j,l}; \sum_k (1 - m_{k,i,j,l}) \sum_r w_{k,i,r}h_{k,j,r}g_{k,l,r}))
 \end{aligned} \tag{6}$$

$$\tag{7}$$

The expected sufficient statistics of the latent components s can be computed in closed form:

$$\mathbb{E}_1(s_{k,i,j,l,r}) = \frac{w_{k,i,r}h_{k,j,r}g_{k,l,r}}{\sum_{r=1}^{R_k} w_{k,i,r}h_{k,j,r}g_{k,l,r}} x_{k,i,j,l} \tag{8}$$

$$\mathbb{E}_0(s_{k,i,j,l,r}) = \frac{w_{k,i,r}h_{k,j,r}g_{k,l,r}}{\sum_{k=1}^K (1 - m_{k,i,j,l}) \sum_{r=1}^{R_k} w_{k,i,r}h_{k,j,r}g_{k,l,r}} \tilde{n}_{i,j,l} \tag{9}$$

When $x_{k,i,j,l}$ is observed, the corresponding latent variables $w_{k,i,:}$, $h_{k,j,:}$ and $g_{k,l,:}$ are conditionally independent from others ($w_{\not{k},i,:}$ and $h_{\not{k},j,:}$, $g_{\not{k},l,:}$). When some $x_{:,i,j,l}$ are missing, however, the latent decomposition variables become coupled, and the residual $\tilde{n}_{i,j,l}$ is shared among the corresponding factor variables.

The EM update equations can be viewed as a simple generalization of KL-NMF [12], or PF [3, 7]. Indeed, when no data is missing, the algorithm identically reduces to the so-called multiplicative updates for KL-NMF. However, with missing data, the model incorporates the residual matrix into the update rules via the second terms. The updates for a maximum a-posteriori (MAP) estimation take the forms given below:

$$w_{k,i,r} \leftarrow \frac{a^w + \sum_{j,l} (m_{k,i,j,l} \mathbb{E}_1(s_{k,i,j,l,r}) + (1 - m_{k,i,j,l}) \mathbb{E}_0(s_{k,i,j,l,r}))}{a^w / b^w + \sum_{j,l} h_{k,j,r} g_{k,l,r}} \quad (10)$$

where a_w and b_w correspond to the shape and mean parameters of the a-priori Gamma distributions for each entry of W

The predictive distribution for $x_{k,i,j,l}$ is a Multinomial with mean:

$$\hat{x}_{k,i,j,l} = n_{i,j,l} \frac{\sum_{r=1}^{R_k} w_{k,i,r} h_{k,j,r} g_{k,l,r}}{\sum_{k=1}^K \sum_{r=1}^{R_k} w_{k,i,r} h_{k,j,r} g_{k,l,r}} \quad (11)$$

3.2 Variational Inference

The hierarchical probability model is conjugate and here we present variational inference equations with mean field approximation. Fully factorized instrumental distribution $q(S, W, H, G)$ is used to approximate posterior distribution $p(S, W, H, G | N, X)$, where $s_{k,i,j,l} | p_{k,i,j,l}$ is the multinomial variational decomposition variable, and $w_{k,i,r} | \alpha_{k,i,r}^w, \beta_{k,i,r}^w$ is the variational gamma factor variable. H and G are devised analogously to W .

The variational objective can be optimized by using coordinate solvers at each iteration due to conjugacy. The updates are shown below:

$$\phi_{k,i,j,l,r} \leftarrow \exp [\mathbb{E}[\log w_{k,i,r}] + \mathbb{E}[\log h_{k,j,r}] + \mathbb{E}[\log g_{k,l,r}]] \quad (12)$$

$$\alpha_{k,i,r}^w \leftarrow a^w + \sum_{j,l} \left[m_{k,i,j,l} x_{k,i,j,l} \frac{\phi_{k,i,j,l,r}}{\sum_{r'} \phi_{k,i,j,l,r'}} + (1 - m_{k,i,j,l}) \tilde{n}_{i,j,l} \frac{\phi_{k,i,j,l,r}}{\sum_{k'} (1 - m_{k',i,j,l}) \sum_{r'} \phi_{k',i,j,l,r'}} \right] \quad (13)$$

$$\beta_{k,i,r}^w \leftarrow \left(a^w / b^w + \sum_{j,l} \mathbb{E}[h_{k,j,r}] \mathbb{E}[\log g_{k,l,r}] \right)^{-1} \quad (14)$$

where $\mathbb{E}[w_{k,i,r}] = \alpha_{k,i,r}^w / \beta_{k,i,r}^w$ and $\mathbb{E}[\log w_{k,i,r}] = \psi(\alpha_{k,i,r}^w) + \log(\beta_{k,i,r}^w)$ with digamma function $\psi(\cdot)$

4 Experiments

In this section, we will illustrate the flexibility of the SCPF in binary and ordinal data analysis tasks. The first experiment with binary synthetic data is on the

Swimmer data set: a synthetically generated sequence of binary images where each image has four “limbs” of that can be in one of 4 positions.¹ In the second set of experiments, we compare the approach to alternative methods on an ordinal Matrix Factorization problem, as applied to the MovieLens data set of 100-K ratings (1–5) from 943 users on 1,682 movies.² Lastly, we evaluate the predictive power of SCPF with link prediction task on Knowledge Graph datasets.

4.1 Binary Matrix Factorization

We define $K = 2$, $x_{1,i,j} = Y(i,j)$, and $x_{2,i,j} = 1 - x_{1,i,j}$ to perform binary matrix factorization with SCPF. The approach is compared to the canonical form alternative, Logistic Matrix Factorization (LMF), i.e. $Y(i,j) \sim \mathcal{BE}(\sigma(\sum_k W(i,k)H(k,j)))$ where $\sigma(\cdot)$ is the sigmoid function and \mathcal{BE} is Bernoulli distribution. Two alternatives are used to factorize the Swimmer data set.

The reshaped basis vectors in the template matrices inferred by LMF and the first component of SCPF are shown in Fig. 2. We observe the effect of moment parametrization for the latter case, resulting in a more interpretable template matrix.

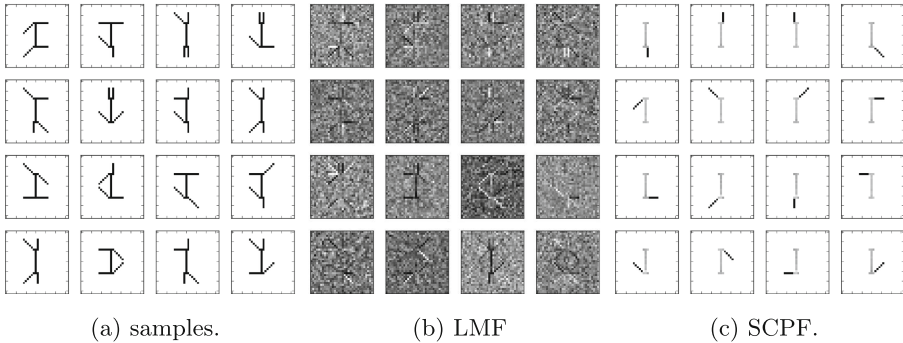


Fig. 2. (a) Each figure is a sample from the data set. (b, c) Each image is constructed by reshaping a basis vector in the template matrix (the first one in the SCPF case) inferred by the models.

4.2 Ordinal Matrix Factorization for Collaborative Filtering

On the movie ratings data set, we define two SCPF models with $K = 2$ and $K = 3$. In both cases $x_{1,i,j} = Y(i,j)$ for observed $Y(i,j)$, and $n_{i,j} = 5$. Conditioning the sum of components to $n_{i,j}$, the maximum allowed rating value, results in binomial and multinomial component (predictive) distributions for $K = 2$ and $K = 3$, respectively. We are interested in completion of the first matrix (which is filled with

¹ <http://www.stanford.edu/~vcs/Data/Y.mat>.

² <https://grouplens.org/datasets/movielens/100k/>.

the actual rating behavior of users), the marginal of X_1 , which is a Binomial in both cases.

As the canonical form alternative for ordinal matrices, we use the ordered probit model described at [17]. We also compare the models with Gaussian [18] and Poisson Matrix Factorizations [3].

Experiment Setup. The models are tested on MovieLens 100-K data set with 5-fold cross validation where we split data set randomly into a training and a test set with 80% and 20% ratings. Each experiment is repeated for the latent ranks $R \in \{20, 50, 100\}$ so $R_k = R$ for each k . In the *MAP*-estimate case, shape A and mean parameter B of SCPF are fixed to 10^{-3} and $1/R$, respectively. For variational inference, the user factors for each component are Gamma random variables a-priori with shape 1 and mean 1, and the movie factors with shape 1 and mean $10/\text{rank}$. While the latent rank of Ordinal Matrix Factorization (OMF) is changed, other parameters are kept as in [17]. The maximum iteration number is taken as 1000 for each algorithm.

Both the maximum a-posteriori estimates and the approximate posterior inference are included in SCPF experimental results. We used the mean of the variational distributions for predictions. Parameter estimation in LMF and Gaussian Matrix Factorization (GMF) is carried out through Stochastic Gradient Descent (SGD) with L_2 regularization. The Gibbs sampler for OMF is used as provided, with a burnin period of 500 iterations.³

Metrics. The models are evaluated for their performance in predicting the test ratings (with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)) and top- K recommendation. Top- K recommendation performance is evaluated with the standard IR metrics Mean Average Precision (MAP) and Recall@ k , where $\text{avg-precision}_i = \sum_{j \in \text{test}_i} \text{Precision}(\text{rank}(i, j)) / |\text{test}_i|$ and $\text{Recall}@k_i = \sum_{j \in \text{test}_i} \mathbf{1}[\text{rank}(i, j) \leq k] / \min(k, |\text{test}_i|)$. Movies that user i rated 5 are included in test_i . $\text{rank}(i, j)$ denotes the position of the item j in the recommendation list for user i .

Results. Table 1 summarizes the Collaborative Filtering (CF) results. We observe both more accurate matrix completion (rating prediction) performance, and higher average precision for SCPF variants, an indicator of ranking performance of the model. Specifically in comparison with PF, this justifies the effect of introducing the second component representing users' dislikes, negatively correlated with the original preference matrix.

When computing MAP, we keep recommending from the predictive regardless of the length of the ranked recommendation list. The users of a recommender system, however, are more likely to observe only a short list of recommendations. We observe higher recalls in recommendation lists of length 10, 20 and 50 produced by SCPF.

³ <http://cogsys.imm.dtu.dk/ordinalmatrixfactorization>.

Table 1. MovieLens 100-K experiment results. SCPF-K denotes the proposed model with K components. R@k is abbreviation for Recall@k.

	Model	RMSE	MAE	MAP	R@10	R@20	R@50
R = 20	SCPF-2	0.961	0.743	0.083	0.113	0.165	0.240
	SCPF-3	0.978	0.753	0.080	0.123	0.150	0.219
	SCPF(VI)-2	0.914	0.718	0.088	0.126	0.184	0.300
	SCPF(VI)-3	0.912	0.720	0.080	0.123	0.178	0.295
	PF	1.330	0.957	0.019	0.009	0.018	0.045
	GMF	0.940	0.744	0.071	0.098	0.132	0.239
	OMF	0.980	0.762	0.041	0.058	0.086	0.174
R = 50	SCPF-2	0.973	0.750	0.086	0.133	0.175	0.255
	SCPF-3	0.999	0.763	0.075	0.122	0.162	0.223
	SCPF(VI)-2	0.918	0.723	0.075	0.112	0.163	0.278
	SCPF(VI)-3	0.915	0.723	0.082	0.124	0.177	0.290
	PF	1.393	1.031	0.020	0.018	0.036	0.099
	GMF	0.926	0.736	0.076	0.112	0.165	0.285
	OMF	0.986	0.766	0.043	0.060	0.092	0.175
R = 100	SCPF-2	0.997	0.763	0.084	0.133	0.180	0.265
	SCPF-3	1.016	0.771	0.074	0.121	0.167	0.266
	SCPF(VI)-2	0.927	0.731	0.065	0.102	0.152	0.262
	SCPF(VI)-3	0.929	0.738	0.065	0.099	0.152	0.268
	PF	1.325	1.011	0.031	0.040	0.067	0.150
	GMF	0.924	0.734	0.068	0.099	0.154	0.276
	OMF	0.999	0.778	0.037	0.054	0.080	0.158

A relevant problem to SCPF for collaborative filtering might be modelling implicit feedback datasets by modelling exposure [13], where we observe positive feedback (likes, bookmarks, etc.) by the users, but explicit negative feedback is absent: a user might dislike, or might be unaware of the existence of the item. Setting $n_{i,j} = 1$ with $K = 3$ might result in 3 competing explanations for $Y(i, j)$ (perhaps when one component is supported by side information as covariates for modelling the non-exposure), representing *like*, *dislike*, and *non-exposure* respectively. We leave further exploration of modeling the described phenomenon within the described framework as a future work.

4.3 Binary Tensor Factorization for Knowledge Graph Completion

We apply SCPF to 3 Knowledge Graph datasets (Nation, UMLS, Kinship)^{4,5}. The input 3-way tensors are treated with a Closed World Assumption (CWA)

⁴ <https://github.com/arongdari/kg-data>.

⁵ <https://github.com/mnick/rescal.py/tree/master/data>.

[15], i.e., known (unknown) facts are considered positive (negative) examples. This makes the problem a binary tensor decomposition, and the canonical form alternative is Logistic Tensor Factorization (LTF) [16] specialized for Knowledge Graph completion: $Y_{ijl} \sim \text{Ber}(Y_{ijl}; \sigma(a_i^T R_l a_j))$ where a_i and a_j are latent feature vectors belong to i^{th} and j^{th} entities respectively, and R_l is latent matrix of l^{th} relation which defines in what way entities interact and generate the observations.

The SCPF model can be setup with $K = 2$, $x_{1,i,j,l} = Y_{ijl}$, $x_{2,i,j,l} = 1 - x_{1,i,j,l}$, and $n_{i,j,l} = 1$.

Experiment Setup and Metrics. For all datasets, We created 50%–50% random splits for training and test. LTF and 2 versions of SCPF (with ML estimation and Variational Inference) are used to make predictions on test data. Expected values of the variational distributions were used for prediction.

Knowledge Graph Completion with CWA is a classification task and the datasets are highly imbalanced in number of positive and negative observations. Area under the ROC (true-positive-rate vs. false-positive-rate) curve created by using different threshold values is used for performance evaluation.

Results. In Fig. 3, we plotted performance of SCPF with Maximum Likelihood and Bayesian inference versus Logistic Tensor Factorization with 20 different train-test splits and for each dataset. For all datasets, SCPF with maximum likelihood outperforms Logistic Tensor Factorization. Predictive performance of SCPF with variational inference is not as high as maximum likelihood, but even so it is superior to LTF in UMLS and Kinship datasets.

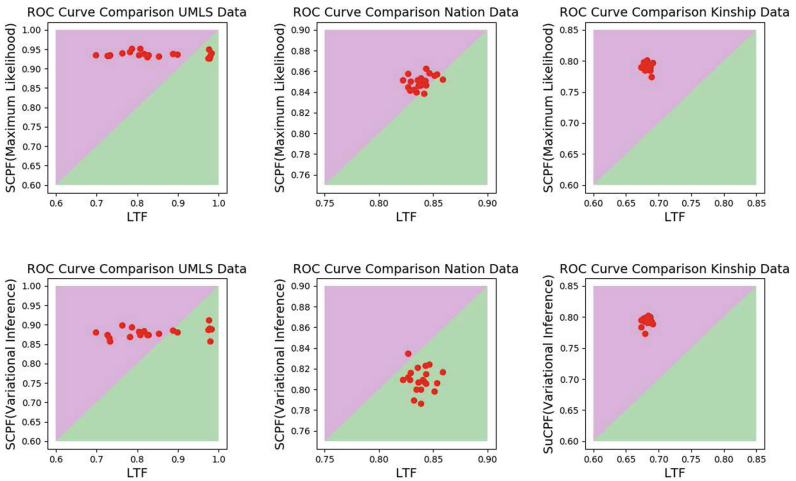


Fig. 3. Performance comparison with area under the ROC curve for Knowledge Graph completion.

LTF with Rescal factorization model is designed for Knowledge Graph problems, specifically. We used the generic SCPF with PARAFAC-type decomposition, and further improvement might be possible with different decomposition types. Considering the effectiveness of SCPF in Swimmer dataset and introduced Bayesian inference, interpretability of entity and relation latent features for Knowledge Graph datasets deserves a further study.

5 Conclusion

In this work, we discuss the moment and canonical parametrization as alternative tensor factorization models and propose a simple method that allows modeling binary, ordinal or categorical data with a moment parametrization. We demonstrate the benefits of our model, which is an extension of Poisson factorization, on various data sets: Swimmer (binary), Knowledge Graphs (binary) and MovieLens 100-K (ordinal). The models are compared with their canonical form alternatives LMF, Logistic Tensor Factorization (LTF) and OMF for binary and ordinal data.

As the toy example on the Swimmer data illustrates, the interpretability of factorization models with a moment parametrization tends to be easier when compared to canonical parametrizations, that are mostly used for binary or bounded discrete data. The experimental results indicate that the proposed models tend to outperform their canonical alternatives and PF in terms of predictive performance for test ratings, classification and top-K recommendation.

We believe that our approach provides a flexible framework for developing fully conjugate tensor decomposition models for binary, ordinal or multinomial data that can be used as a generic building block in hierarchical models for arrays of such data.

References

1. Amari, S.: Information Geometry and Its Applications. Applied Mathematical Sciences. Springer, Japan (2016). <https://doi.org/10.1007/978-4-431-55978-8>
2. Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6**, 1705–1749 (2005)
3. Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.* **2009** (2009)
4. Charlin, L., Ranganath, R., McInerney, J., Blei, D.M.: Dynamic Poisson factorization. In: Proceedings of the 9th ACM Conference on Recommender Systems, pp. 155–162. ACM (2015)
5. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: Nonnegative Matrix and Tensor Factorization. Wiley, Chichester (2009)
6. Collins, M., Dasgupta, S., Schapire, R.E.: A Generalization of Principal Component Analysis to the Exponential Family. *Advances in Neural Information Processing Systems*. MIT Press, Cambridge (2001)
7. Gopalan, P., Hofman, J.M., Blei, D.M.: Scalable recommendation with Poisson factorization. arXiv preprint [arXiv:1311.1704](https://arxiv.org/abs/1311.1704) (2013)

8. Gopalan, P., Ruiz, F.J., Ranganath, R., Blei, D.: Bayesian nonparametric Poisson factorization for recommendation systems. In: *Artificial Intelligence and Statistics*, pp. 275–283 (2014)
9. Hernández-Lobato, J.M., Hounsby, N., Ghahramani, Z.: Probabilistic matrix factorization with non-random missing data. In: *ICML*, pp. 1512–1520 (2014)
10. Hounsby, N., Hernández-Lobato, J.M., Ghahramani, Z.: Cold-start active learning with robust ordinal matrix factorization. In: *ICML*, pp. 766–774 (2014)
11. Kingman, J.F.C.: *Poisson Processes*. Oxford Science Publications, Oxford (1993)
12. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: *NIPS*, vol. 13, pp. 556–562 (2001)
13. Liang, D., Charlin, L., McInerney, J., Blei, D.M.: Modeling user exposure in recommendation. In: *Proceedings of the 25th International Conference on World Wide Web*, pp. 951–961. International World Wide Web Conferences Steering Committee (2016)
14. Mohamed, S., Ghahramani, Z., Heller, K.A.: Bayesian exponential family PCA. In: *Advances in Neural Information Processing Systems*, pp. 1089–1096 (2009)
15. Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E.: A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**(1), 11–33 (2016)
16. Nickel, M., Tresp, V.: Logistic tensor factorization for multi-relational data. arXiv preprint [arXiv:1306.2084](https://arxiv.org/abs/1306.2084) (2013)
17. Paquet, U., Thomson, B., Winther, O.: A hierarchical model for ordinal matrix factorization. *Stat. Comput.* **22**(4), 945–957 (2012)
18. Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 880–887. ACM (2008)
19. Schmidt, M.N.: Linearly constrained Bayesian matrix factorization for blind source separation. In: *NIPS* (2009)
20. Tomé, A., Schachtner, R., Vigneron, V., Puntinet, C., Lang, E.: A logistic non-negative matrix factorization approach to binary data sets. *Multidimens. Syst. Signal Process.* **26**(1), 125–143 (2015)
21. Wainwright, M., Jordan, M.I.: Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1**, 1–305 (2008)



Curve Registered Coupled Low Rank Factorization

Jeremy Emile Cohen^{1(✉)}, Rodrigo Cabral Farias², and Bertrand Rivet³

¹ Department of Mathematics and Operational Research, Faculté polytechnique, Université de Mons, Rue de Houdain 9, Mons, Belgium

jeremy.cohen@umons.ac.be

² Univ. Côte d'Azur, CNRS, I3S, 06900 Sophia-Antipolis, France

³ Univ. Grenoble Alpes, CNRS, GIPSA-lab, 38000 Grenoble, France

Abstract. We propose an extension of the canonical polyadic (CP) tensor model where one of the latent factors is allowed to vary through data slices in a constrained way. The components of the latent factors, which we want to retrieve from data, can vary from one slice to another up to a diffeomorphism. We suppose that the diffeomorphisms are also unknown, thus merging curve registration and tensor decomposition in one model, which we call registered CP. We present an algorithm to retrieve both the latent factors and the diffeomorphism, which is assumed to be in a parametrized form. At the end of the paper, we show simulation results comparing registered CP with other models from the literature.

Keywords: Tensor decompositions · Curve registration · Data fusion

1 Introduction

Joint decomposition models such as the canonical polyadic (CP) tensor decomposition [4] allow to blindly extract patterns of underlying hidden phenomena from a block of data measurements based on their algebraic properties without statistical assumptions. Thanks to their uniqueness properties under mild conditions [4], tensor decompositions have been applied in many domains: neurosciences [1], chemometrics [21] and digital communications [20] to name a few.

To retrieve the latent patterns without statistical assumptions, the number of free parameters must be rather low (*i.e.* the number of latent patterns is small with respect to the data dimensions). For example, in the CP model for a 3-way data block, $\mathcal{M} \in \mathbb{R}^3$, each slice \mathbf{M}_k in one of the dimensions is approximated by a rank R matrix decomposition: $\mathbf{M}_k = \mathbf{A} \text{Diag}(\mathbf{C}(k, :)) \mathbf{B}^T$, where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_R]$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_R]$, and $\text{Diag}(\mathbf{C}(k, :))$ is the diagonal matrix formed with the k -th row of $\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_R]$. Here the columns of these matrices are the latent patterns that we are searching for and the fundamental constraint is that the matrix factors \mathbf{A} and \mathbf{B} are exactly the same as k varies. Clearly, the model for the slices in any of the 3-ways of the CP decomposition corresponds to a coupled matrix decomposition where the \mathbf{A} and \mathbf{B} matrix factors

are shared. Other models with less stringent coupling constraints have been considered in the literature, for example, PARAFAC2 [7], Shift-PARAFAC [6, 13], soft non-negative matrix co-factorization [19] or probabilistic couplings [3].

In this paper, we are also interested in such a less constrained decomposition, where one of the matrix factors, \mathbf{B} for example, is allowed to change over the experimental parameter k : $\mathbf{M}_k = \mathbf{A}\text{Diag}(\mathbf{C}(k, :))\mathbf{B}_k^T$. The components of the factor from one slice to another are all similar up to a diffeomorphism, that is up to local compression and dilations. This can be useful, for example, in ocular-artifact removal in EEG [17] where the coupled latent signals are related to different eye blinks or saccades, or in chromatography [2] where the latent components are time elution responses of chemical compounds on different chromatographic experiments. In both examples, the patterns feature domain variations, that may differ at any given time but are similar after alignment through delay, local dilations and compressions.

Finding the diffeomorphisms, that is, the transformations of the arguments (time or space) of the latent curves, leading to an alignment is known in statistics as curve registration [15] and in signal processing as time warping [18]. In curve registration one may be interested in computing the structured average [10], *i.e.* an aligned mean curve, which serves as a template for trend analysis. In this paper, we are facing a different problem than in curve registration since the curves themselves are unknown latent functions. By merging both curve registration and CP decompositions, we expect that the factors obtained from the joint decomposition of each slice will be retrieved with an increased accuracy when compared with other methods which do not include fully the diffeomorphism coupling information, as in Shift-PARAFAC [6] and PARAFAC2 [11, 13].

In this work we propose to modify the well-known alternating least squares (ALS) algorithm for CP decomposition [4] to include a curve registration step on the factor containing domain variation. Closely related to our work, warped factor analysis (WFA) has been proposed in [8] where curve registration is explicitly carried out using a piecewise linear model for the diffeomorphism. In WFA, the template curve (*i.e.* the structural average which is used as reference) is contained directly in the data, which is a fundamental difference with the proposed approach. In our work we extend WFA (i) to a generalized diffeomorphism model, and (ii) to have a less arbitrary template curve estimated from all latent patterns by searching for a structural average curve. To retrieve this structural average curve and the optimal diffeomorphisms, we follow an alternating approach similar to [22].

Notation: Vectors are denoted in bold symbols \mathbf{a} , matrices as bold capital symbols \mathbf{M} . The (i, j) -th entry of matrix \mathbf{M} is denoted $\mathbf{M}(i, j)$, its i -th column $\mathbf{M}(:, i)$ or \mathbf{m}_i and the i -th row $\mathbf{M}(i, :)$. The transposition operator is denoted as \mathbf{M}^T . \circ is the composition operator: $(f \circ g)(\cdot) = f(g(\cdot))$.

2 Curve Registered Decomposition Models

In this section we present the curve registered decomposition model through a Bayesian estimation perspective. We present it in three steps: Sect. 2.1 develops the measurement model and its corresponding likelihood. Section 2.2 presents the registered CP derived from the maximum *a posteriori* (MAP) estimator of all unknown parameters (*i.e.* both measurement and coupling models). Finally Sect. 2.3 introduces a parametric model for the diffeomorphisms.

2.1 Measurement Model: Low-Rank Matrix Decomposition Model

Without loss of generality, we consider the data block to be a 3-way array, $\mathcal{M} \in \mathbb{R}^{I \times J \times K}$, such that K 2-way measurement arrays ($\mathbf{M}_k \in \mathbb{R}^{I \times J}$) of size $I \times J$ are available. Moreover, we suppose that each matrix is given by a rank- R factorization plus a measurement noise term:

$$\mathbf{M}_k = \sum_{r=1}^R c_{k,r} \mathbf{a}_r \mathbf{b}_{k,r}^T + \mathbf{V}_k = \mathbf{A} \text{Diag}(\mathbf{C}(k, :)) \mathbf{B}_k^T + \mathbf{V}_k \quad (1)$$

where the rank R is supposed to be known and much smaller than the dimensions I , J and K . The factor matrices \mathbf{A} , $\{\mathbf{B}_k\}_{1 \leq k \leq K}$ and \mathbf{C} are the unknown latent patterns to be retrieved and \mathbf{V}_k are noise matrices assumed to be independent from one another and with independent elements. Note that the factor matrix \mathbf{A} is shared across data slices \mathbf{M}_k . The elements $v_{ijk} = \mathbf{V}_k(i, j)$ of the noise matrices are assumed to be independent zero-mean normally distributed with a variance σ_k^2 : $p(v_{ijk}) \propto \exp\{-v_{ijk}^2/2\sigma_k^2\}$. Without further knowledge on a relationship relating factors \mathbf{B}_k , a natural way to retrieve the latent factors is through maximum likelihood estimation. This corresponds to the minimization of the cost function \mathcal{L} w.r.t. the factor matrices:

$$\mathcal{L} = \sum_{k=1}^K \frac{1}{\sigma_k^2} \|\mathbf{M}_k - \mathbf{A} \text{Diag}(\mathbf{C}(k, :)) \mathbf{B}_k^T\|_{\mathbb{F}}^2, \quad (2)$$

where $\|\cdot\|_{\mathbb{F}}$ stands for the Frobenius norm. Minimizing (2) actually corresponds to computing a low rank matrix factorization of the stacked matrices $\mathbf{M}_{1:K} = [\frac{1}{\sigma_1^2} \mathbf{M}_1, \dots, \frac{1}{\sigma_K^2} \mathbf{M}_K]$. Therefore, there is no guarantee that the retrieved patterns will be physically interpretable, since the model is not uniquely identifiable due to rotational ambiguity.

2.2 Registered CP from MAP Formulation

In what follows, factors \mathbf{B}_k are supposed to be similar in shape but with variations on their domain. For example, consider that factors \mathbf{B}_k relate to time and that they are sampled versions of continuous-time signals: $\mathbf{B}_k(r, j) = b_{k,r}(t_j)$. We assume that the sampling grid points t_j , with $j \in \{1, \dots, J\}$, are the same

for all measurement matrices and we consider a normalized time period so that $t_j \in [0, 1]$. For any of the K underlying continuous signals, domain variation can be expressed as

$$\forall(r, k) \in \llbracket 1, R \rrbracket \times \llbracket 1, K \rrbracket, \quad b_{r,k}(t) = b_r^*(\gamma_{r,k}(t)) + w_{r,k}(t), \quad (3)$$

where the functions representing the variation $\gamma_{r,k}(t)$ are diffeomorphisms from $[0, 1]$ to $[0, 1]$. They are non-decreasing functions with $\gamma_{r,k}(0) = 0$ and $\gamma_{r,k}(1) = 1$. Note that the signals $b_r^*(\cdot)$ play the role of common unknown reference shapes, and $w_{r,k}(\cdot)$ are zero mean white Gaussian processes independent for all different r and k . This perturbation in the coupling model may be understood in two ways: (1) As some prior knowledge that the coupling relationship between factors \mathbf{B}_k is not exactly a warping. (2) As a variable splitting that makes the underlying optimization problem easier to solve. Indeed, if additional constraints are imposed on factors \mathbf{B}_k , for instance nonnegativity, we will show below that the estimation process can be cast as constrained least squares problem.

For discrete time samples t_1, \dots, t_J and assuming $\gamma_{r,k}(t_j)$ are known, this approach implies that $b_{r,k}(t_j)$ are independent Gaussian random variables $b_{r,k}(t_j) \sim \mathcal{N}(b_r^*(\gamma_{r,k}(t_j)), \sigma_w^2)$, where σ_w^2 is a known variance. With this prior, criterion (2) can be modified to obtain the following MAP cost function:

$$\mathcal{C} = \mathcal{L} + \frac{1}{\sigma_w^2} \sum_{r,k,j} \left[\mathbf{B}_k(r, j) - b_r^*(\gamma_{r,k}(t_j)) \right]^2, \quad (4)$$

where the coupling term is introduced by the prior. The minimum of \mathcal{C} over all parameters yield the proposed model, coined Registered CP. The main difference with (2) is that the additional constraints are expected to solve the rotational ambiguity intrinsic to matrix factorizations.

It is worth noting that:

- **CP model:** If $\gamma_{r,k}(\cdot)$ are identity and if $\sigma_w^2 \rightarrow 0$, then the model becomes a CP model obtained by stacking matrices \mathbf{M}_k along a third dimension.
- **Indeterminacy:** An indeterminacy remains in determining canonical $b_r^*(\cdot)$ and $\gamma_{r,k}(\cdot)$, since for any given r one can apply a common warping to all $b_{r,k}(\cdot)$ and obtain a different $b_r^*(\cdot)$: $b_{r,k} = (b_r^* \circ \gamma^{-1}) \circ (\gamma \circ \gamma_{r,k})$. In other words, diffeomorphisms $\gamma_{r,k}$ can only be obtained up to a common diffeomorphism.
- **Linear interpolation:** In theory $b_{r,k}$, b_r^* and $\gamma_{r,k}$ are functions of continuous time. In practice we work with discrete time. This means exact time transformations $b_r^*(\gamma_{r,k}(t))$ are not actually computed. Rather, transformed functions are obtained through linear interpolation.

2.3 Parametric Model for the Diffeomorphisms

In their non-parametric continuous-time form, the diffeomorphisms $\gamma_{r,k}(t)$ cannot be handled numerically. While it is possible to use dynamic programming to process these diffeomorphisms as non-parametric functions [16, 22], this is typically very sensitive to the noise and time consuming, specially if the dataset is large.

Therefore, to simplify, we assume that these functions can be modeled with a parametric form, with a small number of parameters. Multiple parametrized forms for these functions exist. Here we focus on exponential maps.

Exponential Maps: Since $\gamma_{r,k}(t)$ are also cumulative distribution functions, they can be defined through their derivatives, which are probability density functions, *i.e.* they are positive and sum to one. We can define easily such functions by applying the exponential map to any function $\phi_{r,k}(t)$ defined on $[0, 1]$. This approach is commonly found in curve registration [9,15] and it is also referred as the log-derivative approach [12]. It leads to the following diffeomorphism:

$$\gamma_{r,k}(t) = \left(\int_0^t e^{\phi_{r,k}(s)} ds \right) / \left(\int_0^1 e^{\phi_{r,k}(s)} ds \right). \quad (5)$$

The main purpose of this representation is that we can parametrize the functions $\phi_{r,k}(t)$ without imposing monotonicity constraints. In particular, we can assume that all $\phi_{r,k}(t)$ are linear combinations of n functions $\psi_i(t)$:

$$\phi_{r,k}(t) = \phi(t, \beta_{r,k}) = \sum_{i=1}^n \beta_{r,k}^i \psi_i(t). \quad (6)$$

where $\beta_{r,k} = [\beta_{r,k}^1 \cdots \beta_{r,k}^N]^\top$ is the vector of parameters characterizing the diffeomorphism. The following particular cases are of interest:

B-splines: Function $\psi_i(t)$ can be a B-splines with a fixed number of knots and degree.

Linear: If a simple linear function is used, with $n = 1$ and $\psi_1 = -t$, then the diffeomorphisms are

$$\gamma_{r,k}(t) = \frac{1 - e^{-\beta_{r,k}t}}{1 - e^{-\beta_{r,k}}}. \quad (7)$$

Constant: If we use a 0-th order B-splines basis then we obtain the parametrization used implicitly in [8].

3 Algorithm

This section describes the alternating algorithm approach to obtain the Registered CP model.

3.1 Multiway Array Decomposition Algorithm

Given a previous update or guess of $\mathbf{b}_r^*(\gamma_{k,r})$, one can minimize w.r.t. \mathbf{A}, \mathbf{C} in an alternating approach using standard linear least squares, while factor \mathbf{B}_k can be retrieved by solving the following least squares problem:

$$\mathbf{B}_k = \underset{\mathbf{B}=[\mathbf{b}_1, \dots, \mathbf{b}_r]}{\operatorname{argmin}} \left\| \mathbf{M}_k - \mathbf{A} \mathbf{D}_k \mathbf{B}^T \right\|_F^2 + \lambda_k \sum_{r=1}^R \left\| \mathbf{b}_{r,k} - \mathbf{b}_r^*[\gamma_{r,k}] \right\|_F^2, \quad (8)$$

where $\mathbf{b}_r^*[\gamma_{r,k}]$ stands for $\mathbf{b}_r^*(\gamma_{r,k}(t))$ taken at sampled times points t_i using linear interpolation

3.2 Shape Alignment Using Exponential Maps

From this point onwards, the diffeomorphisms $\gamma_{r,k}$ are assumed to be well modelled as the previously introduced exponential maps $\gamma_{r,k}(t) = (1 - e^{-\beta_{r,k}t}) / (1 - e^{-\beta_{r,k}})$. Given previous update of latent factors $\mathbf{b}_{r,k}$, what needs to be estimated are both the values of $\beta_{r,k}$ and the underlying \mathbf{b}_r^* . Thus, the following optimization problem needs to be solved for all r :

$$\operatorname{argmin}_{\{\beta_{r,k}\}_k, \mathbf{b}_r^*} \sum_k \frac{1}{\sigma_w^2} \left\| \mathbf{b}_{r,k} - \mathbf{b}_r^*[\gamma_{r,k}] \right\|_F^2. \quad (9)$$

Since estimating both the structured mean and $\gamma_{r,k}(t)$ is cumbersome, as suggested in [22], an alternating strategy is used. The following can be used independently as a very simple alignment algorithm summarized in Algorithm 1:

1. Structure mean estimation \mathbf{b}_r^* : Given the values of $\beta_{r,k}$, the structured averages \mathbf{b}_r^* are computed as the solutions of linear systems, namely

$$\mathbf{b}_r^* = \operatorname{argmin}_{\mathbf{b}} \sum_k \left\| \mathbf{b}_{r,k} - \mathbf{P}_{r,k} \mathbf{b} \right\|_F^2, \quad (10)$$

where $\mathbf{P}_{r,k}$ is the interpolation matrix obtained by linear interpolation from the sampling grid $[t_j]_j$ to the warped sampling grid $[\gamma_{r,k}(t_j)]_j$.

2. Warping parameters estimation: Given \mathbf{b}_r^* , the criterion (9) becomes K one dimensional problems. And even through it is highly non-convex in the general case, good values of $\beta_{r,k}$ can be computed using a grid search. Multiple strategies can then be used to refine the search space once convergence is achieved and we used in particular the Golden Search method [14]. In both cases, the cost of one evaluation is rather low since computing (9) requires a linear interpolation and $K \times R \times J$ multiplications, but evaluating the cost on a grid can be time consuming.

This algorithm should converge since the cost is reduced at each iteration and for each block of parameters. However, we cannot guaranty that the final estimate is a local minimum of the cost function.

3.3 Detailed 3-Way Algorithm

Joining the alternating least squares update of factors \mathbf{A} , \mathbf{B}_k , and \mathbf{C} with the alignment algorithm (Algorithm 1) leads to Algorithm 2, which is given below along with some implementation details. It can be easily adapted for constrained Registered CP by replacing the least squares solver with a constrained one: e.g., for nonnegative least squares, one can use the algorithm described in [5].

Initialization: Due to the highly non-convex behavior of the cost function w.r.t. $\beta_{r,k}$, a good initialization method is required. As a reasonable option, we used the factors given by a standard CP model fitting. Moreover, the initial values of λ_k are also very important, since large values put too much emphasis on

Algorithm 1. Alignment algorithm under parametrized diffeomorphisms.

Input: Initial target \mathbf{b}^* , initial warping parameters β_k , similar-shaped functions $\{\mathbf{b}_k\}_k$, regularization parameters $\{\lambda_k\}_k$.

while residual $\sum_k \lambda_k \|\mathbf{b}_k - \mathbf{b}^*[\gamma_k]\|_F$ is too large **do**

Structure mean estimation: set \mathbf{b}^* as either the

1. $\mathbf{b}_k[\gamma_k^{-1}]$ that minimized the residuals (first iteration)
2. the solution to (10)
3. initial target \mathbf{b}^* (inside a larger optimization scheme)

Warping parameters estimation: $\forall k$

if Residuals are higher than some threshold (coarse estimation) **then**

Compute criterion (9) on a grid to define an interval $[a_k, b_k]$ surrounding the optimum.

else

Find the optimal β_k in interval $[a_k, b_k]$ using Golden Search.

end if

end while

Output: Estimated warping parameters $\{\beta_k\}_k$ and structured mean \mathbf{b}^* .

Algorithm 2. Alternating least squares algorithm for Registered CP under parametrized diffeomorphisms.

Input: Data matrices $\{\mathbf{M}_k\}_k$, initial guesses \mathbf{A} , \mathbf{C} , $\{\mathbf{B}_k\}_k$, initial $\{\lambda_k\}_k$ values.

while Stopping criterion is not met **do**

- Solve $\operatorname{argmin}_{\mathbf{A}} \sum_k \|\mathbf{M}_k - \mathbf{A} \mathbf{D}_k \mathbf{B}_k^T\|_F^2$ and normalize column-wise with the ℓ_2 norm $\Rightarrow \mathbf{A}$
- $\forall k$, solve $\operatorname{argmin}_{\mathbf{D}} \|\mathbf{M}_k - \mathbf{A} \mathbf{D} \mathbf{B}_k^T\|_F^2$, $\Rightarrow \{\mathbf{D}_k\}_k$
- $\forall k$, solve optimization problem (8) and normalize column-wise with the ℓ_∞ norm, $\Rightarrow \{\mathbf{B}_k\}_k$
- Use Algorithm 1 to align the previously estimated $\{\mathbf{B}_k\}_k$, $\Rightarrow \mathbf{B}^*$ and $\{\beta_{r,k}\}_{r,k}$
- If necessary, increase the regularization parameters $\Rightarrow \{\lambda_k\}_k$

end while

Output: Estimated factors \mathbf{A} , $\{\mathbf{B}_k\}_k$ and \mathbf{C} , coupling parameters \mathbf{B}^* and $\{\beta_{r,k}\}_{r,k}$.

the regularization terms, which implies factors B_k not change much and the algorithm mostly fits \mathbf{A} and \mathbf{C} . Empirically, we used the following values for the values of λ_k at the first and second iterations:

$$\lambda_k^0 = 10^{-\frac{\text{SNR}}{10}} \frac{\|M_k - A^0 D_k^0 B_k^{0T}\|_F^2}{\|B_k^0\|_F^2} \quad \text{and} \quad \lambda_k^1 = 10^{-\frac{\text{SNR}}{10}} \frac{\|M_k - A^1 D_k^1 B_k^{1T}\|_F^2}{\|B_k^1 - B^{1*}[\Gamma_k]\|_F^2} \quad (11)$$

where A^0 is the initial value of A , A^1 is the estimate of A after the first iteration, $B^*[\Gamma_k]$ is a matrix containing stacked $b_r^*[\gamma_{r,k}]$ and SNR refers to the expected Signal to Noise ratio of the whole tensor data. We used λ_k^1 in all following iterations.

Normalization: Columns of \mathbf{A} are normalized with ℓ_2 norm, while the columns of \mathbf{B}_k are normalized with ℓ_∞ norm.

Case $\gamma_{kr} = \gamma_k$ for all r : It may happen that all components in \mathbf{B}_k have the same warping, for instance when the variability generating process affects the data uniformly across the sensors. Such an hypothesis is actually exploited also in [8] and is an underlying hypothesis of PARAFAC2. Formally, with parametrized diffeomorphisms, this means that $\beta_{kr} = \beta_k$ for all r . Then the alignment algorithm can be slightly modified to improve estimation accuracy since the number of parameters is reduced.

4 Experiments on Simulated Nonnegative Data

In this section, the Registered CP model is tested on simulated nonnegative data and compared with similar state-of-the-art models, namely the Shift PARAFAC model and the PARAFAC2 model. Many data alignment models have been proposed in the literature, but only those two models align the factors directly inside the optimization process.

Simulation Settings: After setting the rank R , factors \mathbf{A} and \mathbf{C} are drawn entry-wise from uniform distributions over $[0, 1]$. A latent factor \mathbf{B}^* is generated column-wise using the exponential map, which mode is randomly determined but so that all R modes do not overlap. The variances are also randomly determined. Then, $\beta_{k,r}$ are chosen using affine functions of the k variable with random slope depending on the r variable. Thus each component has its own warping range. Finally, the \mathbf{B}_k are generated from \mathbf{B}^* using exponential maps of parameters $\beta_{k,r}$. Additive Gaussian noise variance is determined from a user-defined SNR using $\sigma_n^k = \sqrt{R}10^{-\frac{\text{SNR}}{20}}$.

In the following experiment, the total reconstruction error ε_B on \mathbf{B}_k

$$\varepsilon_B = \left(\sum_{k=1}^K \|\mathbf{B}_k - \Pi_k \widehat{\mathbf{B}}_k\|_F^2 \right) / \left(\sum_{k=1}^K \|\mathbf{B}_k\|_F^2 \right) \quad (12)$$

is monitored over $N = 50$ experiments. Π_k is the best permutation that matches columns of the estimated $\widehat{\mathbf{B}}_k$ with the true \mathbf{B}_k . Note that in (12), the \mathbf{B}_k matrices are normalized column-wise using the ℓ_2 norm. The rank is set to $R = 3$ and data dimensions are $15 \times 200 \times 10$. The Registered CP algorithm is initialized by the result of 100 iterations of standard alternating least squares.

Figure 1 shows ε_B for several SNR values and the various mentioned algorithms. Although PARAFAC2 algorithm should not perform well since it relies on the assumption that $\gamma_{k,r} = \gamma_k$ for all r , it outperforms both the Shift-PARAFAC and the Registered CP model at high SNR values. However, on average, the Registered CP performs the best for medium and low SNR values. All the algorithms feature a high variability in their outputs, thus indicating a high sensibility to the initialization. The fact that PARAFAC2 uses the best of several initializations is probably the reason why it performs best at high SNR.

In order to study the dependence of ε_B on regularization parameters λ for Registered CP, in a second experiment, instead of the values suggested in

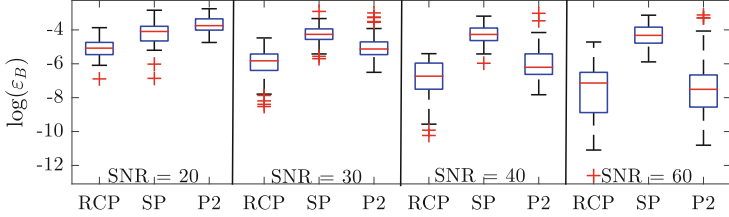


Fig. 1. $\log(\varepsilon_B) \times \text{SNR}$: RCP - registered CP, SP - shift PARAFAC, P2 - PARAFAC2.

Eq. (11), we fixed values $\text{SNR} = 40$ or $\text{SNR} = 60$ gridded over a multiplicative coefficient ρ in front of the initial λ_k values:

$$\lambda_k^0 = \rho \frac{\|M_k - A^0 D_k^0 B_k^{0T}\|_F^2}{\|B_k^0\|_F^2} \text{ and } \lambda_k^1 = \rho \frac{\|M_k - A^1 D_k^1 B_k^{1T}\|_F^2}{\|B_k^1 - B_k^{1*}[\Gamma_k]\|_F^2}. \quad (13)$$

Figure 2 shows the obtained results for $N = 25$ realizations. It can be observed that finding a good set of regularization parameters is important to obtain better results on average. The good performance of the uncoupled matrix factorization algorithm (regularization set to 0) is due to the nonnegativity constraints applied on all factors. Nevertheless, using the Registered CP model, estimation performances on the \mathbf{B}_k are improved at both $\text{SNR} = 40$ and 60. Variability however seems to increase alongside the amount of regularization.

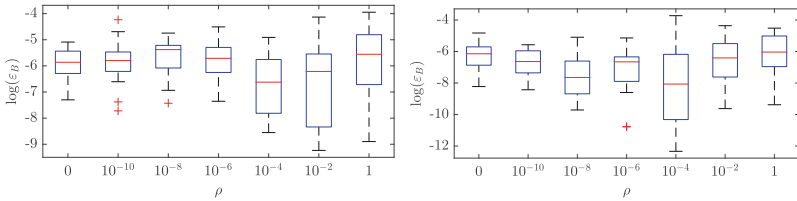


Fig. 2. $\log(\varepsilon_B) \times \rho$: Left - $\text{SNR} = 40$, right - $\text{SNR} = 60$.

5 Conclusion

A new coupled tensor decomposition model is introduced, namely the Registered CP model, where factors on one mode are similar up to time contraction or dilatation. A specific class of diffeomorphisms is used to generate a decomposition algorithm that can identify both the factors and the latent coupling parameters. Simulations on synthetic data show encouraging results, but the Registered CP model is yet to be tested on actual data sets. Furthermore, in future works, the class of allowed diffeomorphisms should be enlarged.

References

1. Becker, H., Albera, L., Comon, P., Gribonval, R., Wendling, F., Merlet, I.: Brain source imaging: from sparse to tensor models. *IEEE Signal Process. Mag.* **32**(6), 100–112 (2015)
2. Bro, R., Andersson, C.A., Kiers, H.A.L.: PARAFAC2-Part II. Modeling chromatographic data with retention time shifts. *J. Chemom.* **13**(3–4), 295–309 (1999)
3. Cabral Farias, R., Cohen, J.E., Comon, P.: Exploring multimodal data fusion through joint decompositions with flexible couplings. *IEEE Trans. Signal Process.* **64**(18), 4830–4844 (2016)
4. Comon, P., Luciani, X., De Almeida, A.L.F.: Tensor decompositions, alternating least squares and other tales. *J. Chemom.* **23**(7–8), 393–405 (2009)
5. Gillis, N., Glineur, F.: Accelerated multiplicative updates and hierarchical als algorithms for nonnegative matrix factorization. *Neural Comput.* **24**(4), 1085–1105 (2012)
6. Harshman, R.A., Hong, S., Lundy, M.E.: Shifted factor analysis–Part I: models and properties. *J. Chemom.* **17**(7), 363–378 (2003)
7. Harshman, R.A.: PARAFAC2: mathematical and technical notes. UCLA working papers in phonetics, vol. 22, no. 3044, p. 122215 (1972)
8. Hong, S.: Warped factor analysis. *J. Chemom.* **23**(7–8), 371–384 (2009)
9. James, G.M.: Curve alignment by moments. *Ann. Appl. Stat.* **1**, 480–501 (2007)
10. Kneip, A., Gasser, T.: Statistical tools to analyze data representing a sample of curves. *Ann. Stat.* **20**(3), 1266–1305 (1992)
11. Marini, F., Bro, R.: Scream: a novel method for multi-way regression problems with shifts and shape changes in one mode. *Chemom. Intell. Lab.* **129**, 64–75 (2013)
12. Marron, J.S., Ramsay, J.O., Sangalli, L.M., Srivastava, A.: Functional data analysis of amplitude and phase variation. *Stat. Sci.* **30**(4), 468–484 (2015)
13. Mørup, M., Hansen, L.K., Arnfred, S.M., Lim, L.-H., Madsen, K.H.: Shift-invariant multilinear decomposition of neuroimaging data. *NeuroImage* **42**(4), 1439–1450 (2008)
14. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. Cambridge University Press, Cambridge (2007)
15. Ramsay, J.O., Li, X.: Curve registration. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**(2), 351–363 (1998)
16. Rivet, B., Cohen, J.E.: Modeling time warping in tensor decomposition. In: *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2016)*, pp. 1–5. IEEE (2016)
17. Rivet, B., Duda, M., Guérin-Dugué, A., Jutten, C., Comon, P.: Multimodal approach to estimate the ocular movements during EEG recordings: a coupled tensor factorization method. In: *Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6983–6986. IEEE (2015)
18. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech Signal Process.* **26**(1), 43–49 (1978)
19. Seichepine, N., Essid, S., Févotte, C., Cappé, O.: Soft nonnegative matrix co-factorization. *IEEE Trans. Signal Process.* **62**(22), 5940–5949 (2014)
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: Applications in the Chemical Sciences*. Wiley, Chichester (2005)
22. Srivastava, A., Wu, W., Kurtek, S., Klassen, E., Marron, J.S.: Registration of functional data using Fisher-Rao metric. *arXiv preprint [arXiv:1103.3817](https://arxiv.org/abs/1103.3817)* (2011)



Source Analysis and Selection Using Block Term Decomposition in Atrial Fibrillation

Pedro Marinho R. de Oliveira^(✉) and Vicente Zarzoso

Université Côte d'Azur, CNRS, I3S Laboratory, CS 40121,
06903 Sophia Antipolis Cedex, France
{marinho,zarzoso}@i3s.unice.fr

Abstract. Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia in clinical practice, and is becoming a major public health concern. To better understand the mechanisms of this arrhythmia an accurate analysis of the atrial activity (AA) signal in electrocardiogram (ECG) recordings is necessary. The block term decomposition (BTD), a tensor factorization technique, has been recently proposed as a tool to extract the AA in ECG signals using a blind source separation (BSS) approach. This paper makes a deep analysis of the sources estimated by BTD, showing that the classical method to select the atrial source among the other sources may not work in some cases, even for the matrix-based methods. In this context, we propose two new automated methods to select the atrial source by considering another novel parameter. Experimental results on ten patients show the validity of the proposed methods.

Keywords: Atrial source selection · Block term decomposition
Atrial fibrillation · Blind source separation

1 Introduction

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia in clinical practice, responsible for up to 25% of strokes and 1/3 of the hospitalizations due to cardiac related disturbances [1]. This arrhythmia is becoming a major public health concern, since about 160 000 new AF cases are discovered every year only in USA, with similar numbers in European countries. This makes AF an increasingly prevalent disease that could become a new epidemic over the years [2]. The mechanisms of this supraventricular arrhythmia are not completely understood, making AF a challenging cardiac condition, considered as the last great frontier of cardiac electrophysiology. During AF, electrical impulses typically generated around the pulmonary veins propagate in a chaotic and irregular way across the atria, replacing the P wave, that corresponds to a normal

P. M. R. de Oliveira—Funded by a Ph.D. scholarship from the IT Doctoral School of the Université Côte d'Azur.

V. Zarzoso—Member of the *Institut Universitaire de France*.

atrial activation, by low-amplitude fibrillatory waves, or f-waves. The f-waves are present through all the electrocardiogram (ECG) recording, but masked by the QRS complex of ventricular activity (VA) in each heartbeat.

To better understand the mechanisms of AF, it is necessary an accurate analysis of the atrial activity (AA), specifically, the f waves. A noninvasive analysis can be made by extracting the AA from the cardiac signals recorded by the standard 12-lead ECG using matrix decompositions techniques for blind source separation (BSS), such as principal component analysis (PCA) and independent component analysis (ICA) [3–5]. This matrix decomposition approach has proven to be useful for AA extraction. However, it has some limitations, since constraints need to be imposed to guarantee the uniqueness of such decompositions, e.g., orthogonality or statistical independence between the sources. Although mathematically coherent, such constraints may lack physiological grounds.

In order to overcome these limitations, a tensor approach has recently been proposed to analyze AF signals [6–9]. As compared to matrix techniques, tensor decompositions present some remarkable features such as essential uniqueness with practically minimal or no constraints. The block term decomposition (BTM) proposed in [12] suits the characteristics of AA in an AF signal, since atrial signals can be approximated by all-pole models and mapped onto Hankel matrices with rank equal to the number of poles [9]. These Hankel matrices that contain the ECG data are stacked in the third dimension of a 3rd-order tensor, and then processed by BTM. Previous experimental results in synthetic and real ECG data showed the potential superiority of BTM as compared to matrix decompositions for short ECG recordings [6–8].

The success of the BSS approach to AA extraction depends on the accurate identification of the atrial signal among the estimated sources. The classical method for atrial source selection consists in selecting the source with the highest spectral concentration (SC) among the sources whose dominant frequency (DF) lies between 3 and 9 Hz [3,4]. The present work makes a deep analysis in the sources estimated by BTM, showing that the classical method may not work in some cases, even when the matrix-based approach is used. Taking this into account, a new parameter to improve the performance of the classical method is proposed. This parameter consists in analyzing the power of the source contribution to the lead V1, a lead that significantly reflects AA. Also, a new automated method for atrial source selection is proposed, using the proposed parameter and another one based on the kurtosis of the signal in the frequency domain. Experimental results using ten patients with persistent AF evaluate the proposed methods, showing their better performance in selecting the atrial source among the sources estimated by BTM and two matrix-based methods previously proposed in literature for AA extraction: RobustICA-f [10] and PCA [11]. It is also shown that BTM can provide a better estimation of the AA signal, outperforming the matrix-based techniques in most cases.

The rest of this paper is organized as follows. Section 2 introduces the notation used in the work. Section 3 recalls the BTM as a tensor approach to solve BSS problems, while Sect. 4 discusses the estimated sources and the atrial source

selection methods. Section 5 presents the experimental results and, finally, Sect. 6 formulates the conclusion of this work, as well as the prospects for future works.

2 Notations

Scalars, vectors, matrices and tensors are represented by lower-case (a, b, c, \dots), boldface lower-case ($\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$), boldface capital ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$) and calligraphic ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$) letters, respectively.

The transpose is represented by $(\cdot)^T$, symbol $\|\cdot\|$ represents the l_2 -norm and \circ represents the outer product. The operator $\text{diag}(\cdot)$ builds a diagonal matrix by placing its arguments along the diagonal. Given a 3rd-order tensor $\mathcal{A} \in \mathbb{C}^{I_1 \times I_2 \times I_3}$, with scalars a_{i_1, i_2, i_3} , its frontal slices are represented by $\mathbf{A}_{\cdot i_3} \in \mathbb{C}^{I_1 \times I_2}$. Given a matrix $\mathbf{A} \in \mathbb{C}^{I_1 \times I_2}$, with scalars a_{i_1, i_2} , its i_1^{th} row and the i_2^{th} column are represented by \mathbf{a}_{i_1} and \mathbf{a}_{i_2} , respectively.

3 Block Term Decomposition

The BTM of an arbitrary 3rd-order tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is written as

$$\mathcal{T} = \sum_{r=1}^R \mathbf{E}_r \circ \mathbf{c}_r, \quad (1)$$

with $\mathbf{c}_r \in \mathbb{R}^{I_3}$. Matrix $\mathbf{E}_r \in \mathbb{R}^{I_1 \times I_2}$ has rank L_r and admits a decomposition $\mathbf{E}_r = \mathbf{A}_r \mathbf{B}_r^T$, where $\mathbf{A}_r \in \mathbb{R}^{I_1 \times L_r}$ and $\mathbf{B}_r \in \mathbb{R}^{I_2 \times L_r}$ have rank L_r . We may then rewrite (1) as

$$\mathcal{T} = \sum_{r=1}^R \left(\mathbf{A}_r \mathbf{B}_r^T \right) \circ \mathbf{c}_r. \quad (2)$$

One can see that the BTM is a decomposition of \mathcal{T} in multilinear rank- $(L_r, L_r, 1)$ terms. If the matrix factors $\mathbf{A} = [\mathbf{A}_1 \mathbf{A}_2 \dots \mathbf{A}_R] \in \mathbb{R}^{I_1 \times \sum_{r=1}^R L_r}$ and $\mathbf{B} = [\mathbf{B}_1 \mathbf{B}_2 \dots \mathbf{B}_R] \in \mathbb{R}^{I_2 \times \sum_{r=1}^R L_r}$ are full-column rank, which requires that $\sum_{r=1}^R L_r \leq I_1, I_2$, and $\mathbf{C} = [\mathbf{c}_1 \mathbf{c}_2 \dots \mathbf{c}_R] \in \mathbb{R}^{I_3 \times R}$ does not contain proportional columns, then the BTM is essentially unique [12, Theorem 2.2]. Milder uniqueness conditions are presented in [12].

The ECG data matrix, with K leads and N samples, can be modeled as

$$\mathbf{Y} = \mathbf{M} \mathbf{S} \in \mathbb{R}^{K \times N}, \quad (3)$$

where $\mathbf{M} \in \mathbb{R}^{K \times R}$ is the mixing matrix, modelling the propagation of the cardiac electrical sources from the heart to the body surface, $\mathbf{S} \in \mathbb{R}^{R \times N}$ is the source matrix that contains the atrial, ventricular and noise sources, and R is the number of sources [5]. The AA extraction in an AF ECG recording can be seen as a BSS problem, since the only data observed is matrix \mathbf{Y} , and we aim to estimate \mathbf{M} and \mathbf{S} from it. In [12], the BTM is proposed as a solution of

a BSS problem like (3), but does not deal with the AA extraction specifically. The idea to obtain a tensor from \mathbf{Y} is to map its k^{th} row onto a Hankel matrix $\mathbf{H}_{\mathbf{Y}}^{(k)} \in \mathbb{R}^{I \times J}$, where $I = J = \frac{N+1}{2}$ if N is odd or $I = \frac{N}{2}$ and $J = \frac{N}{2} + 1$ if N is even, with

$$[\mathbf{H}_{\mathbf{Y}}^{(k)}]_{i,j} \triangleq y_{k,i+j-1}, \quad (4)$$

where $i = 1, \dots, I$, $j = 1, \dots, J$, and $k = 1, \dots, K$. Next, the tensor is built by stacking each Hankel matrix along the third dimension (as frontal slices) of a 3rd-order tensor $\mathcal{Y} \in \mathbb{R}^{I \times J \times K}$, that is

$$\mathbf{Y}_{..k} = \mathbf{H}_{\mathbf{Y}}^{(k)}. \quad (5)$$

The k^{th} matrix slice of the tensor \mathcal{Y} can be represented as

$$\mathbf{Y}_{..k} = \sum_{r=1}^R m_{k,r} \mathbf{H}_{\mathbf{S}}^{(r)}, \quad (6)$$

where $\mathbf{H}_{\mathbf{S}}^{(r)} \in \mathbb{R}^{I \times J}$ is a Hankel matrix built from the r^{th} row of \mathbf{S} . We can see that for each r , the outer product between matrix $\mathbf{H}_{\mathbf{S}}^{(r)}$ and the r^{th} column of \mathbf{M} , i.e., $\mathbf{m}_{.,r}$, is being performed. This way, the tensor \mathcal{Y} can be written as

$$\mathcal{Y} = \sum_{r=1}^R \mathbf{H}_{\mathbf{S}}^{(r)} \circ \mathbf{m}_{.,r}. \quad (7)$$

Comparing Eq.(1) with (7), we can conclude that the tensor ECG data follows a BTB tensor model.

During AF, the AA presents certain harmonicity. Hence, atrial sources can plausibly be represented by the exponential (or all-pole) model as

$$s_{r,n} = \sum_{\ell=1}^{L_r} \lambda_{\ell,r} z_{\ell,r}^{n-1}, \quad (8)$$

where $n = 1, \dots, N$, $r = 1, \dots, R$, L_r is the number of exponential terms, $z_{\ell,r}$ is the ℓ^{th} pole of the r^{th} source, and $\lambda_{\ell,r}$ is the scaling coefficient [6–9]. This way, their associated Hankel matrix accepts the Vandermonde decomposition as in [13].

4 Atrial Source Selection

4.1 Classical Method

To select the AA signal or the source with the most significant AA activity, the classical method considers two parameters. The first one is the DF, that is, the position of the peak frequency in the power spectral density, since the AA during

AF typically has a peak between 3 and 9 Hz. The second parameter, called SC, is the relative amount of energy around the DF, and it is calculated as:

$$SC = \frac{\sum_{0.82f_p}^{1.17f_p} P_{AA}(f_i)}{\sum_0^{F_s/2} P_{AA}(f_i)}, \quad (9)$$

where f_p is the value of DF, F_s is the sampling frequency and P_{AA} is the power spectrum of the AA signal, estimated as in [4]. In this work, the SC is calculated over the first harmonic (fundamental frequency) only.

The classical method of atrial source selection makes the assumption that the atrial source is concentrated in a single source only. This method consists of the following steps:

1. Select all the estimated sources with DF between 3 and 9 Hz. We refer to sources fulfilling this condition as *potential atrial sources*.
2. Select the potential atrial source with the highest SC.

4.2 Proposed Method 1

In the literature, the classical automated method described above has been used to detect the atrial source among the other estimated sources. However, in some cases, this method may not precisely select the atrial source, as will be illustrated later in this work. In Figs. 3 and 4, for example, the atrial source does not correspond to the potential source with the highest SC, despite the fact that they have close values of SC at very close DF positions.

Aiming at a better estimation of the AA signal, this paper proposes two new parameters. The first new parameter is the power contribution to the recording, which is given by

$$P(r) = \frac{1}{N} \|m_r^{(V1)} \mathbf{s}_r\|^2, \quad (10)$$

in mV^2 , where $m_r^{(V1)}$ is the contribution of the r^{th} source to lead V1 (given by the corresponding element of the estimated mixing matrix) and \mathbf{s}_r is the r^{th} source, corresponding to the r^{th} row of matrix \mathbf{S} in Eq. (3). Using the power contribution to the recording as a new parameter the classical method becomes:

1. Select all the estimated sources with DF between 3 and 9 Hz (potential atrial sources).
2. Select all the potential atrial sources with power contribution higher than 10^{-4} mV^2 . We refer to this subset of sources as *likely atrial sources*.
3. Select the likely atrial source with the highest SC.

Selecting the sources with power contribution higher than 10^{-4} mV^2 is needed in order to eliminate all sources that may present AA-like signature but are actually too weak to represent AA components. This threshold is chosen based on initial experiments that showed that sources with power contribution lower than 10^{-4} mV^2 do not present any significant contribution to the original signal. The power contribution is calculated in lead V1 due to the fact that this lead typically reflects AA best in AF ECGs, as its exploring electrode lies close to the right atrium.

4.3 Proposed Method 2

In order to better select the source with the most significant AA content among the other estimated sources, a new automated method is now proposed. The first two steps of this method are the same as those of the proposed method introduced in the previous subsection. The third and last step of this new method is to compute the kurtosis, denoted K , of the signal in the frequency domain acquired by a 4096-point FFT (the second new parameter). As in [10], we use the general expression of kurtosis valid for non-circular complex data. The likely atrial source with the highest kurtosis is related as the atrial source.

In the experiments below, it will be shown that selecting the source with the highest kurtosis provides a better performance than selecting the source with the highest SC. A possible explanation is that kurtosis is computed from the whole signal, while SC is only computed around the DF. Recall that AA in AF is typically a harmonic signal, characterized by a sparse frequency spectrum with few values significantly different from zero. Kurtosis is a measure of peakedness and sparsity of a distribution and, when computed in the frequency domain, it naturally provides a quantitative measure of harmonicity of the signal. Also, kurtosis is parameter free, whereas SC depends on the DF and the definition of a suitable interval for interpretation.

5 Experimental Results

5.1 Real AF ECG Data and Preprocessing

The recordings used in our experiments belong to a database provided by the Cardiology Department of the Princess Grace Hospital Center, Monaco. These recordings were acquired at a 977 Hz sampling rate and preprocessed by a zero-phase forward-backward type-II Chebyshev bandpass filter with cutoff frequencies of 0.5 and 40 Hz, to suppress high-frequency noise and baseline wandering. To analyze the potential atrial sources, we consider a randomly selected heart-beat (QRS-T complex + TQ segment) of a standard 12-lead ECG recording from a persistent AF patient. A single-beat segment of this patient is shown in Fig. 1, where we can see the TQ interval just after the QRS-T complex in lead II. The beat from this patient used to analyze the potential atrial sources is chosen randomly and has 1300 samples.

To assess the atrial selection methods, a population of 10 patients with persistent AF is used in the same way previously described. Similarly, one beat from each of the ten patients is chosen randomly to evaluate atrial source selection performance. The lengths of the chosen beats is between 1000 and 1400 samples (1.02 and 1.43 s, respectively). Due to lack of space, the potential source analysis of all ten patients is not reported in this paper. So only the first patient of the observed population was chosen for source analysis.

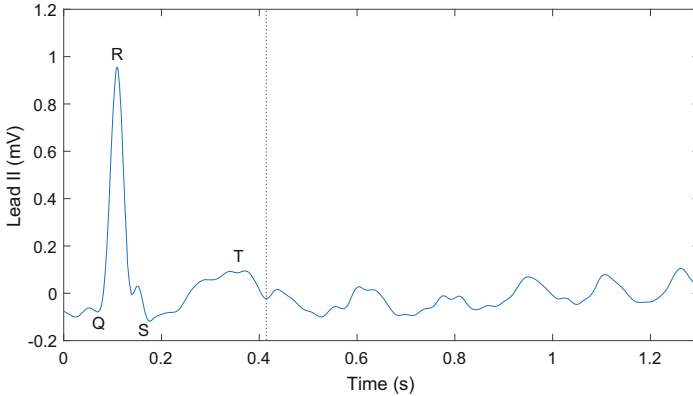


Fig. 1. A single-beat segment of an AF ECG recording of one patient in lead II. The vertical line marks the location of the T-wave offset.

5.2 BTD Setup

The BTD is implemented using the non-linear least squares (NLS) method available in Tensorlab MATLAB toolbox [14] choosing $R = 12$ and $L_r = 95$, for $r = 1, 2, \dots, 12$. This choice is made based on the work [9], which showed that these values provided good results for the heartbeat with the largest TQ segment of one of the patients in the present observed population. The tolerance threshold for convergence is set to 10^{-9} and the maximum number of iterations is set to 1000. BTD is known to be dependent on a suitable initialization of its factors. The experiments reported in this section evaluate the influence of BTD factors initialization on source estimation performance and atrial source selection. Ten Monte Carlo runs, with normalized Gaussian random initialization for the matrix and vector factors at each run, are used to analyze the potential atrial sources found by BTD and compare them with the ones found by the matrix-based methods PCA and RobustICA-f. All the beats are downsampled by a factor of two, since the 3rd-order tensor built from the original 12-lead ECG beat poses some difficulties to Tensorlab.

5.3 Potential Atrial Source Analysis

For the observed patient used to analyze the potential atrial sources, PCA found 6 potential sources, RobustICA-f found 5 potential sources and BTD found a mean of 7.2 potential sources. In 7 out of the 10 independent runs, the BTD found more potential sources than the matrix-based methods, reflecting the ability of the tensor technique to perform undetermined source separation [12]. Finding several potential atrial sources is interesting, since it increases the possibility of finding some features that, although weakly contributing to the AA, may provide useful physiological and clinical information about the arrhythmia. In this work, however, we assume as in the previous literature of this topic that all AA can

be represented by a single source, and we leave the multiple source hypothesis for further works. Due to the lack of space and for the sake of clarity, only two potential atrial sources for PCA, RobustICA-f and BTD are shown in Figs. 2, 3 and 4. The other sources were disregarded for presenting a very weak power contribution.

Looking at Fig. 2, we can see that the atrial source estimated by PCA (located in the second row) has SC equal to 62.5%, while looking at Fig. 3, the estimated atrial source by RobustICA-f (located in the second row) has SC equal to 68.3%. For BTD, 8 out of the 10 independent runs estimated an atrial source with higher SC than PCA and 6 with higher SC than both matrix-based methods, giving an average SC over the 10 runs equal to 67.8%. Figure 4 shows the results for a particular initialization of BTD, where the estimated atrial source (located in the second row) has SC equal to 76.5%. The DF position of both PCA and RobustICA-f are located at 5.96 Hz, while in BTD it lies between 5.72 and 5.96 Hz. For comparison, the DF position obtained from an electrogram simultaneously acquired by a catheter located in the left atrial appendage of the same patient, is equal to 4.77 Hz.

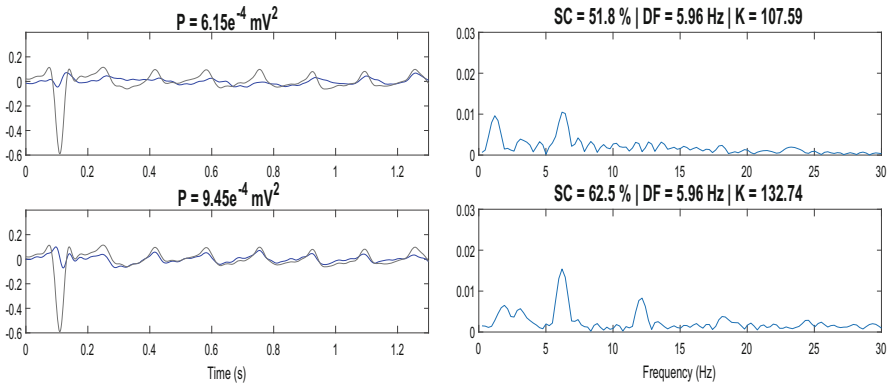


Fig. 2. Potential atrial sources contribution to lead V1 estimated by PCA. Left: time domain (in mV). Right: frequency domain (in $\text{mV}/\sqrt{\text{Hz}}$).

5.4 Atrial Source Selection

As ground truth, the sources were visually analyzed in time and frequency domain with guidance of the parameters previously described. The source with the strongest representation of AA content was taken as the atrial source.

The classical method and the two proposed methods of atrial source selection were assessed in 10 segments of 10 different patients, as previously explained. From a total of 120 runs for the 10 patients (100 for BTD, 10 for PCA and 10 for RobustICA-f) the classical method succeeded only in 45.8% of runs. Applying

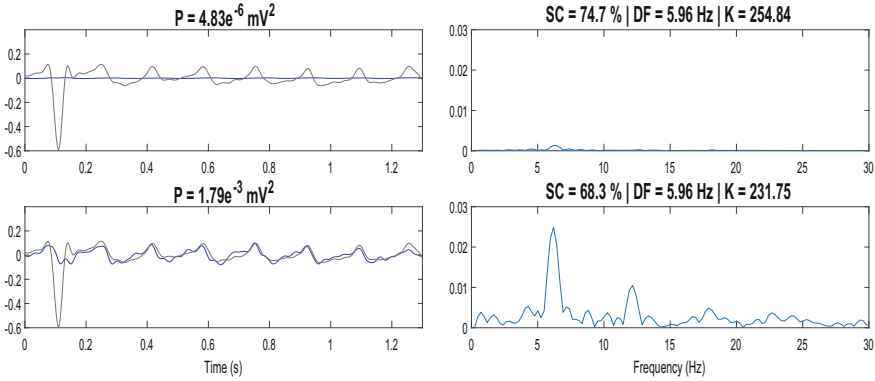


Fig. 3. Potential atrial sources contribution to lead V1 estimated by RobustICA-f. Left: time domain (in mV). Right: frequency domain (in $\text{mV}/\sqrt{\text{Hz}}$).

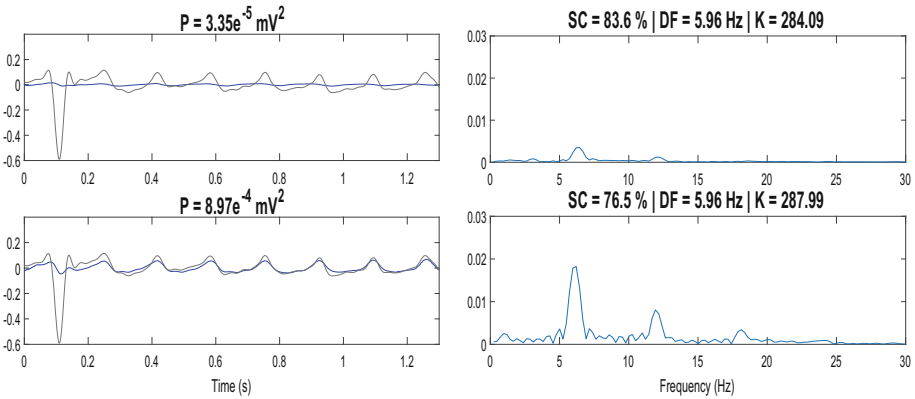


Fig. 4. Potential atrial sources contribution to lead V1 estimated by BTM for a single run. Left: time domain (in mV). Right: frequency domain (in $\text{mV}/\sqrt{\text{Hz}}$).

the first proposed method, the index of success increases to 75%, while the second proposed technique succeeds in 83.7% of the trials. It should be mentioned that in 35.8% of the trials, both the classical and the second proposed method were able to select the source with most AA content. Also, in 12.5% of trials none of the methods were able to select the AA signal. This means that the existing methods are suboptimal regarding the AA source selection. However, from the reported experiments, it is believed that a balanced combination between power contribution and kurtosis may lead to an optimal or at least a better method.

6 Conclusions

The present work has analyzed the potential atrial sources estimated by BTM, showing its satisfactory performance for most initializations in the tested

database. We have shown that the classical method of atrial source selection may not work in some cases, and we have proposed two new automated methods that better select the atrial source among the other potential sources. These methods have been validated in experimental results not only for BTD but also for the matrix-based methods PCA and RobustICA-f in a population of 10 patients with persistent AF. In future works, we aim to assess the proposed methods in a larger database and along consecutive time segments of each patient to analyze intra-patient (temporal) stability.

References

1. January, C.T., Wann, L.S., Alpert, J.S., Calkins, H., Cleveland, J.C., Cigarroa, J.E., Conti, J.B., et al.: 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. *Circulation* **64**(21), 2246–2280 (2014)
2. Mainardi, L., Sörnmo, L., Cerutti, S.: Understanding atrial fibrillation: the signal processing contribution. *Synthesis Lectures on Biomedical Engineering*. Morgan & Claypool Publishers (2008)
3. Rieta, J.J., Castells, F., Sánchez, C., Zarzoso, V., Millet, J.: Atrial activity extraction for atrial fibrillation analysis using blind source separation. *IEEE Trans. Biomed. Eng.* **51**(7), 1176–1186 (2004)
4. Castells, F., Rieta, J.J., Millet, J., Zarzoso, V.: Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias. *IEEE Trans. Biomed. Eng.* **52**(2), 258–267 (2005)
5. Zarzoso, V.: Extraction of ECG characteristics using source separation techniques: exploiting statistical independence and beyond. In: Naït-Ali, A. (ed.) *Advanced Biosignal Processing*, pp. 15–47. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-89506-0_2
6. Ribeiro, L.N., Hidalgo-Muñoz, A.R., Zarzoso, V.: Atrial signal extraction in atrial fibrillation electrocardiograms using a tensor decomposition approach. In: *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, 25–29 August 2015*, pp. 6987–6990 (2015)
7. Ribeiro, L.N., Hidalgo-Muñoz, A.R., Favier, G., Mota, J.C.M., de Almeida, A.L.F., Zarzoso, V.: A tensor decomposition approach to noninvasive atrial activity extraction in atrial fibrillation ECG. In: *Proceedings of the XXIII European Signal Processing Conference, EUSIPCO-2015, Nice, France, 31 August–4 September 2015*, pp. 2576–2580 (2015)
8. Ribeiro, L.N., de Almeida, A.L.F., Zarzoso, V.: Enhanced block term decomposition for atrial activity extraction in atrial fibrillation ECG. In: *Proceedings of the 9th IEEE Sensor Array and Multichannel Signal Processing Workshop, SAM-2016, Rio de Janeiro, Brazil, 10–13 July 2016*
9. Zarzoso, V.: Parameter estimation in block term decomposition for noninvasive atrial fibrillation analysis. In: *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP-2017, Curaçao, Dutch Antilles, 10–13 December 2017*

10. Zarzoso, V., Comon, P.: Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size. *IEEE Trans. Neural Netw.* **21**(2), 248–261 (2010)
11. Jolliffe, I.: *Principal Component Analysis*. Wiley Online Library (2005)
12. 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)
13. Boley, D.L., Luk, F.T., Vandevoorde, D.: Vandermonde factorization of a Hankel matrix. In: *Proceedings of the Workshop on Scientific Computing*, Hong Kong, March 1997
14. Vervliet, N., Debals, O., Sorber, L., Van Barel, M., De Lathauwer, L.: *Tensorlab 3.0*, March 2016. <https://www.tensorlab.net/>



Some Issues in Computing the CP Decomposition of NonNegative Tensors

Mohamad Jouni^(✉), Mauro Dalla Mura, and Pierre Comon

Univ. Grenoble Alpes, CNRS, Grenoble INP, Gipsa-Lab, 38000 Grenoble, France
{mohamad.jouni,maurodalla.mura,pierre.comon}@gipsa-lab.fr,
<http://www.gipsa-lab.grenoble-inp.fr>

Abstract. Tensor decompositions are still in the process of study and development. In this paper, we point out a problem existing in nonnegative tensor decompositions, stemming from the representation of decomposable tensors by outer products of vectors, and propose approaches to solve it. In fact, a scaling indeterminacy appears whereas it is not inherent in the decomposition, and the choice of scaling factors has an impact during the execution of iterative algorithms and should not be overlooked. Computer experiments support the interest in the greedy algorithm proposed, in the case of the CP decomposition.

1 Introduction

Tensors of order d are represented by data arrays with d indices, ($d = 2$ for matrices). They provide unique features as they are a suitable data structure for representing multimodal or multisource data, in which each diversity is represented by one of the ways of the tensor. One of the most interesting applications of tensors is the Canonical Polyadic (CP) decomposition defined below, which aims at representing a tensor as a sum of *decomposable* rank one tensors, revealing relationships among its d ways.

CP Decomposition. In this paper, we shall focus our attention on the CP decomposition of third order tensors. To begin with, a tensor \mathcal{D} is *decomposable* if it can be expressed as the outer product of vectors, *i.e.*: $\mathcal{D}_{ijk} = a_i b_j c_k$, which will be denoted compactly as $\mathcal{D} = \mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$, where \otimes is the outer (tensor) product. Next, every real tensor \mathcal{T} of order $d = 3$ and dimensions $I \times J \times K$, admits a CP Decomposition of the following form:

$$\mathcal{T} = \sum_{r=1}^R \lambda_r \mathcal{D}(r), \quad (1)$$

where $\mathcal{D}(r) \stackrel{\text{def}}{=} \mathbf{a}(r) \otimes \mathbf{b}(r) \otimes \mathbf{c}(r)$, $\mathbf{a}(r)$, $\mathbf{b}(r)$ and $\mathbf{c}(r)$ being real vectors, which can be stored in the so-called factor matrices, \mathbf{A} , \mathbf{B} and \mathbf{C} respectively, of size

This work was supported in part by ERC Advanced Grant 2013-320594 “DECODA”.

$I \times R$, $J \times R$, and $K \times R$ respectively, and λ_r are real positive scalars. The CP decomposition reveals *tensor rank* when R is minimal, which will be assumed from now on; for instance, tensors $\mathcal{D}(r)$ are of rank 1. Note that another writing of (1) in terms of factor matrices is $T_{ijk} = \sum_{r=1}^R \lambda_r A_{ir} B_{jr} C_{kr}$. In addition, because of the presence of λ_r , the columns of factor matrices may be normalized to 1.

At this stage, it is important to stress that *there is no scaling ambiguity* in the CP decomposition (1), contrary to what is sometimes claimed in the literature. Only the *representation* of tensors $\mathcal{D}(r)$ by triplets of vectors is subject to this indeterminacy. In fact, by definition, tensors are precisely *equivalence classes* with respect to scaling [1–4]: the triplets $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ and $(\alpha\mathbf{a}, \beta\mathbf{b}, \gamma\mathbf{c})$ represent *the same tensor* provided that $\alpha\beta\gamma = 1$.

The rank R of the CP decomposition (1) is of particular interest in applications since it is related to the intrinsic dimensionality of multilinear data. Furthermore, the CP decomposition, contrary to other tensor decompositions, *e.g.*, Tucker’s or High-Order Singular Value Decomposition (HOSVD), enjoys uniqueness if the rank is not too large [5–7]. Uniqueness is of utmost importance since it eventually allows physical interpretation of relationships among the ways of a tensor.

2 Motivation

NonNegativity. When the observation tensor \mathcal{T} contains only real nonnegative entries, it is suitable to impose decomposable tensors $\mathcal{D}(r)$ to also be nonnegative. By doing this, we define a *nonnegative rank*, R^+ , which may be larger than R . This is actually already true for matrices (tensors of order 2). In fact, Herbert E. Robbins exhibited a simple example of a 5×5 matrix having rank 3 but nonnegative rank 4; see [4, 8] for its expression. It is thus necessary to define the nonnegative CP decomposition of a nonnegative tensor as:

$$\mathcal{T} = \sum_{r=1}^{R^+} \lambda_r \mathbf{a}(r) \otimes \mathbf{b}(r) \otimes \mathbf{c}(r), \quad (2)$$

where $a_i(r) \in \mathbb{R}^+$, $b_j(r) \in \mathbb{R}^+$ and $c_k(r) \in \mathbb{R}^+$, $\forall(i, j, k, r)$.

There are many applications where nonnegativity is relevant, as to provide better interpretable results when dealing with variables related to physical quantities such as luminance in images, spectra or chemical concentrations [9, 10]. There exist many algorithms aiming at computing the CP decomposition of nonnegative tensors [9, 11]. However, due to measurement noise or modeling errors, the tensor to decompose may not be nonnegative or may have a too large rank, hence requiring to be approximated. It turns out that, given any real tensor \mathcal{T} of rank R , it is fortunately always possible to find a best nonnegative approximation of \mathcal{T} of given nonnegative rank R^+ . This problem is indeed well-posed [12, 13] (which would not be the case in \mathbb{R} instead of \mathbb{R}^+).

Projection onto the NonNegative Orthant: In the nonnegative CP decomposition (2), all quantities are nonnegative. For instance, vector $\mathbf{a}(r)$ belongs to the nonnegative orthant $(\mathbb{R}^+)^I$. In iterative algorithms, this constraint is ensured at each iteration by projecting a computed value onto the nonnegative orthant. This is where the problem shows up. In fact, projecting $\mathcal{D}(r)$ or its building vectors $\{\mathbf{a}(r), \mathbf{b}(r), \mathbf{c}(r)\}$ do not yield the same result. Since this observation is already true for matrices, a simple example will be most convincing.

Example. Take the matrix \mathbf{M} below, of rank 1. Now its projection \mathbf{M}^+ has rank 2. So it is preferred to project its supporting vectors $\{\mathbf{a}, \mathbf{b}\}$ instead. The obtained vectors are $\{\mathbf{a}^+, \mathbf{b}^+\}$ and yield a matrix of nonnegative rank equal to 1:

$$\mathbf{M} = \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \otimes \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \mathbf{a} \otimes \mathbf{b}, \quad \mathbf{M}^+ = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{a}^+ \otimes \mathbf{b}^+ = \begin{pmatrix} 4 & 0 \\ 0 & 0 \end{pmatrix}.$$

The problem is that vectors $\{\mathbf{a}, \mathbf{b}\}$ are not uniquely defined. We could have taken $\{-\mathbf{a}, -\mathbf{b}\}$ without changing \mathbf{M} . Should we do that, we obtain instead:

$$\mathbf{M} = \begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \mathbf{a} \otimes \mathbf{b}, \quad \mathbf{M}^+ = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{a}^+ \otimes \mathbf{b}^+ = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We see that the projected matrix of nonnegative rank 1 is not the same. This issue comes from the fact that no care has been taken of the scaling indeterminacies (which reduce to sign indeterminacies thanks to the use of factors λ_r) inherent in the representation of a rank-1 tensor by a triplet of vectors.

Algorithms resorting to projection include Alternating Nonnegative Least Squares (ANLS) [9], Projected and Compressed ANLS (ProCo) [14], or Alternating Direction Method of Multipliers (ADMM) [11], among others. Hard thresholding is the procedure in which it is the easiest to illustrate the occurring of the problem.

Algorithm 1. Alternating Nonnegative Least Squares (ANLS)

Require: \mathcal{T} , $\mathbf{B} = \mathbf{B}[0]$, $\mathbf{C} = \mathbf{C}[0]$

$t = 0$;

while stopping criterion is not met, **do**

$t = t + 1$

 compute \mathbf{A} from $\mathbf{B}[t - 1]$ and $\mathbf{C}[t - 1]$; $\mathbf{A}[t] \leftarrow \mathbf{A}^+$

 compute \mathbf{B} from $\mathbf{C}[t - 1]$ and $\mathbf{A}[t]$; $\mathbf{B}[t] \leftarrow \mathbf{B}^+$

 compute \mathbf{C} from $\mathbf{A}[t]$ and $\mathbf{B}[t]$; $\mathbf{C}[t] \leftarrow \mathbf{C}^+$

 normalize columns: $\mathbf{A}[t] \leftarrow \mathbf{A}[t]\mathbf{A}_A^{-1}$; $\mathbf{B}[t] \leftarrow \mathbf{B}[t]\mathbf{A}_B^{-1}$; $\mathbf{C}[t] \leftarrow \mathbf{C}[t]\mathbf{A}_A\mathbf{A}_B$;

end while

normalize columns: $\mathbf{C}[t] \leftarrow \mathbf{C}[t]\mathbf{A}^{-1}$

return $\mathbf{A}[t]$, $\mathbf{B}[t]$, $\mathbf{C}[t]$, \mathbf{A}

ANLS. One algorithm that has been widely used to compute CP decomposition (1) is the Alternating Least Squares (ALS) algorithm. ALS minimizes with respect to matrices \mathbf{A} , \mathbf{B} , \mathbf{C} in an alternating fashion, the loss:

$$\Phi = \sum_{ijk} [\mathcal{T}_{ijk} - \sum_{r=1}^R \lambda_r A_{ir} B_{jr} C_{kr}]^2. \quad (3)$$

Factor matrices are updated in turns during each iteration until a certain condition is attained (*e.g.* the number of iterations or a certain threshold on the reconstruction error). When a nonnegative decomposition is sought, each factor matrix can be projected onto the nonnegative orthant right after its calculation; this is the ANLS algorithm [9, p.47]. The pseudo-code is given in Algorithm 1.

3 Proposed Approach

We illustrate the problem with hard thresholding (cf. Sect. 4), but our solution could also reveal useful in soft thresholding as well. The problem is worse when all entries in a column vector are set to zero; this prevents its normalization (as it would lead to a division by zero) or imposes an erroneous reduction of the rank (due to the arbitrary removal of the null columns). The solution we describe overcomes these two difficulties most of the time, up to negligible extraneous computation load. We propose to implement this in a procedure to be executed before projection. The concept goes as follows. Because of normalization, the scaling indeterminacy reduces merely to signs. In fact, in every decomposable tensor $\mathcal{D}(r)$, we have two variables, $\epsilon, \eta \in \{-1, +1\}$, which are to be used as sign flippers for the columns $\mathbf{a}(r)$, $\mathbf{b}(r)$ and $\mathbf{c}(r)$ that are together involved in an outer product term, without changing the result of the outer product given by:

$$\mathbf{a}(r) \otimes \mathbf{b}(r) \otimes \mathbf{c}(r) = (\epsilon\eta\mathbf{a}(r)) \otimes (\epsilon\mathbf{b}(r)) \otimes (\eta\mathbf{c}(r)), \quad \forall (\epsilon, \eta) \in \{-1, +1\}. \quad (4)$$

This formula covers all 4 combinations of sign flipping of vectors, without affecting the result of the original outer product. Now denote by $\mathbf{a}'(r) = \epsilon\eta\mathbf{a}(r)$, $\mathbf{b}'(r) = \epsilon\mathbf{b}(r)$, and $\mathbf{c}'(r) = \eta\mathbf{c}(r)$, and:

$$\mathbf{a}^-(r) = \mathbf{a}'(r) \text{ where } \mathbf{a}'(r) < 0, \text{ and } 0 \text{ elsewhere} \quad (5)$$

$$\mathbf{b}^-(r) = \mathbf{b}'(r) \text{ where } \mathbf{b}'(r) < 0, \text{ and } 0 \text{ elsewhere} \quad (6)$$

$$\mathbf{c}^-(r) = \mathbf{c}'(r) \text{ where } \mathbf{c}'(r) < 0, \text{ and } 0 \text{ elsewhere.} \quad (7)$$

Vectors $\mathbf{a}^+(r)$, $\mathbf{b}^+(r)$ and $\mathbf{c}^+(r)$ are defined in a similar manner, with positive entries. In particular, $\mathbf{a}^+(r) + \mathbf{a}^-(r) = \mathbf{a}'(r)$.

Given a triplet of vectors, $(\mathbf{a}, \mathbf{b}, \mathbf{c})$, there are 4 possibilities to construct a non-negative decomposable tensor $\mathcal{D}^{[l]}$ by just flipping their signs without changing the value of $(\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c})$ and by setting to zero negative values in each vector (refer to Table. 1).

Table 1. The 4 possibilities created by sign flipping.

(ϵ, η)	(+, +)	(-, -)	(+, -)	(-, +)
ℓ	1	2	3	4
$\mathcal{D}^{[\ell]}$	$\mathcal{D}^{[1]} = \mathbf{a}^+ \otimes \mathbf{b}^+ \otimes \mathbf{c}^+$	$\mathcal{D}^{[2]} = \mathbf{a}^+ \otimes \mathbf{b}^- \otimes \mathbf{c}^-$	$\mathcal{D}^{[3]} = \mathbf{a}^- \otimes \mathbf{b}^+ \otimes \mathbf{c}^-$	$\mathcal{D}^{[4]} = \mathbf{a}^- \otimes \mathbf{b}^- \otimes \mathbf{c}^+$

where for the sake of convenience, \mathbf{a}^+ stands for vector $\mathbf{a}^{[\ell]}(r)^+$, and similarly for \mathbf{a}^- , \mathbf{b}^+ , \mathbf{b}^- , \mathbf{c}^+ and \mathbf{c}^- .

We are interested to know which combination would yield the minimal number of resets. Ultimately, we are concerned about (i) avoiding to set a whole vector to zero, which would lead to decrease the rank. This goal can mean “set as few entries to zero as possible”. And we also aim at (ii) minimizing the distance between the original tensor and its nonnegative approximation.

We explored several criteria. The first is to minimize $\Phi_0 = \|\mathcal{T} - \sum_r \mathcal{D}^{[\ell]}(r)\|_2$. This criterion is very costly to optimize, due to the large number of combinations. In fact, for every r , there are 4 possibilities to assign (ϵ, η) , and this assignment can be different for each r . This would result in 4^R possibilities to explore. This is why we propose two greedy algorithms searching for the optimal solution $\mathcal{D}^{[\ell]}(r)$ independently for every r . One possibility is to minimize w.r.t. ℓ the following product **for every r independently**, and for the L^2 norm:

$$\Phi_1(\ell, r) = \|\mathcal{D}(r) - \mathcal{D}^{[\ell]}(r)\|_2^2. \quad (8)$$

Let us express this criterion for $\ell = 1$, without loss of generality. We have for any fixed r (that we drop for the sake of convenience):

$$\Phi_1(1, r) = \|\mathcal{D}(r)\|_2^2 + \|\mathcal{D}^{[1]}(r)\|_2^2 - 2 \sum_{ijk} a_i a_i^+ b_j b_j^+ c_k c_k^+. \quad (9)$$

The last term can be rewritten as $2(\mathbf{a}^\top \mathbf{a}^+)(\mathbf{b}^\top \mathbf{b}^+)(\mathbf{c}^\top \mathbf{c}^+)$. Next, it is also equal to $2\|\mathbf{a}^+\|^2 \|\mathbf{b}^+\|^2 \|\mathbf{c}^+\|^2$, since \mathbf{a}^+ and \mathbf{a}^- are orthogonal and $\mathbf{a} = \mathbf{a}^+ - \mathbf{a}^-$. This suggests another criterion to minimize w.r.t. ℓ :

$$\Phi_2(\ell, r) = \|\mathbf{a}^-\| \cdot \|\mathbf{b}^-\| \cdot \|\mathbf{c}^-\| \quad (10)$$

Criteria Φ_1 and Φ_2 are easy to optimize w.r.t. (ϵ, η) , *i.e.* w.r.t. ℓ , and need negligible extraneous computation load.

4 A Toy Example

Consider the factor matrices:

$$\mathbf{A} = \begin{bmatrix} 0.8025 & 0.1914 \\ 0.0089 & 0.9106 \\ 0.5966 & 0.3662 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.0088 & 0.7495 \\ 1 & 0.6620 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0.7071 & 0 \\ 0.7071 & 0 \end{bmatrix}.$$

Algorithm 2. Minimization of Φ_1 w.r.t. to ϵ and η for fixed r

Require: $\mathbf{a}(r)$, $\mathbf{b}(r)$, $\mathbf{c}(r)$
 initialize $\epsilon(\ell)$, $\eta(\ell)$ for each possible $\mathcal{D}^{[\ell]}$; (according to Table.1)
 normalize: $\mathbf{a} \leftarrow \frac{\mathbf{a}(r)}{\|\mathbf{a}(r)\|}$; $\mathbf{b} \leftarrow \frac{\mathbf{b}(r)}{\|\mathbf{b}(r)\|}$; $\mathbf{c} \leftarrow \frac{\mathbf{c}(r)}{\|\mathbf{c}(r)\|}$;
 $\ell = 0$;
while $\ell \leq 4$, **do**
 $\ell = \ell + 1$;
 $\mathbf{a}' \leftarrow \epsilon(\ell)\eta(\ell)\mathbf{a}$; $\mathbf{b}' \leftarrow \epsilon(\ell)\mathbf{b}$; $\mathbf{c}' \leftarrow \eta(\ell)\mathbf{c}$;
 compute \mathbf{a}^+ , \mathbf{b}^+ , and \mathbf{c}^+ ;
 $\Phi_1(\ell) \leftarrow \|\mathcal{D} - \mathcal{D}^{[\ell]}\|_2^2$;
end while
 Find $\ell_o = \arg \min_{\ell} \Phi_1(\ell)$;
 $\mathbf{a}(r) \leftarrow \epsilon(\ell_o)\eta(\ell_o)\mathbf{a}(r)$; $\mathbf{b}(r) \leftarrow \epsilon(\ell_o)\mathbf{b}(r)$; $\mathbf{c}(r) \leftarrow \eta(\ell_o)\mathbf{c}(r)$;
return $\mathbf{a}(r)$, $\mathbf{b}(r)$, $\mathbf{c}(r)$;

Algorithm 3. Modified ANLS

Require: \mathcal{T} , $\mathbf{B} = \mathbf{B}[0]$, $\mathbf{C} = \mathbf{C}[0]$
 $t = 0$;
while stopping criterion is not met, **do**
 $t = t + 1$
 compute \mathbf{A} from $\mathbf{B}[t-1]$ and $\mathbf{C}[t-1]$; $\mathbf{A}[t]$
 compute \mathbf{B} from $\mathbf{C}[t-1]$ and $\mathbf{A}[t]$; $\mathbf{B}[t]$
 compute \mathbf{C} from $\mathbf{A}[t]$ and $\mathbf{B}[t]$; $\mathbf{C}[t]$
 $r = 0$;
 while $r < R$ **do**
 update $\mathbf{a}(r)$, $\mathbf{b}(r)$, and $\mathbf{c}(r)$ using Alg.2;
 $r = r + 1$;
 end while
 $\mathbf{A}[t] \leftarrow \mathbf{A}^+$; $\mathbf{B}[t] \leftarrow \mathbf{B}^+$; $\mathbf{C}[t] \leftarrow \mathbf{C}^+$;
 replace null columns in \mathbf{A} , \mathbf{B} or \mathbf{C} by random values using e.g. absolute value of standard Gaussian distribution; (see Sect.5)
 normalize columns: $\mathbf{A}[t] \leftarrow \mathbf{A}[t]\mathbf{A}_A^{-1}$; $\mathbf{B}[t] \leftarrow \mathbf{B}[t]\mathbf{A}_B^{-1}$; $\mathbf{C}[t] \leftarrow \mathbf{C}[t]\mathbf{A}_A\mathbf{A}_B$;
end while
 normalize columns: $\mathbf{C}[t] \leftarrow \mathbf{C}[t]\mathbf{A}^{-1}$
return $\mathbf{A}[t]$, $\mathbf{B}[t]$, $\mathbf{C}[t]$, \mathbf{A}

When computing the CP Decomposition with Algorithm 1, after one update of \mathbf{A} , one of its columns became negative, and hence one of its columns got discarded, and the rank was decreased by 1. Note that, for the sake of conciseness, during the loop of updates only the columns of \mathbf{A} and \mathbf{B} are normalized and their norms multiply \mathbf{C} ; after the loop ends, \mathbf{C} is normalized and its column norms (containing that of \mathbf{A} and \mathbf{B}) form the values of \mathbf{A} (cf. Algorithm 1).

Algorithm 1: Standard ALS

$$\mathbf{A} : \begin{bmatrix} 0.2311 & -0.0464 \\ 0.1891 & -0.0627 \\ 0.2178 & -0.0498 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2311 & 0 \\ 0.1891 & 0 \\ 0.2178 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.6252 & \text{Undefined} \\ 0.5118 & \text{Undefined} \\ 0.5893 & \text{Undefined} \end{bmatrix} \rightarrow \begin{bmatrix} 0.6252 \\ 0.5118 \\ 0.5893 \end{bmatrix}$$

$$\mathbf{B} : \begin{bmatrix} 0.2962 & 0 \\ 1.0561 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2962 & 0 \\ 1.0561 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2701 & \text{Undefined} \\ 0.9628 & \text{Undefined} \end{bmatrix} \rightarrow \begin{bmatrix} 0.2701 \\ 0.9628 \end{bmatrix}$$

$$\mathbf{C} : \begin{bmatrix} 0 & 0 \\ 0.4978 & 0 \\ 0 & 0 \\ 1.3779 & 0 \\ 1.3779 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0.4978 & 0 \\ 0 & 0 \\ 1.3779 & 0 \\ 1.3779 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0.2018 & 0 \\ 0 & 0 \\ 0.5585 & 0 \\ 0.5585 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 0.2018 \\ 0 \\ 0.5585 \\ 0.5585 \end{bmatrix}$$

At the end the output of standard ANLS results in:

$$\mathbf{A} = \begin{bmatrix} 0.8004 \\ 0.0249 \\ 0.5990 \end{bmatrix}; \mathbf{B} = \begin{bmatrix} 0.0165 \\ 0.9999 \end{bmatrix}; \mathbf{C} = \begin{bmatrix} 0 \\ 0.0835 \\ 0 \\ 0.7046 \\ 0.7046 \end{bmatrix}$$

Algorithm 2: Modified ANLS (using Φ_1)

$$\mathbf{A} : \begin{bmatrix} 0.2311 & -0.0464 \\ 0.1891 & -0.0627 \\ 0.2178 & -0.0498 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2311 & 0.0464 \\ 0.1891 & 0.0627 \\ 0.2178 & 0.0498 \end{bmatrix} \rightarrow \begin{bmatrix} 0.6252 & 0.5015 \\ 0.5118 & 0.6772 \\ 0.5893 & 0.5384 \end{bmatrix}$$

$$\mathbf{B} : \begin{bmatrix} 0.3420 & 0.4425 \\ 1.1203 & 0.6211 \end{bmatrix} \rightarrow \begin{bmatrix} 0.3420 & 0.4425 \\ 1.1203 & 0.6211 \end{bmatrix} \rightarrow \begin{bmatrix} 0.2919 & 0.5802 \\ 0.9564 & 0.8145 \end{bmatrix}$$

$$\mathbf{C} : \begin{bmatrix} 0 & 0 \\ -0.4978 & -7.6392 \\ 0 & 0 \\ 2.9691 & 11.1635 \\ 2.9691 & 11.1635 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 7.6392 \\ 0 & 0 \\ 2.9691 & 0 \\ 2.9691 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 \\ 0 & 0.5392 \\ 0 & 0 \\ 1.2853 & 0 \\ 1.2853 & 0 \end{bmatrix}$$

At the end the output of Modified ANLS results in:

$$\mathbf{A} = \begin{bmatrix} 0.8025 & 0.1914 \\ 0.0089 & 0.9106 \\ 0.5966 & 0.3662 \end{bmatrix}; \mathbf{B} = \begin{bmatrix} 0.0088 & 0.7495 \\ 1 & 0.6620 \end{bmatrix}; \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0.7071 & 0 \\ 0.7071 & 0 \end{bmatrix}$$

5 Computer Results

500 realizations of 10×5 matrices $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\}$ are drawn. The rank of the tensor that is tested is hence $R = 5$. Entries of factor matrices are the absolute value

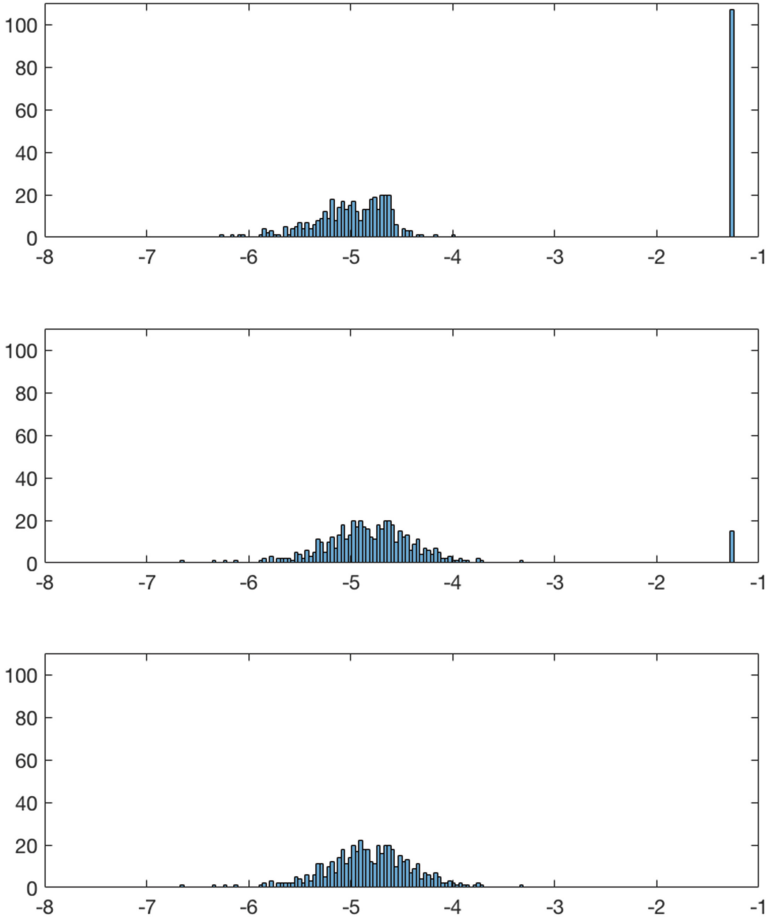


Fig. 1. Histograms of the error, in \log_{10} scale, obtained after 500 iterations. Top: ANLS. Middle: ANLS modified with Φ_1 . Bottom: ANLS modified with Φ_1 and with column reinitialization.

of i.i.d. drawn from a standard Gaussian distribution. On each realization, both ANLS and a modified version based on the minimization of Φ_1 are run.

As can be seen in Fig. 1, 107 realizations out of 500 are unsuccessful with ANLS, that is, 107 realizations generate one fully negative column in a factor matrix which is then zeroed due to hard thresholding. This eventually leads to a decrease of the rank down to 4 and hence to a large reconstruction error (close to 10^{-1}). Among those 107 pathological cases, our simple function described by Algorithm 2 could cope with 92 of them without a significant increase in complexity. However, 15 realizations remain unsolved, because they correspond to either one of two particular cases: (i) either one column, say $\mathbf{a}(r_o)$, is fully

negative, and the two others, namely $\mathbf{b}(r_o)$ and $\mathbf{c}(r_o)$ are fully positive, or (ii) all the three columns are fully negative.

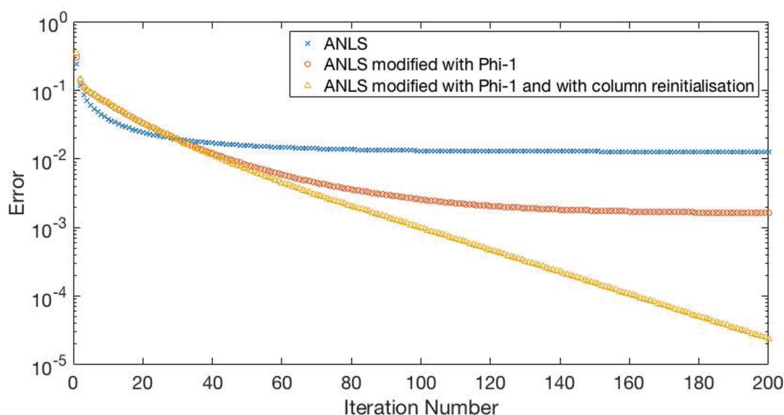


Fig. 2. The average error obtained after 500 iterations, as a function of the number of iterations in ANLS (blue crosses), ANLS modified with Φ_1 (red circles), and ANLS modified with Φ_1 and with column reinitialization (yellow triangles). (Color figure online)

In order to cope with the latter cases, a straightforward improvement was brought in Algorithm 3, by drawing a fresh column vector (also the absolute value of i.i.d drawn from a standard Gaussian distribution) to replace null vectors when generated in the unsolved pathological cases, before normalizing the columns. The results can be seen in Figs. 1 (bottom) and 2, where all 15 cases were solved and the rank was preserved.

6 Concluding Remarks

In this paper, we have emphasized the fact that rank-1 tensors should not be treated as a collection of vectors without care, and showed an illustration in the case of ANLS using hard thresholding. In the latter case, two modifications have been proposed to fix the problem. In future works, we plan to investigate applications to other algorithms such as ADMM, and/or using soft thresholding. The influence of noise would also deserve to be further addressed.

References

1. Ruiz-Tolosa, J.R., Castillo, E.: From Vectors to Tensors. Universitext. Springer, Heidelberg (2005). <https://doi.org/10.1007/b138560>
2. Landsberg, J.M.: Tensors: Geometry and Applications, Graduate Studies in Mathematics, vol. 128. AMS Publications (2012)

3. Hackbusch, W.: *Tensor Spaces and Numerical Tensor Calculus*. Series in Computational Mathematics. Springer, Heidelberg (2012)
4. Comon, P.: Tensors: a brief introduction. *IEEE Sig. Proc. Mag.* **31**(3), 44–53 (2014). Special issue on BSS. hal-00923279
5. Sidiropoulos, N.D., Bro, R.: On the uniqueness of multilinear decomposition of N-way arrays. *J. Chemometr.* **14**, 229–239 (2000)
6. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions. *Linear Algebra Appl.* **18**, 95–138 (1977)
7. Chiantini, L., Ottaviani, G., Vannieuwenhoven, N.: An algorithm for generic and low-rank specific identifiability of complex tensors. *SIAM J. Matrix Anal. Appl.* **35**(4), 1265–1287 (2014)
8. Cohen, J., Rothblum, U.: Nonnegative ranks, decompositions and factorizations of nonnegative matrices. *Linear Algebra Appl.* **190**, 149–168 (1993)
9. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.: *Nonnegative Matrix and Tensor Factorizations*. Wiley, Chichester (2009)
10. Comon, P., Jutten, C. (eds.): *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, Oxford UK, Burlington USA (2010). hal-00460653
11. Huang, K., Sidiropoulos, N., Liavas, A.P.: A flexible and efficient algorithmic framework for constrained matrix and tensor factorization. *IEEE Trans. Sig. Process.* **64**(19), 5052–5065 (2016)
12. Lim, L.-H., Comon, P.: Nonnegative approximations of nonnegative tensors. *J. Chemometr.* **23**, 432–441 (2009). hal-00410056
13. Qi, Y., Comon, P., Lim, L.H.: Uniqueness of nonnegative tensor approximations. *IEEE Trans. Inf. Theory* **62**(4), 2170–2183 (2016). [arXiv:1410.8129](https://arxiv.org/abs/1410.8129)
14. Cohen, J.E., Farias, R.C., Comon, P.: Fast decomposition of large nonnegative tensors. *IEEE Sig. Proc. Lett.* **22**(7), 862–866 (2015). hal-01069069

Matrix and Tensor Factorizations



A Grassmannian Minimum Enclosing Ball Approach for Common Subspace Extraction

Emilie Renard^{1(✉)}, Kyle A. Gallivan², and P.-A. Absil¹

¹ ICTEAM Institute, Université catholique de Louvain,
1348 Louvain-la-Neuve, Belgium

{emilie.renard, pa.absil}@uclouvain.be

² Department of Mathematics, Florida State University,
Tallahassee, FL 32306-4510, USA
kgallivan@fsu.edu

Abstract. We study the problem of finding a subspace representative of multiple datasets by minimizing the maximal dissimilarity between this subspace and all the subspaces generated by those datasets. After arguing for the choice of the dissimilarity function, we derive some properties of the corresponding formulation. We propose an adaptation of an algorithm used for a similar problem on Riemannian manifolds. Experiments on synthetic data show that the subspace recovered by our algorithm is closer to the true common subspace than the solution obtained using an SVD.

Keywords: Common subspace extraction · Total grassmannian
Minimal enclosing ball

1 Introduction

We address the problem of extracting common information from multiple datasets. In recent years data has become increasingly easy to generate and store for analysis to guide decision making, and it is not uncommon to have access to datasets representing similar but not exactly equivalent phenomena. A typical example can be found in bioinformatic, where datasets usually have a few tens to at most a few hundreds of samples for a few (tens of) thousands of features. However there usually exist various datasets measuring the same disease on different sets of patients, but corresponding to different studies and different experimental conditions that should be taken into account in further analysis. Considering all those similar datasets at once can be very useful to deal with the high number of features since statistical inferences require a large number of samples to be robust enough and generalizable to other data.

Beside the basic possibility to simply concatenate all the datasets X_1, \dots, X_m into a larger dataset $X = [X_1 \ \dots \ X_m]$ and apply usual methods such as principal components analysis on X , more specific approaches exist to extract common

components present in the datasets. A method to factorize two datasets with a common factor was proposed in [1] with a closed-form solution, and an extension to more than two datasets was proposed in [2]. However, such methods assume that the common dimension of the datasets is full-rank, which is not the case if we consider datasets with a higher number of variables than samples, such as gene expression datasets. The best known method is probably canonical correlation analysis (CCA) [3], which aims to find a linear combination of the initial features for both datasets maximizing the correlation between those two combinations. When dealing with two datasets only, an exact solution can be computed based on the covariance matrix. In order to find more than one pair of correlated combination of features, deflation is usually used: the same procedure (CCA) is repeated on the data from which the previous components were removed. Another well known method, partial least square regression [4], aims to find linear combinations of features for the two datasets such that the covariance between those two new representations is maximal. As in CCA a closed-form solution exists, and deflation can be used to compute the following components. Another variation is co-inertia analysis (CIA) and its extension multiple CIA [5] that maximizes a sum of weighted squared covariances between linear combination of the datasets features and a reference vector. Consensus principal component analysis is very similar to CIA, the main difference being in the deflation process [6]. Different extensions of those methods to more than two datasets have been proposed, with various criteria to optimize (see for example [7, 8] and references within): maximizing a sum on all pairs of datasets of covariances or correlations, possibly squared or in absolute value, and with different constraints. In such cases, a closed-form solution does not always exist.

A central question when using more than two datasets is the importance to give to those different (pairs of) datasets. Common approaches are to give all datasets the same importance or, as in [7], to consider if a pair of datasets is connected or not and to give to the corresponding term a weight of 1 or 0. If we are dealing with a set of datasets all very similar except one (for example, because measured using another technology), those kind of choices can lead to components representing very well all the similar datasets but being not representative at all of the last one. Here, we want to avoid this situation, and in order to take all X_i into account we propose to minimize the maximal dissimilarity d between the common component $U \in \mathbb{R}^{p \times K}$ and all datasets $X_i \in \mathbb{R}^{p \times n_i}$:

$$U^* = \arg \min_U \max_i d(U, X_i). \quad (1)$$

This formulation can be viewed as looking for the center of the smallest-radius sphere enclosing all X_i , and can be linked to the minimum enclosing ball, 1-center problem or minimax optimization problem. However, since here U represents a subspace, we are really interested in the subspace generated by the columns of U . So we want to solve problem (1) such that $d(U, X_i) = d(\mathcal{U}, \mathcal{X}_i)$ is a dissimilarity measure between \mathcal{U} and \mathcal{X}_i , the subspaces generated by the columns of U and X .

The problem of finding the smallest enclosing ball of a finite point set $\mathbb{X} = \{x_1, \dots, x_m\}$ has been already thoroughly investigated in Euclidean space, and an efficient approximation algorithm has been proposed in [9]. An adaptation of the algorithm presented in [9] to Riemannian geometry is proposed in [10] with a study of the convergence rate, and in [11] to compute Riemannian L_1 and L_∞ center of mass of structure tensor images in order to denoise those images.

In this paper we assume that each point \mathcal{X}_i represents a subspace of dimension n_i in \mathbb{R}^p , that is \mathcal{X}_i belongs to the Grassmannian manifold $\mathcal{G}(n_i, p)$ and so $\mathbb{X} = \{\mathcal{X}_1, \dots, \mathcal{X}_m\}$ is included in the total Grassmannian $\cup \mathcal{G}(n_i, p)$. The proposed approach to solve problem (1) is inspired by [10]. The main difference is that each data point \mathcal{X}_i belongs to a different Grassmannian $\mathcal{G}(n_i, p)$, which prevents us by using the usual Grassmannian distance. Instead we use an adaptation based on principal angles, which allows us to measure the dissimilarity between any pair of subspaces of different dimension, and to project $\mathcal{G}(n_i, p)$ on $\mathcal{G}(K, p)$ in order to return to a common manifold.

The paper is organized as followed. We first discuss the choice of the dissimilarity measure and the resulting problem in Sect. 2, then details of the proposed approach are presented in Sect. 3. Section 4 describes the results obtained on synthetic data, and we conclude in Sect. 5.

2 Problem Formulation

Let $X_i \in \mathbb{R}^{p \times n_i}$ be a matrix of p variables times n_i samples, for $i = 1, \dots, m$. Our goal is to find a subspace \mathcal{U} of dimension K representative of all the subspaces \mathcal{X}_i , where \mathcal{X}_i is the subspace generated by the columns of X_i . In other words we are looking for a $U \in \mathbb{R}^{p \times K}$ minimizing $d(U, X_i)$ for all i , where $d(U, X) = d(\mathcal{U}, \mathcal{X})$ is a dissimilarity measure between the span of U and the span of X .

2.1 Dissimilarity Measure

Different dissimilarities are possible to quantify $d(U, X)$, we detail some of them below. For $K = 1$, a possible choice to evaluate if a vector $u \in \mathbb{R}^p$ is close to \mathcal{X} is the angle between u and its orthogonal projection on \mathcal{X} . A vector u is close to the subspace \mathcal{X} if the (positive) angle between them is small. If we define ϕ as the angle between u and \mathcal{X} (in $[-\frac{\pi}{2}, \frac{\pi}{2}]$), we have

$$u^\top \tilde{X} \tilde{X}^\top u = \cos^2 \phi$$

where $\|u\| = 1$ and \tilde{X} is an orthonormal basis of \mathcal{X} . The term $u^\top \tilde{X} \tilde{X}^\top u$ can then be seen as a similarity measure evaluating how close u is to \mathcal{X} , with a value of 1 when u is in the subspace \mathcal{X} , and 0 when they are orthogonal. We can then define a dissimilarity:

$$d(u, X) = 1 - u^\top \tilde{X} \tilde{X}^\top u = \sin^2 \phi$$

with $d(u, X) = 0$ if and only if $u \in \mathcal{X}$.

This can be extended to a more general $U \in \mathbb{R}^{p \times K}$ with $p \geq n \geq K \geq 1$ (with n the dimension of \mathcal{X}) by summing the dissimilarities obtained for each element of an orthonormal basis \check{U} of \mathcal{U} :

$$d_a(U, X) = \sum_k 1 - \check{U}(:, k)^\top \check{X} \check{X}^\top \check{U}(:, k) = K - \text{Tr}(\check{U}^\top \check{X} \check{X}^\top \check{U}) = \sum_k \sin^2 \phi_k(U, X)$$

with $\cos \phi_k(U, X)$ the singular values of $\check{U}^\top \check{X}$. Note that this quantity does not depend on the \check{U} or \check{X} chosen.

Another possible dissimilarity is [12]:

$$\begin{aligned} d_b(U, X) &= \frac{1}{\sqrt{2}} \|\check{U} \check{U}^\top - \check{X} \check{X}^\top\|_F = \sqrt{\frac{K+n}{2} - \text{Tr}(\check{U}^\top \check{X} \check{X}^\top \check{U})} \\ &= \sqrt{\frac{n-K}{2} + \sum_k \sin^2 \phi_k(U, X)}. \end{aligned}$$

Similarly, we can consider the norm between \check{X} and its projection onto the common subspace \mathcal{U} (termed *chordal metric* in [13, Table 3]):

$$\begin{aligned} d_c(U, X) &= \|(I - \check{U} \check{U}^\top) \check{X}\|_F = \sqrt{n - \text{Tr}(\check{U}^\top \check{X} \check{X}^\top \check{U})} \\ &= \sqrt{n - K + \sum_k \sin^2 \phi_k(U, X)}. \end{aligned}$$

Another possibility is to consider the principal angles ϕ_k between both subspaces:

$$d_d(U, X) = \sqrt{\sum_k \phi_k^2(U, X)}$$

See [13] for other possible dissimilarity measures.

Letting $\sigma_k = \cos \phi_k(U, X)$ denote the k th singular value of $\check{U}^\top \check{X}$, we can compare the different dissimilarities in Table 1 (with n_u and n_x dimensions of subspaces \mathcal{U} and \mathcal{X}). When using those dissimilarities in $\min_U \max_i d(U, X_i)$, d_b and d_c will give more importance to datasets X_i with a higher n_i . All dissimilarities except d_d can be directly expressed in terms of $\check{U}^\top X X^\top \check{U}$. As d_a and d_d respect $\mathcal{U} \subsetneq \mathcal{X} \Rightarrow d(U, X) = 0$, they are not distances. Note that if $n_x = n_u$, we have $\sqrt{d_a} = d_b = d_c$.

In the context of (1), it is natural to require that $d(U, X) = 0$ when $\mathcal{U} \subset \mathcal{X}$ or $\mathcal{X} \subset \mathcal{U}$. We opt for d_a , since it yields a simpler objective function than d_d . Hence, (1) becomes:

$$\min_{U \in \mathbb{R}^{p \times K}} \max_i K - \text{Tr}(\check{U}^\top \check{X}_i \check{X}_i^\top \check{U}).$$

Since K is fixed and \check{U} verifies $\check{U}^\top \check{U} = I_K$, this is equivalent to

$$\max_{U^\top U = I} \min_i \text{Tr}(U^\top \check{X}_i \check{X}_i^\top U). \quad (2)$$

Table 1. Summary of the dissimilarities

Formulation	Distance	$\mathcal{U} \subset \mathcal{X} \Rightarrow d(U, X) = 0$
$d_a(X, U) = \min(n_x, n_u) - \sum_k^{\min(n_x, n_u)} \cos^2(\phi_k)$	-	✓
$d_b(X, U) = \sqrt{\frac{n_x + n_u}{2} - \sum_k^{\min(n_x, n_u)} \cos^2(\phi_k)}$	✓	-
$d_c(X, U) = \sqrt{\max(n_x, n_u) - \sum_k^{\min(n_x, n_u)} \cos^2(\phi_k)}$	✓	-
$d_d(X, U) = \sqrt{\sum_k^{\min(n_x, n_u)} \phi_k^2}$	-	✓

Since $\max_U \min_i f_i(U)$ is equivalent to $\max_{U, \tau} \tau$ subject to $\tau \leq f_i(U)$ for all i , (2) is equivalent to:

$$\begin{aligned} & \max_{U, \tau} \tau \\ & \text{s.t. } \tau - \sum_{k=1}^K u_k^\top \check{X}_i \check{X}_i^\top u_k \leq 0 \quad \forall i = 1, \dots, m \end{aligned} \quad (3a)$$

$$u_j^\top u_j - 1 = 0 \quad \forall j = 1, \dots, K \quad (3b)$$

$$u_j^\top u_k = 0 \quad \forall k \neq j, j = 1, \dots, K ; k = 1, \dots, K \quad (3c)$$

with u_i the i th column of U . Observe that (3) is an optimization problem with a linear objective function and quadratic (in)equality constraints.

2.2 KKT Conditions

We derive the first order necessary conditions of optimality for problem (3). Associating Lagrange multipliers γ_i 's with constraints (3a), M_{jj} 's with constraints (3b) and M_{jk} 's with constraints (3c), the KKT conditions, see e.g., [14] can be written as:

$$\sum_i \gamma_i = 1 \quad (4a)$$

$$\left(\sum_i \gamma_i \check{X}_i \check{X}_i^\top \right) U = UM \quad (4b)$$

$$U^\top U = I \quad (4c)$$

$$\tau - \text{Tr}(U^\top \check{X}_i \check{X}_i^\top U) \leq 0 \quad \forall i = 1, \dots, m \quad (4d)$$

$$\gamma_i \geq 0 \quad \forall i = 1, \dots, m \quad (4e)$$

$$\gamma_i (\tau - \text{Tr}(U^\top \check{X}_i \check{X}_i^\top U)) = 0 \quad \forall i = 1, \dots, m \quad (4f)$$

The M_{ij} 's correspond to the Lagrange multipliers associated with constraints $u_i^\top u_j = 0$ and the M_{ii} 's to $u_i^\top u_i - 1 = 0$, so M is symmetric. Therefore there exist a diagonal matrix D and an orthogonal matrix Q such that $M = QDQ^\top$. We have then $(\sum_i \gamma_i \check{X}_i \check{X}_i^\top) U Q = U Q D$ which means that $U Q$ is a matrix of

eigenvectors of $\sum_i \gamma_i \check{X}_i \check{X}_i^\top$. The γ_i 's can be interpreted as the importance given to the corresponding subspaces, and are positive only for those subspaces that achieve the max of problem (3).

Let $U_Y D_Y V_Y^\top$ be the singular value decomposition of

$$Y = [\sqrt{\gamma_1} \check{X}_1, \sqrt{\gamma_2} \check{X}_2, \dots, \sqrt{\gamma_m} \check{X}_m] \in \mathbb{R}^{p \times N}.$$

Observe that $U_Y D_Y^2 U_Y^\top$ is then an eigendecomposition of $Y Y^\top$. A candidate solution of problem (3) would then be, for fixed γ_i respecting condition (4f):

$$\begin{aligned} M_{ij} &= 0 \quad \forall i \neq j & U &= U_Y \\ M_{ii} &= D_Y^2(i, i) & \tau &= \text{Tr}(U^\top Y Y^\top U). \end{aligned}$$

The last equality results from the combination of conditions (4a) and (4f):

$$\tau = \sum_i \gamma_i \tau = \sum_i \gamma_i \text{Tr}(U^\top X_i X_i^\top U).$$

To maximize τ , we should consider the K first singular values of Y . The difficulty is then to find γ_i such that condition (4f) is respected.

We can easily see that unless the optimal U belongs to all subspaces \mathcal{X}_i , more than one γ_i is nonzero. To see this, observe that if $\gamma_i = 0$ for all $i \neq j$, constraint (4b) would imply that U belongs to subspace \mathcal{X}_j , which means that $\text{Tr}(U^\top \check{X}_j \check{X}_j^\top U) = K$ and $\tau = K$ by condition (4f). Since for all i, k we have $0 \leq u_k^\top \check{X}_i \check{X}_i^\top u_k \leq 1$ and $\text{Tr}(U^\top \check{X}_i \check{X}_i^\top U) \geq \tau = K$ by condition (4d), we have $\text{Tr}(U^\top \check{X}_i \check{X}_i^\top U) = K$ for all i , and U belongs to all the other \mathcal{X}_i 's. As a result, any candidate solution should have at least two \mathcal{X}_i 's realizing the optimum.

3 Proposed Approach

In [9], a fast and simple procedure is proposed to find an approximation of the minimum enclosing ball center of a finite-dimensional Euclidean space. The procedure is extended to arbitrary Riemannian manifolds in [10]:

- Initialize the candidate solution $U^{(t)}$ with a point in the set
- Iteratively update as $U^{(t+1)} = \text{Geodesic}\left(U^{(t)}, X_f^{(t)}, \frac{1}{t+1}\right)$, where $X_f^{(t)}$ is the farthest point to $U^{(t)}$, and $\text{Geodesic}(p, q, t)$ represents the intermediate point m on the geodesic passing through p and q such that $\text{dist}(p, m) = \text{dist}(p, q)$.

Since we are interested in finding the best subspace of dimension K in \mathbb{R}^p , our solution U belongs to the Grassmann manifold $\mathcal{G}(K, p)$. The main difference with [10] is that we are dealing with points representing subspaces of different dimensions n_i and therefore belonging to different manifolds $\mathcal{G}(n_i, p)$. The first consequence is that the usual Grassmannian distance cannot be used to determine the farthest point $X_f^{(t)}$. Since we want to preserve $d(U, X_i) = 0$ when

$\mathcal{U} \subset \mathcal{X}_i$, we used a dissimilarity which is not a metric except if the two subspaces belongs to the same Grassmannian. The second consequence is that to update the current iterate $U^{(t)}$ using a geodesic, $X_f^{(t)}$ should be first projected on $\mathcal{G}(K, p)$. The next proposition shows how, given $\mathcal{X}_f \in \mathcal{G}(n_f, p)$ and $\mathcal{U} \in \mathcal{G}(K, p)$ with $n_f \geq K$, we can compute $\mathcal{Y}_f \in \mathcal{G}(K, p)$ included in \mathcal{X}_f that minimizes the distance to \mathcal{U} . We can then update U using the corresponding geodesic.

Proposition 1. *Let $\mathcal{Y}, \mathcal{U} \in \mathcal{G}(K, p)$ and $\mathcal{X} \in \mathcal{G}(n, p)$ where $n \geq K$, with \check{X} and \check{U} orthonormal basis of \mathcal{X} and \mathcal{U} . Let $A_1 D_1 B_1^T$ be an SVD of $\check{U}^T \check{X}$, then we have*

$$\min_{\mathcal{Y} \subset \mathcal{X}} d_a(\mathcal{Y}, \mathcal{U}) = d_a(\mathcal{X}, \mathcal{U}) = d_a(\text{Col}(X B_1), \mathcal{U}).$$

Those equalities hold also for d_d .

Algorithm 1. Heuristic to extract a subspace minimizing the maximal dissimilarity with the X_i .

Require: tol, ϵ , X_1, \dots, X_N

```

1:  $t \leftarrow 1$ 
2:  $U^{(0)} D^{(0)} V^{(0)T} \leftarrow \text{SVD}([\check{X}_1 \dots \check{X}_m], K)$ 
3:  $\text{err}^{(0)} \leftarrow \text{tol} + 1$ 
4: while  $\text{err}^{(t)} \geq \text{tol}$  do
5:   for all  $\check{X}_j$  do
6:      $d_j^{(t)} \leftarrow d_a(U^{(t)}, \check{X}_j)$ 
7:   end for
8:    $i_{\max} \leftarrow \arg \max_i d_i^{(t)}$ 
9:    $U_c D_c V_c^T \leftarrow \text{SVD}(\check{X}_{i_{\max}}^T \check{U}^{(t)}, K)$ 
10:   $S_0 \leftarrow U^{(t)} V_c$ 
11:   $S_1 \leftarrow \check{X}_{i_{\max}} U_c$ 
12:   $\Theta \leftarrow \arccos \text{diag } D_c$ 
13:   $\Gamma_\alpha \leftarrow \text{diag } \cos \alpha \Theta$ 
14:   $\Sigma_\alpha \leftarrow \text{diag } \sin \alpha \Theta$ 
15:   $\delta \leftarrow \frac{1}{t+1}$ 
16:   $U^{(t+1)} \leftarrow S_0 \Gamma_\delta + (S_1 - S_0 \Gamma_1) \Sigma_1^{-1} \Sigma_\delta$ 
17:   $d_{\text{sorted}} \leftarrow \text{sort}_{\text{decreasing}}(d^{(t)})$ 
18:   $\text{err}^{(t)} \leftarrow d_{\text{sorted}}(1) - d_{\text{sorted}}(2)$ 
19:   $t \leftarrow t + 1$ 
20: end while

```

An adaptation is proposed in Algorithm 1, integrating results obtained from the KKT conditions analysis. We initialize using a K -truncated SVD of $Y = [\check{X}_1, \check{X}_2, \dots, \check{X}_m]$, corresponding to the case where all the γ_i 's are equal (line 2), and stop when the two farthest subspaces have close dissimilarity values (line 18). As explained in Subsect. 2.2, this is a necessary, but not sufficient, condition at optimality. The farthest X_i from current $U^{(t)}$ is determined using the chosen dissimilarity based on the principal angles (lines 5 to 8). The associated

orthonormal basis S_0 and S_1 of \mathcal{U} and \mathcal{X}_{imax} are computed (lines 9 to 11) to update $U^{(t)}$ in the direction of X_{imax} with a step $\frac{1}{t+1}$ along the Grassmannian geodesic [15] (lines 12 to 16).

4 Experiments

We generated synthetic data to represent a case where datasets are unevenly distributed in space and the minimax approach is justified. We first generated a common subspace $U_c \in \mathbb{R}^{p \times K_c} \sim N(0, 1)$. We then perturbed it to generate two different noisy versions $U_j = U_c + N(0, s_j \mu_{U_c})$, $j \in \{1, 2\}$, with $\mu_{U_c} = \text{mean}(|U_c|)$, from which we generated two groups of data. For each U_j , $j \in \{1, 2\}$, we generated different datasets X_i :

$$X_i = [U_j \ A_i] \begin{bmatrix} V_i^\top \\ B_i^\top \end{bmatrix}$$

where $B_i \in \mathbb{R}^{n_i \times K_i}$ is distributed $\sim U_{[0,1]}$, and $A_i \in \mathbb{R}^{p \times K_i} \sim N(0, 1)$. Each column of matrices U_j , A_i and B_i is normalized (using the L_2 norm) to give the same importance to each component within the dataset. Each column $V_i(:, j)$ of $V_i \in \mathbb{R}^{n_i \times K_c}$ is distributed $\sim U_{[0, \frac{3w_{ij}}{p}]}$, where w_{ij} represents the importance of the common component j within dataset i . Finally, Gaussian noise $\epsilon_i \sim N(0, \sigma_i * \mu_{X_i})$ is added to each datasets: $X_i \leftarrow X_i + N(0, \sigma_i * \mu_{X_i})$ with $\mu_{X_i} = \text{mean}(|X_i|)$.

We generated datasets in two groups: the first, based on U_1 , contains more datasets but with higher noise, while the second group, based on U_2 , contains fewer less noisy datasets. The first group contains 17 datasets with $s_1 = 1$, while the second contains 3 datasets with $s_2 = 0.1$. We took $K_c = 3$ common components and $K_i = 5$ additional components, $p = 1000$ features and $n_i \sim U_{[20 \ 220]}$ samples for each dataset X_i . The weights w_{ij} were randomly generated as $\sim U_{[0.05 \ 0.5]}$ to 'hide' the common components in the datasets. The final added noise has $\sigma_i = 0.1$.

We compared our Grassmannian Minimum Enclosing Ball approach GMEB_{da} described in Algorithm 1 to a K -truncated SVD on $X = [X_1 \dots X_n]$ (SVD) and $\check{X} = [\check{X}_1 \dots \check{X}_n]$ (SVD_o). Working with \check{X} instead of X improves the recovery of components that are (weakly) present in all X_i 's. For each subspace obtained, we computed its maximal dissimilarity to \check{X}_i , but also to the background truth U_c and the two noisy U_j . Mean results on 100 randomly generated datasets are shown on Fig. 1, where we also give results when using dissimilarities d_b , d_c or d_d in Algorithm 1.

When computing dissimilarities to the U 's, we logically have $\sqrt{d_a} = d_b = d_c$ since, in these cases, n_x and n_u of Table 1 are equivalent. Results obtained for d_b and d_c with \check{X}_i are similar for all methods, due to the influence of n_i in the dissimilarities. Since we have $d_c(U, X_i) \in [\sqrt{n_i - K}, \sqrt{n_i}]$ and $n_i > K$, the results are mainly influenced by $\max_i n_i$. On the criterion minimized (d_a on \check{X}_i), the common subspace approach is the best one. As expected, SVD_o recovers very well the noisy components U_1 , but the common subspace approach recovers better the U_2 . The original U_c is then recovered better by the subspace approach than by SVD_o .

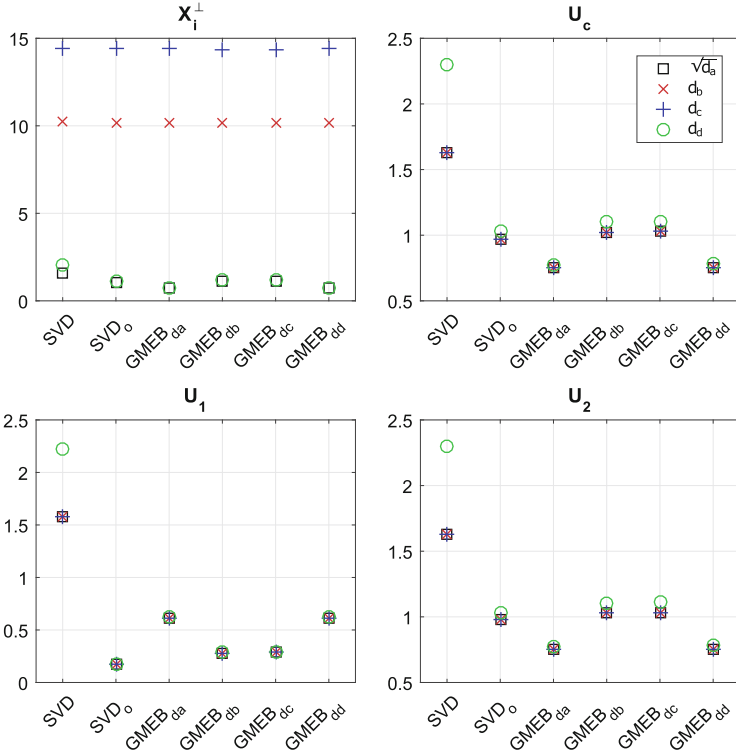


Fig. 1. Mean on 100 tests of maximal dissimilarity, for different dissimilarities and methods. Observe that methods GMEB_{da} and GMEB_{dd} perform best at recovering the ground truth U_c .

5 Conclusion

In this paper, we examined the problem of finding a subspace representative of multiple datasets by minimizing the maximal dissimilarity between this subspace and all the subspaces generated by those datasets. After arguing for a particular choice of dissimilarity measure, we derived some properties of the corresponding formulation. Based on those properties, we proposed an adaptation of an algorithm used for a similar problem on a Riemannian manifold. We then tested the proposed algorithm on synthetic data. Compared to SVD, the subspace recovered by our algorithm is closer to the true common subspace. Based on these promising results, the next step is to analyze properly the convergence of the proposed algorithm. Other approaches to solve the problem should also be investigated, for example based on the KKT conditions or on linearization.

Acknowledgments. Part of this work was performed while the second author was a visiting professor at Université catholique de Louvain.

References

1. Alter, O., Brown, P.O., Botstein, D.: Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci.* **100**(6), 3351–3356 (2003)
2. Ponnappalli, S.P., Saunders, M.A., Van Loan, C.F., Alter, O.: A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PloS one* **6**(12), e28072 (2011)
3. Hotelling, H.: Relations between two sets of variates. *Biometrika* **28**(3/4), 321–377 (1936)
4. Wold, H.: Partial least squares. *Encycl. Stat. Sci.* **6**, 581–591 (1985)
5. Meng, C., Kuster, B., Culhane, A.C., Gholami, A.M.: A multivariate approach to the integration of multi-omics datasets. *BMC Bioinf.* **15**(1), 162 (2014)
6. Hanafi, M., Kohler, A., Qannari, E.M.: Connections between multiple co-inertia analysis and consensus principal component analysis. *Chemometr. Intell. Lab. Syst.* **106**(1), 37–40 (2011)
7. Tenenhaus, A., Tenenhaus, M.: Regularized generalized canonical correlation analysis. *Psychometrika* **76**(2), 257–284 (2011)
8. Westerhuis, J.A., Kourti, T., MacGregor, J.F.: Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometr.* **12**(5), 301–321 (1998)
9. Badoiu, M., Clarkson, K.L.: Smaller core-sets for balls. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, pp. 801–802 (2003)
10. Arnaudon, M., Nielsen, F.: On approximating the Riemannian 1-center. *Comput. Geom.* **46**(1), 93–104 (2013)
11. Angulo, J.: Structure tensor image filtering using Riemannian L_1 and L_∞ center-of-mass. *Image Anal. Stereol.* **33**(2), 95–105 (2014)
12. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**(2), 303–353 (1998)
13. Ye, K., Lim, L.H.: Schubert varieties and distances between subspaces of different dimensions. *SIAM J. Matrix Anal. Appl.* **37**(3), 1176–1197 (2016)
14. Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006). <https://doi.org/10.1007/978-0-387-40065-5>
15. Gallivan, K.A., Srivastava, A., Liu, X., Van Dooren, P.: Efficient algorithms for inferences on grassmann manifolds. In: *IEEE Workshop on Statistical Signal Processing*, pp. 315–318 (2003)



Decoupling Multivariate Functions Using Second-Order Information and Tensors

Philippe Dreesen^(✉), Jeroen De Geeter, and Mariya Ishteva

Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium
{philippe.dreesen, jeroen.de.geeter, mariya.ishteva}@vub.be

Abstract. The power of multivariate functions is their ability to model a wide variety of phenomena, but have the disadvantages that they lack an intuitive or interpretable representation, and often require a (very) large number of parameters. We study decoupled representations of multivariate vector functions, which are linear combinations of univariate functions in linear combinations of the input variables. This model structure provides a description with fewer parameters, and reveals the internal workings in a simpler way, as the nonlinearities are one-to-one functions. In earlier work, a tensor-based method was developed for performing this decomposition by using first-order derivative information. In this article, we generalize this method and study how the use of second-order derivative information can be incorporated. By doing this, we are able to push the method towards more involved configurations, while preserving uniqueness of the underlying tensor decompositions. Furthermore, even for some non-identifiable structures, the method seems to return a valid decoupled representation. These results are a step towards more general data-driven and noise-robust tensor-based framework for computing decoupled function representations.

Keywords: Tensor · CPD · Function decomposition
Tensor decomposition · Waring decomposition · Polynomial

1 Introduction

1.1 Towards Interpretability of Nonlinear Models

Nonlinear models are used in a wide variety of science and engineering fields, such as data analytics, signal processing, system identification, and control engineering. While nonlinear models are able to capture wild nonlinear effects, this often comes at the cost of high parametric complexity, and a lack of ‘model interpretability’.

This paper studies the question how a given nonlinear multivariate vector function $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ can be decomposed into a simpler structure, as in [5, 9, 18, 19, 21]. In particular, we investigate a structure of the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{g}(\mathbf{V}^T \mathbf{x}), \quad (1)$$

where \mathbf{W} and \mathbf{V} are transformation matrices, and the vector function $\mathbf{g}(\mathbf{z}) = [g_1(z_1) \cdots g_r(z_r)]^T$ is composed of univariate functions $g_i(z_i)$ in its r components. The decoupled representation is visualized in Fig. 1.

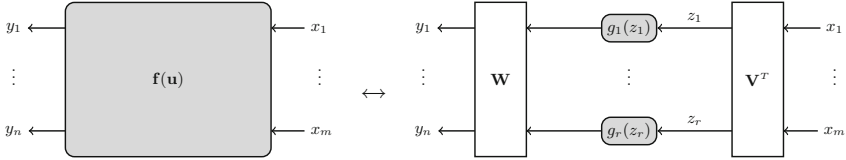


Fig. 1. A multivariate nonlinear vector function $\mathbf{f}(\mathbf{x})$ can be represented in a decoupled representation $\mathbf{f}(\mathbf{x}) = \mathbf{W}\mathbf{g}(\mathbf{V}^T\mathbf{x})$. A decoupled representation typically has fewer parameters, and reveals in an intuitive way, the internal nonlinearities of $\mathbf{f}(\mathbf{x})$.

The decomposition (1) provides a representation that is easier to comprehend, as the nonlinearity is contained in a set of univariate components. Moreover, it typically has a lower parametric complexity, and could hence be viewed as a form of nonlinear model order reduction.

1.2 Tensorization Methods for Decoupling Polynomials

When \mathbf{f} is polynomial, the decomposition (1) has connections to the canonical polyadic decomposition (CPD) of a partially symmetric tensor (possibly a joint decomposition), see [20]. Indeed, typical tensorization methods make use of the connection between homogeneous polynomials and their coefficient tensors [4, 20]. The (homogeneous) scalar polynomial case, *i.e.*, $n = 1$, is known as the Waring decomposition [2, 13], and is a fundamental problem in algebraic geometry. The problem (1) we study is hence very reminiscent of the classical Waring problem, but we consider the *non-homogeneous case of several polynomials*. The non-homogeneous Waring problem is studied in [1, 14]. The simultaneous Waring problem for several homogeneous polynomials is studied in [2, 18].

1.3 Contributions and Organization of This Paper

We start from the tensorization method of [9]. In this framework, the function \mathbf{f} and its first-order derivative information are evaluated in a number of sampling points. A tensor is constructed from the set of corresponding Jacobian matrices, which admits a CPD that allows for the reconstruction of the decomposition (1). This approach has the following advantages: (i) the order and size of the constructed tensor do not increase with the degree of \mathbf{f} , and, (ii) the approach is not limited to the use of polynomials.

In the current paper, we generalize the method [9] to incorporate second-order derivative information. Since second-order derivative information leads to partially symmetric tensors, we are ultimately able to formulate the decomposition as a partially symmetric joint tensor decomposition. By involving the

second-order derivatives, we impose additional constraints on the (joint) tensor decompositions, hence it is expected to enjoy more relaxed uniqueness conditions. In the article, we assume that an exact and uniquely identifiable [5] representation of $\mathbf{f}(\mathbf{x})$ exists. Nevertheless, the resulting joint tensor decomposition will ultimately be phrased as an optimization problem, and provides a natural starting point for studying both the noisy decoupling problem, as well as a model reduction interpretation, but this is beyond the scope of the current paper.

The current article is organized as follows. Section 2 outlines the tensor-based decoupling method, leading to a joint tensor decomposition formulation. We illustrate how the uniqueness properties improve by including second-order derivatives. Section 3 validates the method on three simulation examples. Section 4 summarizes the results and points out a few future research directions.

1.4 Notation

Scalars are denoted by lowercase or uppercase letters and vectors are denoted by lowercase bold-face letters. Elements of a vector are denoted by lowercase letters with an index as subscript, *e.g.*, $\mathbf{x} = [x_1 \dots x_m]^T$. Matrices are denoted by uppercase bold-face letters, *e.g.*, $\mathbf{V} \in \mathbb{R}^{m \times r}$. The entry in the i -th row and j -th column of the matrix \mathbf{V} is denoted by v_{ij} . A matrix $\mathbf{V} \in \mathbb{R}^{m \times r}$ has columns \mathbf{v}_i as in $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_r]$. The transpose of a matrix \mathbf{V} is denoted by \mathbf{V}^T . A diagonal matrix with diagonal elements a_1, a_2, a_3 is denoted by $\text{diag}(a_1, a_2, a_3)$ or $\text{diag}(a_i)$. Higher-order tensors are denoted by bold-face uppercase caligraphical letters, *e.g.*, $\mathcal{J} \in \mathbb{R}^{n \times m \times N}$. For scalar, vector, matrix and higher-order tensor *functions*, we employ the same conventions. The outer product is denoted by \circ and defined as follows: For $\mathcal{X} = \mathbf{u} \circ \mathbf{v} \circ \mathbf{w}$, the entry in position (i, j, k) is $u_i v_j w_k$. The Frobenius norm of a tensor \mathcal{X} is denoted as $\|\mathcal{X}\|_F$. The Euclidean norm of a vector \mathbf{x} is denoted as $\|\mathbf{x}\|$. The first-order and second-order derivatives of a univariate function $g(z)$ are denoted by $g'(z)$ and $g''(z)$, respectively.

2 Decoupling Multivariate Functions Using Tensors

2.1 The Canonical Polyadic Decomposition

The canonical polyadic decomposition (CPD) [3, 10, 11] is the decomposition of a tensor into a minimal sum of rank-one components. For instance, a third-order tensor \mathcal{T} has a CPD of the form

$$\mathcal{T} = \sum_{i=1}^R \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i, \quad (2)$$

or in a short-hand notation $\mathcal{T} = \llbracket \mathbf{A}, \mathbf{B}, \mathbf{C} \rrbracket$, where $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_R]$ (similar for \mathbf{B} and \mathbf{C}). The CPD is a celebrated tensor decomposition, which has found a variety of applications in signal processing and data sciences. One of the attractive properties of the CPD is its more relaxed uniqueness conditions. In contrast with

matrix factorization, where uniqueness is only possible by imposing additional constraints (*e.g.*, orthogonality in the singular value decomposition (SVD)), the CPD has milder uniqueness properties [6–8, 11, 12].

In our framework, the uniqueness conditions of the CPD will allow us to ensure that the proposed decoupling method retrieves the uniquely identifiable model structure: for identifiable models, uniqueness of the CPD is a sufficient condition for uniqueness of the model.

2.2 Decoupling Functions Using First-Order Information

Consider the function $\mathbf{f} : \mathbb{R}^m \rightarrow \mathbb{R}^n$ that admits a decoupled representation (1). The Jacobian $\mathbf{J}(\mathbf{x})$ is represented with an $n \times m$ matrix function defined as

$$\mathbf{J}_{ij}(x) = \frac{\partial f_i(\mathbf{x})}{\partial x_j}. \quad (3)$$

The chain rule for derivation shows that the Jacobian matrix $\mathbf{J}(\mathbf{x})$ can be factorized as $\mathbf{J}(\mathbf{x}) = \mathbf{W} \text{diag}(g'_i(\mathbf{v}_i^T \mathbf{x})) \mathbf{V}^T$, or alternatively, in a CPD formulation $\mathcal{J} = \llbracket \mathbf{W}, \mathbf{V}, (\mathbf{g}'(\mathbf{V}^T \mathbf{x}))^T \rrbracket$. Then an $n \times m \times N$ tensor \mathcal{J} , built from evaluating the Jacobian matrix $\mathbf{J}(\mathbf{x}^{(k)})$ in a set of N sampling points $\mathbf{x}^{(k)}, k = 1, \dots, N$, admits a CPD of the form

$$\mathcal{J} = \llbracket \mathbf{W}, \mathbf{V}, \mathbf{G}' \rrbracket, \quad (4)$$

where \mathbf{G}' contains the first-order derivatives of the functions g_i in the N points, *i.e.*, $\mathbf{G}'_{ki} = g'_i(\mathbf{v}_i^T \mathbf{x}^{(k)})$. In this way, the decoupled representation can be reconstructed from a simultaneous matrix diagonalization, or a CPD (Fig. 2).

2.3 Decoupling Functions Using Second-Order Information

Along the same lines, we may represent the Hessian $\mathcal{H}(\mathbf{x})$ by an $n \times m \times m$ tensor function defined as

$$\mathcal{H}_{ijk}(\mathbf{x}) = \frac{\partial^2 f_i(\mathbf{x})}{\partial x_j \partial x_k}, \quad (5)$$

which is symmetric in the second and third mode since $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$. The Hessian tensor function has a CPD representation of the form

$$\mathcal{H}(\mathbf{x}) = \llbracket \mathbf{W}, \mathbf{V}, \mathbf{V}, (\mathbf{g}''(\mathbf{V}^T \mathbf{x}))^T \rrbracket. \quad (6)$$

By evaluating the Hessian in a set of N sampling points $\mathbf{x}^{(k)}, k = 1, \dots, N$, we find the CPD of the $n \times m \times m \times N$ tensor \mathcal{H} as

$$\mathcal{H} = \llbracket \mathbf{W}, \mathbf{V}, \mathbf{V}, \mathbf{G}'' \rrbracket, \quad (7)$$

where the columns of \mathbf{G}'' contain the second-order derivatives of the functions g_i in the N points, *i.e.*, $\mathbf{G}''_{ki} = g''_i(\mathbf{v}_i^T \mathbf{x}^{(k)})$.

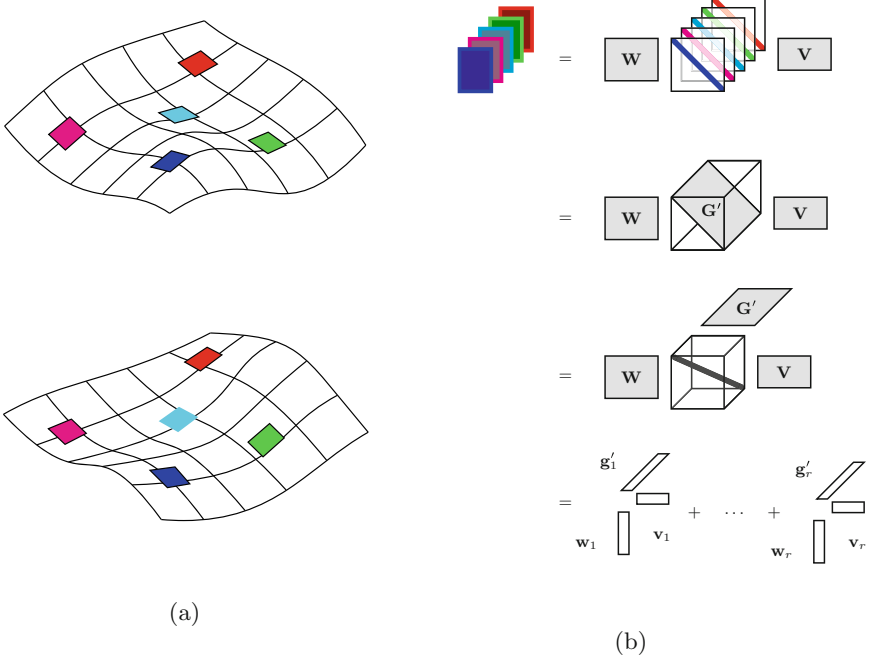


Fig. 2. The first-order information of \mathbf{f} is collected in a set of sampling points $\mathbf{x}^{(k)}$, with $k = 1, \dots, N$ (indicated by the colored patches on the surfaces in (a)). The corresponding Jacobian matrices $\mathbf{J}(\mathbf{x}^{(k)})$ are arranged into a three-way tensor (b). Each Jacobian matrix can be written as $\mathbf{J}(\mathbf{x}^{(k)}) = \mathbf{W} \text{diag}(g'_i(\mathbf{v}_i^T \mathbf{x}^{(k)})) \mathbf{V}^T$. This results in a simultaneous matrix diagonalization problem, which is computed by the CPD.

2.4 A Joint Tensor Decomposition with Partial Symmetry

The first-order and second-order derivative information can be combined into a joint tensor decomposition with partial symmetry. This can be phrased into the Structured Data Fusion framework [15] and is implemented in tensorlab [22] for MATLAB. The underlying optimization problem is

$$\underset{\mathbf{W}, \mathbf{V}, \mathbf{G}', \mathbf{G}''}{\text{minimize}} \quad \alpha_1 \|\mathcal{J} - \llbracket \mathbf{W}, \mathbf{V}, \mathbf{G}' \rrbracket\|_F^2 + \alpha_2 \|\mathcal{H} - \llbracket \mathbf{W}, \mathbf{V}, \mathbf{V}, \mathbf{G}'' \rrbracket\|_F^2, \quad (8)$$

where the two terms in the cost function can be given different weights α_1 and α_2 . The factor matrices \mathbf{W} and \mathbf{V} are shared among both decompositions. The partial symmetry in the Hessian tensor can be recognized in the fact that the factor \mathbf{V} occurs twice. The joint decomposition approach is visualized in Fig. 3.

2.5 Some Remarks on Uniqueness

Our framework contains two notions that relate to ‘uniqueness’ or ‘identifiability’, which could be confusing. Therefore, it is useful to elaborate

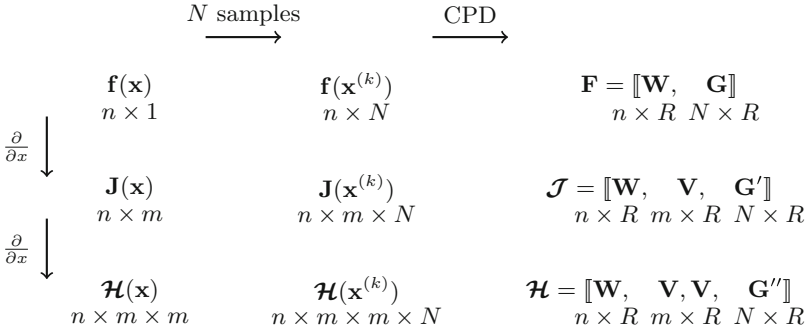


Fig. 3. The proposed method takes into account a combination of first-order and second-order derivatives evaluations. These evaluations are organized into two tensors which admit a joint canonical polyadic decomposition with partial symmetry. The corresponding cost function is $\alpha_1 \|\mathcal{J} - [\mathbf{W}, \mathbf{V}, \mathbf{G}']\|_F^2 + \alpha_2 \|\mathcal{H} - [\mathbf{W}, \mathbf{V}, \mathbf{V}, \mathbf{G}'']\|_F^2$.

briefly on this matter. There is a notion of uniqueness at the level of the function decomposition (1), as well as at the level of the CPD of a (corresponding) tensor. It is important to realize that in both cases, we consider so-called ‘essential uniqueness’, which makes abstraction of trivial scaling and permutation invariances.

On the one hand, in the polynomial case of (1), the problem at hand has a rich algebraic structure, which has recently been studied in the X-rank framework, leading to novel identifiability results [5]. These results assert that, for certain choices for m , n and r , the function (1) has a single ‘unique’ representation for generic choices of \mathbf{W} , \mathbf{V} and $\mathbf{g}(\mathbf{z})$. In other words, there does not exist an equivalent representation (1) having different \mathbf{W} , \mathbf{V} and $\mathbf{g}(\mathbf{z})$.

On the other hand, tensor decompositions have uniqueness properties themselves [6–8, 11, 12], which ensure that if the tensor decomposition has converged to a (numerical) zero error, then the (essentially) unique underlying factorization has been retrieved. Observe that uniqueness of the tensor decomposition is a sufficient condition for uniqueness of the function decomposition (1).

Uniqueness properties of joint CPDs have only been studied recently in [16, 17]. Intuitively, it can be expected that, by imposing additional constraints on a decomposition, it is likely that uniqueness conditions are more easily met.

3 Numerical Examples

In the current section, we will illustrate the proposed method on a number of examples. The joint decomposition (8) is implemented in tensorlab [22].

3.1 Second-Order Derivatives for Tackling the Waring Decomposition

In the single-output case, the Jacobian of the multivariate scalar function $f(\mathbf{x})$ is a vector function rather than a matrix function. Its representation simplifies to $f(\mathbf{x}) = \mathbf{w}^T \mathbf{g}(\mathbf{V}^T \mathbf{x})$. Notice that \mathbf{w} could be absorbed into the function \mathbf{g} , but we have chosen to keep it explicitly in the formula to show the resemblance to the vector function case. For a scalar function $f(\mathbf{x})$, the Jacobian reduces to a vector function, rather than a matrix function, *i.e.*, $\mathbf{j}^T(\mathbf{x}) = \llbracket \mathbf{w}^T, \mathbf{V}, (\mathbf{g}'(\mathbf{V}^T \mathbf{x}))^T \rrbracket$. Evaluating the Jacobian in a set of sampling points then gives rise to a matrix, rather than a tensor. Summarizing, the $n \times m \times N$ Jacobian tensor \mathcal{J} reduces in this situation to an $m \times N$ matrix \mathbf{J} , and we obtain a matrix factorization question, rather than the third-order tensor CPD (4). It is easy to understand that there is no unique solution, since one can insert $\mathbf{M}\mathbf{M}^{-1}$, with \mathbf{M} an invertible $R \times R$ matrix, to obtain an equivalent factorization $\mathbf{J} = (\mathbf{V}\mathbf{M})(\mathbf{G}'\mathbf{M}^{-T}) = \tilde{\mathbf{V}}\tilde{\mathbf{G}}'^T$.

A possible solution to resolve this lack of uniqueness is to consider second-order derivative information of $f(\mathbf{x})$. We evaluate the $m \times m$ Hessian function

$$\mathbf{H}_{ij}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}, \quad (9)$$

in a set of N sampling points $\mathbf{x}^{(k)}$. This gives an $m \times m \times N$ tensor \mathcal{H} with

$$\mathcal{H} = \llbracket \mathbf{w}^T, \mathbf{V}, \mathbf{V}, \mathbf{G}'' \rrbracket, \quad (10)$$

as in (7). Now, we have again a case in which uniqueness of the CPD is attainable.

For instance, we consider the function

$$f(x_1, x_2) = -37x_1^3 - 213x_1^2x_2 - 399x_1x_2^2 + 5x_1 - 239x_2^3 + 9x_2 - 2, \quad (11)$$

which can be decomposed with

$$\mathbf{w}^T = [1 \ 1], \quad \mathbf{V} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad (12)$$

and

$$\begin{aligned} g_1(z_1) &= 3z_1^3 - z_1 + 5, \\ g_2(z_2) &= -5z_2^3 + 3z_2 - 7. \end{aligned} \quad (13)$$

We draw an i.i.d. set of sampling points $\mathbf{x}^{(k)}$, $k = 1, \dots, 200$ from a uniform distribution between -10 and 10 in both components. The Jacobian matrix is hence a 2×200 matrix \mathbf{J} , which admits a non-unique rank-two factorization. If we take into account the Hessian information, the problem again becomes a uniquely defined tensor question. The CPD of \mathcal{H} is computed using tensorlab [22], and retrieves up to a scaling and permutation invariance, the true factors \mathbf{V} and \mathbf{G}'' .

3.2 Second-Order Derivatives Improve Uniqueness of the CPD

For a general vector function $\mathbf{f}(\mathbf{x})$, including the Hessian information can improve the uniqueness properties beyond the Jacobian tensor method. For instance, in the case $m = n = 2$, the bound by [9] ensures uniqueness up to $r \leq 2$. It can be verified that the Jacobian-based CPD method is not able to retrieve the underlying model. However, [5] asserts that $r = 3$ is still identifiable for polynomial models (with degree $d \geq 3$). Considering the second-order information then leads to a CPD that is generically unique.

We consider the function $\mathbf{f}(\mathbf{x})$ which is defined as

$$\begin{aligned} f_1(x_1, x_2) &= 24x_1^3 + 36x_1^2x_2 - 4x_1^2 + 18x_1x_2^2 - 4x_1x_2 + 84x_2^3 - x_2^2 - 6x_2 + 7, \\ f_2(x_1, x_2) &= -43x_1^3 - 72x_1^2x_2 + 8x_1^2 - 36x_1x_2^2 + 8x_1x_2 - 3x_1 \\ &\quad + 75x_2^3 + 2x_2^2 - 6x_2 - 1, \end{aligned} \quad (14)$$

and admits a representation with

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 1 \\ -2 & -1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}, \quad (15)$$

and

$$\begin{aligned} g_1(z_1) &= 3z_1^3 - z_1^2 + 5, \\ g_2(z_2) &= -5z_2^3 + 3z_2 - 7, \\ g_3(z_3) &= 3z_3^3 - 2z_3 + 2. \end{aligned} \quad (16)$$

A set of sampling points $\mathbf{x}^{(k)}$, $k = 1, \dots, 200$ is sampled again uniformly on $[-10, 10]^2$. Applying the Jacobian method results in an $2 \times 2 \times 200$ tensor which does not satisfy the uniqueness conditions. Indeed, although the CPD has an error of the order 10^{-11} , it does not return the correct factors. However, if we compute the CPD of the Hessian tensor, the underlying representation is found.

3.3 Can We Go Beyond Identifiable Structures?

In our experiments, we have observed a number of cases where adding Hessian info ensures interpretability, while the underlying model structure does not seem to be identifiable. For instance consider the $m = n = 2$ and $r = 4$ case of polynomials of degree $d = 3$. We consider a function $\mathbf{f}(\mathbf{x})$ of the form (1) with

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 1 & 2 \\ -2 & -1 & 1 & 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 0 & 3 & -1 \end{bmatrix}, \quad (17)$$

and

$$\begin{aligned} g_1(z_1) &= 3z_1^3 - z_1^2 + 5, \\ g_2(z_2) &= -5z_2^3 + 3z_2 - 7, \\ g_3(z_3) &= 3z_3^3 - 2z_3 + 2, \\ g_4(z_4) &= z_4^3 - 2z_4^2 + 1. \end{aligned} \quad (18)$$

We observe that by considering the second-order derivatives information, we are able to retrieve a decomposition having (numerical) zero error. However, the linear transformations \mathbf{W} and \mathbf{V} are *not equal* (up to scaling and permutation) to the underlying factors. Nevertheless, when investigating the factor \mathbf{G}'' , we see that the retrieved factor does have in its components a set of linear relations as expected from $\mathbf{G}''_{ki} = g''_i(\mathbf{v}_i^T \mathbf{x}^{(k)})$. The fact that an interpretable model is obtained is a surprising result: it seems to suggest that, considering the second-order information enforces that only ‘interpretable’ models are retrieved. However, we should mention that we have observed that this effect does not always hold for other cases.

4 Conclusions and Perspectives

In this article, we generalized a tensor-based method for finding a decoupled representation of a given nonlinear multivariate vector function. The method works by evaluating second-order derivatives in a set of sampling points. First-order and second-order information can be combined in this way into a simultaneous higher-order tensor decomposition task with partial symmetry. We illustrated the promising abilities of this approach on a number of simulation examples. The method was shown to outperform the existing approach: uniquely identifiable structures are recovered in a greater number of configurations, such as the single-output case, and cases in which the Jacobian tensor method was not able to ensure uniqueness. We also observed that the method seems to extract meaningful representations even in some cases when the model structure is not uniquely identifiable, which is a property that merits further investigation.

In future work, we want to investigate how function evaluations can be taken into account: this seems to make sense when the corresponding (matrix) factorization is a low-rank approximation, which occurs when $r < n$. Also the use of higher-order derivatives is a possible extension, which might improve the uniqueness conditions even further. In this sense, the results in this article are a step towards a more general data-driven and noise-robust tensor-based framework for decoupling function representations.

Acknowledgments. This work was supported in part by the Flemish Government (Methusalem), and by the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no 30468160 and research projects G.0280.15N and G.0901.17N.

References

1. Białynicki-Birula, A., Schinzel, A.: Representations of multivariate polynomials as sums of polynomials in linear forms. *Colloq. Math.* **112**(2), 201–233 (2008)
2. Carlini, E., Chipalkatti, J.: On Waring’s problem for several algebraic forms. *Comment. Math. Helv.* **78**, 494–517 (2003)
3. Carroll, J., Chang, J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika* **35**(3), 283–319 (1970)

4. Comon, P., Golub, G., Lim, L.H., Mourrain, B.: Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.* **30**(3), 1254–1279 (2008)
5. Comon, P., Qi, Y., Usevich, K.: Identifiability of an X-rank decomposition of polynomial maps. *SIAM J. Appl. Algebra and Geom.* **1**(1), 388–414 (2017)
6. De Lathauwer, L.: A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM J. Matrix Anal. Appl.* **28**(3), 642–666 (2006)
7. Domanov, I., De Lathauwer, L.: On the uniqueness of the canonical polyadic decomposition of third-order tensors – part I: basic results and uniqueness of one factor matrix. *SIAM J. Matrix Anal. Appl.* **34**(3), 855–875 (2013)
8. Domanov, I., De Lathauwer, L.: On the uniqueness of the canonical polyadic decomposition of third-order tensors – part II: uniqueness of the overall decomposition. *SIAM J. Matrix Anal. Appl.* **34**(3), 876–903 (2013)
9. Dreesen, P., Ishteva, M., Schoukens, J.: Decoupling multivariate polynomials using first-order information and tensor decompositions. *SIAM J. Matrix Anal. Appl.* **36**(2), 864–879 (2015)
10. Harshman, R.A.: Foundations of the PARAFAC procedure: model and conditions for an “explanatory” multi-mode factor analysis. *UCLA Working Pap. Phon.* **16**(1), 1–84 (1970)
11. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**(3), 455–500 (2009)
12. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Lin. Algebra Appl.* **18**, 95–138 (1977)
13. Oeding, L., Ottaviani, G.: Eigenvectors of tensors and algorithms for Waring decomposition. *J. Symb. Comput.* **54**, 9–35 (2013)
14. Schinzel, A.: On a decomposition of polynomials in several variables. *Journal de Théorie des Nombres de Bordeaux* **14**(2), 647–666 (2002)
15. Sorber, L., Van Barel, M., De Lathauwer, L.: Structured data fusion. *IEEE J. Sel. Top. Sig. Process.* **9**(4), 586–600 (2015)
16. Sørensen, M., De Lathauwer, L.: Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank- $(L_{r,n}, L_{r,n}, 1)$ terms—part I: uniqueness. *SIAM J. Matrix Anal. Appl.* **36**(2), 496–522 (2015)
17. Sørensen, M., De Lathauwer, L.: Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank- $(L_{r,n}, L_{r,n}, 1)$ terms—part II: algorithms. *SIAM J. Matrix Anal. Appl.* **36**(3), 1015–1045 (2015)
18. Tiels, K., Schoukens, J.: From coupled to decoupled polynomial representations in parallel Wiener-Hammerstein models. In: *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*, Florence, Italy, pp. 4937–4942 (2013)
19. Usevich, K.: Decomposing multivariate polynomials with structured low-rank matrix completion. In: *Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems (MTNS 2014)*, Groningen, The Netherlands, pp. 1826–1833 (2014)
20. Usevich, K., Dreesen, P., Ishteva, M.: Decoupling multivariate polynomials: interconnections between tensorizations (2017). Preprint [arXiv:1703.02493](https://arxiv.org/abs/1703.02493)
21. Van Mulders, A., Vanbeylen, L., Usevich, K.: Identification of a block-structured model with several sources of nonlinearity. In: *Proceedings of the 13th European Control Conference* (2014)
22. Vervliet, N., Debals, O., Sorber, L., Van Barel, M., De Lathauwer, L.: Tensorlab 3.0 (2016). <http://www.tensorlab.net/>



Nonnegative PARAFAC2: A Flexible Coupling Approach

Jeremy E. Cohen^{1,2} and Rasmus Bro^{1,2}

¹ Département de Mathématiques et Opérationnel Recherche, Faculté polytechnique, Université de Mons, Rue de Houdain 9, Mons, Belgium

jeremy.cohen@umons.ac.be

² Département of Food Science, University of Copenhagen, Rolighedsvej 26 1958 Frederiksberg C, Copenhagen, Denmark

Abstract. Modeling variability in tensor decomposition methods is one of the challenges of source separation. One possible solution to account for variations from one data set to another, jointly analysed, is to resort to the PARAFAC2 model. However, so far imposing constraints on the mode with variability has not been possible. In the following manuscript, a relaxation of the PARAFAC2 model is introduced, that allows for imposing nonnegativity constraints on the varying mode. An algorithm to compute the proposed flexible PARAFAC2 model is derived, and its performance is studied on both synthetic and chemometrics data.

Keywords: PARAFAC2 · Nonnegativity constraints
Flexible coupling

1 Introduction

The PARAFAC2 model is an interesting alternative to the more widespread PARAFAC model [7]. As opposed to PARAFAC, it allows for non-linearities such that the data need not behave according to a low-rank trilinear model. In fact, it can even handle sub-matrices (slabs) of varying length. This is often useful for example when one of the modes is a time mode [13, 14]. One of the prime uses of PARAFAC2 is in the resolution of chromatographic data [1, 5, 9].

The PARAFAC2 model has shown to have a remarkable ability to resolve complicated chromatographic data. A typical three-way dataset will have one mode made up of the various physical samples measured. These could e.g. be different milk samples. Another mode will be reflecting elution time which is a physical separation of the constituents of the sample over time. The last mode refers to the spectral detection such as mass spectrometry—which represents the actual measurement of a mass spectrum at each time point for each sample. A successful PARAFAC2 model will provide parameters for each mode, so-called loading matrices, that will ideally represent the concentrations of the

Research funded by F.R.S.-FNRS incentive grant for scientific research n° F.4501.16.

chemical compounds measured, the corresponding elution time profiles and the corresponding pure analyte mass spectra. Usually, the PARAFAC2 model is only applied to a narrow time interval as for example a timespan of a few overlapping peaks that are hard to separate without the use of PARAFAC2.

In the context of chromatographic data, the ‘Selling point’ of PARAFAC2 is that it allows the elution profile of a given chemical to be different in each experiment. If chromatographic data would be modeled with a PARAFAC model and most other conventional curve resolution methods, they would require a given chemical to have the same elution shape in every sample. Unfortunately, that is almost never the case. There will often be retention time shifts and other shape changes that makes it impossible to model the data with a conventional approach. The PARAFAC2 model, though, can handle this type of artefacts quite well.

In many cases, it is desired that the parameters are constrained to be nonnegative. Most notably because ideally, concentrations, elution profiles and spectra are nonnegative. Unfortunately, it is not hitherto possible to constrain all the parameters to be nonnegative. The ‘elution time’ mode of the PARAFAC2 model is estimated implicitly as a product of two matrices and so far no algorithms have been presented that allows imposing nonnegativity on the product of those two matrices. In this paper, we will develop such an algorithm. In the first section, the PARAFAC2 model is cast as a coupled matrix factorization model, which is used in Sect. 4 to derive an algorithm for computing Flexible PARAFAC2 with nonnegativity constraints. Finally, Sect. 5 shows the performance of the proposed method on both synthetic and real world data.

2 Reminders on the PARAFAC2 Model

The PARAFAC2 model was first introduced by Harshman in the context of phonetics [7]. In his work, Harshman looked for a way to factorize simultaneously several matrices given that one factor was almost the same, but not exactly, in all those matrices. He thus imposed a linear transformation as a coupling relationship between the similar factors. However using a generic linear coupling model adds too many parameters, and to ensure identifiability of both the factors and the coupling matrices, orthogonality constraints were imposed. This leads to the following PARAFAC2 model:

$$M_k = AD_k B_k^T + E_k, \quad B_k = P_k B^*, \quad P_k^T P_k = I_r, \quad (1)$$

where B^* is a $r \times r$ matrix of coefficients, D_k is a $r \times r$ diagonal matrix, P_k is a $m_k \times r$ left-orthogonal matrix and E_k is a $n \times m_k$ residual error matrix. Here the coupled matrices are the B_k , and the coupling matrices, the P_k .

Another way to understand PARAFAC2, more widely used in the tensor community, is to cast it as a relaxation of the PARAFAC model. Indeed, stacking matrices M_k into one large tensor \mathcal{T} , the PARAFAC2 model yields:

$$\begin{aligned}
 T_{ijk} &= \sum_{p=1}^r A_{ip} B_{jp}^{(k)} C_{kp} \\
 \sum_{p=1}^r B_{j_1 p}^{(k)} B_{j_2 p}^{(k)} &= \sum_{p=1}^r B_{j_1 p}^* B_{j_2 p}^* \quad \forall j_1, j_2, k,
 \end{aligned} \tag{2}$$

where C is obtained by stacking the diagonals of D_k in rows. One can observe that contrary to the PARAFAC model, the B factor is allowed to vary for each slice k . This variation is controlled by the inner products stored in B^* and kept constant through k . As a matter of fact, the orthogonality constraints on the P_k matrices are equivalent to imposing a shared Gramian matrix for all B_k , that is $B_k^T B_k = B^{*T} B^*$ for all k . The power of PARAFAC2 comes from the fact that this constraint is implicit, and may give birth to a wide range of variability among the B_k while maintaining an overall coupling structure. In contrast, other similar models like Shift-PARAFAC impose a coupling constraint in an explicit fashion that may be too specific and difficult to implement [8, 11].

To identify the parameters of the (unconstrained) PARAFAC2 model, the following optimization problem needs to be solved:

$$\begin{aligned}
 &\underset{A, D_k, P_k, B^*}{\operatorname{argmin}} \sum_{k=1}^K \|M_k - A D_k (P_k B^*)^T\|_F^2 \\
 &\text{so that } P_k^T P_k = I
 \end{aligned} \tag{3}$$

An efficient alternating algorithm to solve (3) has been introduced in [10]. It relies on the fact that if the P_k matrices are known, then multiplying each data slice M_k by P_k on the right, the PARAFAC2 model becomes a PARAFAC model with second mode factor B^* . Therefore, an alternating algorithm may first estimate P_k fixing the other parameters, then pre-process the data by multiplying each slice with P_k^T , and then use a few step of an algorithm to compute PARAFAC, for instance Alternating Least Squares [4]. The estimation of the orthogonal coupling matrices is easily obtained with SVD, knowing that the solution of

$$\begin{aligned}
 &\underset{P \in \mathbb{R}^{m \times r}}{\operatorname{argmin}} \|M - PX\|_F^2 \\
 &\text{such that } P^T P = I
 \end{aligned} \tag{4}$$

is given by $P = U(:, 1:r)V(:, 1:r)^T$, where $[U, S, V]$ is the Singular Value Decomposition of MX^T .

3 About Exact Nonnegative PARAFAC2

Imposing nonnegativity on the B mode in the PARAFAC2 model is known to be a difficult problem and no solver actually implements it currently. Let us show rapidly why it is not straightforward, but still feasible, to impose nonnegativity within the algorithmic framework described above, that is when estimating P_k, A, B^* and C alternatively.

Clearly, imposing nonnegativity on B^* —which would be possible since nonnegativity is well understood for PARAFAC—does not guarantee that the reconstructed $B_k = P_k B^*$ are themselves nonnegative. Therefore, the following set of constraints has to be imposed on P_k and B^* in the PARAFAC2 model:

$$P_k B^* \geq 0 \quad \forall k \in [1, l], \quad (5)$$

which requires to modify the estimation procedures of both P_k and B^* .

3.1 Estimating the Orthogonal Coupling Matrices

The estimation of the orthogonal matrices P_k is a crucial step in the ALS algorithm which can be done slice by slice. The following optimization problem is solved:

$$\begin{aligned} & \underset{P_k \in \mathbb{R}^{m_k \times r}}{\operatorname{argmin}} \quad \|M_k - A D_k (P_k B^*)^T\|_F^2 \\ & \text{so that } P_k^T P_k = I, \quad P_k B^* \geq 0 \end{aligned} \quad (6)$$

Without nonnegativity constraints, P_k is computed using the Singular Value Decomposition (SVD). Sadly such a simple procedure cannot be used anymore in order to build a converging optimization algorithm because of the nonnegativity constraints. This optimization problem is reminiscent of the Orthogonal Nonnegative Matrix Factorization problem [12] which is difficult to solve.

3.2 Estimating the Latent Factor

Supposing matrices P_k have been computed in a previous step, after the data matrices M_k have been processed by multiplying them with P_k^T , the second mode variable in the PARAFAC model becomes B^* .

Within the framework of alternating optimization that we develop here¹, knowing the current estimates for A and C , the following optimization problem is to be solved:

$$\begin{aligned} & \underset{B^* \in \mathbb{R}^{r \times r}}{\operatorname{argmin}} \quad \frac{1}{2} \|T_{[2]} - B^* (A \odot C)^T\|_F^2 \\ & \text{s.t. } P_k B^* \geq 0 \quad \forall k \in [1, l] \end{aligned} \quad (7)$$

A possible approach to our problem would be to solve the exact nonnegative least squares using the Kronecker structure of the problem. This is by no means an easy task, and we could find no other work related to this issue. Another approach would be to use a projected gradient, but a projector on the constraint space would then be needed, which is not known in closed form.

As a consequence, since both the estimation of P_k and B^* are cumbersome, the algorithm implementing the methods described above proved to be quite slow and very sensitive to initialization, making it mostly useless in practice. That is the reason why the flexibly coupled PARAFAC2 is introduced in the next section.

¹ Alternating optimization may be avoided using an all-at-once method but the problem of satisfying the nonnegativity constraints still remains.

4 A Flexible PARAFAC2 Model

As described in Sect. 2, the PARAFAC2 model can be understood as a coupled matrix low rank factorization, where the coupled factors B_k are constrained to have the same inner products. The difficulty of working with constrained PARAFAC2 is that, by parameterizing each B_k as $P_k B^*$, constraints on the coupled mode are imposed on a product of two blocks of variables. In particular the P_k matrices are already constrained to be orthogonal.

Moreover, even though PARAFAC2 is less constrained than PARAFAC and has therefore been applied in context of subject variability, it makes the important underlying assumption that all the columns of B_k are transformed similarly, by opposition to component by component transformation found in other related models like Shift-PARAFAC. For instance, in the context of Gas Chromatography—Mass Spectroscopy, from one batch to another, elution profiles change in a slightly unpredictable manner, and their inner products are not exactly constant over the batches. Relaxing the hard coupling constraint in PARAFAC2 could allow for a better fitting of the PARAFAC2 in difficult cases.

For both those reasons, it makes sense to introduce a Flexible PARAFAC2 model, where the coupled factors B_k are no longer parameterized, but instead constrained to be close to $P_k B^*$. Formally, the Flexible PARAFAC2 model can be cast as follows:

$$\begin{aligned} M_k &= AD_k B_k^T + N_k, & B_k &= P_k B^* + \Gamma_k, & P_k^T P_k &= I_r, \\ \|A(:, i)\|_2 &= 1 \text{ and } \|B^*(:, i)\|_2 &= 1, & \forall i \in \{1..r\} \end{aligned}, \quad (8)$$

where Γ_k is an coupling error matrix. This kind of flexible coupling have been introduced in [3]. Under Gaussianity assumption for both model and coupling errors, a Maximum A Priori estimator of the different variables can be easily obtained by solving an optimization problem, here cast with nonnegativity constraints:

$$\begin{aligned} \operatorname{argmin}_{A, B_k, B^*, P_k, D_k} & \sum_{k=1}^K \|M_k - AD_k B_k^T\|_F^2 + \mu_k \|B_k - P_k B^*\|_F^2, \\ \text{so that } A \geq 0, B_k \geq 0, D_k \geq 0, & \|A(:, i)\|_2 = 1, \|B^*(:, i)\|_2 = 1 \quad \forall i \in \{1, \dots, r\} \end{aligned}, \quad (9)$$

where μ_k is a collection of regularization parameters controlling the distance between the factors B_k and their coupled counterparts $P_k B^*$. If noise levels on each data slice M_k are available, they can be added as a normalization constant in front of the data fitting terms. Note that the normalization of A and B^* in Eq. (8) is important, otherwise the regularization parameters μ_k and the latent factor B^* are defined up to scaling and that makes the coupling terms difficult to interpret.

The main advantage of solving (9) over (3) is that the nonnegativity constraints now apply directly on factors B_k . In an alternating optimization scheme, alternating over variables A, D_k, B_k, P_k and B^* , the coupled factors can be estimated with a simple nonnegative least squares algorithm, for instance [6]. The estimates for P_k can be obtained using SVD, and computing B^* is a least squares

problem. Therefore deriving an alternating optimization algorithm as the suggested Algorithm 1 is straightforward. Moreover, because each sub-problem in Algorithm 1 is optimally solved, given that the parameters μ_k are kept constant, the cost function is guaranteed to decrease after each iteration. Therefore, the proposed algorithm for computing Flexible PARAFAC2 is guaranteed to converge, although little can be said about whether the final estimate is a stationary point or not.

At this stage, the Flexible PARAFAC2 model can be thought of as a relaxation of the PARAFAC2 model, but it is also possible to interpret (9) as a relaxed optimization problem to solve the exact PARAFAC2 model, see for instance Chap.17 in [15]. Then by increasing the values of μ_k during the optimization algorithm, asymptotically, minimizing (9) yields an exactly coupled PARAFAC2 model. As a consequence, introducing flexibility may be understood as an optimization trick that makes constrained PARAFAC2 easier to compute. Practically, the residual relative coupling errors $\frac{\|B_k - P_k B^*\|_F^2}{\|B_k\|_F^2}$ can be monitored so that when a low value of such error is reached, the regularization parameter μ_k may stop increasing to ensure final convergence.

Algorithm 1. Alternating nonnegative least squares algorithm for solving Flexible PARAFAC2 with nonnegativity constraints.

INPUT: Data slices M_k , initial guesses for factors A, D_k, B_k, P_k, B^* .

1. Set small initial values for μ_k using (10) and normalize M_k with the total ℓ_2 norm of all slices.

while Stopping criterion is not met **do**

2. For all k , increase μ_k if necessary

3. For all k , P_k estimation: $P_k = U(:, 1:r)V(:, 1:r)^T$
where $[U, S, V^T] = \text{SVD}(B_k B^{*T})$

4. B^* estimation: $B^* = \frac{1}{\sum_{k=1}^K \mu_k} \sum_{k=1}^K \mu_k P_k^T B_k$ normalized columnwise.

5. A estimation: $A = \underset{A \geq 0}{\text{argmin}} \sum_{k=1}^K \|M_k - A D_k B_k^T\|_F^2$ solved by nonnegative least squares, then normalized columnwise.

6. For all k , B_k estimation: $B_k = \underset{B_k \geq 0}{\text{argmin}} \|M_k - A D_k B_k^T\|_F^2 + \mu_k \|B_k - P_k B^*\|_F^2$
solved by nonnegative least squares.

7. For all k , D_k estimation: $D_k = \underset{D_k \geq 0}{\text{argmin}} \|M_k - A D_k B_k^T\|_F^2$ solved by nonnegative least squares after vectorization.

8. If this is the first iteration, for all k , choose μ_k so that regularization is a certain percent of cost function using (10).

end while

OUTPUT: Estimated nonnegative factors A, D_k, B_k and coupling factors P_k, B^* .

Remark. If the parameters μ_k increase too fast at the beginning of the algorithm, then the updates of B_k are mostly driven by the regularization term. In that case, we observed that the values of B_k do not change much, and the algorithm ends up in a local minimum where only A and C are optimized. Therefore, it is important to not increase parameters μ_k over some reasonable threshold that

depends on the data fitting terms. For the initial value of μ_k and their values after the first iteration, we used respectively

$$\mu_k^0 = 10^{-1} \frac{\|M_k - A^0 D_k^0 B_k^{0T}\|_F^2}{\|B_k^0\|_F^2} \text{ and } \mu_k^1 = 10^{-SNR/10} \frac{\|M_k - A^1 D_k^1 B_k^{1T}\|_F^2}{\|B_k^1 - P_k^1 B_k^{1*}\|_F^2} \quad (10)$$

where A^0 is the initial value of A , A^1 is the estimate of A after the first iteration and SNR refers to the expected Signal to Noise ratio of the whole tensor data. These values can be tuned by the user if necessary. The increase of μ_k at each iteration is implemented as a multiplication of the current value by 1.02 if $\mu_k \leq 10$. 10 is the maximal value used in the simulation sections, and seemed to work well in our case. Larger maximal values can be used to obtain a stronger final coupling constraint depending on the problem at hand.

Initialization. In the experiments conducted in the next section, we used random factors for initialization. Another possible choice for the factors initial values is to use the output of a PARAFAC model or to compute independent nonnegative matrix factorizations for each slice. In our experiments, all these methods provided good initialization to the flexible PARAFAC2 model, yet this claim will be rigorously studied in later research. A good choice of initial P_k in any case is the zero-padded identity matrix.

5 Experiments on Synthetic Data

In this section, we provide experimental proof that the proposed Flexible PARAFAC2 model allows for imposing nonnegativity constraints on the B mode while showing performance at least similar to the state-of-the-art PARAFAC2 algorithm introduced in [10]. Also, we show that the proposed algorithm exhibits better robustness to random initialization, which in practice means a reduced number of initial trials is required.

The synthetic data are generated as follows. The dimensions are set to $[20 \times 30 \times 20]$ and the rank to $R = 3$. The entries of factors A are Gaussian with unit variance, then clipped to zero to have a sparse factor matrix. The entries of factor C are drawn from a uniform distribution on $[0, 1]$. Both A and C are then normalized column-wise using the ℓ_2 norm. In the experiments above, an i.i.d. Gaussian noise of variance σ^2 is added to each entry of the obtained tensor, where σ is a parameter of the experiments.

Generating nonnegative factors B_k that have the same Gramian matrix is not straightforward. In this manuscript, we used a particular coupling between the B_k for which the inner products are trivially kept constant over the third mode. Namely, a first factor B_1 is drawn entry-wise using a Gaussian unitary distribution, clipped to 0 and normalized column-wise, then factors B_k are obtained by circularly shifting B_1 along the grid of indexes. The obtained model is then actually a Shifted PARAFAC model, which is a particular case of PARAFAC2 that can be easily generated for simulation purpose. The general case will be studied in further extended work.

The maximum number of iterations is set to 1000, and a stopping criterion based on the relative error decrease is used.

The experiments are conducted to check the performance of the nonnegative flexible PARAFAC2 proposed algorithm with respect to the state-of-the-art algorithm [10], which does not implement nonnegativity on the coupled mode. To this end, the following relative error on factors B_k is computed for $N = 50$ simulated tensor data drawn with various noise values σ ranging from 5×10^{-2} to 10^{-4} :

$$\frac{1}{K} \sum_{k=1}^K \frac{\|B_k - [\widehat{B}_k]^+\|_F^2}{\|B_k\|_F^2}, \quad (11)$$

where all B_k and \widehat{B}_k have been normalized column-wise.

To study robustness to initialization, Fig. 1 exhibits the error on B_k with both one random initialization and the best out of five initializations.

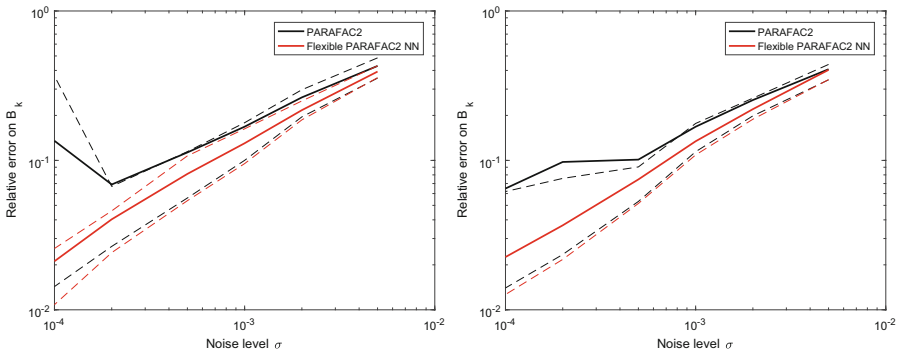


Fig. 1. Relative error on B_k for one initialization (left) and best of five initializations (right), showing in dotted lines the 20% and 80% quantiles.

From the results shown in Fig. 1, it can be concluded that, for the specific choice of shifted coupled factor B_k , the flexible PARAFAC2 best performance is similarly to the state-of-the-art PARAFAC2 algorithm best performance, with slightly lower estimation error due to the nonnegativity constraints applied on the B factor. Also, the average performances are significantly better for the flexible PARAFAC2, and the worst results are also much closer to the best ones. Therefore, it seems that the flexible algorithm is more robust to random initialization.

6 Experiments on Chromatography Data

To further assess the performance of the proposed flexible PARAFAC2 model with nonnegativity constraints, a Gas Chromatography Mass Spectroscopy (GC-MS) time interval is deconvolved, for which the usual PARAFAC2 model produces poor results. The data come from an analysis of various types of red wine

of the type Cabernet Sauvignon. The analysis was done using headspace GC-MS analysis on a Hewlett Packard 6890 GC coupled with an Agilent (Santa Clara, California, United States) 5973 Mass Selective Detector. More details can be found in the publication by Ballabio et al. [2].

The chosen time interval is difficult to decompose since there is supposedly a double peak in the time elution factors, meaning that there are two columns of factor B_k that are highly colinear. The rank is expected to be either 3 or 4, so both values were used in the comparisons below. Initial factors for the PARAFAC2 decompositions were drawn from uniform distributions on $[0,1]$. We picked the best results out of ten runs for both the unconstrained and flexible PARAFAC2 algorithms, based on the reconstruction error.

Results are presented in Fig. 2. First it can be observed that the elution profiles obtained using PARAFAC2 and flexible PARAFAC2 with nonnegativity constraints are different. Only the flexible PARAFAC2 outputs make sense in terms of elution profiles, for both three and four components models, in the sense that they are nonnegative. Having negative elution profiles is not physically meaningful.

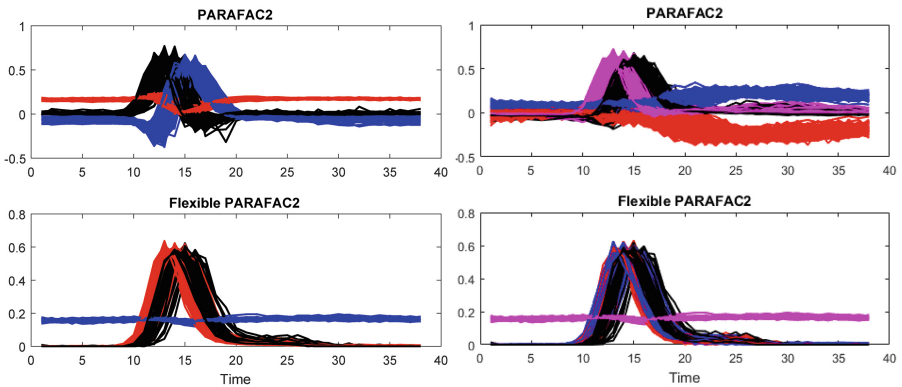


Fig. 2. Elution profiles obtained by (top) PARAFAC2 and (bottom) Flexible PARAFAC2 with nonnegativity constraints. The rank is set to (left) 3 and (right) four. The profiles represent all the B_k matrices. E.g., in the three-component model, each of the B_k matrices consist of a blue, a black and a red profile. (Color figure online)

7 Conclusion

The difficult problem of imposing nonnegative constraints on the coupled mode in the PARAFAC2 model is tackled in this manuscript. Using a flexible coupling formalism, the coupled variables and their latent representation are split, which leads to a simple constrained alternating least squares algorithm that is easily shown to converge for fixed regularization parameters. Through the decomposition of both simulated and gas chromatography mass spectroscopy data, it is

shown that the proposed flexible PARAFAC2 model behaves at worse similarly to the state-of-the-art PARAFAC2 model, but implementing nonnegativity constraints on all modes and featuring more robustness to random initialization. Further works will focus on a more precise analysis of the flexible PARAFAC2 model for solving various problems, and an extension for imposing any off-the-shelf constraints on the coupled mode.

Acknowledgements. The authors wish to thank Nicolas Gillis for helpful discussions on alternatives to the flexible coupling approach for computing nonnegative PARAFAC2.

References

1. Amigo, J.M., Skov, T., Bro, R., Coello, J., Maspoch, S.: Solving GC-MS problems with PARAFAC2. *TrAC Trends Anal. Chem.* **27**(8), 714–725 (2008)
2. Ballabio, D., Skov, T., Leardi, R., Bro, R.: Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques. *J. Chemometr.* **22**(8), 457–463 (2008)
3. Farias, R.C., Cohen, J.E., Comon, P.: Exploring multimodal data fusion through joint decompositions with flexible couplings. *IEEE Trans. Sign. Process.* **64**(18), 4830–4844 (2016)
4. Comon, P., Luciani, X., de Almeida, A.L.F.: Tensor decompositions, alternating least squares and other tales. *J. Chemometr.* **23**(7–8), 393–405 (2009)
5. García, I., Sarabia, L., Ortiz, M.C., Aldama, J.M.: Building robust calibration models for the analysis of estrogens by gas chromatography with mass spectrometry detection. *Anal. Chim. Acta* **526**(2), 139–146 (2004)
6. Gillis, N., Glineur, F.: Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Comput.* **24**(4), 1085–1105 (2012)
7. Harshman, R.A.: PARAFAC2: Mathematical and technical notes. *UCLA Work. Pap. phonetics*, **22**, pp. 30–44 (1972). 122215
8. Harshman, R.A., Hong, S., Lundy, M.E.: Shifted factor analysis—part I: models and properties. *J. Chemometr.* **17**(7), 363–378 (2003)
9. Johnsen, L.G., Skou, P.B., Khakimov, B., Bro, R.: Gas chromatography-mass spectrometry data processing made easy. *J. Chromatogr. A* **1503**, 57–64 (2017)
10. Kiers, H.A.L., Ten Berge, J.M.F., Bro, R.: PARAFAC2-part I. a direct fitting algorithm for the PARAFAC2 model. *J. Chemometr.* **13**(3–4), 275–294 (1999)
11. Mørup, M., Hansen, L.K., Arnfred, S.M., Lim, L.-H., Madsen, K.H.: Shift-invariant multilinear decomposition of neuroimaging data. *NeuroImage* **42**(4), 1439–1450 (2008)
12. Pompili, F., Gillis, N., Absil, P.-A., Glineur, F.: Two algorithms for orthogonal nonnegative matrix factorization with application to clustering. *Neurocomputing* **141**, 15–25 (2014)
13. Skov, T., Bro, R.: A new approach for modelling sensor based data. *Sens. Actuators B: Chem.* **106**(2), 719–729 (2005)
14. Wise, B.M., Gallagher, N.B., Martin, E.B.: Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch. *J. Chemometr.* **15**(4), 285–298 (2001)
15. Nocedal, J., Wright, S.J.: *Sequential quadratic programming*. Springer, New York (2006)



Applications of Polynomial Common Factor Computation in Signal Processing

Ivan Markovsky¹(✉), Antonio Fazzi², and Nicola Guglielmi²

¹ Department ELEC, Vrije Universiteit Brussel (VUB), Pleinlaan 2, Building K, 1050 Brussels, Belgium

imarkovs@vub.ac.be

² Gran Sasso Science Institute (GSSI), 67010 L' Aquila, Italy
antonio.fazzi@gssi.it, guglielm@univaq.it

Abstract. We consider the problem of computing the greatest common divisor of a set of univariate polynomials and present applications of this problem in system theory and signal processing. One application is blind system identification: given the responses of a system to unknown inputs, find the system. Assuming that the unknown system is finite impulse response and at least two experiments are done with inputs that have finite support and their Z-transforms have no common factors, the impulse response of the system can be computed up to a scaling factor as the greatest common divisor of the Z-transforms of the outputs. Other applications of greatest common divisor problem in system theory and signal processing are finding the distance of a system to the set of uncontrollable systems and common dynamics estimation in a multi-channel sum-of-exponentials model.

Keywords: Blind system identification
Sum-of-exponentials modeling · Distance to uncontrollability
Approximate common factor · Low-rank approximation

1 Introduction

Finding the *greatest common divisor* of a set of univariate polynomials is a classic problem in algebra, which is still an active research topic. Numerically it is an ill-conditioned problem: small perturbations in the input data (the polynomials' coefficients) may result in large changes in the solution (the greatest common divisor coefficients). This requires computing an *approximate* common divisor.

There are two different formulations of the approximate common divisor problem. In the first one, the degree of the common divisor is a priori specified and the smallest perturbation on the polynomial coefficients that leads to polynomials with common divisor of such a degree is sought [6, 9, 13, 15, 18]. In the second formulation, referred to as *ϵ -common divisor*, the size of the maximum perturbation is given and perturbed polynomials with maximal degree common divisor is sought [1, 14, 17]. The two problems are dual to each other. In fact,

they are two different scalarizations of the biobjective problem where the size of the perturbation is minimized while the degree of the perturbed polynomials common divisor is maximized. The two formulations trace the same Pareto optimal trade-off curve.

The approximate greatest common divisor problem is a non-convex optimization problem, for which there are no efficient global solution methods. The existing methods can be classified as local optimization methods and convex relaxations. Local optimization methods require an initial approximation and are in general computationally more expensive than the relaxation methods, however, the local optimization methods explicitly optimize the desired criterion (size of the coefficient perturbation), which ensures that they produce at least as good result as a relaxation method, provided the solution of the relaxation method is used as an initial approximation for the local optimization method. For a recent overview of computational approaches, we refer the reader to [15].

Applications of the greatest common divisor in systems, control, and signal processing, however, are surprisingly missing from the broad literature on the theoretical and computational aspects of the problem. We present here applications that are directly solvable by a greatest common divisor computation. Subsequently existing greatest common divisor methods, algorithms and software can be used in the applications. Vice versa, methods, algorithms and software developed for the applications can be viewed as and used for greatest common divisor computation.

In this paper, we consider the following approximate common factor computation problem: given polynomials p^1, \dots, p^N and a natural number \mathbf{d} ,

$$\begin{aligned} & \text{minimize} && \text{over } \hat{p}^1, \dots, \hat{p}^N && \sum_{i=1}^N \|p^i - \hat{p}^i\|_2^2 && (1) \\ & \text{subject to} && \deg(\gcd(\hat{p}^1, \dots, \hat{p}^N)) \geq \mathbf{d}. \end{aligned}$$

($\gcd(p^1, \dots, p^N)$ is the greatest common divisor of the polynomials $p^1(z), \dots, p^N(z)$.) Sect. 2 shows application of (1) for blind finite impulse response system identification. Section 3 shows application of (1) for computing the distance of a given linear time-invariant system to the set of uncontrollable systems. Section 4 shows application of (1) for estimation of common dynamics across multiple channels of an autonomous linear time-invariant system.

2 Blind Finite Impulse Response System Identification

The identification problem considered in this section is defined as follows.

Problem 1 (Blind finite impulse response system identification). Given output observations y^1, \dots, y^N of a finite impulse response system, generated by unknown signals u^1, \dots, u^N , find the impulse response h of the system.

The Z-transform of a finite duration time-domain signal y^i is a polynomial $y^i(z)$. (We use the argument z , as in $y^i(z)$, to indicate that the signal is in the Z-domain).

Theorem 1. *Assuming that at least $N = 2$ responses y^1, \dots, y^N of a finite impulse response system are given,*

1. *the inputs u^1, \dots, u^N have finite support, and*
2. *$\gcd(u^1(z), \dots, u^N(z)) = 1$,*

the impulse response h of the system is up to a scaling factor $\alpha \in \mathbb{R}$ the greatest common factor of $y^1(z), \dots, y^N(z)$,

$$h(z) = \alpha \gcd(y^1(z), \dots, y^N(z)).$$

Proof. Let \star be the convolution operator. We have,

$$y^i = h \star u^i, \quad \text{for } i = 1, \dots, N.$$

Since the system is finite impulse response $h(z) := Z(h)$ is a polynomial. Under assumption 1, $u^i(z) := Z(u^i)$ are also polynomials. Therefore, $y^i(z) := Z(y^i)$

$$y^i(z) = h(z)u^i(z), \quad \text{for } i = 1, \dots, N \quad (2)$$

are polynomials. It follows from (2) that $h(z)$ is a common factor of y^1, \dots, y^N . By assumption 2, $h(z)$ is the greatest common factor of y^1, \dots, y^N . \square

With noisy data

$$y_{\text{d}}^i = \bar{y}^i + \tilde{y}^i, \quad \text{for } i = 1, \dots, N$$

(the subscript index “d” stands for “data”), where \bar{y}^i is the “true value” and \tilde{y}^i is the measurement noise, $y_{\text{d}}^1, \dots, y_{\text{d}}^N$ are generically co-prime, *i.e.*, they have no nontrivial common factor. Assuming that the noise \tilde{y} is zero mean, white, Gaussian, the maximum-likelihood estimator of the “true impulse response” \bar{h} is given by the following problem

$$\begin{aligned} &\text{minimize over } \hat{y}^1, \dots, \hat{y}^N, \hat{u}^1, \dots, \hat{u}^N, \text{ and } \hat{h} \quad \sum_{i=1}^N \|y_{\text{d}}^i - \hat{y}^i\|_2^2 \\ &\text{subject to } \hat{y}^i = \hat{h} \star \hat{u}^i, \quad \text{for } i = 1, \dots, N. \end{aligned} \quad (3)$$

Note that since we include \hat{h} as an optimization variable, we assume that its length (or equivalently the order $\bar{n} = \dim(\bar{h}) - 1$ of the true system) is a priori known.

Problem (3) is an approximate common factor computation problem.

Theorem 2. *Problems (3) and (1) are equivalent.*

3 Distance to Uncontrollability

Verifying whether a given linear time-invariant system is controllable involves rank computation. Arbitrary small perturbations of the system's parameters can switch the property. This issue is addressed by the notion of distance to uncontrollability, which is quantitative rather than qualitative measure of controllability. The definition of distance to uncontrollability, considered in the literature [10], is a property of the parameters A and B in a state space representation of the system. Using the notion of controllability in the behavioral setting [12] we define a representation invariant measure of distance to uncontrollability and propose an algorithm for computing it.

Consider a linear time-invariant system \mathcal{B} with a state space representation

$$\mathcal{B} = \mathcal{B}_{i/s/o}(A, B, C, D) := \{ w = (u, y) \mid \sigma x = Ax + Bu, y = Cx + Du \}, \quad (4)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{p \times m}$ are parameters of \mathcal{B} ; and σ is the shift operator $(\sigma x)(t) = x(t+1)$ (in discrete-time) or the derivative operator $\sigma x = dx/dt$ (in continuous-time).

We adopt the behavioral setting [12], *i.e.*, a system is viewed as a set of trajectories. For a given system \mathcal{B} , the parameters A , B , and C of the state space representation (4) of \mathcal{B} are not unique due to the fact that for any change of basis $x' = Vx$ of the state space, $\mathcal{B}(VAV^{-1}, VB, CV^{-1}, D)$ is the same model as $\mathcal{B}(A, B, C, D)$, *i.e.*,

$$\mathcal{B}_{i/s/o}(A, B, C, D) = \mathcal{B}_{i/s/o}(VAV^{-1}, VB, CV^{-1}, D).$$

In addition, the parameters A , B , and C are not unique due to nonminimality of the state dimension; for example

$$\mathcal{B}_{i/s/o}(A, B, C, D) = \mathcal{B} \left(\begin{bmatrix} A & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B \\ 0 \end{bmatrix}, [C \ 0], D \right),$$

for any $A_{12} \in \mathbb{R}^{n \times \Delta n}$, $A_{21} \in \mathbb{R}^{\Delta n \times n}$, and $A_{22} \in \mathbb{R}^{\Delta n \times \Delta n}$.

A state space representation with parameters A and B is *state controllable* if and only if the matrix

$$\mathcal{C}(A, B) := [A \ AB \ \dots \ A^{n-1}B]$$

is full rank. Note that this classical notion of controllability is a property of the pair of matrices (A, B) and is not a property of a system \mathcal{B} due to the nonuniqueness of a state space representation. The question of whether a given state space representation is state controllable is a rank test problem for the structured matrix $\mathcal{C}(A, B)$. A corresponding quantitative measure is the distance of $\mathcal{C}(A, B)$ to rank deficiency, *i.e.*, the smallest $(\Delta A, \Delta B)$, such that

$$\mathcal{C}(\widehat{A}, \widehat{B}) := \mathcal{C}(A, B) + \mathcal{C}(\Delta A, \Delta B)$$

is rank deficient.

Consider the set of $m \times n$ structured matrices \mathcal{S} and define the distance measure

$$d_r(A) := \min_{\Delta A \in \mathcal{S}} \|\Delta A\| \quad \text{subject to} \quad A + \Delta A \text{ has rank } r,$$

where $\|\cdot\|$ is a matrix norm. With $\mathcal{S} = \mathbb{R}^{m \times n}$, $d_r(A)$ is a distance to *unstructured* rank- r matrices. In the special cases of spectral and Frobenius norms, the unstructured distance $d_r(A)$ can be computed using the singular value decomposition of A .

Motivated by the issues of computing the numerical rank of a matrix, C. Paige defined in [10] the distance to uncontrollability

$$d_{\text{unctr}}(A, B) := \text{minimize} \quad \text{over } \widehat{A}, \widehat{B} \quad \left\| \begin{bmatrix} A & B \end{bmatrix} - \begin{bmatrix} \widehat{A} & \widehat{B} \end{bmatrix} \right\|_{\text{F}}$$

subject to $(\widehat{A}, \widehat{B})$ is uncontrollable.

This problem falls into a broader category of *distance problems* [4], such as distance to instability, distance to positive definiteness, etc. There is a big volume of literature devoted on the problem of computing $d_{\text{unctr}}(A, B)$, see, e.g., [2, 3, 5, 7, 8]. The measure $d_{\text{unctr}}(A, B)$, however, is not invariant of the state space representation because it depends on the choice of basis. This issue is resolved in the behavioral setting, where controllability is defined as a property of the system rather than a property of a particular representation.

Definition 1 ([16]). *A time-invariant system \mathcal{B} is controllable if for any two trajectories $w_p, w_f \in \mathcal{B}$, there is $\Delta t > 0$ and a trajectory $w_c \in \mathcal{B}$, such that $w_p(t) = w_c(t)$, for all $t < 0$, and $w_f(t) = w_c(t)$, for all $t \geq \Delta t$.*

Checking the controllability property in practice is done by performing a numerical test on the parameters of a specific representation of the system. For a single-input single-output linear time-invariant system with an input/output representation

$$\mathcal{B}_{i/o}(p, q) := \left\{ \begin{bmatrix} u \\ y \end{bmatrix} \mid p(\sigma)y = q(\sigma)u \right\} \tag{5}$$

is controllable if and only if the polynomials are co-prime.

Theorem 3 ([12]). *Consider the polynomials $p(z)$ and $q(z)$ and let the degree of p be higher than or equal to the degree of q . The single-input single-output system $\mathcal{B}_{i/o}(p, q)$ is controllable if and only if p and q are co-prime.*

By Theorem 3, the system $\mathcal{B}_{i/o}(p, q)$ is controllable if and only if p and q have no common factors of degree one or more.

Let $\overline{\mathcal{L}_{\text{ctrb}}}$ be the set of uncontrollable linear time-invariant systems:

$$\overline{\mathcal{L}_{\text{ctrb}}} = \{ \mathcal{B} : \mathcal{B} \text{ is linear time-invariant and uncontrollable} \}$$

and consider the distance measure

$$\text{dist}(\mathcal{B}_{i/o}(p, q), \mathcal{B}_{i/o}(\widehat{p}, \widehat{q})) := \left\| \begin{bmatrix} q \\ p \end{bmatrix} - \begin{bmatrix} \widehat{q} \\ \widehat{p} \end{bmatrix} \right\|_2. \tag{6}$$

The representation invariant notion of distance to uncontrollability proposed is: Given a controllable system $\mathcal{B}_{i/o}(p, q)$, find:

$$d_{\text{unctr}}(\mathcal{B}) := \min_{\widehat{\mathcal{B}} \in \mathcal{L}_{\text{ctrb}}} \text{dist}(\mathcal{B}, \widehat{\mathcal{B}}). \quad (7)$$

We refer to $d_{\text{unctr}}(\mathcal{B})$ as the *uncontrollability radius*.

Theorem 4. *Problems (7) and (1) with $\mathbf{d} = 1$ are equivalent.*

Proof. Follows directly from Theorem 3. □

4 Common Dynamics Estimation

The problem considered in this section is defined as follows.

Problem 2. Given a set of N scalar autonomous linear time-invariant systems $\mathcal{B}_1, \dots, \mathcal{B}_N$, find their “common dynamics”, defined as $\mathcal{B} := \mathcal{B}_1 \cap \dots \cap \mathcal{B}_N$.

Let the systems be represented by their kernel representations $\mathcal{B}_i = \ker(p^i(\sigma))$, where

$$\ker(p(\sigma)) := \{y \mid p_0 y + p_1 \sigma y + \dots + p_n \sigma^n y = 0\}. \quad (8)$$

Then, the kernel representation $\ker(p(\sigma))$ of the common dynamics \mathcal{B} is given by the greatest common divisor $p = \text{gcd}(p^1, \dots, p^N)$. In the case when $\mathcal{B}_1, \dots, \mathcal{B}_N$ have no common dynamics ($\mathcal{B} = \{0\}$), a problem of finding approximate common dynamics of a specified dimension is considered. The approximate common dynamics problem is equivalent to the approximate common divisor Problem (1).

A variation of the common dynamic’s estimation problem is considered in [11]. In this case, which we call “data-driven” in order to distinguish it from the “model-based” Problem 2, the aiming is to model a set of scalar time series y^1, \dots, y^N by sums of, respectively, $\mathbf{n}_1, \dots, \mathbf{n}_N$ damped exponentials, which have $\mathbf{n}_c \leq \min(\mathbf{n}_1, \dots, \mathbf{n}_N)$ common exponents. The given time series

$$y^i = (y^i(1), \dots, y^i(T_i))$$

are approximated by time series \widehat{y} satisfying the model equation

$$\widehat{y}^i(t) = \sum_{j=1}^{\mathbf{n}_i - \mathbf{n}_c} \alpha_{ij} \lambda_{ij}^t + \sum_{j=1}^{\mathbf{n}_c} \beta_{ij} \mu_j^t, \quad t = 1, \dots, T_i. \quad (9)$$

Here, $\mu_1, \dots, \mu_{\mathbf{n}_c}$ are the exponents common to all signals and $\lambda_{i1}, \dots, \lambda_{i\mathbf{n}_i}$ are the remaining exponents of the i th signal.

In [11], a subspace-type method for common dynamics estimation is proposed. Assuming that the data is generated in the output error setup, *i.e.*, $y^i = \bar{y}^i + \tilde{y}^i$, where the true values \bar{y}^i satisfy the model (9) and \tilde{y}^i is the measurement noise that is zero mean, white, Gaussian, the maximum-likelihood estimator is

$$\begin{aligned} &\text{minimize} \quad \text{over } \hat{y}^i \in \mathbb{R}^{T_i}, \lambda_{ij} \in \mathbb{C}, \mu_j \in \mathbb{C}, \alpha_{ij} \in \mathbb{C}, \text{ and } \beta_{ij} \in \mathbb{C} \quad \sqrt{\sum_{i=1}^N \|y^i - \hat{y}^i\|_2^2} \quad (10) \\ &\text{subject to} \quad (9). \end{aligned}$$

The following result shows equivalent optimization problems to (10) based on the kernel and state space representations of the model.

Theorem 5. *Problem (10) is equivalent to the following problems:*

– *kernel representation*

$$\begin{aligned} &\text{minimize} \quad \text{over } \hat{y}^i \in \mathbb{R}^{T_i}, R_{s,i}, R_c \quad \|y - \hat{y}\|_2 \quad (11) \\ &\text{subject to} \quad (R_{s,i} \star R_c) \mathcal{H}_{n_i+1}(\hat{y}^i) = 0, \quad \text{for } i = 1, \dots, N. \end{aligned}$$

– *state-space representation*

$$\begin{aligned} &\text{minimize} \quad \text{over } \hat{y}^i \in \mathbb{R}^{T_i}, \lambda_i, \mu, c_i, c' \quad \|y - \hat{y}\|_2 \\ &\text{subject to} \quad \hat{y} \in \mathcal{B} \left(\text{diag}(\lambda_1, \dots, \lambda_N, \mu), \begin{bmatrix} c_1 & & c'_1 \\ & \ddots & \vdots \\ & & c_N \ c'_N \end{bmatrix} \right), \quad (12) \end{aligned}$$

where $\lambda_i \in \mathbb{C}^{1 \times (\ell_i - \ell_c)}$, $c_i \in \mathbb{C}^{1 \times (\ell_i - \ell_c)}$, and $c' \in \mathbb{C}^{1 \times \ell_c}$.

Although these problems are not equivalent to the approximate common divisor problem (1), the solution methods are closely related. Indeed (11) is a Hankel structured low-rank approximation problem. As shown in [15], Problem (1) is a Sylvester structured low-rank approximation problem.

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC Grant 258581 “Structured low-rank approximation: Theory, algorithms, and applications”; Fund for Scientific Research (FWO-Vlaanderen), FWO projects G028015N “Decoupling multivariate polynomials in nonlinear system identification”; G090117N “Block-oriented nonlinear identification using Volterra series”; and FWO/FNRS Excellence of Science project 30468160 “Structured low-rank matrix/tensor approximation: numerical optimization-based algorithms and applications”.

References

1. Bini, D.A., Boito, P.: A fast algorithm for approximate polynomial GCD based on structured matrix computations. In: Bini, D.A., Mehrmann, V., Olshevsky, V., Tyrtyshnikov, E.E., van Barel, M. (eds.) *Numerical Methods for Structured Matrices and Applications. Operator Theory: Advances and Applications*, vol. 199, pp. 155–173. Birkhäuser, Basel (2010). https://doi.org/10.1007/978-3-7643-8996-3_6
2. Eising, R.: Distance between controllable and uncontrollable. *Control Lett.* **4**, 263–264 (1984)
3. Gu, M., Mengi, E., Overton, M., Xia, J., Zhu, J.: Fast methods for estimating the distance to uncontrollability. *SIAM J. Matrix Anal. Appl.* **28**, 477–502 (2006)
4. Higham, N.: Matrix nearness problems and applications. In: Gover, M., Barnett, S. (eds.) *Applications of Matrix Theory*, pp. 1–27. Oxford University Press, Oxford (1989)
5. Hu, G., Davison, E.: Real controllability/stabilizability radius of LTI systems. *IEEE Trans. Automat. Control* **49**, 254–258 (2004)
6. Kaltofen, E., Corless, R.M., Jeffrey, D.J.: Challenges of symbolic computation: my favorite open problems. *J. Symbolic Comput.* **29**(6), 891–919 (2000)
7. Karow, M., Kressner, D.: On the structured distance to uncontrollability. *Control Lett.* **58**, 128–132 (2009)
8. Khare, S., Pillai, H., Belur, M.: Computing the radius of controllability for state space systems. *Control Lett.* **61**, 327–333 (2012)
9. Markovsky, I., Van Huffel, S.: An algorithm for approximate common divisor computation. In: *Proceedings of the 17th Symposium on Mathematical Theory of Networks and Systems*, Kyoto, Japan, pp. 274–279 (2006)
10. Paige, C.C.: Properties of numerical algorithms related to computing controllability. *IEEE Trans. Automat. Control* **26**, 130–138 (1981)
11. Papy, J.M., Lathauwer, L.D., Huffel, S.V.: Common pole estimation in multi-channel exponential data modeling. *Signal Process.* **86**(4), 846–858 (2006)
12. Polderman, J., Willems, J.C.: *Introduction to Mathematical Systems Theory*. Springer, New York (1998). <https://doi.org/10.1007/978-1-4757-2953-5>
13. Qiu, W., Hua, Y., Abed-Meraim, K.: A subspace method for the computation of the GCD of polynomials. *Automatica* **33**(4), 741–743 (1997)
14. Rupprecht, D.: An algorithm for computing certified approximate GCD of n univariate polynomials. *J. Pure Appl. Algebra* **139**(1–3), 255–284 (1999)
15. Usevich, K., Markovsky, I.: Variable projection methods for approximate (greatest) common divisor computations. *Theor. Comput. Sci.* **681**, 176–198 (2017)
16. Willems, J.C.: Paradigms and puzzles in the theory of dynamical systems. *IEEE Trans. Automat. Control* **36**(3), 259–294 (1991)
17. Zeng, Z., Dayton, B.: The approximate GCD of inexact polynomials. Part I: a univariate algorithm. In: *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, pp. 320–327 (2004)
18. Zhi, L., Yang, Z.: Computing approximate GCD of univariate polynomials by structure total least norm. In: Wang, D., Zhi, L. (eds.) *International Workshop on Symbolic-Numeric*, Xian, China, pp. 188–201 (2005)



Joint Nonnegative Matrix Factorization for Underdetermined Blind Source Separation in Nonlinear Mixtures

Ivica Kopriva^(✉)

Division of Electronics, Ruder Bošković Institute, Bijenička cesta 54,
10000 Zagreb, Croatia
Ivica.Kopriva@irb.hr

Abstract. An approach is proposed for underdetermined blind separation of nonnegative dependent (overlapped) sources from their nonlinear mixtures. The method performs empirical kernel maps based mappings of original data matrix onto reproducible kernel Hilbert spaces (RKHSs). Provided that sources comply with probabilistic model that is sparse in support and amplitude nonlinear underdetermined mixture model in the input space becomes overdetermined linear mixture model in RKHS comprised of original sources and their mostly second-order monomials. It is assumed that linear mixture models in different RKHSs share the same representation, i.e. the matrix of sources. Thus, we propose novel sparseness regularized joint nonnegative matrix factorization method to separate sources shared across different RKHSs. The method is validated comparatively on numerical problem related to extraction of eight overlapped sources from three nonlinear mixtures.

Keywords: Underdetermined blind source separation · Nonlinear mixtures
Empirical kernel map · Joint nonnegative matrix factorization · Sparseness

1 Introduction

Blind source separation (BSS) refers to extraction of source signals from observed mixture signals only [1]. When the sources and mixing matrix are nonnegative algorithms of nonnegative matrix factorization (NMF) are shown to be effective solving the BSS problem [2–4]. In particular, when nonnegativity is combined with sparseness underdetermined BSS problems, characterized with more sources than mixtures available, can be solved [5, 6]. That, as an example, is relevant to mass spectrometry [7] or nuclear magnetic resonance (NMR) spectroscopy [8] based metabolic profiling where large number of sources (a.k.a. pure components or analytes) needs to be separated from the small number of available mixture spectra [7]. However, in a large number of cases algorithms for BSS address separation of sources from their linear mixtures. As opposed to them the number of methods that address the nonlinear BSS problem is rather small, see chapter 14 in [1]. Thus, we propose a method for separation of nonnegative mutually dependent (overlapped) but individually independent and identically distributed (i.i.d.) sources from smaller number of their nonlinear mixtures. Compared with proposed

method existing methods either: (i) address determined case, where the number of sources equals the number of mixtures [9–18]; (ii) do not take into account nonnegativity constraint [9–21]; (iii) assume that sources [10–12, 15–17, 19–22] or their derivatives [18] are statistically independent or that sources are individually correlated [16, 19–21]. In particular, we map data matrix from the input space onto reproducible kernel Hilbert spaces (RKHSs) by means of empirical kernel maps (EKM) [23]. We treat mapped data as they are coming from different views and propose a linear mixture model (LMM)-based representation such that all models have different mixing matrices but share the same source (representation) matrix. Thus, we propose an algorithm for joint NMF such that LMMs of mixture data mapped in multiple RKHSs share the same source matrix. That is different from joint NMF approach to multi-view clustering [24], where LMMs comprised of view dependent mixing and source matrices are assumed such that source matrices are forced to converge towards common consensus. We introduce nonlinear BSS problem in Sect. 2. Section 3 presents new joint NMF-based approach to nonlinear underdetermined BSS problem. Results of comparative performance analysis on numerical problem are presented in Sect. 4. Section 5 concludes the paper.

2 Problem Formulation

Nonlinear BSS problem with nonnegative dependent sources is formulated as:

$$\mathbf{X} = \mathbf{f}(\mathbf{S}) + \mathbf{E} \quad (1)$$

where $\mathbf{X} \in \mathbb{R}_{0+}^{N \times T}$ stands for nonnegative matrix of N nonlinear mixtures at T observations, $\mathbf{S} \in \mathbb{R}_{0+}^{M \times T}$ stands for matrix of M unknown nonnegative sources, $\mathbf{f} : \mathbb{R}_{0+}^M \rightarrow \mathbb{R}_{0+}^N$ stands for unknown nonlinear mapping $\mathbf{f} := [f_1(\mathbf{S}) \dots f_N(\mathbf{S})]^T$ acting observation-wise, $\mathbf{E} \in \mathbb{R}_{0+}^{N \times T}$ stands for an error term and \mathbb{R}_{0+} stands for the set of real nonnegative numbers.

The symbol “:=” means “by definition”. We also assume that $\left\{ \|\mathbf{s}_t\|_0 \leq K \right\}_{t=1}^T$, where $\|\mathbf{s}_t\|_0$ is indicator function that counts number of non-zero entries of \mathbf{s}_t , and K denotes maximal number of sources that can be present (active) at any observation coordinate t . The nonlinear BSS problem implies that sources $\{\mathbf{s}_m \in \mathbb{R}_{0+}^{1 \times T}\}_{m=1}^M$ have to be inferred from the mixture data matrix \mathbf{X} only. Herein, we impose assumptions on the sources and nonlinear mixture model (1):

- (A1) $0 \leq s_{mt} \leq 1 \forall m = 1, \dots, M$ and $\forall t = 1, \dots, T$,
- (A2) Amplitude s_{mt} is i.i.d. random variable that obeys exponential distribution on $(0, 1]$ interval and discrete distribution at zero, see Eq. (2),
- (A3) Components of the vector-valued function $\mathbf{f}(\mathbf{S}) : \{\mathbf{f}_n(\mathbf{S}) : \mathbb{R}_{0+}^{M \times T} \rightarrow \mathbb{R}_{0+}^{1 \times T}\}_{n=1}^N$ are differentiable up to second-order,
- (A4) $M > N$.

Assumptions A1 to A4 are shown in [7] to be relevant for separation of pure components from nonlinear mixtures of mass spectra. They are expected to hold for

separation of pure components from amplitude NMR spectra as well [8]. To be useful solution of any BSS problem is expected to be essentially unique [1]. However, even for linear underdetermined BSS problem hard (sparseness) constraints ought to be imposed on sources [7, 25] to obtain essentially unique solution. The quality of separation heavily depends on degree of sparseness, i.e. the value of K . To make nonlinear underdetermined BSS problem tractable we assume, as in [27], that amplitudes of the source signals comply with sparse probabilistic model [25, 26]:

$$p(s_{mt}) = \rho_m \delta(s_{mt}) + (1 - \rho_m) \delta^*(s_{mt}) g(s_{mt}) \quad \forall m = 1, \dots, M \text{ and } \forall t = 1, \dots, T \quad (2)$$

where $\delta(s_{mt})$ is an indicator function and $\delta^*(s_{mt}) = 1 - \delta(s_{mt})$ is its complementary function, $\rho_m = \{P(s_{mt} = 0)\}_{t=1}^T$. Thus, $\{P(s_{mt} > 0)\} = 1 - \rho_m\}_{t=1}^T$. The nonzero state of s_{mt} is distributed according to probability density function $g(s_{mt})$. Exponential distribution $g(s_{mt}) = (1/\mu_m) \exp(-s_{mt}/\mu_m)$ is selected in which case the most probable outcome is equal to μ_m . To emphasize practical relevance of probabilistic model (2) we point out [7]. Another modality to which model (2) can be relevant is NMR spectroscopy [8]. It has been verified in [7] that mass spectra of pure components obey (2) with exponential distribution selected for $g(s_{mt})$. Thereby $\hat{\rho}_m \in [0.27, 0.74]$ and $\hat{\mu}_m \in [0.0012, 0.0014]$. Under such priors the nonlinear mixture model (1) simplifies to [27]:

$$\mathbf{X} = \mathbf{J}\mathbf{S} + \frac{1}{2}\mathbf{H}_{(1)} \begin{bmatrix} \mathbf{s}_1^2 \\ \vdots \\ \mathbf{s}_M^2 \\ \vdots \\ \{\mathbf{s}_i \mathbf{s}_j\}_{i,j=1}^M \end{bmatrix} + HOT = \mathbf{B} \begin{bmatrix} \mathbf{S} \\ \mathbf{s}_1^2 \\ \vdots \\ \mathbf{s}_M^2 \\ \vdots \\ \{\mathbf{s}_i \mathbf{s}_j\}_{i,j=1}^M \end{bmatrix} + HOT \quad (3)$$

where \mathbf{J} stands for Jacobian matrix, $\mathbf{H}_{(1)}$ stands for mode-1 unfolded third-order Hessian tensor, $\mathbf{B} = [\mathbf{J} \frac{1}{2} \mathbf{H}_{(1)}]$ stands for the overall mixing matrix and HOT stands for higher order terms. Since original nonlinear problem (1) is underdetermined the equivalent linear problem (3) is even more underdetermined because it is comprised of the same number of mixtures, N , but of the $P = 2M + M(M - 1)/2$ dependent sources. When degree of the overlap of the sources in (1) is K degree of the overlap of new sources in (3) is $Q \approx 2K + K(K - 1)/2$. Uniqueness of the solution of (3) depends on the triplet (N, P, Q) . For deterministic mixing matrix \mathbf{B} the necessary condition for uniqueness is $N = O(Q^2)$ [28]. Thus, it becomes virtually impossible to obtain an essentially unique solution of the underdetermined nonlinear BSS problem (1) with overlapped sources. Separation quality can, however, be increased through nonlinear mapping of mixture data $\{\mathbf{x}_t \in \mathbb{R}_{0+}^{N \times 1} \rightarrow \phi(\mathbf{x}_t) \in \mathbb{R}_{0+}^{\tilde{N} \times 1}\}_{t=1}^T$ where explicit feature map (EFM) $\phi(\mathbf{x}_t)$ maps data into, in principle, infinite dimensional feature space. To make calculations in mapped space computationally tractable, $\phi(\mathbf{X}) := \{\phi(\mathbf{x}_t)\}_{t=1}^T$ needs to be projected to a low-dimensional subspace of induced space spanned by $\phi(\mathbf{V}) := \{\phi(\mathbf{v}_d)\}_{d=1}^D$. Thereby, the basis $\mathbf{V} := \{\mathbf{v}_d \in \mathbb{R}_{0+}^{N \times 1}\}_{d=1}^D$ spans the input space: $span\{\mathbf{v}_d\}_{d=1}^D \approx span\{\mathbf{x}_t\}_{t=1}^T$ and it is estimated from \mathbf{X} by k -means clustering

algorithm. Projection known as EKM, see Definition 2.15 in [23], maps data from input space onto RKHS:

$$\Psi(\mathbf{V}, \mathbf{X}) = \phi(\mathbf{V})^T \phi(\mathbf{X}) = \mathbf{K}(\mathbf{V}, \mathbf{X}) \quad (4)$$

where $\mathbf{K}(\mathbf{V}, \mathbf{X}) \in \mathbb{R}_{0+}^{D \times T}$ denotes Gram or kernel matrix with the elements $\{\kappa(\mathbf{v}_d, \mathbf{x}_t) = \phi(\mathbf{v}_d)^T \phi(\mathbf{x}_t)\}_{d,t=1}^{D,T}$. It is shown in [7] that under sparse probabilistic prior (2) Eq. (4) becomes:

$$\Psi(\mathbf{V}, \mathbf{X}) \approx \mathbf{A} \begin{bmatrix} \mathbf{0}_{1 \times T} \\ \mathbf{S} \\ \{\mathbf{s}_i \mathbf{s}_j\}_{i,j=1}^M \end{bmatrix} + \bar{\mathbf{E}} \quad (5)$$

where \mathbf{A} denotes a nonnegative mixing matrix of appropriate dimensions, $\mathbf{0}_{1 \times T}$ stands for row vector of zeros and $\bar{\mathbf{E}}$ stands for approximation error. The uniqueness condition for system (5) becomes: $D = O(Q^2)$, [28]. When $D \gg N$ that can be fulfilled with greater probability than uniqueness condition for system (3): $N = O(Q^2)$, [7, 27]. Thus, the role of nonlinear EKM-based mapping is to “increase number of mixtures”.

3 Joint Nonnegative Matrix Factorization in Reproducible Kernel Hilbert Spaces

It has been demonstrated that sparseness constrained NMF in an EKM-induced RKHS enables separation of nonnegative dependent sources from smaller number of their nonlinear mixtures [27]. However, the fundamental issue is how to select the kernel function, i.e. its parameters in (4)/(5). The common choice is a Gaussian kernel $\kappa(\mathbf{v}_d, \mathbf{x}_t) = \exp(-\|\mathbf{v}_d - \mathbf{x}_t\|^2 / \sigma^2)$. That is justified by its universal approximation property [29]. However, proper selection of the kernel variance σ^2 requires a priori knowledge of the signal-to-noise (SNR) ratio. When dealing with experimental data that is often hard to know in practice. Herein, we propose to map data \mathbf{X} into multiple RKHSs using EKMs with Gaussian kernel with the values for variance that cover wide enough range: $\sigma^2 \in \{\sigma_1^2, \dots, \sigma_{n_v}^2\}$. Hence, we obtain n_v data matrices in induced RKHSs with representations as follow:

$$\Psi_i(\mathbf{V}, \mathbf{X}) = \mathbf{A}_i \bar{\mathbf{S}} + \bar{\mathbf{E}}_i \quad i = 1, \dots, n_v \quad (6)$$

where meaning of $\bar{\mathbf{S}}$ is clear from direct comparison between (6) and (5). To establish weak analogy with the multi-view clustering, [24], we denoted mixture matrices in RKHSs as data arising from multiple views. Also, without loss of generality, to enable fair comparison with multi-view NMF algorithm [24] we assume that mixture matrices in each “view i ” satisfy $\{\|\Psi_i(\mathbf{V}, \mathbf{X})\|_1 = 1\}_{i=1}^{n_v}$. The difference between our model (6) and multi-view NMF model [24] is that our model (6) assumes that all the views share the same source matrix $\bar{\mathbf{S}}$, while in [24] source matrices are different for each view and

are enforced to converge towards a common consensus. To derive the NMF update rule on the level of “view i ” we assume Gaussian distribution for the error term in (6) and minimize the loss function under constrains $\mathbf{A}_i \geq \mathbf{0}$, $\bar{\mathbf{S}} \geq \mathbf{0}$:

$$L(\mathbf{A}_i, \bar{\mathbf{S}}) = \frac{1}{2} \|\Psi_i(\mathbf{V}, \mathbf{X}) - \mathbf{A}_i \bar{\mathbf{S}}\|_2^2 + \alpha \|\bar{\mathbf{S}}\|_1 \quad (7)$$

where α stands for sparseness regularization constant. Minimization yields the following update rules for \mathbf{A}_i and $\bar{\mathbf{S}}$, see also Table 1 in [3]:

$$\begin{aligned} \mathbf{A}_i &= \mathbf{A}_i \otimes \frac{\Psi_i(\mathbf{V}, \mathbf{X}) \bar{\mathbf{S}}^T}{\mathbf{A}_i \bar{\mathbf{S}} \bar{\mathbf{S}}^T + \varepsilon \mathbf{1}_{DP}} \\ \bar{\mathbf{S}} &= \bar{\mathbf{S}} \otimes \frac{[\mathbf{A}_i^T \Psi_i(\mathbf{V}, \mathbf{X}) - \alpha \mathbf{1}_{PT}]_+}{\mathbf{A}_i^T \mathbf{A}_i \bar{\mathbf{S}} + \varepsilon \mathbf{1}_{PT}} \end{aligned} \quad (8)$$

In (8) \otimes denotes entry-wise multiplication, $\mathbf{1}_{DP}$ and $\mathbf{1}_{PT}$ stand for matrices of all ones, ε is a small constant and $[x]_+$ stands for $\max(0, x)$ operator. At each iteration the algorithm cycles through all the views $1, \dots, n_v$. It is clear that representation (6) automatically resolves the permutation indeterminacy issue that is problematic for joint NMF across multiple views [24]. We coin our method multi-view NMF (mvNMF). The joint NMF method [24] is coined multi-view consensus NMF (mvCNMF). Even though our method is developed for separation of sources from nonlinear underdetermined mixtures it can be applied directly to multi-view clustering in the same spirit as joint NMF method in [24]. In that case $\Psi_i(\mathbf{V}, \mathbf{X})$ ought to be replaced with the data matrix at view i : \mathbf{X}_i . Furthermore, when BSS problem is linear, $\mathbf{X} = \mathbf{A}\mathbf{S}$, with one view only, i.e. $n_v = 1$, Eq. (8) with the appropriate substitutions represents standard sparseness constrained NMF [3].

4 Numerical Evaluation

To validate proposed mvNMF method we generate three nonlinear mixtures of eight overlapped sources according to:

$$\begin{aligned} f_1(\mathbf{s}) &= s_1^3 + s_2^2 + \tan^{-1}(s_3) + s_4^2 + s_5^3 + s_6^3 + \tanh(s_7) + \sin(s_8) + \mathbf{e}_1 \\ f_2(\mathbf{s}) &= \tanh(s_1) + s_2^3 + s_3^3 + \tan^{-1}(s_4) + \tanh(s_5) + \sin(s_6) + s_7^2 + s_8^2 + \mathbf{e}_2 \\ f_3(\mathbf{s}) &= \sin(s_1) + \tan^{-1}(s_2) + s_3^2 + s_4^3 + \tanh(s_5) + \sin(s_6) + s_7^3 + \tan^{-1}(s_8) + \mathbf{e}_3 \end{aligned}$$

We generated eight source signals in $T = 1000$ observations with degrees of overlap equal to $K \in \{1, 3, 5\}$. According to probabilistic prior (2) we set $\{\rho_m = 0.6\}_{m=1}^8$ and $\{\mu_m = 0.15\}_{m=1}^8$. Thus, generated sources correspond with the real world signals such as mass spectra. Furthermore, noise was added to the mixtures with SNR = 0 dB. We use the Gaussian kernel with the variance $\sigma^2 \in \{1000, 100, 10, 1, 0.1, 0.01, 0.001\}$. That covers wide range of possible SNRs. The basis matrix \mathbf{V} for

EKM-based mappings was estimated from \mathbf{X} by *k-means* algorithm with $D = 100$ cluster centers. We compare the proposed mvNMF algorithm with the mvCNMF algorithm [24], with ordinary NMF algorithm [3] applied directly to mixture data matrix \mathbf{X} and with algorithm (8) applied to each “view”, mapped data matrix $\{\Psi_i(\mathbf{V}, \mathbf{X})\}_{i=1}^{n_v}$, separately. We coined the last algorithm as single view NMF (svNMF) and point out that it coincides with the algorithm [27] applied in each RKHS separately. We set sparseness related regularization constant in (8) to $\alpha = 0.2$. In case of mvCNMF algorithm we use the result for consensus matrix to be compared with the result of mvNMF. For each value of K we repeated the comparison 100 times. In each experiment we separated eight sources from the mixtures and annotated them with the true sources using mean normalized correlation as criterion:

$$\text{mean correlation} = \left(\sum_{i \in I_c} c_i(\hat{s}_i, s_i) \right) / M \quad (9)$$

where I_c denotes index set of correctly assigned sources, \hat{s}_i denotes the separated and s_i the true source and $0 \leq c_i(\hat{s}_i, s_i) \leq 1$ stands for the normalized correlation coefficient. Thus, if more than one separated source was assigned to the same true source that was counted as assignment error and reduced value of the mean correlation.

Figure 1a shows mean values of assignment errors (with the variance as error bar) for NMF, mvNMF and mvCNMF algorithms. Figures 1b shows assignment errors for the svNMF algorithm. Corresponding correlation coefficients (9) are shown in Fig. 2a and b. The largest mean value of assignment error is 36% for mvNMF, 34.63% for NMF, 45.37% for mvCNMF and 35.37% for svNMF. The largest values of mean correlation coefficient for the algorithms in respective order are 8.12%, 4.23%, 7.44% and 4.36%. Thus, proposed mvNMF method increased correlation coefficient in comparison with NMF and svNMF methods having similar assignment error. In comparison with mvCNMF the mvNMF method has similar correlation coefficient but smaller assignment error. The mvCNMF method extracted typically two or three unique sources with the “highest” value of correlation coefficient. That explains its

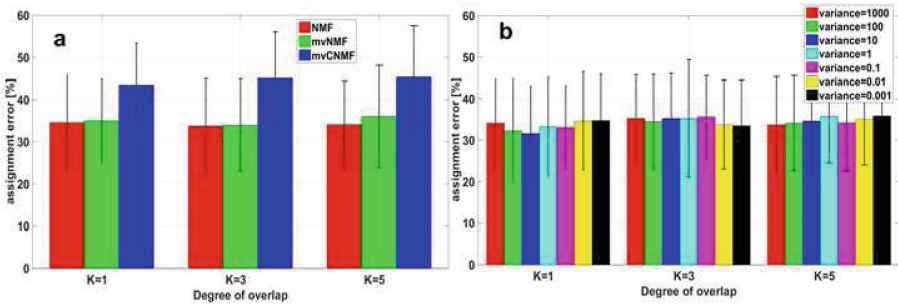


Fig. 1. Assignment error. (a) NMF, mvNMF and mvCNMF algorithms. (b) svNMF algorithm, i.e. NMF algorithm applied to each “view” separately. (Color figure online)

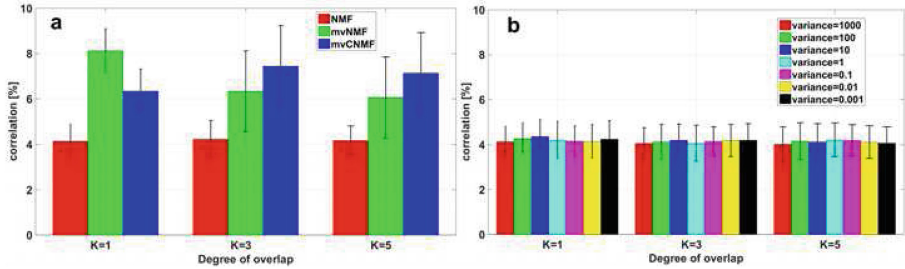


Fig. 2. Mean correlation coefficients (9) of correctly assigned sources separated with: (a) NMF, mvNMF and mvCNMF algorithms. (b) svNMF algorithm, i.e. NMF algorithm applied to each “view” separately. (Color figure online)

“good” performance in terms of correlation and poor in terms of assignment error. Figures 1b and 2b show that performing source separation in each induced RKHS separately yields results worse than when all the RKHSs are used. Although the separation quality of proposed mvNMF method could be considered low we comment that described nonlinear BSS problem is hard and for it, to the best of our knowledge, no method is developed yet.

5 Conclusion

Blind separation of nonnegative dependent (overlapped) sources from smaller number of nonlinear mixtures represents a hard problem with, arguably, no algorithm proposed to solve it. Herein, we propose method for separation of sparse dependent sources by joint NMF on mixture matrices mapped in multiple RKHSs. RKHSs were induced by mappings based on Gaussian kernel with variances that cover a wide range of possible SNR values. Mixtures in induced RKHSs were represented with the linear mixture models comprised of different mixing matrices and common matrix of sources. That is justified by the fact that mixtures in mapped data space are obtained from the same mixture matrix in input data space. Thus, a novel joint NMF method is proposed to separate common source matrix from multiple mixtures. On numerical experiment the proposed method achieved competitive performance. In addition for nonlinear BSS proposed joint NMF method could be also used for clustering data from multiple views in the spirit of [24].

Acknowledgments. This work has been supported in part by the Grant IP-2016-06-5253 funded by Croatian Science Foundation and in part by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

1. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, New York, NY, USA (2010)
2. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. Wiley, Chichester, UK (2009)
3. Cichocki, A., Zdunek, R., Amari, S.: Csiszár's divergences for non-negative matrix factorization: family of new algorithms. In: Rosca, J., Erdogmus, D., Principe, J.C., Haykin, S. (eds.) ICA 2006. LNCS, vol. 3889, pp. 32–39. Springer, Heidelberg (2006). https://doi.org/10.1007/11679363_5
4. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. *Nature* **401**, 788–791 (1999)
5. Peharz, R., Pernkopf, F.: Sparse nonnegative matrix factorization with ℓ^0 -constraints. *Neurocomputing*, **80**, 38–46 (2012)
6. Gillis, N., Glineur, F.: Using underapproximations for sparse nonnegative matrix factorization. *Pattern Rec.* **43**, 1676–1687 (2010)
7. Kopriva, I., Jerić, I., Brkljačić, L.: Nonlinear mixture-wise expansion approach to underdetermined blind separation of nonnegative dependent sources. *J. of Chemometr.* **27**, 189–197 (2013)
8. Kopriva, I., Jerić, I.: Blind separation of analytes in nuclear magnetic resonance spectroscopy: improved model for nonnegative matrix factorization. *Chemometr. Int. Lab. Syst.* **137**, 47–56 (2014)
9. Zhang, K., Chan, L.: Minimal nonlinear distortion principle for nonlinear independent component analysis. *J. Mach. Learn. Res.* **9**, 2455–2487 (2008)
10. Levin, D.N.: Using state space differential geometry for nonlinear blind source separation. *J. Appl. Phys.* **103**(044906), 1–12 (2008)
11. Levin, D.N.: Performing nonlinear blind source separation with signal invariants. *IEEE Trans. Sig. Proc.* **58**, 2132–2140 (2010)
12. Taleb, A., Jutten, C.: Source separation in post-nonlinear mixtures. *IEEE Trans. Sig. Proc.* **47**, 2807–2820 (1999)
13. Duarte, L.T., Suyama, R., Rivet, B., Attux, R., Romano, J.M.T., Jutten, C.: Blind compensation of nonlinear distortions: applications to source separation of post-nonlinear mixtures. *IEEE Trans. Sig. Proc.* **60**, 5832–5844 (2012)
14. Filho, E.F.S., de Seixas, J.M., Calôba, L.P.: Modified post-nonlinear ICA model for online neural discrimination. *Neurocomputing* **73**, 2820–2828 (2010)
15. Nguyen, V.T., Patra, J.C., Das, A.: A post nonlinear geometric algorithm for independent component analysis. *Digit. Sig. Proc.* **15**, 276–294 (2005)
16. Ziehe, A., Kawanabe, M., Harmeling, S., Müller, K.R.: Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation. *J. Mach. Learn. Res.* **4**, 1319–1338 (2003)
17. Zhang, K., Chan, L.W.: Extended gaussianization method for blind separation of post-nonlinear mixtures. *Neural Comput.* **17**, 425–452 (2005)
18. Ehsandoust, B., Babaie-Zadeh, M., Rivet, B., Jutten, C.: Blind source separation in nonlinear mixtures: separability and a basic algorithm. *IEEE Trans. Sig. Proc.* **65**, 4352–4399 (2017)
19. Harmeling, S., Ziehe, A., Kawanabe, M.: Kernel-based nonlinear blind source separation. *Neural Comput.* **15**, 1089–1124 (2003)
20. Martinez, D., Bray, A.: Nonlinear blind source separation using kernels. *IEEE Trans. Neural Net.* **14**, 228–235 (2003)

21. Yu, H.-G., Huang, G.-M., Gao, J.: Nonlinear blind source separation using kernel multi-set canonical correlation analysis. *Int. J. Comput. Netw. Inf. Secur.* **1**, 1–8 (2010)
22. Almeida, L.: MISEP-linear and nonlinear ICA based on mutual information. *J. Mach. Learn. Res.* **4**, 1297–1318 (2003)
23. Schölkopf, B., Smola, A.: *Learning With Kernels*. The MIT Press, Cambridge, MA, USA (2002)
24. Liu, L., Wang, C., Gao, J., Han, J.: Multi-view clustering via joint nonnegative matrix factorization. In: 2013 Proceedings of SIAM International Conference on Data Mining (SDM 2013), pp. 252–260 (2013). <https://doi.org/10.1137/1.9781611972832.28>
25. Caifa, C., Cichocki, A.: Estimation of sparse nonnegative sources from noisy overcomplete mixtures using MAP. *Neural Comput.* **21**, 3487–3518 (2009)
26. Bouthemy, P., Piriou, C.H.G., Yao, J.: Mixed-state auto-models and motion texture modeling. *J. Math Imag. Vis.* **25**, 387–402 (2006)
27. Kopriva, I., Jerić, I., Filipović, M., Brkljačić, L.: Empirical kernel map approach to nonlinear underdetermined blind separation of sparse nonnegative dependent sources: pure components extraction from nonlinear mixtures mass spectra. *J. of Chemometr.* **28**, 704–715 (2014)
28. DeVore, R.A.: Deterministic constructions of compressed sensing matrices. *J. Complex.* **27**, 918–925 (2007)
29. Micchelli, C.A., Xu, Y., Zhang, H.: Universal kernels. *J. Mach. Learn. Res.* **7**, 2651–2667 (2006)



Image Completion with Nonnegative Matrix Factorization Under Separability Assumption

Tomasz Sadowski^(✉)  and Rafał Zdunek^(✉) 

Faculty of Electronics, Wrocław University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{tomasz.sadowski, rafal.zdunek}@pwr.edu.pl

Abstract. Nonnegative matrix factorization is a well-known unsupervised learning method for part-based feature extraction and dimensionality reduction of a nonnegative matrix with a variety of applications. One of them is a matrix completion problem in which missing entries in an observed matrix is recovered on the basis of partially known entries. In this study, we present a geometric approach to the low-rank image completion problem with separable nonnegative matrix factorization of an incomplete data. The proposed method recursively selects extreme rays of a simplicial cone spanned by an observed image and updates the latent factors with the hierarchical alternating least-squares algorithm. The numerical experiments performed on several images with missing entries demonstrate that the proposed method outperforms other algorithms in terms of computational time and accuracy.

Keywords: Nonnegative Matrix Factorization · Image completion
Geometric NMF · Separable matrix factorization

1 Introduction

Nonnegative Matrix Factorization (NMF) [1] is an unsupervised method for extracting a latent structure from an input matrix which contains only nonnegative entries. The basic model of NMF assumes an approximate decomposition of an input nonnegative matrix into lower-rank nonnegative factors. Due to the huge flexibility of NMF and easy interpretation of its factors, this model has already found many applications in various areas of research and engineering [1, 2].

NMF was popularized by Lee and Seung [3, 4] who proposed simple multiplicative algorithms for updating the factors. Nowadays, there are plenty of computational strategies for updating the factors in various NMF models [1, 2], and thousands of publications about NMF. Among them, an emerging group consists of geometry-based NMF, such as the XRAY [5], Hottopix [6], Successive Nonnegative Projection Algorithm (SNPA) [7], Hierarchical Convex-Hull NMF (HCH-NMF) [8], SimplexMax [9]. These algorithms find the feature vectors by searching the extreme rays of a simplicial cone spanned by observations.

If the extreme rays are represented by a subset of columns (or rows) of an input matrix, then such an NMF model satisfies the separability assumption [5, 9]. If so, any other data point can be presented by the conic combination of the extreme rays. The geometry-based NMF algorithms work very efficiently in many applications, especially for a blind unmixing hyperspectral problem [7, 10, 11] or textual document representation [5, 12, 13].

NMF has been also applied to a matrix completion problem [14], where a subset of its known entries is used to restore missing entries in a given incomplete matrix. The problem can be expressed by the following model:

$$\min_{\mathbf{Y}} \text{rank}(\mathbf{Y}), \quad \text{st.} \quad y_{it} = m_{it}, \forall (i, t) \in \Omega \quad (1)$$

where $\mathbf{M} = [m_{it}] \in \mathbb{R}_+^{I \times T}$ is the original incomplete matrix, $\mathbf{Y} = [y_{it}] \in \mathbb{R}_+^{I \times T}$ is the recovered matrix, and Ω is the set of indexes of the known elements in \mathbf{M} . As expressed by (1), the matrix completion problem boils down to the problem of finding such a minimum-rank matrix \mathbf{Y} that has the same set of elements as the matrix \mathbf{M} among the items indicated by the set Ω .

One of the recently studied models for the above-mentioned problem is Nonnegative Matrix Completion model under Separability Assumption (NMCSA) which was proposed by Yu *et al.* in [15]. The NMCSA combines a geometric approach to NMF with coordinate-descent gradient optimization for updating the latent factors. The columns of \mathbf{M} generate a convex cone $\mathcal{C}(\mathbf{M}) \subset \mathbb{R}_+^I$, and under the separability condition, the selected columns from \mathbf{M} are regarded as its extreme rays. The NMCSA selects the extreme rays with random projections. Motivated by the efficiency of the SimplexMax algorithm [9] for initialization of latent factors in the standard NMF model, we propose to use the concept of the SimplexMax to select the extreme rays of $\mathcal{C}(\mathbf{M})$ in the NMCSA. Furthermore, the missing entries in \mathbf{M} may occur in every column, including the columns selected for representing the extreme rays. The missing entries in these columns cannot be recovered in one run of the NMCSA. To tackle this problem, we propose to run the NMCSA recursively, and in further recursive steps to randomly select the candidates for the extreme rays. We applied the proposed method to the image completion problem with a few incomplete images.

The paper is organized in following way: in the next section, we briefly discuss the NMCSA [15]. The proposed methods are described in Sect. 3. The numerical experiments performed for various image completion problems are presented in Sect. 4. The last section contains the summary and conclusions.

Notations: boldface uppercase letters (e.g. \mathbf{X}) denotes for matrices; lowercase boldface ones stand for vectors (e.g. \mathbf{x}); not bold letters are scalars. For a matrix \mathbf{X} , $x_{i,j}$ denotes the (i, j) -th element, \mathbf{x}_j or $\underline{\mathbf{x}}_j$ stand for the j -th column or row, respectively. The symbol $\|\cdot\|_F$ denotes the Frobenius norm of a matrix; $\|\cdot\|_2$ denotes the 2-nd norm. The set of nonnegative real numbers is denoted by \mathbb{R}_+ . For a matrix, $\mathbf{X} \in \mathbb{R}_+^{I \times J}$ or $\mathbf{X} \geq 0$ means that all elements in \mathbf{X} are nonnegative. Let Γ be a subset of an arbitrary set, then $\bar{\Gamma}$ is its complement, and $|\Gamma|$ is the cardinality of Γ . The submatrix created from the columns of \mathbf{X} indexed by Γ is

denoted by \mathbf{X}_Γ . The symbol $\mathcal{C}(\mathbf{X})$ stands for the convex set generated by the columns of \mathbf{X} .

2 NMCSA Algorithm

In the NMCSA, the matrix \mathbf{Y} in (1) is assumed to satisfy the separability assumption. Hence, the NMCSA model takes the form:

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{P}, \mathbf{I}, \mathbf{X}, \Gamma} \quad & \frac{1}{2} \|\mathbf{Y}\mathbf{P}\mathbf{I} - \mathbf{Y}_\Gamma [\mathbf{I} \ \mathbf{X}]\|_F^2, \quad \text{s.t. } \mathbf{Y} \in \mathbb{R}_+^{I \times T}, y_{it} = m_{it}, \forall (i, t) \in \Omega, \\ & \mathbf{X}^T \mathbf{1}_J = \mathbf{1}_{T-J}, \mathbf{X} \in \mathbb{R}_+^{J \times (T-J)}, |\Gamma| = J, \end{aligned} \quad (2)$$

where J is a given rank of factorization, $\mathbf{P}\mathbf{I}$ is a permutation matrix, $\mathbf{Y}_\Gamma \in \mathbb{R}_+^{I \times J}$ contains the features represented by the anchors selected from \mathbf{Y} , and the columns of \mathbf{X} contain coefficients of a conic combination of the features. The set Γ contains the indices of the anchors. From the separability condition, the columns of \mathbf{Y}_Γ represent the extreme rays of the simplicial cone $\mathcal{C}(\mathbf{Y}) \subset \mathbb{R}_+^I$, $\mathcal{C}(\mathbf{Y}) \subseteq \mathcal{C}(\mathbf{Y}_\Gamma)$, and $\forall t: \mathbf{y}_t \in \mathcal{C}(\mathbf{Y}_\Gamma)$.

If the set Γ is known, the problem (2) can be reformulated as:

$$\begin{aligned} \min_{\mathbf{Q}, \mathbf{A}, \mathbf{X}} \quad & \frac{1}{2} \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2, \quad \text{s.t. } \mathbf{Y} \in \mathbb{R}_+^{I \times T}, y_{it} = m_{it}, \forall (i, t) \in \Omega, \\ & \mathbf{X}^T \mathbf{1}_J = \mathbf{1}_{T-J}, \mathbf{X} \in \mathbb{R}_+^{J \times (T-J)}, |\Gamma| = J, \end{aligned} \quad (3)$$

where $\mathbf{Q} = \mathbf{Y}_{\bar{\Gamma}} \in \mathbb{R}_+^{I \times (T-J)}$, $\mathbf{A} = \mathbf{Y}_\Gamma \in \mathbb{R}_+^{I \times J}$, and $\mathbf{Y} = [\mathbf{Q} \ \mathbf{A}]\mathbf{P}\mathbf{I} \in \mathbb{R}_+^{I \times T}$.

To update \mathbf{A} and \mathbf{X} , the following subproblems created from (3) are solved:

$$\min_{\mathbf{A}} \frac{1}{2} \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2, \quad \text{s.t. } \mathbf{Y} \in \mathbb{R}_+^{I \times T}, y_{it} = m_{it}, \forall (i, t) \in \Omega, \quad (4)$$

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{Q} - \mathbf{A}\mathbf{X}\|_F^2, \quad \text{s.t. } \mathbf{X}^T \mathbf{1}_J = \mathbf{1}_{T-J}, \mathbf{X} \in \mathbb{R}_+^{J \times (T-J)}. \quad (5)$$

The matrix \mathbf{Q} in (3) is updated according to the rule: $\mathbf{Q} = \mathbf{Y}_{\bar{\Gamma}} = (\mathbf{A}\mathbf{X})_{\bar{\Gamma}}$.

In the NMCSA, the problems (4) and (5) are solved with nonnegatively constrained coordinate-descent gradient algorithms. The algorithm for updating \mathbf{A} is closely related to the Hierarchical Alternating Least-Squares (HALS) algorithm which was proposed by Cichocki *et al.* [16] for updating latent factors in NMF/NTF models. The problem (5) is solved with a similar algorithm to the HALS, however, instead of rank-one update, it performs two-rank updates, i.e. two rows of \mathbf{X} are updated simultaneously in each inner iteration.

To solve the problems (4) and (5), the set Γ must be known. In the NMCSA, the set Γ is estimated with the random projections expressed by Algorithm 1. According to the considerations in [15], Algorithm 1 selects the anchors with a high probability according to the statistics defined via T_p times projections on randomly selected standard basis vector \mathbf{e} of \mathbb{R}^I .

Algorithm 1. Random Projections

Input : $M \in \mathbb{R}^{I \times T}$ – incomplete matrix, J – rank of factorization, T_p – number of random projections

Output: Γ – indexes of anchors

1 Initialization: $\mathcal{I} = \emptyset$;

2 **for** $t = 1, \dots, T_p$ **do**

3 randomly select standard basis vector \mathbf{e} ;

4 $i_* = \arg \max_{t \in \{1, \dots, T\}} \mathbf{m}_t^T \mathbf{e}$;

5 $\mathcal{I} = \mathcal{I} \cup i_*$;

6 $\Gamma = J$ unique elements of \mathcal{I} with largest occurrences;

3 Improved NMCSA for Image Completion Problem

The random projections in Algorithm 1 select possible dense columns of \mathbf{Y} , which is justified by the fact that dense vectors are not incomplete. However, such vectors may be far away from the true extreme rays of $\mathcal{C}(\mathbf{Y})$. There are many methods for estimating the extreme rays of a convex cone, which are mentioned in Sect. 1. Many experiments confirmed that the SimplexMax [9] is one of the most efficient algorithms for pursuing this task. Hence, we also use the concept of the SimplexMax to estimate the set Γ in (3).

To apply the SimplexMax, the columns of \mathbf{Y} are scaled to unit l_1 -norm, i.e. $\bar{\mathbf{Y}} = \mathbf{Y}\mathbf{D}$, where $\mathbf{D} \in \mathbb{R}_+^T$ is a diagonal scaling matrix. The scaling means that $\mathcal{C}(\mathbf{Y})$ is cut with the hyperplane determined by the unit vectors in \mathbb{R}_+^I . The intersection forms the convex hull $\mathcal{H}(\mathbf{Y})$. If the factorization model $\mathbf{Y} = \mathbf{A}\mathbf{X}$ satisfies the separability condition, the vertices of the convex hull $\mathcal{H}(\mathbf{Y})$ are the anchors of $\bar{\mathbf{Y}}$ which span a polytope of the maximal volume [17]. The set Γ in (3) can be obtained by solving the problem:

$$\Gamma = \arg \max_{\mathcal{T} \subset \{1, \dots, T\}} \text{vol } \mathcal{H}(\bar{\mathbf{Y}}_{\mathcal{T}}) = \arg \max_{\mathcal{T} \subset \{1, \dots, T\}} \det \{(\bar{\mathbf{Y}}_{\mathcal{T}})^T (\bar{\mathbf{Y}}_{\mathcal{T}})\}, \quad (6)$$

where $\text{vol } \mathcal{H}(\bar{\mathbf{Y}}_{\mathcal{T}})$ is the volume of the polytope $\mathcal{H}(\bar{\mathbf{Y}}_{\mathcal{T}})$ generated by the vectors in $\bar{\mathbf{Y}}_{\mathcal{T}}$ with $|\mathcal{T}| = |\Gamma| = J$.

The problem (6) belongs to a class of combinatorial problems, however, under the separability condition of NMF, it can be well approximated by the following recursive algorithm. In the first step, we attempt to find such the vector $\bar{\mathbf{y}}_t$ from $\bar{\mathbf{Y}}$ that is located in the furthest distance from the central point $\bar{\mathbf{y}}_m = \frac{1}{T} \sum_{t=1}^T \bar{\mathbf{y}}_t \in \mathbb{R}_+^I$. Such a vector determines one of the vertices of $\mathcal{H}(\bar{\mathbf{Y}})$, i.e. the vector \mathbf{a}_1 of \mathbf{A} . The index t is the first entry of Γ . In the next step, another vector $\bar{\mathbf{y}}_s$ from $\bar{\mathbf{Y}}$ is searched that maximizes the area of the parallelogram formed by the vectors $\bar{\mathbf{y}}_t$ and $\bar{\mathbf{y}}_s$. As a result, $\Gamma = \{t, s\}$. In each recursive step, the new vector from $\bar{\mathbf{Y}}$ is added to the basis of the previously found vertex vectors, and the corresponding index is added to the set Γ . After performing J recursive steps, the matrix \mathbf{A} and the corresponding set Γ are found. In real applications, p vectors from $\bar{\mathbf{Y}}$ with the highest impact on the solution to (6) is found in each

Algorithm 2. ICSA Algorithm

Input : $M \in \mathbb{R}^{I \times T}$ - incomplete matrix, J - rank of factorization, S - number of recursive steps

Output: $Y \in \mathbb{R}^{I \times T}$ - completed matrix

- 1 **Initialization**: $\mathcal{S}^{(0)} = \mathcal{U} = \{1, \dots, T\}$, $s = 1$;
 - 2 **while** $s \leq S$, and $|\mathcal{S}^{(s-1)}| > J$ **do**
 - 3 Select Γ from $\mathcal{S}^{(s-1)}$ using the SimplexMax or random projections;
 - 4 Solve the problems (4) and (5) using the NMCSA updating rules, and update Q , where $\bar{\Gamma} = \mathcal{U} \setminus \Gamma$;
 - 5 Update: $\mathcal{S}^{(s)} = \mathcal{S}^{(s-1)} \setminus \Gamma$ and $s \leftarrow s + 1$;
 - 6 **Return** Y ;
-



Fig. 1. Original images: (first) *Lena* (256×256 pixels); (second) *boats* (512×512 pixels); (third) *mountain* (384×254 pixels); (fourth) *ship* (316×466 pixels)

recursive step. Then the vectors are averaged to form a rough estimator of the desired vertex.

Regardless of the algorithm used for finding the extreme rays, it is highly probable that each column vector of Y is incomplete, and such anchors remain incomplete if the problems (4) and (5) are solved. To tackle this deficiency, we propose to re-run the NMCSA with different extreme rays. In the first step, Γ is determined from the set $\mathcal{U} = \{1, \dots, T\}$. In the next step, the columns from Y_Γ go to $Y_{\bar{\Gamma}^{(s)}}$, and $\Gamma^{(s)}$ is selected from the difference $\mathcal{U} \setminus \Gamma$. The procedure can be repeated S times, where $S \leq \lfloor \frac{T}{J} \rfloor$, or until the set $|\mathcal{S}^{(s)}| < J$. The concept of re-selection of the extreme rays is illustrated by Algorithm 2.

4 Experiments

The numerical experiments are performed on image completion problems with four original images that are illustrated in Fig. 1. The incomplete data are obtained by removing a certain number of pixels from original images.

The discussed algorithms are run in PLGRID¹ queues on the distributed cluster server in Wroclaw Center for Networking and Supercomputing (WCSS)²

¹ <http://www.plgrid.pl/en>

² <https://www.wcss.pl/en/>

using Matlab2016 parallel workers. Our resources are limited to 8 cores (ncpus) and 8 GB RAM (mem). Due to the non-convex nature of NMF algorithms, the tests were repeated 100 times.

The results are evaluated quantitatively with the Signal-to-Interference Ratio (SIR) measure, defined as $\text{SIR} = 20 \log_{10} \frac{\|\mathbf{M}_0\|_F}{\|\mathbf{M}_0 - \mathbf{Y}\|_F}$, where \mathbf{M}_0 is an original grey-scale image or a colormap. For RGB images, the SIRs are averaged over colormaps. The selected images are shown in the form of 2D plots. The tested algorithms are also validated in terms of the averaged runtime per iteration.

A couple of numerical experiments are conducted. In all the tests, we set $S = 2$, and $J = 50$, which is motivated by the upper bound for the rank in [18]. In the first experiment, we compare the performance of the proposed algorithms, i.e. ICSA with random projections (ICSA-rand) and ICSA with the SimplexMax (ICSA-SI), with the baseline NMCSA (with random projections), by applying them to the image completion problem with well-known images. The aim of this test is to select the most efficient algorithm for further, more extensive tests. Two well-known images from Fig. 1 were selected: the **Lena** in a gray-scale and the **boats**, and then degraded to an incomplete version with about 50 % of missing pixels. The choice of images is also motivated by their different resolution and details of the scenery. The averaged SIR results are presented in Table 1.

Table 1. SIR-values [dB] for various algorithms

Image	NMCSA	ICSA-rand	ICSA-SI
Lena	10.60 ± 0.09	17.36 ± 0.10	17.63 ± 0.10
Boats	14.12 ± 0.07	18.23 ± 0.10	18.96 ± 0.09

As we can see from Fig. 2 and Table 1, the NMCSA does not provide a satisfactory solution; the basis lines (extreme rays) are not updated properly. The ICSA-SI provides much better results. Since the ICSA-SI outperforms the ICSA-rand, the further experiments show only a comparison of the ICSA-SI with the NMCSA and other algorithms.

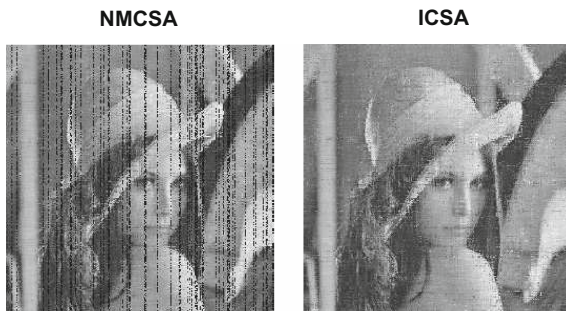


Fig. 2. NMCSA vs ICSA-SI.

In the second experiment, we compare the algorithms on following images: **boats** and **mountain**. Our choice was motivated by content-dependent image completion. The **boats** image presents the detailed scenery with a non-smooth horizontal structure. The other image contains such a structure which might be easier represented by a conic combination of its few columns in each color map. For such images, the NMCSA-based algorithms should work better. The incomplete images were generated by removing from the original image: (a) randomly selected 50%, 70% and 90% of pixels, (b) a single line of pixels forming a regular grid of 10 pixels wide. For RGB images, the completion process is performed separately for each color map. The following algorithms are tested and compared: NMCSA [15], SVT [19], LMaFit [20], and ICSA-SI. The boxplots of SIR values are demonstrated in Fig. 3, together with the selected completed images. The averaged runtime/iteration is listed in Table 2.

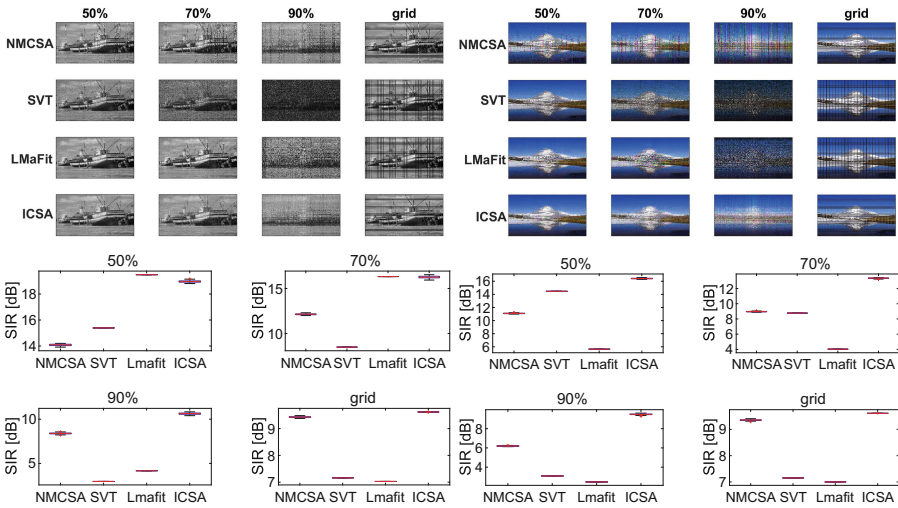
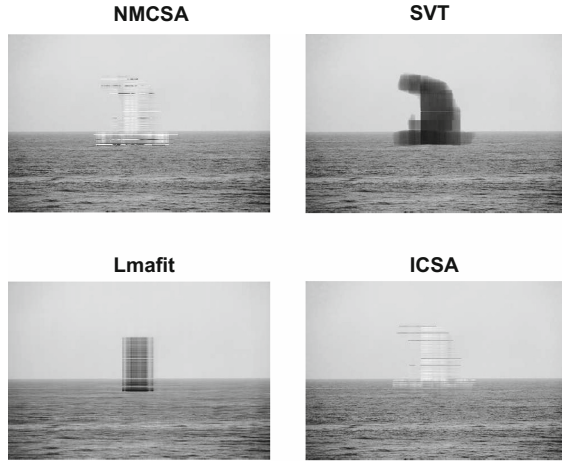


Fig. 3. Results of images completion: ● *upper left*: recovered **boats** image, ● *upper right*: recovered **mountain** image, ● *bottom left*: boxplots of SIR-values for **boats** image, ● *bottom right*: boxplots of SIR-values for **mountain** image.

The third experiment, performed on the *ship* image [21] from Fig. 1, is aimed to show how the discussed algorithms deal with a large block of missing pixels. In this case, the block was a ship on the sea. The best method should replace the ship with the background. The recovered images are shown in Fig. 4, and the corresponding SIR values and the runtime/iteration are listed in Table 3.

Table 2. Runtime/iteration ratio [in milliseconds] for the algorithms used in the second experiment.

Boats				
	NMCSA	SVT	LMaFit	ICSA-SI
50%	91.38 \pm 31.0	184.21 \pm 9.05	6.73 \pm 0.91	94.45 \pm 16.85
70%	110.49 \pm 42.45	178.86 \pm 16.14	5.13 \pm 1.27	119.75 \pm 48.47
90%	118.83 \pm 23.07	174.05 \pm 8.73	3.55 \pm 0.47	128.76 \pm 23.09
Grid	46.84 \pm 4.41	157.65 \pm 6.55	9.92 \pm 1.52	43.43 \pm 2.09
Mountain				
50%	40.57 \pm 2.47	43.24 \pm 2.12	2.85 \pm 0.29	42.01 \pm 3.06
70%	53.72 \pm 4.83	43.36 \pm 2.23	2.08 \pm 0.24	55.28 \pm 5.0
90%	70.28 \pm 9.04	44.82 \pm 3.63	1.51 \pm 0.29	72.46 \pm 10.51
Grid	25.56 \pm 3.27	43.01 \pm 4.61	4.69 \pm 0.72	21.17 \pm 1.47

**Fig. 4.** Block distortion removal.**Table 3.** SIR-values and the runtime/iteration obtained in the third experiment.

	NMCSA	SVT	LMaFit	ICSA-SI
SIR [dB]	20.4 \pm 0.7	14.0 \pm 0	19.0 \pm 0	21.0 \pm 1.0
Time [ms/iter]	16.4 \pm 0.6	47.6 \pm 1.5	4.0 \pm 0	22.0 \pm 1.0

5 Conclusions

We proposed the ICSA-SI – the modified version of the NMCSA for solving image completion problems. The experiments confirmed that the ICSA-SI outperforms

the other algorithms in terms of the quality of recovered images, especially for highly incomplete data (in which at least 70% of pixels is missing). Only the ICSA-SI is able to remove the grid (with some distortions) from the grid-disturbed images, leading to the highest SIR values. The SIR results listed in Table 1 shows that the most promising approach is to repeat the procedure for selecting the extreme rays. The experiments confirmed that initially-selected extreme rays contain missing pixels, and hence the recursive strategy in the ICSA is important to relax this problem. Comparing the ICSA-`rand` and ICSA-SI, we can conclude that the `SimplexMax`-based strategy may lead to slightly better results than the random projections but we did not observe the case where the difference is large. This strategy probably would be very efficient for fully-separable images, however natural images are at most near-separable, and hence they might be only roughly modeled by a low-order polyhedral cone generated by the selected columns from the underlying image. The quality of recovery missing entries with the ICSA depends on the scenery of a completed image. For example, the `mountain` image in Fig. 1 demonstrates a slow-varying horizontal smoothness, and hence it may be better represented by such a geometric model, whereas the `boats` contains more details in its horizontal direction, so its geometric representation might be worse. Indeed, this statement is confirmed in Fig. 3. The quality of the reconstructed `mountain` image is better (in terms of the SIR measure) than for the `boats` image.

In the third experiment with the `ship` image, the ICSA-SI also gives the best solution, although not as efficient as presented in [21], introducing some disturbances.

The runtime/iteration of ICSA-SI depends on the number of recovered entries (see Table 2), and is comparable with the NMCSA, and usually shorter than for SVT. However, it is order-value longer than for LMaFit.

Summing up, the proposed method outperforms the other methods in terms of the quality of results for all the tested cases. Further research is needed to optimize the computational time.

Acknowledgment. This work was partially supported by the grant 2015/17/B/ST6/01865 funded by National Science Center (NCN) in Poland. Calculations were performed at the Wroclaw Centre for Networking and Supercomputing under grant no. 127.

References

1. 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, Chichester (2009)
2. Wang, Y.X., Zhang, Y.J.: Nonnegative matrix factorization: a comprehensive review. *IEEE Trans. Knowl. Data Eng.* **25**(6), 1336–1353 (2013)
3. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999)

4. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. In: *Advances in Neural Information Processing, NIPS*, vol. 13, pp. 556–562. MIT Press (2001)
5. Kumar, A., Sindhvani, V., Kambadur, P.: Fast conical hull algorithms for near-separable non-negative matrix factorization. In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*, Atlanta, Georgia, USA, vol. 28, pp. 231–239 (2013)
6. Bittorf, V., Recht, B., Re, C., Tropp, J.: Factoring nonnegative matrices with linear programs. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 25, pp. 1214–1222 (2012)
7. Gillis, N.: Successive nonnegative projection algorithm for robust nonnegative blind source separation. *SIAM J. Imaging Sci.* **7**(2), 1420–1450 (2014)
8. Kersting, K., Wahabzada, M., Thureau, C., Bauckhage, C.: Hierarchical convex NMF for clustering massive data. In: Sugiyama, M., Yang, Q. (eds.) *Proceedings of 2nd Asian Conference on Machine Learning Research*, PMLR, Tokyo, Japan, vol. 13, pp. 253–268, 08–10 November 2010
9. Zdunek, R.: Initialization of nonnegative matrix factorization with vertices of convex polytope. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) *ICAISC 2012. LNCS (LNAI)*, vol. 7267, pp. 448–455. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29347-4_52
10. Bioucas-Dias, J.M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., Chanussot, J.: Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**(2), 354–379 (2012)
11. Chan, T.H., Ma, W.K., Ambikapathi, A.M., Chi, C.Y.: A simplex volume maximization framework for hyperspectral endmember extraction. *IEEE Trans. Geosci. Remote Sens.* **49**(11), 4177–4193 (2011)
12. Arora, S., Ge, R., Halpern, Y., Mimno, D.M., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. In: *Proceedings of ICML, JMLR Workshop and Conference Proceedings*, JMLR.org, vol. 28, pp. 280–288 (2013)
13. Ding, W., Rohban, M.H., Ishwar, P., Saligrama, V.: Topic discovery through data dependent and random projections. In: *Proceedings of ICML*, vol. 28, pp. 471–479 (2013)
14. Guo, X., Ma, Y.: Generalized tensor total variation minimization for visual data recovery. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015
15. Yu, X., Bian, W., Tao, D.: Scalable completion of nonnegative matrix with separable structure. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, 12–17 February 2016, pp. 2279–2285 (2016)
16. Cichocki, A., Zdunek, R., Amari, S.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) *ICA 2007. LNCS*, vol. 4666, pp. 169–176. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74494-8_22
17. Wang, F.Y., Chi, C.Y., Chan, T.H., Wang, Y.: Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(5), 875–888 (2010)
18. Yokota, T., Zhao, Q., Cichocki, A.: Smooth PARAFAC decomposition for tensor completion. *IEEE Trans. Signal Process.* **64**(20), 5423–5436 (2016)

19. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20**(4), 1956–1982 (2010)
20. Wen, Z., Yin, W., Zhang, Y.: Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Math. Program. Comput.* **4**(4), 333–361 (2012)
21. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans. Image Process.* **16**(11), 2649–2661 (2007)



Feature Selection in Weakly Coherent Matrices

Stéphane Chrétien¹(✉) and Olivier Ho²

¹ National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK

stephane.chretien@npl.co.uk

² Université de Franche Comté, 16 route de Gray, 25000 Besançon, France

zhen_wai_olivier.ho@univ-fcomte.fr

<https://sites.google.com/view/stephanchretien/home>

Abstract. A problem of paramount importance in both pure (Restricted Invertibility problem) and applied mathematics (Feature extraction) is the one of selecting a submatrix of a given matrix, such that this submatrix has its smallest singular value above a specified level. Such problems can be addressed using perturbation analysis. In this paper, we propose a perturbation bound for the smallest singular value of a given matrix after appending a column, under the assumption that its initial coherence is not large, and we use this bound to derive a fast algorithm for feature extraction.

Keywords: Restricted invertibility · Coherence · Null space property

1 Introduction

In this paper, all considered matrices will be assumed to have their columns ℓ_2 -normalised.

1.1 Background on Singular Value Perturbation

Spectrum perturbation after appending a column has been addressed recently in the literature as a key ingredient in the study of graph sparsification [3], control of pinned systems of ODE's [26], the spiked model in statistics [25]; it can also be useful in Compressed Sensing [16] or for the column selection problem [17]. It is also connected to column selection problems in pure mathematics (Grothendieck and Pietsch factorisation and the Bourgain-Tzafriri restricted invertibility problem) [30].

The goal of the present paper is to study this particular perturbation problem in the special context of column subset selection. The column selection problem was proved essential in High Dimensional Data Analysis [4, 22, 23, 31, 33], etc. Different criteria for column subset selection have been studied [8]. The need for efficient column selection in the era of Big Data is more pressing than ever. Moreover, deterministic techniques are often preferred over randomised techniques in industrial applications due to repeatability constraints.

1.2 Previous Approaches to Column Selection

Several approaches have been extensively discussed in the literature. Other *deterministic* approaches have been studied recently in the pure mathematics literature, namely [28,32]. However, these approaches are computationally expensive because of the necessity to perform a matrix inversion at each step. The method of [30] combines randomness with semi-definite programming and although very elegant, is not computationally efficient in practice. A quite efficient technique is the rank-revealing QR decomposition. Table 1 in [9] provides the performance of this approach and compares it with various other methods. Randomized sampling-based approaches sometimes prove to be faster than the deterministic approaches. For instance methods based e.g. on leverage scores is often giving satisfactory results in practice. Note also that CUR decomposition is much related to the Column Selection tasks and the associated methods can be relevant in practice. A very interesting and efficient approach is the simple greedy algorithm presented in [20,20]. However, the method of [21] does not allow for control on the smallest singular value of the selected submatrix, a criterion which is often considered important for selecting sufficiently decorrelated features.

1.3 Coherence

The coherence of a matrix X , usually denoted by $\mu(X)$, is defined as

$$\mu(X) = \max_{1 \leq k < l \leq p} |\langle X_k, X_l \rangle|. \quad (1.1)$$

If the coherence is equal to zero, then the matrix is orthogonal. On the other hand, small coherence does not mean that X is close to square and orthogonal. Indeed, as easy computations show, e.g. i.i.d. Gaussian matrices with values in $\mathbb{R}^{n \times p}$ and normalised columns can have a coherence of order $\log(p)^{-1}$ even for n of order $\log(p)^3$; see [13, Sect. 1.1]. Situations where small coherence holds arise often in practice, especially in signal processing [11] and statistics [13]. The coherence of a matrix has attracted renewed interest recently due to its prominent role in Compressed Sensing [14], Matrix Completion [27], Robust PCA [12] and Sparse Estimation in general. The relationship between coherence and how many columns one can extract uniformly at random which build up a robustly invertible submatrix are studied in [15]. When the coherence is not sufficiently small, the results in [15] are not so much useful anymore and we should turn to the problem of extracting one submatrix with largest possible number of columns with smallest possible correlation. Using coherence information in the study of fast column selection procedures is one interesting question to address in this field.

1.4 Contribution of the Paper

We propose a greedy algorithm for column subset selection and apply this algorithm to some practical problems. Our contribution to the perturbation and the

column selection problems focuses on the special setting where the matrix under study has low coherence. Interestingly, standard perturbation results, e.g. [5] do not take into account the potential incoherence of the matrix under study. The results presented in this paper seem to be the first to incorporate such prior information into the analysis of a column subset selection procedure.

Our approach here is based on a new eigenvalue perturbation bound for matrices with small coherence. Previous bounds have been obtained using the famous Gershgorin's circles theorem [1] but Gershgorin's bound is often too crude as demonstrated in [18] and recent advances have been obtained in this direction in [28, 32].

2 Main Results

Our main result is a bound on the smallest singular value after appending a column of a given data matrix with potentially small coherence. Our approach is based on a new result about eigenvalue perturbation. Perturbation after appending a column is a special type of perturbation [16]. The goal of the next subsections is to prove refined results of this type for this problem.

Theorem 2.1 is our first main result on perturbation. This result gives a perturbation bound on the spectrum of a submatrix X_{T_0} of a matrix X . Corollary 2.2 takes into account the fact that the coherence of a submatrix can be smaller by a factor α than the coherence of the full matrix. This factor α is crucial in the study of e.g. greedy algorithms for column selection where at each step, the selected submatrix has much better coherence than the full matrix from which it is extracted. Corollary 2.3 proves a bound on the smallest singular value after successively appending several columns. An example where this result will be useful is the application to greedy column selection algorithms where it can provide a relevant stopping criterion.

2.1 Appending One Vector: Perturbation of the Smallest Non Zero Eigenvalue

If we consider a subset T_0 of $\{1, \dots, p\}$ and a submatrix X_{T_0} of X , the problem of studying the eigenvalue perturbations resulting from appending a column X_j to X_{T_0} , with $j \notin T_0$ can be studied using Cauchy's Interlacing Lemma as in the following result.

Theorem 2.1. *Let $T_0 \subset \{1, \dots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X . Let $\lambda_1(X_{T_0}X_{T_0}^t) \geq \dots \geq \lambda_{s_0}(X_{T_0}X_{T_0}^t)$ be the eigenvalues of $X_{T_0}X_{T_0}^t$. We have*

$$\lambda_{s_0+1}(X_{T_0}X_{T_0}^t + X_jX_j^t) \geq \lambda_{s_0}(X_{T_0}X_{T_0}^t) - \min \left(\|X_{T_0}^t X_j\|_2, \frac{\|X_{T_0}^t X_j\|_2^2}{1 - \lambda_{s_0}(X_{T_0}X_{T_0}^t)} \right). \quad (2.2)$$

Proof. Setting $v = X_j$

$$A = X_{T_0} X_{T_0}^t$$

we obtain from Proposition A.1 that the smallest nonzero eigenvalue of $X_{T_0} X_{T_0}^t + X_j X_j^t$ is the smallest root of

$$f(x) = 1 - \sum_{i=1}^n \frac{\langle v, u_i \rangle^2}{x - \lambda_i (X_{T_0} X_{T_0}^t)}.$$

We can decompose this function into two terms

$$f(x) = 1 - \sum_{i=1}^{s_0} \frac{\langle v, u_i \rangle^2}{x - \lambda_i (X_{T_0} X_{T_0}^t)} - \sum_{i=s_0+1}^n \frac{\langle v, u_i \rangle^2}{x - \lambda_i (X_{T_0} X_{T_0}^t)}.$$

Since $\lambda_i (X_{T_0} X_{T_0}^t) = 0$ for $i = s_0 + 1, \dots, n$, we get

$$f(x) = 1 + \sum_{i=1}^{s_0} \frac{\langle v, u_i \rangle^2}{\lambda_i (X_{T_0} X_{T_0}^t) - x} - \sum_{i=s_0+1}^n \frac{\langle v, u_i \rangle^2}{x}.$$

Notice that

$$\sum_{i=1}^{s_0} \langle v, u_i \rangle^2 \leq \frac{1}{\lambda_{s_0} (X_{T_0} X_{T_0}^t)} \sum_{i=1}^{s_0} \lambda_i (X_{T_0} X_{T_0}^t) \langle v, u_i \rangle^2 = \frac{1}{\lambda_{s_0} (X_{T_0} X_{T_0}^t)} \|X_{T_0}^t v\|_2^2.$$

Since f is increasing on the set $]0, \lambda_{s_0} (X_{T_0} X_{T_0}^t) [$, the smallest root of f is larger than the smallest positive root of \tilde{f} with

$$\tilde{f}(x) = 1 + \frac{\|X_{T_0}^t X_j\|_2^2}{\lambda_{s_0} (X_{T_0} X_{T_0}^t) (\lambda_{s_0} (X_{T_0} X_{T_0}^t) - x)} - \frac{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t)^{-1} \|X_{T_0}^t X_j\|_2^2}{x}.$$

Thus, after some easy calculations, we find that

$$\lambda_{s_0+1} (X_{T_0} X_{T_0}^t + X_j X_j^t) \geq \frac{1 + \lambda_{s_0} (X_{T_0} X_{T_0}^t) - \sqrt{(1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t))^{-2} + 4 \|X_{T_0}^t X_j\|_2^2}}{2}$$

which, using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $\sqrt{1+a} \leq 1 + \frac{a}{2}$, easily gives (2.2).

This theorem is useful in the case where μ small enough so that $\|X_{T_0}^t X_j\|_2^2 \leq 1$. In practice, the submatrices X_{T_0} of X have better coherence than X , up to a factor α . Moreover, we have $\|X_{T_0} X_j\|_2^2 \leq s_0 \mu^2$. The following corollary rephrases Theorem Sect. 2 using the parameter α .

Corollary 2.2. *Let X and T_0 be defined as in Theorem 2.1 and assume*

$$\|X_{T_0}^t X_j\|_2^2 \leq \alpha s_0 \mu^2.$$

Then

$$\lambda_{s_0+1} (X_{T_0} X_{T_0}^t + X_j X_j^t) \geq \lambda_{s_0} (X_{T_0} X_{T_0}^t) - \min \left(\sqrt{\alpha s_0 \mu^2}, \frac{\alpha s_0 \mu^2}{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t)} \right). \tag{2.3}$$

2.2 Successive Perturbations

If we append s_1 columns successively to the matrix X_{T_0} , we obtain the following result

Corollary 2.3. *Let $T_0 \subset \{1, \dots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X . Let $T_1 \subset \{1, \dots, p\}$ with $|T_1| = s_1$ and $T_0 \cap T_1 = \emptyset$. Let*

$$\varepsilon_{min} = \min \left(\sqrt{\alpha \mu^2} \sum_{i=s_0}^{s_0+s_1} \sqrt{i}, \frac{\alpha \mu^2 s_0}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t)} + \frac{2(1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t))}{s_0} \sum_{i=s_0+1}^{s_0+s_1} \frac{i}{i-1} \right). \quad (2.4)$$

Then

$$\lambda_{s_0+s_1}(X_{T_0 \cup T_1}^t X_{T_0 \cup T_1}) \geq \lambda_{s_0}(X_{T_0} X_{T_0}^t) - \varepsilon_{min} \quad (2.5)$$

3 A Greedy Algorithm for Column Selection

The analysis in Sect. 2 suggest that a greedy algorithm can be easily devised for efficient column extraction. The idea is quite simple: append the column which minimises the norm of the scalar products with the columns selected up to the current iteration. This algorithm is described with full details in Algorithm 1 below.

Algorithm 1. Greedy column selection

1: **procedure** GREEDY COLUMN SELECTION

2: Set $s = 1$ and choose a random singleton $T = \{j^{(1)}\} \subset \{1, \dots, p\}$. Set $\eta^{(1)} = 1$.

3: **while** $\eta^{(s)} \geq 1 - \varepsilon$ **do**

4: Set

$$j^{(s)} \in \operatorname{argmin}_{j \in \{1, \dots, p\} \setminus T} \|X_T^t X_j\|_2.$$

5: Set

$$\alpha^{(s)} = \|X_T^t X_{j^{(s)}}\|_2^2 / (s \mu^2).$$

6: Set $T = T \cup \{j^{(s)}\}$.

7:

$$\eta^{(s+1)} = \eta^{(s)} - \min \left(\sqrt{\alpha^{(s)}} s \mu, \frac{\alpha \mu^2 s}{1 - \lambda_s(X_T^t X_T)} \right)$$

8: $s \leftarrow s + 1$

Note that Algorithm 1 requires the computation of the smallest eigenvalue at each step, which might be computationally expensive in large dimensional settings.

4 Numerical Experiments

4.1 Extracting Representative Time Series

Time series are ubiquitous in a world where so many phenomena are monitored via sensor networks. One interesting application of greedy column selection is to

- extract representative time series among large datasets and
- understand the intrinsic “dimension” of the dataset, i.e. the maximum number of different dynamics that are present.
- extract potential outliers.

In this experiment, we considered a set of 1479 times series of length 39 which consist in non-linear transformation of satellite InSAR data¹. Then, starting from a random time series, we extracted 150 times series sequentially minimizing $\|X_T^t X_j\|_2, j \notin T$ at each step. Figure 1 shows the behavior of our algorithm over time. For large μ , we see that the bound provided by Corollary 2.3 are worse than the Gershgorin bound and successive applications of Theorem 2.1 provides again a better bound.

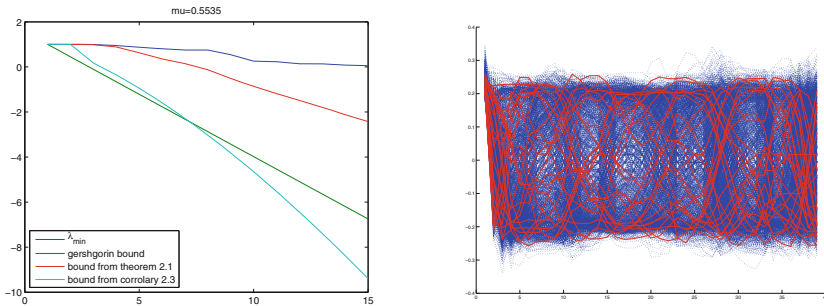


Fig. 1. Left: Evolution of the smallest singular value in the greedy column selection Algorithm 1. Right: Main extracted Features.

4.2 Extracting Representative Images from a Dataset

Extracting representative objects in a dataset is of great importance in data analytics. It can be used to detect outliers or clusters. In this example, we applied our technique to the Yale Faces database shown in Fig. 2 (Left). In order to cluster the set of images, we performed a preliminary scattering transform [10, 24] of the images in the dataset. We then reshaped the resulting scattering transform matrices into column vectors that we further concatenated into a single matrix X . We selected 9 faces using our column selection algorithm and we obtained the result shown in Fig. 2 (Right). The total time for this computation was .07 s. Larger Pictures are given in the associated report [18].

¹ A non-linear transformation was performed in order to make the time-series locations and sources impossible to identify.

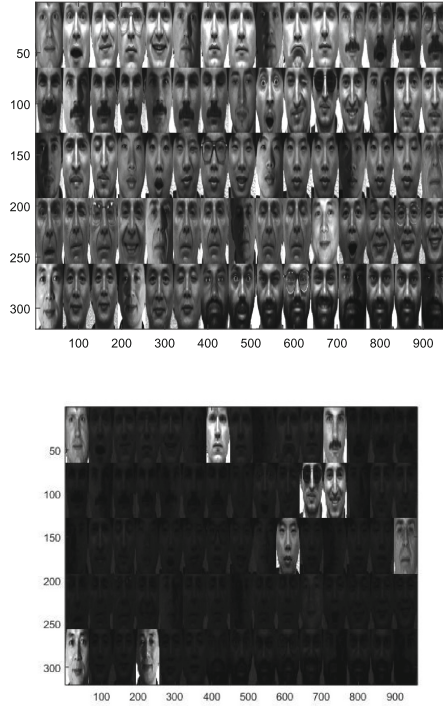


Fig. 2. Left: Faces from the Yale database. Right: Faces selected by our algorithm.

4.3 Comparison with CUR

We compared the behavior of our method with the CUR algorithm proposed in [9]. We generated 100 matrices with i.i.d. standard Gaussian entries, with 100 rows and 10000 columns and performed both Algorithm 1 from the present paper and the CUR method. We restricted the study to the case of 10 columns to be extracted. The following histograms in Fig. 3 show the relative performance of our method as compared to CUR [9]².

The Monte Carlo experiments shown in Fig. 3 prove that our method performs better than the CUR method, both from the viewpoint of providing submatrices with larger singular values on average and for a much smaller computational effort (our method was around 50 times faster for these experiments). These experiments are extracted from a more extensive set of experiments, including comparison with other methods, proposed in [18].

² We used the Matlab implementation provided on Christos Boutsidis webpage.

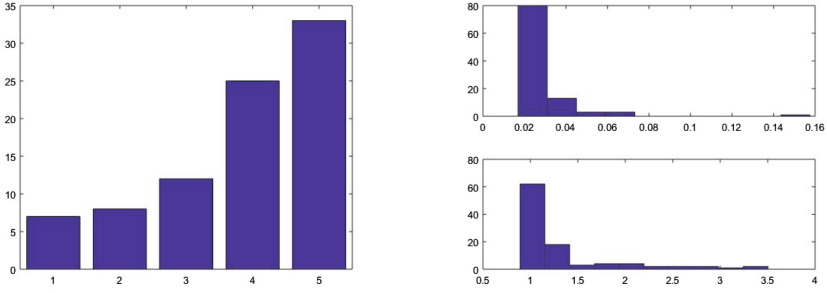


Fig. 3. Left: counts of the number of singular values of the submatrix extracted with Algorithm 1 larger than for CUR among the 5 smallest singular spectrum for 100 independent Monte Carlo trials. Right-top: histogram of the computation time for Algorithm 1. Right-bottom: histogram of the computation time for the CUR method [9].

5 Conclusion and Perspectives

In this paper, we established a relationship between the coherence and a perturbation bound for incoherent matrices. Our approach is based on perturbation theory and no randomness assumption on the design matrix is used to establish this property. Coherence plays an important role in many pure and applied mathematical problems and perturbation results may help go significantly further. Two such problems for which we are planning further investigations are the following.

- **Random submatrices are well conditioned.** Matrices with small coherence have a very nice property: most submatrices with s columns have their eigenvalues concentrated around 1 for s of the order $n/\log(p)$. This was first studied in [29], [13, Theorem 3.2 and following comments] and then improved in [15]. The study of such properties is of tremendous importance in the study of designs for sparse recovery [13]. An interesting potential application of studying spectrum perturbations after appending a column is the one of spectrum concentration via the bounded difference inequality [6]. Such concentration bounds should also appear essential in understanding the behavior of random column sampling algorithms [8, 19].
- **The restricted invertibility problem.** Given any matrix X , the Restricted Invertibility problem of Bourgain and Tzafriri is the one of extracting the largest number of columns X_j , $j \in T$ from X while ensuring that the smallest singular value of X_T stays away from zero. Different procedures have been proposed for this problem. Some of them are randomised and some are deterministic. The original results obtained by Bourgain and Tzafriri were based on random selection [7]. The current best results were recently obtained by Youssef in [32] based on an remarkable inequality discovered by Batson, Spielman and Srivastava in [2]. In [17], using an elementary perturbation approach,

the first author and S. Darses recently obtained a very short proof of a weaker version of the Bourgain-Tzafriri theorem (up to a $\log(s)$ multiplicative term). Our next goal is to refine these types of perturbation results in the small coherence setting and extend the applicability to Big Data analytics.

Acknowledgements. The work of the first author was funded by The National Measurement Office of the UK’s Department for Business, Energy and Industrial Strategy supported this work as part of its Materials and Modelling programme.

A Interlacing and the Characteristic Polynomial

Recall that for a matrix A in $\mathbb{R}^{n \times n}$, p_A denotes the characteristic polynomial of A .

Proposition A.1. Cauchy’s Interlacing theorem. *If $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix with eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$ and associated eigenvectors v_1, \dots, v_n , and $v \in \mathbb{R}^n$, then*

$$p_{A+vv^t}(x) = p_A(x) \left(1 - \sum_{i=1}^n \frac{\langle v, u_i \rangle^2}{x - \lambda_i} \right). \tag{A.6}$$

The previous lemma states in particular that the eigenvalues of A interlace those of $A + vv^t$.

B Proof of Corollary 2.3

Define $\lambda_{s_0+s,\min}$ by

$$\begin{cases} \lambda_{s_0,\min} = \lambda_{s_0}(X_{T_0}X_{T_0}^t) \\ \lambda_{s_0+s+1,\min} = \lambda_{s_0+s}(X_{T_0 \cup T}X_{T_0 \cup T}^t) - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha\mu^2(s_0+s)}{1-\lambda_{s_0+s,\min}}\right) \end{cases}$$

There are two step to prove for the theorem. The first step set up the basis for some recurrence relation. We show that, for $s \geq 0$, to obtain a lower-bound of λ_{s_0+s+1} , it is enough to use $\lambda_{s_0+s,\min}$ as the basis for Corollary 2.2. Or simply that we have

$$\begin{aligned} & \lambda_{s_0+s,\min} - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s,\min}}\right) \\ & \leq \lambda_{s_0+s}(X_{T_0 \cup T}X_{T_0 \cup T}^t) - \min\left(\sqrt{\alpha\mu^2(s_0+s)}, \frac{\alpha(s_0+s)\mu^2}{1-\lambda_{s_0+s}(X_{T_0 \cup T}X_{T_0 \cup T}^t)}\right) \\ & \leq \lambda_{s_0+s+1}(X_{T_{s+1}}X_{T_{s+1}}^t). \end{aligned}$$

It is obvious that the case where one minimum is equal to $\sqrt{\alpha\mu(s_0+s)}$ satisfy the property. Therefore, we study the following inequality

$$\lambda_{s_0+s, \min} - \frac{\alpha(s_0 + s)\mu^2}{1 - \lambda_{s_0+s, \min}} \leq \lambda_{s_0+s} - \frac{\alpha(s_0 + s)\mu^2}{1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t)}$$

It is easily verified that the property is true for $s = 0$. Denote

$$\varepsilon = \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t) - \lambda_{s_0+s+1} (X_{T_{s+1}} X_{T_{s+1}}^t). \tag{B.7}$$

Then the recursion step is equivalent to proving that

$$\alpha\mu^2 \frac{s_0 + s}{1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t)} + \alpha\mu^2 \frac{s_0 + s + 1}{1 - \lambda_{s_0+s+1, \min}} \geq \varepsilon + \alpha\mu^2 \frac{s_0 + s + 1}{1 - \lambda_{s_0+s} (X_{T_0} X_{T_0}^t) + \varepsilon}. \tag{B.8}$$

This inequality can be interpreted as the sum of errors obtained by applying Corollary 2.2 twice is greater than the sum of errors obtained if we knew the true value after one perturbation then apply Corollary 2.2.

Let g be defined by

$$g_{s_0+s}(x) = x + \alpha\mu^2 \frac{s_0 + s + 1}{1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t) + x}.$$

Since $\varepsilon \leq \alpha\mu^2 \frac{s_0+s}{1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t)}$ by Corollary (2.2), it is enough to prove g increasing.

A simple analysis show that g is strictly increasing if

$$\alpha\mu^2 \frac{s_0 + s + 1}{(1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t))^2} < \frac{3}{4}.$$

In the case $\alpha\mu^2 \frac{s_0+s+1}{(1 - \lambda_{s_0+s} (X_{T_0 \cup T} X_{T_0 \cup T}^t))^2} > \frac{3}{4}$, we can show that the left side of in Eq. (B.8) is larger than $1 - \lambda_{s_0+s} (T_0 \cup T)$ and this means that we obtain the trivial bound 0 and therefore of not relevant interest.

This solves the problem of not knowing the true value $\lambda_{s_0+s}(T_0 \cup T_1)$.

For the second part, we aim at bounding the sum of errors. We have

$$\sum_{i=s_0}^{s_0+s} \min \left(\sqrt{\alpha\mu^2 i}, \frac{\alpha\mu^2 i}{1 - \lambda_{i, \min}} \right) \leq \min \left(\sum_{i=s_0}^{s_0+s} \sqrt{\alpha\mu^2 i}, \sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1 - \lambda_{i, \min}} \right).$$

The second sum writes

$$\begin{aligned} & \sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t) + \sum_{j=s_0}^{i-1} \frac{\alpha\mu^2 j}{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t)}} \\ &= \sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t) + \frac{\alpha\mu^2}{1 - \lambda_{s_0} (X_{T_0} X_{T_0}^t)} \sum_{j=s_0}^{i-1} j} \end{aligned}$$

This is equal to

$$\begin{aligned} & \sum_{i=s_0}^{s_0+s} \frac{\alpha\mu^2 i}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t) + \sum_{j=s_0}^{i-1} \frac{\alpha\mu^2 j}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t)}} \\ &= \sum_{i=s_0+1}^{s_0+s} \frac{\alpha\mu^2 i}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t) + \frac{\alpha\mu^2 s_0(i-1)}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t)}} + \frac{\alpha\mu^2 s_0}{1 - \lambda_{s_0}(X_{T_0} X_{T_0}^t)} \end{aligned}$$

Simple computations lead to the result.

Therefore applying s_1 times Corollary 2.2 and each time upper-bounding, we have (2.5).

References

1. Bandeira, A.S., Fickus, M., Mixon, D.G., Wong, P.: The road to deterministic matrices with the restricted isometry property. *J. Fourier Anal. Appl.* **19**(6), 1123–1149 (2013)
2. Batson, J., Spielman, D.A., Srivastava, N.: Twice-Ramanujan sparsifiers. *SIAM J. Comput.* **41**(6), 1704–1721 (2012)
3. Batson, J., Spielman, D.A., Srivastava, N., Teng, S.-H.: Spectral sparsification of graphs: theory and algorithms. *Commun. ACM* **56**(8), 87–94 (2013)
4. Ben-Hur, A., Guyon, I.: Detecting stable clusters using principal component analysis. In: Brownstein, M.J., Khodursky, A.B. (eds.) *Functional Genomics*, pp. 159–182. Springer, Heidelberg (2003)
5. Bhatia, R.: *Perturbation Bounds for Matrix Eigenvalues*. SIAM, Philadelphia (2007)
6. Boucheron, S., Lugosi, G., Massart, P.: *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford (2013)
7. Bourgain, J., Tzafriri, L.: Invertibility of ‘large’ submatrices with applications to the geometry of banach spaces and harmonic analysis. *Isr. J. Math.* **57**(2), 137–224 (1987)
8. Boutsidis, C., Drineas, P., Magdon-Ismail, M.: Near-optimal column-based matrix reconstruction. *SIAM J. Comput.* **43**(2), 687–717 (2014)
9. Boutsidis, C., Mahoney, M.W., Drineas, P.: An improved approximation algorithm for the column subset selection problem. In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, pp. 968–977 (2009)
10. Bruna, J., Stéphane, M.: Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1872–1886 (2013)
11. Candes, E., Romberg, J.: Sparsity and incoherence in compressive sampling. *Inverse Probl.* **23**(3), 969 (2007)
12. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? *J. ACM (JACM)* **58**(3), 11 (2011)
13. Candès, E.J., Plan, Y., et al.: Near-ideal model selection by ℓ_1 minimization. *Ann. Stat.* **37**(5A), 2145–2177 (2009)
14. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**(2), 21–30 (2008)

15. Chrétien, S., Darses, S.: Invertibility of random submatrices via tail-decoupling and a matrix chernoff inequality. *Stat. Probab. Lett.* **82**(7), 1479–1487 (2012)
16. Chretien, S., Darses, S.: Perturbation bounds on the extremal singular values of a matrix after appending a column, arXiv preprint [arXiv:1406.5441](https://arxiv.org/abs/1406.5441) (2014)
17. Chretien, S., Darses, S.: An elementary approach to the problem of column selection in a rectangular matrix, arXiv preprint [arXiv:1509.00748](https://arxiv.org/abs/1509.00748) (2015)
18. Chretien, S., Ho, Z.-W.O.: Feature selection in weakly coherent matrices, In preparation (2018)
19. Deshpande, A., Rademacher, L.: Efficient volume sampling for row, column subset selection. In: *IEEE 2010 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 329–338 (2010)
20. Farahat, A.K., Ghodsi, A., Kamel, M.S.: An efficient greedy method for unsupervised feature selection. In: *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pp. 161–170. IEEE (2011)
21. Farahat, A.K., Ghodsi, A., Kamel, M.S.: Efficient greedy feature selection for unsupervised learning. *Knowl. Inf. Syst.* **35**(2), 285–310 (2013)
22. Krzanowski, W.J.: Selection of variables to preserve multivariate data structure, using principal components. *Appl. Stat.* **36**, 22–33 (1987)
23. Mahoney, M.W., Drineas, P.: Cur matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci.* **106**(3), 697–702 (2009)
24. Mallat, S.: Group invariant scattering. *Commun. Pure Appl. Math.* **65**(10), 1331–1398 (2012)
25. Nadler, B.: Finite sample approximation results for principal component analysis: a matrix perturbation approach. *Ann. Stat.* **36**, 2791–2817 (2008)
26. Porfiri, M., Di Bernardo, M.: Criteria for global pinning-controllability of complex networks. *Automatica* **44**(12), 3100–3106 (2008)
27. Recht, B.: A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12**, 3413–3430 (2011)
28. Spielman, D.A., Srivastava, N.: An elementary proof of the restricted invertibility theorem. *Isr. J. Math.* **190**(1), 83–91 (2012)
29. Tropp, J.A.: Norms of random submatrices and sparse approximation. *Comptes Rendus Math.* **346**(23), 1271–1274 (2008)
30. Tropp, J.A.: Column subset selection, matrix factorization, and eigenvalue optimization. In: *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics*, pp. 978–986 (2009)
31. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach. *J. Mach. Learn. Res.* **6**, 1855–1887 (2005)
32. Youssef, P.: Restricted invertibility and the banach-mazur distance to the cube. *Mathematika* **60**(01), 201–218 (2014)
33. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 1151–1157. ACM (2007)



Variable Projection Applied to Block Term Decomposition of Higher-Order Tensors

Guillaume Olikier¹(✉), P.-A. Absil¹, and Lieven De Lathauwer^{2,3}

¹ ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium
guillaume.olikier@uclouvain.be

² Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

³ KU Leuven Campus Kortrijk, Kortrijk, Belgium

Abstract. Higher-order tensors have become popular in many areas of applied mathematics such as statistics, scientific computing, signal processing or machine learning, notably thanks to the many possible ways of decomposing a tensor. In this paper, we focus on the best approximation in the least-squares sense of a higher-order tensor by a block term decomposition. Using variable projection, we express the tensor approximation problem as a minimization of a cost function on a Cartesian product of Stiefel manifolds. The effect of variable projection on the Riemannian gradient algorithm is studied through numerical experiments.

Keywords: Numerical multilinear algebra · Higher-order tensor
Block term decomposition · Variable projection method
Riemannian manifold · Riemannian optimization

1 Introduction

Higher-order tensors have found numerous applications in signal processing and machine learning thanks to the many tensor decompositions available [1–4]. In this paper, we focus on a recently introduced tensor decomposition called block term decomposition (BTD) [5–7]. The usefulness of BTD in blind source separation was outlined in [8, 9] and further examples are discussed in [10–14].

This work was supported by (1) “Communauté française de Belgique - Actions de Recherche Concertées” (contract ARC 14/19-060), (2) Research Council KU Leuven: C1 project C16/15/059-nD, (3) F.W.O.: project G.0830.14N, G.0881.14N, (4) Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS Project no. 30468160 (SeLMA), (5) EU: The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC Advanced Grant: BIOTENSORS (no. 339804). This paper reflects only the authors’ views and the Union is not liable for any use that may be made of the contained information.

The BTD unifies the two most well known tensor decompositions which are the Tucker decomposition and the canonical polyadic decomposition (CPD). It also gives a unified view on how the basic concept of rank can be generalized from matrices to tensors. While in CPD, as well as in classical matrix decompositions, the components are rank-one terms, i.e., “atoms” of data, the terms in a BTD have “low” (multilinear) rank and can be thought of as “molecules” (consisting of several atoms) of data. Rank-one terms can only model data components that are proportional along columns, rows, . . . and this assumption may not be realistic. On the other hand, block terms can model multidimensional sources, variations around mean activity, mildly nonlinear phenomena, drifts of setting points, frequency shifts, mildly convolutive mixtures, and so on. Such a molecular analysis is not possible in the matrix setting. Furthermore, it turns out that, like CPDs, BTDs are still unique under mild conditions [6,10].

In practice, it is more frequent to approximate a tensor by a BTD than to compute an exact BTD. More precisely, the problem of interest is to compute the best approximation in the least-squares sense of a higher-order tensor by a BTD. Only a few algorithms are currently available for this task. The Matlab toolbox Tensorlab [15] proposes the two following functions: (i) `btd_minf` uses L-BFGS with dogleg trust region (a quasi-Newton method), (ii) `btd_nls` uses nonlinear least squares by Gauss–Newton with dogleg trust region. Another available algorithm is the alternating least squares algorithm introduced in [7]. This algorithm is not included in Tensorlab and does not work better than `btd_nls` in general.

In this paper, we show that the performance of numerical methods can be improved using variable projection. Variable projection consists in exploiting the fact that, when the optimal value of some of the optimization variables is easy to find when the others are fixed, this optimal value can be injected in the objective function, yielding a new optimization problem where only the other variables appear. This technique has already been applied to the Tucker decomposition in [16] and exploited in [17,18]. Here we extend it to the BTD approximation problem which is then expressed as a minimization of a cost function on a Cartesian product of Stiefel manifolds. Numerical experiments show that variable projection modifies the performance of the Riemannian gradient algorithm for BTDs of two terms by either increasing or decreasing its running time and/or its reliability. Preliminary results can be found in the short conference paper [19]. The present paper gives a detailed derivation of the variable projection technique and presents numerical experiments for noised BTDs. We focus on third-order tensors for simplicity but the generalization to tensors of any order is straightforward.

2 Preliminaries and Notation

We let $\mathbb{R}^{I_1 \times I_2 \times I_3}$ denote the set of real third-order tensors of size (I_1, I_2, I_3) . In order to improve readability, vectors are written in bold-face lower-case (e.g., \mathbf{a}), matrices in bold-face capitals (e.g., \mathbf{A}), and higher-order tensors in calligraphic letters (e.g., \mathcal{A}). For $n \in \{1, 2, 3\}$, the mode- n vectors of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$

are obtained by varying the n th index while keeping the other indices fixed. The mode- n rank of \mathcal{A} , denoted $\text{rank}_n(\mathcal{A})$, is the dimension of the linear space spanned by its mode- n vectors. The multilinear rank of \mathcal{A} is the triple of the mode- n ranks. The mode- n product of \mathcal{A} by $\mathbf{B} \in \mathbb{R}^{J_n \times I_n}$, denoted $\mathcal{A} \cdot_n \mathbf{B}$, is obtained by multiplying all the mode- n vectors of \mathcal{A} by \mathbf{B} . We endow $\mathbb{R}^{I_1 \times I_2 \times I_3}$ with the standard inner product, defined by

$$\langle \mathcal{A}, \mathcal{B} \rangle := \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \sum_{i_3=1}^{I_3} \mathcal{A}(i_1, i_2, i_3) \mathcal{B}(i_1, i_2, i_3),$$

and we let $\|\cdot\|$ denote the induced norm, i.e., the Frobenius norm. It is sometimes convenient to represent a tensor as a vector (vectorization) or as a matrix (matricization). The vectorization of $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, denoted $\text{vec}(\mathcal{A})$, is the vector of length $I_1 I_2 I_3$ defined as follows:

$$(\text{vec}(\mathcal{A})) ((i_1 - 1)I_2 I_3 + (i_2 - 1)I_3 + i_3) := \mathcal{A}(i_1, i_2, i_3).$$

We define the following matrix representations of \mathcal{A} :

$$\begin{aligned} \mathcal{A}(i_1, i_2, i_3) &= (\mathbf{A}_{(1)})(i_1, I_3(i_2 - 1) + i_3) \\ &= (\mathbf{A}_{(2)})(i_2, I_1(i_3 - 1) + i_1) \\ &= (\mathbf{A}_{(3)})(i_3, I_2(i_1 - 1) + i_2). \end{aligned}$$

One can check that if $\mathcal{A} = \mathcal{S} \cdot_1 \mathbf{U} \cdot_2 \mathbf{V} \cdot_3 \mathbf{W}$, then

$$\text{vec}(\mathcal{A}) = (\mathbf{U} \otimes \mathbf{V} \otimes \mathbf{W}) \text{vec}(\mathcal{S}), \quad (1)$$

$$\mathbf{A}_{(1)} = \mathbf{U} \mathbf{S}_{(1)} (\mathbf{V} \otimes \mathbf{W})^T, \quad (2)$$

$$\mathbf{A}_{(2)} = \mathbf{V} \mathbf{S}_{(2)} (\mathbf{W} \otimes \mathbf{U})^T, \quad (3)$$

$$\mathbf{A}_{(3)} = \mathbf{W} \mathbf{S}_{(3)} (\mathbf{U} \otimes \mathbf{V})^T. \quad (4)$$

Vectorization and matricization are linear mappings which preserve the norm.

3 Variable Projection

Let $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. Consider positive integers R and R_i such that $R_i \leq \text{rank}_i(\mathcal{A})$ for each $i \in \{1, 2, 3\}$ and $m := I_1 I_2 I_3 \geq R R_1 R_2 R_3 =: n$. The approximation of \mathcal{A} by a BTM of R terms of multilinear rank (R_1, R_2, R_3) is a nonconvex minimization problem which can be expressed using variable projection as

$$\min_{\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W}} \left\| \underbrace{\mathcal{A} - \sum_{r=1}^R \mathcal{S}_r \cdot_1 \mathbf{U}_r \cdot_2 \mathbf{V}_r \cdot_3 \mathbf{W}_r}_{=: f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})} \right\|^2 = \min_{\mathbf{U}, \mathbf{V}, \mathbf{W}} \underbrace{\min_{\mathcal{S}} f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}_{=: g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})}$$

for the variables $\mathcal{S} \in (\mathbb{R}^{R_1 \times R_2 \times R_3})^R$, $\mathbf{U} \in (\mathbb{R}^{I_1 \times R_1})^R$, $\mathbf{V} \in (\mathbb{R}^{I_2 \times R_2})^R$ and $\mathbf{W} \in (\mathbb{R}^{I_3 \times R_3})^R$ subject to the constraints $\mathbf{U} \in \text{St}(R_1, I_1)^R$, $\mathbf{V} \in \text{St}(R_2, I_2)^R$

and $\mathbf{W} \in \text{St}(R_3, I_3)^R$, where given integers $p \geq q \geq 1$ we let $\text{St}(q, p)$ denote the *Stiefel manifold*, i.e.,

$$\text{St}(q, p) := \{\mathbf{X} \in \mathbb{R}^{p \times q} : \mathbf{X}^T \mathbf{X} = \mathbf{I}_q\}.$$

A schematic representation of the BTD approximation problem is given in Fig. 1. Each term in a BTD is a Tucker term. The tensors $\mathcal{S}_r \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ are called the core tensors while the matrices $\mathbf{U}_r, \mathbf{V}_r, \mathbf{W}_r$, which can be assumed to be in the Stiefel manifold without loss of generality, are referred to as the factor matrices.

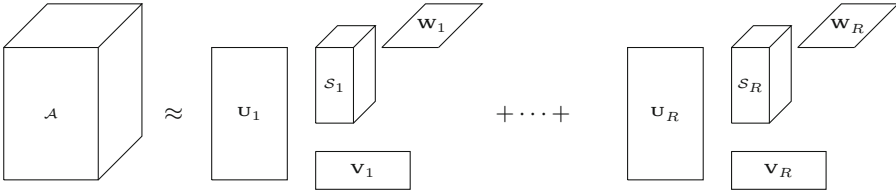


Fig. 1. Schematic representation of the BTD approximation problem.

Computing $g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is a least squares problem. Indeed, using (1), if we define $\mathbf{a} := \text{vec}(\mathcal{A}) \in \mathbb{R}^m$, $\mathbf{P}(\mathbf{U}, \mathbf{V}, \mathbf{W}) := [\mathbf{U}_j \otimes \mathbf{V}_j \otimes \mathbf{W}_j]_{i,j=1}^{1,R} \in \mathbb{R}^{m \times n}$ and $\mathbf{s} := [\text{vec}(\mathcal{S}_i)]_{i,j=1}^{R,1} \in \mathbb{R}^n$, then

$$g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W}) = \min_{\mathbf{s} \in \mathbb{R}^n} \|\mathbf{a} - \mathbf{P}(\mathbf{U}, \mathbf{V}, \mathbf{W})\mathbf{s}\|^2.$$

We let $\mathcal{S}^*(\mathbf{U}, \mathbf{V}, \mathbf{W})$ denote the minimizer of this least squares problem.¹ Thus,

$$g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W}) = f_{\mathcal{A}}(\mathcal{S}^*(\mathbf{U}, \mathbf{V}, \mathbf{W}), \mathbf{U}, \mathbf{V}, \mathbf{W}).$$

Computing the partial derivatives of $g_{\mathcal{A}}$ reduces to the computation of partial derivatives of $f_{\mathcal{A}}$. Indeed, using the first-order optimality condition

$$\left. \frac{\partial f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathcal{S}} \right|_{\mathcal{S}=\mathcal{S}^*(\mathbf{U}, \mathbf{V}, \mathbf{W})} = \mathbf{0} \tag{5}$$

and the chain rule yields

$$\frac{\partial g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial(\mathbf{U}, \mathbf{V}, \mathbf{W})} = \left. \frac{\partial f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial(\mathbf{U}, \mathbf{V}, \mathbf{W})} \right|_{\mathcal{S}=\mathcal{S}^*(\mathbf{U}, \mathbf{V}, \mathbf{W})}. \tag{6}$$

It remains to compute those partial derivatives of $f_{\mathcal{A}}$. In order to make the derivation convenient, we first recall some basic facts on differentiation. Given two vector spaces X and Y over a same field, we let $\text{Lin}(X, Y)$ denote the vector space of linear mappings from X to Y .

¹ The minimizer is unique if and only if the matrix $\mathbf{P}(\mathbf{U}, \mathbf{V}, \mathbf{W})$ has full column rank which is the case almost everywhere (with respect to the Lebesgue measure) since $m \geq n$.

Total Derivative and Gradient. Let $(X, \langle \cdot, \cdot \rangle)$ be a pre-Hilbert space and let $\|\cdot\|$ denote the norm induced by the inner product $\langle \cdot, \cdot \rangle$. A function $f : X \rightarrow \mathbb{R}$ is *differentiable* at $x \in X$ if and only if there is $L \in \text{Lin}(X, \mathbb{R})$ such that

$$\lim_{h \rightarrow 0} \frac{f(x+h) - f(x) - L(h)}{\|h\|} = 0,$$

which means that for every $\epsilon > 0$, there is $\delta > 0$ such that for any $h \in X$, $\|h\| \leq \delta$ implies

$$\frac{|f(x+h) - f(x) - L(h)|}{\|h\|} \leq \epsilon.$$

If such a L exists, it is unique, denoted by $Df(x)$, and called the *total derivative* of f at x . The *gradient* of f at x is the only $g \in X$ such that

$$Df(x)[h] = \langle g, h \rangle$$

for all $h \in X$; it is denoted by $\text{grad } f(x)$. If f is differentiable at $x \in X$, then

$$Df(x)[h] = \lim_{t \rightarrow 0} \frac{f(x+th) - f(x)}{t}.$$

for every $h \in X$.

Gradient of the Squared Norm. Let $f : X \rightarrow \mathbb{R} : x \mapsto f(x) := \|x\|^2$. For any $x, h \in X$ and any real $t \neq 0$,

$$\frac{f(x+th) - f(x)}{t} = \frac{2t\langle x, h \rangle + t^2 \|h\|^2}{t} = 2\langle x, h \rangle + t \|h\|^2.$$

It follows that $Df(x)[h] = 2\langle x, h \rangle$ and so that $\text{grad } f(x) = 2x$.

Affine Transformation. Let $(X, \langle \cdot, \cdot \rangle_X)$ and $(Y, \langle \cdot, \cdot \rangle_Y)$ be two pre-Hilbert spaces, $g : Y \rightarrow \mathbb{R}$ be differentiable, $L \in \text{Lin}(X, Y)$, $b \in Y$, $A : X \rightarrow Y : x \mapsto A(x) := L(x) + b$, and $f := g \circ A$. For any $x, h \in X$,

$$\begin{aligned} \langle \text{grad } f(x), h \rangle_X &= \lim_{t \rightarrow 0} \frac{f(x+th) - f(x)}{t} \\ &= \lim_{t \rightarrow 0} \frac{g(L(x) + b + tL(h)) - g(L(x) + b)}{t} \\ &= \langle \text{grad } g(L(x) + b), L(h) \rangle_Y. \end{aligned}$$

From now on, let us assume that X and Y have finite dimension so that L has an *adjoint*, which means that there is a (unique) $L^* \in \text{Lin}(Y, X)$ such that

$$\langle y, L(x) \rangle_Y = \langle L^*(y), x \rangle_X$$

for any $x \in X$ and $y \in Y$. This allows us to conclude that for any $x \in X$,

$$\text{grad } f(x) = L^*(\text{grad } g(L(x) + b)).$$

Adjoint of the Matrix Product. Let $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{q \times n}$. The adjoint of

$$L : \mathbb{R}^{p \times q} \rightarrow \mathbb{R}^{m \times n} : \mathbf{X} \mapsto \mathbf{A}\mathbf{X}\mathbf{B}$$

is

$$L^* : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{p \times q} : \mathbf{Y} \mapsto \mathbf{A}^T \mathbf{Y} \mathbf{B}^T.$$

Partial derivatives of $f_{\mathcal{A}}$. Using the matricization formulas (2)–(4) yields

$$\begin{aligned} f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W}) &= \left\| \sum_{r=1}^R \mathbf{U}_r (\mathbf{S}_r)_{(1)} (\mathbf{V}_r \otimes \mathbf{W}_r)^T - \mathbf{A}_{(1)} \right\|^2 \\ &= \left\| \sum_{r=1}^R \mathbf{V}_r (\mathbf{S}_r)_{(2)} (\mathbf{W}_r \otimes \mathbf{U}_r)^T - \mathbf{A}_{(2)} \right\|^2 \\ &= \left\| \sum_{r=1}^R \mathbf{W}_r (\mathbf{S}_r)_{(3)} (\mathbf{U}_r \otimes \mathbf{V}_r)^T - \mathbf{A}_{(3)} \right\|^2. \end{aligned}$$

Applying the results of the preceding paragraphs to these three equations gives the three following ones for every $i \in \{1, \dots, R\}$:

$$\begin{aligned} \frac{\partial f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{U}_i} &= 2 \left(\sum_{j=1}^R \mathbf{U}_j (\mathbf{S}_j)_{(1)} (\mathbf{V}_j \otimes \mathbf{W}_j)^T - \mathbf{A}_{(1)} \right) (\mathbf{V}_i \otimes \mathbf{W}_i) (\mathbf{S}_i)_{(1)}^T, \\ \frac{\partial f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{V}_i} &= 2 \left(\sum_{j=1}^R \mathbf{V}_j (\mathbf{S}_j)_{(2)} (\mathbf{W}_j \otimes \mathbf{U}_j)^T - \mathbf{A}_{(2)} \right) (\mathbf{W}_i \otimes \mathbf{U}_i) (\mathbf{S}_i)_{(2)}^T, \\ \frac{\partial f_{\mathcal{A}}(\mathcal{S}, \mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{W}_i} &= 2 \left(\sum_{j=1}^R \mathbf{W}_j (\mathbf{S}_j)_{(3)} (\mathbf{U}_j \otimes \mathbf{V}_j)^T - \mathbf{A}_{(3)} \right) (\mathbf{U}_i \otimes \mathbf{V}_i) (\mathbf{S}_i)_{(3)}^T. \end{aligned}$$

4 Riemannian Gradient Algorithm

We have shown in the preceding section that the approximation of \mathcal{A} by a BTD reduces to the minimization of a real-valued function defined on a Riemannian manifold, namely, the restriction of $g_{\mathcal{A}}$ on $\prod_{i=1}^3 \text{St}(R_i, I_i)^R$. In this section, we briefly introduce the Riemannian gradient algorithm which we shall use to solve our problem; our reference is [20].

Line-search methods to minimize a real-valued function F defined on a Riemannian manifold \mathcal{M} are based on the update formula

$$x_{k+1} = R_{x_k}(t_k \eta_k),$$

where η_k is selected in the tangent space to \mathcal{M} at x_k , denoted $\mathbb{T}_{x_k} \mathcal{M}$, R_{x_k} is a retraction on \mathcal{M} at x_k , and $t_k \in \mathbb{R}$. The algorithm is defined by the choice of three ingredients: the retraction R_{x_k} , the search direction η_k and the step size t_k .

The gradient method consists of choosing $\eta_k := -\text{grad } F(x_k)$ where $\text{grad } F$ is the Riemannian gradient of F . In the case where \mathcal{M} is an embedded submanifold of a linear space \mathcal{E} and F is the restriction on \mathcal{M} of some function $\bar{F} : \mathcal{E} \rightarrow \mathbb{R}$, $\text{grad } F(x)$ is simply the projection of the usual gradient of \bar{F} at x on $\text{T}_x \mathcal{M}$. For instance, $\text{St}(q, p)$ is an embedded submanifold of $\mathbb{R}^{p \times q}$ and the projection of $\mathbf{Y} \in \mathbb{R}^{p \times q}$ on $\text{T}_{\mathbf{X}} \text{St}(q, p)$ is given by [20, Eq. (3.35)]

$$(\mathbf{I}_p - \mathbf{X}\mathbf{X}^T)\mathbf{Y} + \mathbf{X}\text{skew}(\mathbf{X}^T\mathbf{Y}) \quad (7)$$

where $\text{skew}(\mathbf{A}) := \frac{1}{2}(\mathbf{A} - \mathbf{A}^T)$ is the skew-symmetric part of \mathbf{A} . Our cost function, the restriction of $g_{\mathcal{A}}$ on $\prod_{i=1}^3 \text{St}(R_i, I_i)^R$, is defined on a Cartesian product of Stiefel manifolds; this is not an issue since the tangent space of a Cartesian product is the Cartesian product of the tangent spaces and the projection can be performed componentwise. We are now able to compute the Riemannian gradient of the restriction of $g_{\mathcal{A}}$. Starting from the first-order optimality condition (5) written in matrix forms (2)–(4), we can show that for each $i \in \{1, \dots, R\}$,

$$\mathbf{U}_i^T \frac{\partial g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{U}_i} = \mathbf{V}_i^T \frac{\partial g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{V}_i} = \mathbf{W}_i^T \frac{\partial g_{\mathcal{A}}(\mathbf{U}, \mathbf{V}, \mathbf{W})}{\partial \mathbf{W}_i} = \mathbf{0}.$$

Therefore, in view of the projection formula (7), the Riemannian gradient of the restriction of $g_{\mathcal{A}}$ is equal to the (usual) gradient of $g_{\mathcal{A}}$ given by (6).

A popular retraction on $\text{St}(q, p)$, which we shall use in our problem, is the qf retraction [20, Eq. (4.8)]:

$$R_{\mathbf{X}}(\mathbf{Y}) := \text{qf}(\mathbf{X} + \mathbf{Y})$$

where $\text{qf}(\mathbf{A})$ is the \mathbf{Q} factor of the decomposition of $\mathbf{A} \in \mathbb{R}^{p \times q}$ with $\text{rank}(\mathbf{A}) = q$ as $\mathbf{A} = \mathbf{Q}\mathbf{R}$ where $\mathbf{Q} \in \text{St}(q, p)$ and \mathbf{R} is an upper triangular $q \times q$ matrix with positive diagonal elements. Again, the manifold in our problem is a Cartesian product of Stiefel manifolds and in this case the retraction can be performed componentwise.

At this point, it remains to specify the step size t_k . For that purpose, we will use the backtracking strategy presented in [20, Sect. 4.2]. Assume we are at the k th iteration. We want to find $t_k > 0$ such that $F(R_{x_k}(-t_k \text{grad } F(x_k)))$ is sufficiently small compared to $F(x_k)$. This can be achieved by the Armijo rule: given $\bar{\alpha} > 0$, $\beta, \sigma \in (0, 1)$ and $\tau_0 := \bar{\alpha}$, we iterate $\tau_i := \beta\tau_{i-1}$ until

$$F(R_{x_k}(-\tau_i \text{grad } F(x_k))) \leq F(x_k) - \sigma\tau_i \|\text{grad } F(x_k)\|^2$$

and then set $t_k := \tau_i$. In our implementation, we set $\bar{\alpha} := 0.2$, $\sigma := 10^{-3}$, $\beta := 0.2$ and we perform at most 10 iterations in the backtracking loop.

The procedure described in the preceding paragraph corresponds to [20, Algorithm 1] with $c := 1$ and equality in [20, Eq. (4.12)], except that the number of iterations in the backtracking loop is limited. In our problem, the domain of the cost function is compact since it is a Cartesian product of Stiefel manifolds. Therefore, [20, Corollary 4.3.2] applies and ensures that

$$\lim_{k \rightarrow \infty} \|\text{grad } F(x_k)\| = 0,$$

except if at some iteration the backtracking loop needs more than 10 iterations. In view of this result, it seems natural to stop the algorithm as soon as the norm of the Riemannian gradient becomes smaller than a given quantity $\epsilon > 0$.

5 Numerical Results

In this section, we perform numerical experiments to study the effect of variable projection on the Riemannian gradient algorithm applied to the BTD problem. To this end, we evaluate the ability of this algorithm, both with and without variable projection, to recover known BTDs possibly corrupted by some noise. Thus, in this experiment, we try to recover a structure that is really present.

First, we explain how we build BTDs for this test. We set $R := 2$ and we select the parameters (I_1, I_2, I_3) and (R_1, R_2, R_3) . Then, for each $r \in \{1, \dots, R\}$, we select $\mathcal{S}_r \in \mathbb{R}^{R_1 \times R_2 \times R_3}$, $\mathbf{U}_r \in \text{St}(R_1, I_1)$, $\mathbf{V}_r \in \text{St}(R_2, I_2)$ and $\mathbf{W}_r \in \text{St}(R_3, I_3)$ according to the standard normal distribution, i.e., $\mathcal{S}_r := \text{randn}(R_1, R_2, R_3)$ and $\mathbf{U}_r := \text{qf}(\text{randn}(I_1, R_1))$ in Matlab. Then, we set

$$\mathcal{A} := \sum_{r=1}^R \mathcal{S}_r \cdot_1 \mathbf{U}_r \cdot_2 \mathbf{V}_r \cdot_3 \mathbf{W}_r. \quad (8)$$

Finally, we select $\mathcal{N} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ according to the standard normal distribution, i.e., $\mathcal{N} := \text{randn}(I_1, I_2, I_3)$ in Matlab, and define

$$\mathcal{A}_\sigma := \frac{\mathcal{A}}{\|\mathcal{A}\|} + \sigma \frac{\mathcal{N}}{\|\mathcal{N}\|} \quad (9)$$

for some real value of the parameter σ which controls the noise level on the BTD.

Now, we describe the test itself. For 100 different \mathcal{A}_σ as in (9), we ran the Riemannian gradient algorithm with variable projection (i.e., on the cost function $g_{\mathcal{A}_\sigma}$) and without variable projection (i.e., on the cost function $f_{\mathcal{A}_\sigma}$) using for each \mathcal{A}_σ a randomly selected starting iterate. Representative results are given in Table 1 for $\sigma := 0$ and $\sigma := 0.3$, which corresponds to a signal-to-noise ratio of about 10 dB, both for $(I_1, I_2, I_3) := (5, 5, 5)$ and $(R_1, R_2, R_3) := (2, 2, 2)$.²

The success ratios are not equal to one because the number of iterations that can be performed by the algorithm was (arbitrarily) limited to 10^4 . When variable projection is used, on one hand, the mean running time is multiplied by about 0.86 for $\sigma := 0$ and 0.78 for $\sigma := 0.3$, and on the other hand, the success ratio is multiplied by about 0.89 for both $\sigma := 0$ and $\sigma := 0.3$.

The same test with $(I_1, I_2, I_3) := (10, 10, 10)$ and $(R_1, R_2, R_3) := (2, 2, 3)$, still with $\sigma := 0$ and $\sigma := 0.3$, has been conducted.³ For both values of σ , we observed that variable projection multiplies the running time by about 1.1 on one hand, and multiplies the success ratio by about 1.4 on the other hand.

² The Matlab code that produced the results is available at <https://sites.uclouvain.be/absil/2018.01>.

³ With these parameters, the BTD \mathcal{A} in (8) is *essentially unique* by [6, Theorem 5.3].

Table 1. By “success”, we mean for $\sigma = 0$ that the norm of the (Riemannian) gradient is brought below $5 \cdot 10^{-14}$ and that the objective function is brought below 10^{-25} within 10^4 iterations; for $\sigma = 0.3$, we mean that the norm of the gradient is brought below 10^{-7} still within 10^4 iterations; the algorithm was not able to bring the norm of the gradient as low as in the noise-free case. Notation: “iter” refers to the number of iterations performed by the gradient algorithm while “backtracking iter” refers to the number of iterations performed in the backtracking loops. Running times are given in seconds. The information in each column is computed based only on the successful runs.

	$\sigma := 0$		$\sigma := 0.3$	
	With VP	Without VP	With VP	Without VP
successes	39	44	41	46
min(iter)	2047	2069	995	891
mean(iter)	5644	5966	4119	4740
max(iter)	9509	9960	9498	9958
mean(backtracking iter)	1	1	1.004	1
min(time)	2.11	2.36	1.05	1.02
mean(time)	5.85	6.83	4.25	5.44
max(time)	9.79	11.35	9.77	11.35

6 Conclusion

In this paper, we applied variable projection to the BTD problem and discussed its effect on the Riemannian gradient algorithm. Our numerical experiments showed that variable projection may either increase or decrease the running time and/or the reliability of the algorithm depending on the particular data tensor considered.

References

1. Cichocki, A., Mandic, D., Phan, A.H., Caiafa, C., Zhou, G., Zhao, Q., De Lathauwer, L.: Tensor decompositions for signal processing applications: from two-way to multi-way component analysis. *IEEE Signal Process. Mag.* **32**(2), 145–163 (2015)
2. Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. *IEEE Trans. Signal Process.* **65**(13), 3551–3582 (2017)
3. Cichocki, A., Lee, N., Oseledets, I., Phan, A.H., Zhao, Q., Mandic, D., et al.: Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Found. Trends Mach. Learn.* **9**(4–5), 249–429 (2016)
4. Cichocki, A., Phan, A.H., Zhao, Q., Lee, N., Oseledets, I., Sugiyama, M., Mandic, D., et al.: Tensor networks for dimensionality reduction and large-scale optimization: Part 2 applications and future perspectives. *Found. Trends Mach. Learn.* **9**(6), 431–673 (2017)

5. 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)
6. De Lathauwer, L.: Decompositions of a higher-order tensor in block terms-Part II: Definitions and uniqueness. *SIAM J. Matrix Anal. Appl.* **30**(3), 1033–1066 (2008)
7. 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**(3), 1067–1083 (2008)
8. Lathauwer, L.: Block component analysis, a new concept for blind source separation. In: Theis, F., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) *LVA/ICA 2012*. LNCS, vol. 7191, pp. 1–8. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28551-6_1
9. Yang, M., Kang, Z., Peng, C., Liu, W., Cheng, W.: On block term tensor decompositions and its applications in blind signal separation. http://archive.ymsc.tsinghua.edu.cn/pacm_paperurl/20160105102343471889031
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. Debals, O., Van Barel, M., De Lathauwer, L.: Löwner-based blind signal separation of rational functions with applications. *IEEE Trans. Signal Process.* **64**(8), 1909–1918 (2016)
12. Hunyadi, B., Camps, D., Sorber, L., Van Paesschen, W., De Vos, M., Van Huffel, S., De Lathauwer, L.: Block term decomposition for modelling epileptic seizures. *EURASIP J. Adv. Signal Process.* **2014**(1), 139 (2014)
13. Chatzichristos, C., Kofidis, E., Kopsinis, Y., Moreno, M.M., Theodoridis, S.: Higher-order block term decomposition for spatially folded fMRI data. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) *LVA/ICA 2017*. LNCS, vol. 10169, pp. 3–15. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_1
14. Chatzichristos, C., Kofidis, E., Theodoridis, S.: PARAFAC2 and its block term decomposition analog for blind fMRI source unmixing. In: *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2081–2085, August 2017
15. Vervliet, N., Debals, O., Sorber, L., Van Barel, M., De Lathauwer, L.: Tensorlab 3.0, March 2016. <https://www.tensorlab.net>
16. De Lathauwer, L., De Moor, B., Vandewalle, J.: On the best rank-1 and rank- (R_1, R_2, \dots, R_N) approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21**(4), 1324–1342 (2000)
17. Ishteva, M., Absil, P.-A., Van Huffel, S., De Lathauwer, L.: Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme. *SIAM J. Matrix Anal. Appl.* **32**(1), 115–135 (2011)
18. Savas, B., Lim, L.-H.: Quasi-Newton methods on grassmannians and multilinear approximations of tensors. *SIAM J. Sci. Comput.* **32**(6), 3352–3393 (2010)
19. Olikier, G., Absil, P.-A., De Lathauwer, L.: A variable projection method for block term decomposition of higher-order tensors. Accepted for ESANN 2018
20. Absil, P.-A., Mahony, R., Sepulchre, R.: *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton (2008)

ICA Methods



Accelerating Likelihood Optimization for ICA on Real Signals

Pierre Ablin¹(✉), Jean-François Cardoso², and Alexandre Gramfort¹

¹ Inria, Université Paris-Saclay, Palaiseau, France
pierreablin@gmail.com

² Institut d’Astrophysique de Paris/CNRS, Paris, France

Abstract. We study optimization methods for solving the maximum likelihood formulation of independent component analysis (ICA). We consider both the problem constrained to white signals and the unconstrained problem. The Hessian of the objective function is costly to compute, which renders Newton’s method impractical for large data sets. Many algorithms proposed in the literature can be rewritten as quasi-Newton methods, for which the Hessian approximation is cheap to compute. These algorithms are very fast on simulated data where the linear mixture assumption really holds. However, on real signals, we observe that their rate of convergence can be severely impaired. In this paper, we investigate the origins of this behavior, and show that the recently proposed Preconditioned ICA for Real Data (Picard) algorithm overcomes this issue on both constrained and unconstrained problems.

Keywords: Independent component analysis
Maximum likelihood estimation · Preconditioning · Optimization

1 Introduction

Linear Independent Component Analysis (ICA) [1] is an unsupervised data exploration technique, which models the set of observed signals as a linear instantaneous mixture of independent sources. Several methods have been proposed in the literature for recovering the sources and mixing matrix. When formulated as a maximum likelihood estimation task, ICA becomes an optimization problem where the negative log-likelihood has to be minimized. ICA may constitute a bottleneck in practical data processing pipelines, for example due to very long signals, high number of sources or bootstrapping techniques [2]. It is hence crucial to maximize the likelihood as quickly as possible.

Several approaches are found in the literature. Infomax [3] can be seen as a stochastic gradient descent [4]. Several second order methods have also been proposed. In [5], the author propose a quasi-Newton method dubbed “Fast Relative Newton” method, which we will refer to as “FR-Newton” in the following. In [6], a trust-region technique is used. AMICA [7] also uses a quasi-Newton approach. Although it is formulated as a fixed point algorithm, FastICA [8] is a

maximum likelihood estimator under whiteness constraint of the signals [9], and also behaves like a quasi-Newton method close to convergence [10].

The aforementioned algorithms all share the following property: the Hessian approximation that they use (implicitly or explicitly) stems from the expression that the true Hessian takes when the problem is solved, *i.e.* when the signals are truly independent. Unfortunately, in most practical cases, the assumption that the observed signals are a mixture of independent signals is false to some extent. There might be fewer/more sources than observed signals, the sources might not be i.i.d. or stationary, they might be partially correlated, or there might be some convolutive mixture.

In the following, we demonstrate that this can lead to large differences between the true Hessian and its approximations, often leading to slow convergence on real data. We then show that the recently proposed Preconditioned ICA for Real Data (Picard) algorithm [10, 11] overcomes this problem and is able to build a better Hessian approximation.

This article is organized as follows. In Sect. 2, we recall the maximum likelihood formulation of ICA, study the objective function, and derive a classical Hessian approximation. In Sect. 3, we give some classical results about quasi-Newton algorithms, and show how the convergence speed is linked with the distance between the true Hessian and the approximation. Section 4 contains a brief description of the Picard algorithm. Finally, we illustrate the previous result with experiments in Sect. 5. We show that Picard builds a much better Hessian approximation than those used in previous algorithms. Through extensive experiments, we show that this leads indeed to faster convergence.

Notation. The mean of a time-indexed sequence $x(t)_{t=1..T}$ is noted $\hat{E}[x(t)] \triangleq \frac{1}{T} \sum_{t=1}^T x(t)$, and its expectation is noted $\mathbb{E}[x]$. When M is a square $N \times N$ matrix, $\exp(M)$ denotes its matrix exponential, defined as $\exp(M) \triangleq \sum_{n=0}^{\infty} \frac{M^n}{n!}$. For two $N \times N$ matrices M and M' , we use the Frobenius scalar product: $\langle M|M' \rangle \triangleq \sum_{i,j} M_{ij}M'_{ij}$. We denote by $\|M\| \triangleq \sqrt{\langle M|M \rangle}$ the associated norm. For a fourth order tensor H of size $N \times N \times N \times N$, the scalar product with respect to H is defined as $\langle M|H|M' \rangle \triangleq \sum_{i,j,k,l} H_{ijkl}M_{ij}M'_{kl}$. The *spectrum* $\text{Sp}(B)$ of a linear symmetric operator B is the set of its eigenvalues. The Kronecker symbol δ_{ij} is equal to 1 when $i = j$ and to 0 otherwise.

2 Maximum-Likelihood ICA

In this section, we derive the maximum-likelihood formulation of ICA, and study the underlying objective function.

2.1 Objective Function

One observes N temporal signals $x_1(t), \dots, x_N(t)$ of T samples each. The signal matrix is $X = [x_1(t), \dots, x_N(t)]^\top \in \mathbb{R}^{N \times T}$.

For the rest of this article, we assume without loss of generality that X is white, *i.e.* the covariance $C \triangleq \frac{1}{T}XX^\top = I_N$. This can be enforced by a preprocessing whitening step: multiplying X by a square root inverse of C .

The linear ICA model considered here is the following [1]: there are N statistically independent and identically distributed signals, $s_1(t), \dots, s_N(t)$, which are noted as $S \in \mathbb{R}^{N \times T}$ in matrix form, and an invertible matrix $A \in \mathbb{R}^{N \times N}$ such that $X = AS$. The s_i are referred to as sources, and A is called the mixing matrix. The aim is to estimate A and S given X . In the following, p_i denotes the probability density function (p.d.f.) of the i -th source s_i .

The likelihood of A writes [12]:

$$p(X|A) = \prod_{t=1}^T \frac{1}{|\det(A)|} \prod_{i=1}^N p_i([A^{-1}X]_{it}). \quad (1)$$

It is more practical to work with the averaged negative log-likelihood, and the variable $W = A^{-1}$ called the *unmixing matrix*. In the following, $Y \triangleq WX$ denotes the current estimated sources. We define $\mathcal{L}(W) \triangleq -\frac{1}{T} \log(p(X|W^{-1}))$. It writes:

$$\mathcal{L}(W) = -\log|\det W| + \sum_{i=1}^N \hat{E}[-\log(p_i(Y_{it}))], \quad (2)$$

where \hat{E} denotes the time-averaging operation. FastICA attempts to minimize $\mathcal{L}(W)$ under whiteness constraint $WW^\top = I_N$.

2.2 Relative Gradient and Hessian

To study the variations of \mathcal{L} , it is convenient to work in a relative framework [13], where the gradient G and Hessian H are given by the Taylor expansion of $\mathcal{L}(\exp(\mathcal{E})W)$ where \mathcal{E} is a small $N \times N$ matrix. G and H are implicitly defined by the equation:

$$\mathcal{L}(\exp(\mathcal{E})W) = \mathcal{L}(W) + \langle G|\mathcal{E} \rangle + \frac{1}{2} \langle \mathcal{E}|H|\mathcal{E} \rangle + \mathcal{O}(\|\mathcal{E}\|^3). \quad (3)$$

G is a square $N \times N$ matrix, and H is a linear operator from matrices to matrices, which can be seen as a $N \times N \times N \times N$ tensor. In the following, $\psi_i \triangleq -\frac{p'_i}{p_i}$ is referred to as the *score function*. Simple computations yield (see [10] for details):

$$G(W)_{ij} = \hat{E}[\psi_i(y_i)y_j] - \delta_{ij} \text{ for } 1 \leq i, j \leq N \quad (4)$$

$$H(W)_{ijkl} = \delta_{il}\delta_{jk}\hat{E}[\psi_i(y_i)y_i] + \delta_{ik}\hat{E}[\psi'_i(y_i)y_jy_l] \text{ for } 1 \leq i, j, k, l \leq N \quad (5)$$

The Hessian is sparse since it has of the order of N^3 non-zero coefficients. Still, its evaluation requires computing $O(N^3)$ sample averages $\hat{E}[\psi'_i(y_i)y_jy_l]$, making the standard Newton's method impractical for large data sets.

Algorithm 1. Quasi-Newton method for likelihood optimization

input : Set of white mixed signals X , boolean “whiteness constraint”
Set $W = I_N$;
Set $Y = X$;
repeat
 Compute the gradient G using (4);
 if *whiteness constraint* **then**
 | Project G on the antisymmetric matrices: $G \leftarrow \frac{1}{2}(G - G^\top)$;
 end
 Compute a Hessian approximation \hat{H} ;
 Compute the search direction $D = -\hat{H}^{-1}G$;
 if *whiteness constraint* **then**
 | Project D on the antisymmetric matrices: $D \leftarrow \frac{1}{2}(D - D^\top)$;
 end
 Compute the step size $\alpha = \arg \min_{\alpha} \mathcal{L}(\exp(\alpha D)W)$ using line-search ;
 Set $W \leftarrow \exp(\alpha D)W$;
 Set $Y = WX$;
output: Unmixing matrix W , unmixed signals Y .

2.3 The Hessian Approximation

If the signals $(y_1(t), \dots, y_N(t))$ are independent, then $\mathbb{E}[\psi'_i(y_i)y_j y_l] = \delta_{jil}\mathbb{E}[\psi'_i(y_i)y_j^2]$. A natural approximation of H is then:

$$\tilde{H}(W)_{ijkl} = \delta_{il}\delta_{jk}\hat{E}[\psi'_i(y_i)y_l] + \delta_{ik}\delta_{jl}\hat{E}[\psi'_i(y_i)y_j^2]. \quad (6)$$

This approximation matches the true Hessian **if the number of samples goes to infinity and the (y_i) are independent**. If the linear ICA model holds, i.e. if there exists independent signals S and a mixing matrix A such that $X = AS$, then, for $W^* = A^{-1}$, $\tilde{H}(W^*) = H(W^*) + \mathcal{O}(\frac{1}{\sqrt{T}})$. As the number of samples is generally large, the approximation is very good in that case.

However, in a practical case, ICA is performed on real data for which the ICA model does not hold exactly. In that case, even for $W^* = \arg \min \mathcal{L}(W)$, one does not necessarily have $\mathbb{E}[\psi'_i(y_i)y_j y_l] = \delta_{jil}\mathbb{E}[\psi'_i(y_i)y_j^2]$, and $\tilde{H}(W^*)$ may be quite far from $H(W^*)$.

3 Speed of Convergence of Quasi-Newton Methods

In the following, we consider a general relative quasi-Newton method to minimize \mathcal{L} , described in Algorithm 1. It takes as input the set of mixed signals X , which are assumed white for simplicity, and a boolean “whiteness constraint” which determines if the algorithm works under whiteness constraint. Note that the policy to compute the approximation \hat{H} is not specified: one could use $\hat{H} = \tilde{H}$, but other choices are possible. To keep the analysis simple, we assume that the line-search is perfect, i.e. that the objective function is always minimized in the search direction.

3.1 Theoretical Results

Let us recall some results on the convergence speed of such method. These results mostly come from Numerical Optimization [14], Chap. 3.3.

First, the following theorem shows that under mild assumptions, the sequence of unmixing matrices produced by Algorithm 1 converges to a local minimum of \mathcal{L} .

Theorem 1. *Assume that the sequence of Hessian approximations \hat{H} used in Algorithm 1 is positive definite, of spectrum lower bounded by some constant $\lambda_{min} > 0$. Then, the sequence of unmixing matrices generated by the algorithm converges towards a matrix W^* such that $G(W^*) = 0$ and $H(W^*)$ is positive definite.*

This theorem is a direct consequence of Zoutendijk's result (see [14], Theorem 3.2). Interestingly, it implies that the algorithm cannot converge to a saddle point (where $H(W^*)$ is not positive), but only towards local minima, as guaranteed for gradient based methods.

Quasi-Newton methods typically aim at finding a direction close to Newton's direction $-H^{-1}G$, and ideally have the same quadratic convergence rate. By Theorem 3.6 in [14], this happens if and only if at convergence, the Hessian approximation matches the true Hessian in the search direction. As we have seen before, even when the ICA model holds, the simple approximation \tilde{H} only matches asymptotically the true Hessian, meaning that the above theorem never practically applies. Thus, the convergence of Algorithm 1 can only be linear. The following algorithm gives the rate of convergence.

Theorem 2. *Assume that the condition of Theorem 1 holds. Assume that the sequence of approximate Hessians \hat{H} converges towards \hat{H}^* . Let λ_m (resp. λ_M) be the smallest (resp. largest) eigenvalue of $\hat{H}^{*-1/2} H \hat{H}^{*-1/2}$ and define the condition number:*

$$\kappa \triangleq \frac{\lambda_M}{\lambda_m}. \quad (7)$$

Then, for all $r < \frac{1}{\kappa}$ and n large enough, the sequence W_n of unmixing matrices produced by Algorithm 1 satisfies $\mathcal{L}(W_{n+1}) - \mathcal{L}(W^) \leq (1-r)[\mathcal{L}(W_n) - \mathcal{L}(W^*)]$.*

We now give a brief sketch of proof.

Proof. For simplicity, the proof is made in a non-relative framework, where the update rule is $W_{n+1} = W_n - \alpha \hat{H}_n^{-1} \nabla \mathcal{L}(W_n)$. First, we make the useful change of variable $U_n = \hat{H}^{*1/2} W_n$, and define the new objective function $L(U_n) = \mathcal{L}(\hat{H}^{*-1/2} U_n)$. Simple computations show that U_n verifies $U_{n+1} = U_n - \alpha B_n \nabla L(U_n)$, where $B_n \triangleq \hat{H}^{*1/2} \hat{H}_n^{-1} \hat{H}^{*1/2}$. This sequence tends towards identity, meaning that the behavior of U_n is asymptotically the same as a gradient descent. One has $\nabla^2 L(U) = \hat{H}^{*-1/2} [\nabla^2 \mathcal{L}(W)] \hat{H}^{*-1/2}$.

Let $\varepsilon > 0$ be a small number. Since $\text{Sp}(B_n) \rightarrow \{1\}$ and $\text{Sp}(\nabla^2 L(U_n)) \subset [\lambda_m, \lambda_M]$ as n goes to infinity, for n large enough we have that $\text{Sp}(B_n) \subset [1 - \varepsilon, 1 + \varepsilon]$ and $\text{Sp}(\nabla^2 L(U_n)) \subset [(1 - \varepsilon)\lambda_m, (1 + \varepsilon)\lambda_M]$. This means that the iterates U_n are in a set where L is $(1 + \varepsilon)\lambda_M$ -smooth and $(1 - \varepsilon)\lambda_m$ -strongly convex. The smoothness implies the following convexity inequality:

$$L(V) \leq L(U) + \langle \nabla L(U) | V - U \rangle + \frac{(1 + \varepsilon)\lambda_M}{2} \|U - V\|^2 \quad (8)$$

and the strong convexity enforces the Polyak-Lojasiewicz conditions [15]:

$$\frac{1}{2} \|\nabla f(U)\|^2 \geq (1 - \varepsilon)\lambda_m [L(U) - L(U^*)] \quad (9)$$

Let β be a positive scalar. For an exact line-search, we have $L(U_{n+1}) \leq L(U_n - \beta B_n \nabla L(U_n))$. Using $U = U_n$ and $V = U_n - \beta B_n \nabla L(U_n)$ in inequality (8), we obtain:

$$L(U_{n+1}) - L(U_n) \leq -\beta \langle \nabla L(U_n) | B_n \nabla L(U_n) \rangle + \beta^2 \frac{(1 + \varepsilon)\lambda_M}{2} \|B_n \nabla L(U_n)\|^2 \quad (10)$$

The condition on the spectrum of B_n implies $\langle \nabla L(U_n) | B_n \nabla L(U_n) \rangle \geq (1 - \varepsilon) \|\nabla L(U_n)\|^2$ and $\|B_n \nabla L(U_n)\|^2 \leq (1 + \varepsilon)^2 \|\nabla L(U_n)\|^2$. Replacing in Eq. (10) yields:

$$L(U_{n+1}) - L(U_n) \leq \left(-\beta(1 - \varepsilon) + \beta^2 \frac{(1 + \varepsilon)^3 \lambda_M}{2} \right) \|\nabla L(U_n)\|^2 \quad (11)$$

This holds for any β , in particular for $\beta = \frac{1 - \varepsilon}{(1 + \varepsilon)^3 \lambda_M}$ (which minimizes the scalar factor in front of $\|\nabla L(U_n)\|^2$). We obtain:

$$L(U_{n+1}) - L(U_n) \leq -\frac{(1 - \varepsilon)^2}{2(1 + \varepsilon)^3 \lambda_M} \|\nabla L(U_n)\|^2 \quad (12)$$

Using Eq. (9) then gives:

$$L(U_{n+1}) - L(U_n) \leq -\frac{(1 - \varepsilon)^3 \lambda_m}{(1 + \varepsilon)^3 \lambda_M} [L(U_n) - L(U^*)] \quad (13)$$

Rearranging the terms, we obtain the desired result for $r = \left(\frac{1 - \varepsilon}{1 + \varepsilon}\right)^3 \frac{1}{\kappa}$.

3.2 Link with Maximum Likelihood ICA

There are many ICA algorithms closely related to the minimization of \mathcal{L} and similar to Algorithm 1. For instance, Infomax is a stochastic version of Algorithm 1 without whiteness constraint and with $\hat{H} = Id$. In [5], the author proposes to use $\hat{H} = \tilde{H}$ in Algorithm 1, without the whiteness constraint. The algorithm is denoted as ‘‘Fast Relative Newton method’’, or FR-Newton for short. The same approach is used in AMICA [7]. In [10], it is shown that close to convergence,

FastICA’s iterations are similar to those of Algorithm 1 with the whiteness constraint, and where the Hessian approximation has the same properties as \tilde{H} : it coincides asymptotically with H when the underlying signals (y_i) are independent, but may differ otherwise. Thus, the previous results apply for a wide range of popular ICA methods.

4 Preconditioned ICA for Real Data

Let us now introduce the Preconditioned ICA for Real Data (Picard) algorithm, which finds a better Hessian approximation than \tilde{H} . The algorithm is an adaptation of the L-BFGS algorithm [16]. It has a memory of size m which stores the m previous iterates W and gradients G . From these values, it recursively builds a Hessian approximation starting from \tilde{H} . In the following, H_P denotes that approximation. It does so in an uninformed fashion, without any prior on the local geometry. L-BFGS has been shown to perform well on a wide variety of problems. Here, we have the advantage of having \tilde{H} as a good initialization for the approximate Hessian. Another asset of this method is that the Hessian approximation never has to be computed, because there is an efficient way of computing the direction $-H_P^{-1}G$. Picard can handle both constrained and unconstrained problems. For further details for the practical implementation, see [10, 11].

Python and Matlab/Octave code for Picard is available online.¹

5 Experiments

5.1 Comparison of the Condition Numbers

In this section, we show how close the Hessian approximations \tilde{H} and H_P are to H on simulated and real data. We consider two different datasets X of $N = 8$

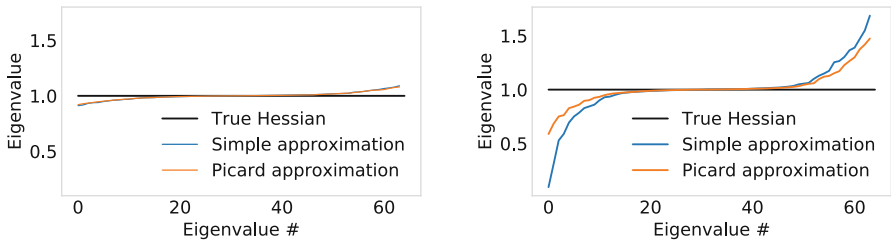


Fig. 1. A measure of the closeness of the approximate Hessians to the true Hessian at the maximum likelihood: sorted spectrum of $\hat{H}^{-\frac{1}{2}}H\hat{H}^{-\frac{1}{2}}$. Left: simulated data where the ICA model holds. Right: real data. On the simulated data, we find $\kappa = 1.2$ for both $\hat{H} = \tilde{H}$ and $\hat{H} = H_P$. For that example on real data, we find $\kappa = 29$ for \tilde{H} and a significantly smaller $\kappa = 2.6$ for H_P .

¹ <https://github.com/pierreablin/picard>.

signals of length $T = 20000$. The first one is obtained by simulating a source matrix S of independent signals, and a random mixing matrix A . We take $X = AS$. For that dataset, the linear ICA model holds by construction. The second one is obtained by extracting 20000 square patches of size $(8, 8)$ from a natural image. PCA is then applied to reduce to 8 the number of signals.

First, we find a local minimum W^* of $\mathcal{L}(W)$ by running one of the algorithms on this dataset. Then, the simple approximation $\hat{H}(W^*)$, the Picard approximation $H_P(W^*)$ and the true Hessian $H(W^*)$ are computed. As explained by

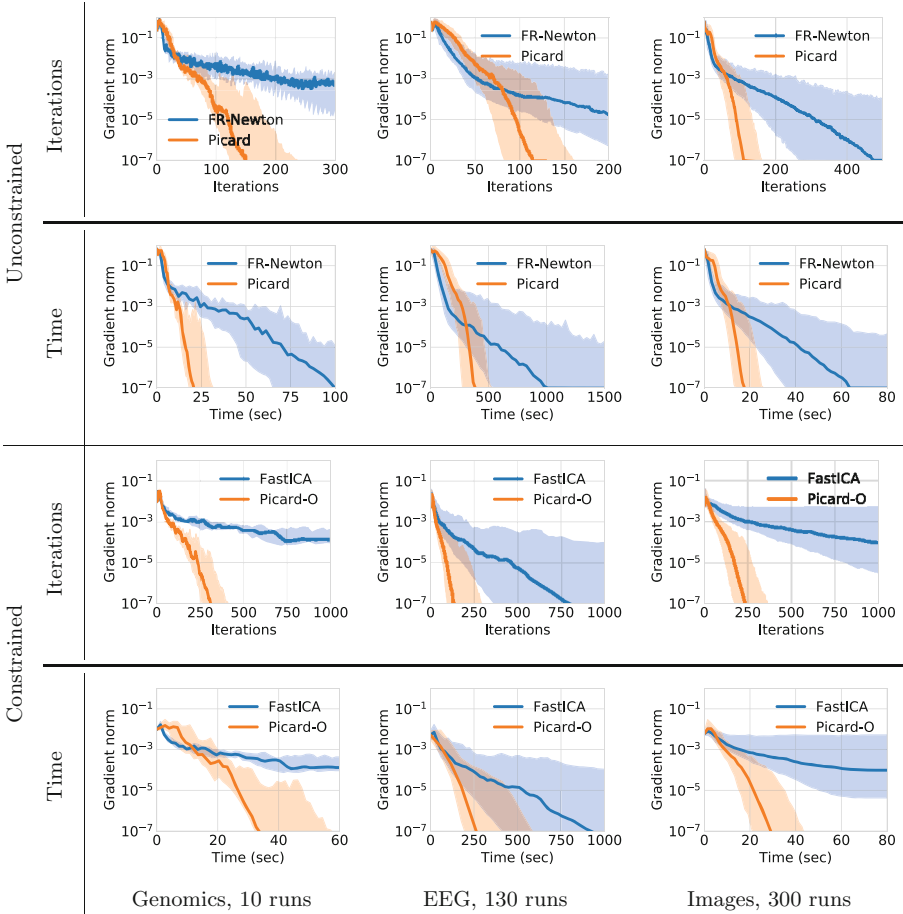


Fig. 2. Convergence speed of several ICA algorithms on 3 real data sets. Each column corresponds to a type of data. The first two rows correspond to the unconstrained algorithms, the last two to the constrained algorithms. The first row of each pair displays the evolution of gradient across iterations, the second one displays the evolution of gradient against time. Bold lines correspond to the medians of the gradient norms, and the shading displays the 10–90%.

theorem 2, what drives the convergence speed of the algorithms is the spectrum of $\hat{H}^{-\frac{1}{2}}H\hat{H}^{-\frac{1}{2}}$ where \hat{H} is the approximation. Figure 1 displays these spectrum for the two datasets.

We observe that H_P and \tilde{H} are very similar on the simulated dataset, and that the resulting condition numbers are close to 1, which explains the fast convergence of the two algorithms. On the real dataset, the results are different: the spectrum obtained with H_P is flatter than the one obtained with \tilde{H} , which means that Picard builds a Hessian approximation which is significantly better than \tilde{H} .

5.2 Convergence Speed on Real Datasets

We now compare the convergence speed of Picard/Picard-O with FR-Newton from [5] and FastICA [9] on three types of data on which ICA is widely used.

The first is a cancer genomics dataset generated by the TCGA Research Network: <http://cancergenome.nih.gov>, of initial size $N \simeq 2000$ and $T \simeq 20000$ for which the dimension has been reduced to $N = 60$ by PCA. The second consists of 13 EEG recordings datasets [17] of size $N = 71$ and $T \simeq 300000$. The last one is 30 datasets of $T = 20000$ extracted image patches of size (8, 8), flattened to obtain $N = 64$ signals. We run the aforementioned algorithms 10 times on each datasets. We keep track of the evolution of the gradient norm across iterations and time. Figure 2 displays the median and 10–90% of the trajectories.

As expected regarding the previous results on the Hessian spectrum, Picard and Picard-O converge faster than their counterparts relying purely on \tilde{H} as Hessian approximation.

6 Conclusion

This article considers quasi-Newton methods for maximum likelihood ICA using approximated Hessian matrices. We argue that while the standard Hessian approximation works very well on simulated data, it differs a lot from the true Hessian on most applied problems. As a consequence, quasi-Newton algorithms which model the curvature of the objective function with such an approximation can have poor convergence rates. We advocate the L-BFGS method to refine ‘on the fly’ the approximation of the Hessian. This is supported by experiments on 3 types of real signals which clearly demonstrate that this approach leads to faster convergence.

References

1. Comon, P.: Independent component analysis, a new concept? *Sig. Process.* **36**(3), 287–314 (1994)
2. Himberg, J., Hyvärinen, A., Esposito, F.: Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* **22**(3), 1214–1222 (2004)

3. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**(6), 1129–1159 (1995)
4. Cardoso, J.-F.: Infomax and maximum likelihood for blind source separation. *IEEE Sig. Process. Lett.* **4**(4), 112–114 (1997)
5. Zibulevsky, M.: Blind source separation with relative Newton method. In: *Proceedings of the ICA*, vol. 2003, pp. 897–902 (2003)
6. Choi, H., Choi, S.: A relative trust-region algorithm for independent component analysis. *Neurocomputing* **70**(7), 1502–1510 (2007)
7. Palmer, J.A., Kreutz-Delgado, K., Makeig, S.: AMICA: an adaptive mixture of independent component analyzers with shared components. Technical report (2012)
8. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**(3), 626–634 (1999)
9. Hyvärinen, A.: The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Process. Lett.* **10**(1), 1–5 (1999)
10. Ablin, P., Cardoso, J.-F., Gramfort, A.: Faster ICA under orthogonal constraint. In: *Proceedings of the IEEE ICASSP* (2018)
11. Ablin, P., Cardoso, J.-F., Gramfort, A.: Faster independent component analysis by preconditioning with hessian approximations, Arxiv Preprint (2017)
12. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Sig. Process.* **45**(7), 1712–1725 (1997)
13. Cardoso, J.-F., Laheld, B.H.: Equivariant adaptive source separation. *IEEE Trans. Sig. Process.* **44**(12), 3017–3030 (1996)
14. Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer, New York (1999)
15. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In: Frascioni, P., Landwehr, N., Manco, G., Vreeken, J. (eds.) *ECML PKDD 2016. LNCS (LNAI)*, vol. 9851, pp. 795–811. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46128-1_50
16. Nocedal, J.: Updating Quasi-Newton matrices with limited storage. *Math. Comput.* **35**(151), 773–782 (1980)
17. Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S.: Independent EEG sources are dipolar. *PLoS ONE* **7**(2), e30135 (2012)



Orthogonally-Constrained Extraction of Independent Non-Gaussian Component from Non-Gaussian Background Without ICA

Zbyněk Koldovský¹(✉), Petr Tichavský², and Nobutaka Ono³

¹ Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic

`zbynek.koldovsky@tul.cz`

² The Czech Academy of Sciences, Institute of Information Theory and Automation,
Pod vodárenskou věží 4, P.O. Box 18, 182 08 Praha 8, Czech Republic

`tichavsk@utia.cas.cz`

³ Faculty of System Design, Tokyo Metropolitan University, 6-6 Asahigaoka,
Hino-shi, Tokyo 191-0065, Japan

`onono@tmu.ac.jp`

Abstract. We propose a new algorithm for Independent Component Extraction that extracts one non-Gaussian component and is capable to exploit the non-Gaussianity of background signals without decomposing them into independent components. The algorithm is suitable for situations when the signal to be extracted is determined through initialization; it shows an extra stable convergence when the target component is dominant. In simulations, the proposed method is compared with Natural Gradient and One-unit FastICA, and it yields improved results in terms of the Signal-to-Interference ratio and the number of successful extractions.

Keywords: Independent Component Analysis
Blind source separation · Non-Gaussian distribution · Score function
Independent Vector Analysis

1 Introduction

The Blind Source Extraction (BSE) problem where the goal is to extract one particular component from a linear mixture

$$\mathbf{x} = \mathbf{A}\mathbf{u}, \tag{1}$$

has been a live topic for decades, also before the birth of Independent Component Analysis (ICA) [3, 4, 7]. In the mixture, \mathbf{u} and \mathbf{x} are $d \times 1$ vectors, respectively,

This work was supported by The Czech Science Foundation through Project No. 17-00902S and partially supported by JSPS KAKENHI Grant Number 16H01735.

of d original and mixed signals, and \mathbf{A} is a $d \times d$ non-singular mixing matrix. The components of \mathbf{u} are assumed to be *mutually independent*. Let, without any loss of generality, the desired component be u_1 , which will be referred to as SOI (the source of interest); the other signals will be briefly called *background*.

By information theory, it is possible to extract an independent component through finding a direction having minimum entropy (maximum non-Gaussianity). However, methods extracting one non-Gaussian independent component in this way (from here referred to as “one-unit” methods) are known to have a limited asymptotic accuracy compared to methods performing the whole ICA decomposition of (1). Performance analyses of several one-unit methods showed that they perform as if background components were all Gaussian [5, 12, 14].

Specifically, let \mathbf{W} be an unbiased estimate of \mathbf{A}^{-1} (a de-mixing matrix) up to the order and scales of its rows, and $\mathbf{G} = \mathbf{W}\mathbf{A} \approx \mathbf{P}\mathbf{A}$, where \mathbf{P} and \mathbf{A} is, respectively, a permutation and a diagonal matrix. The Cramér-Rao bound (CRLB) for ICA says that [14, 15]

$$\mathbb{E}[G_{ij}^2] \geq \frac{1}{N} \frac{\kappa_j}{\kappa_i \kappa_j - 1}, \quad i \neq j, \quad (2)$$

where $\mathbb{E}[\cdot]$ stands for the expectation operator, N is the number of samples of \mathbf{x} (assuming identically and independently distributed samples), and $\kappa_i = \mathbb{E}[\psi_i^2]$ where $\psi_i(x) = -\partial/\partial x \log p_i(x)$, which is the score function of p_i where p_i is the pdf of the i th original signal u_i .

For normalized variables with unit variance it holds that $\kappa_i \geq 1$ where $\kappa_i = 1$ if and only if the i th pdf is Gaussian. Let \mathbf{w} be the first row of \mathbf{W} corresponding to the extracted SOI, and let \mathbf{u} have all unit variance. The asymptotic accuracy (for $N \rightarrow +\infty$) of one-unit methods (when the true score function is used in the algorithm’s contrast function) was shown to be characterized by [5, 12, 14]

$$\mathbb{E}[g_j^2] \approx \frac{1}{N} \frac{1}{\kappa_1 - 1}, \quad j \neq 1, \quad (3)$$

where $\mathbf{g} = \mathbf{w}^T \mathbf{A}$. The right-hand side coincides with the CRLB in (2) for $i = 1$ when $\kappa_j = 1$ for $j = 2, \dots, d$, which is the case when u_2, \dots, u_d are Gaussian (for which case the CRLB (2) formally does not exist unless $d = 2$).

Recently, we have revised the BSE problem through Independent Component Extraction (ICE) [10, 11]. Here, the mixing model (1) is re-parameterized for the extraction of the SOI in the way that the rest of the mixture is not object of any particular decomposition, as compared to ICA. In the statistical model, s is assumed to be non-Gaussian while the other components are assumed to be Gaussian. Under these conditions, the CRLB for ICE has been confirmed to correspond to the right-hand side of (3); see [8]. In [10], orthogonally-constrained gradient learning algorithms for ICE have been proposed based on the maximum likelihood principle.¹ An appealing property of these algorithms resides in their

¹ A particular variant of these algorithms (OGICE_w) coincides with a method proposed earlier by Pham in [12], which was derived based on a simplified form of mutual information that is valid for Gaussian background.

ability to keep converging to the desired source, e.g., to a dominant SOI. Using methods that guarantee the extraction of the SOI with a high probability, the complete ICA decomposition and the subsequent component selection due to the random order can be avoided, which brings significant computational savings.

In this paper, our goal is to overcome the accuracy limitation given by (3). We derive a new gradient ICE algorithm using the maximum likelihood approach. The method takes into account possible non-Gaussianity of background. For simplicity, real-valued mixing scenario and signals will be considered, although a complex-valued extension is possible.

The rest of this paper is organized as follows. The ICE mixing model and the statistical model of signals are described in Sect. 2. In Sect. 3, the novel algorithm is proposed and described in details. Section 4 is devoted to simulations and comparisons, and Sect. 5 concludes the paper.

Notation: Plain letters denote scalars; bold letters denote vectors; bold capital letters denote matrices. The Matlab convention for matrix/vector concatenation and indexing will be used, e.g., $[1; \mathbf{g}] = [1, \mathbf{g}^T]^T$, and $(\mathbf{A})_{j,:}$ is the j th row of \mathbf{A} . Symbolic scalar and vector random variables will be denoted by lower case letters, e.g. s and \mathbf{x} , \mathbf{z} , while the quantities collecting their N samples will be denoted by bold (capital) letters, e.g. \mathbf{s} (a row vector $1 \times N$) and \mathbf{X} , \mathbf{Z} . Estimated values of signals will be denoted by hat, e.g., \hat{s} , $\hat{\mathbf{Z}}$.

2 Problem Formulation

2.1 Algebraic Mixing Model

Let the SOI be $s = u_1$ and \mathbf{a} be the first column of \mathbf{A} , so \mathbf{A} can be partitioned as $\mathbf{A} = [\mathbf{a}, \mathbf{A}_2]$. Then, \mathbf{x} can be written as

$$\mathbf{x} = \mathbf{a}s + \mathbf{y}, \quad (4)$$

where $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ and $\mathbf{u}_2 = [u_2, \dots, u_d]^T$. The fact that $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ means that the mixture consists of the same number of sources as that of input channels.

Let the new parameterization of the mixing matrix and of its inverse matrix be denoted by \mathbf{A}_{ICE} and \mathbf{W}_{ICE} , respectively. In ICE, the identification of \mathbf{A}_2 or the decomposition of \mathbf{y} into independent signals \mathbf{u}_2 is *not* the goal. Therefore, we assume that $\mathbf{A}_{\text{ICE}} = [\mathbf{a}, \mathbf{Q}]$ where \mathbf{Q} is, for now, arbitrary with full column-rank. Then, (4) can be written as

$$\mathbf{x} = \mathbf{A}_{\text{ICE}}\mathbf{v}, \quad (5)$$

where $\mathbf{v} = [s; \mathbf{z}]$, and $\mathbf{y} = \mathbf{Q}\mathbf{z}$. Hence, the subspace spanned by \mathbf{z} is the same as that of \mathbf{u}_2 . To complete the mixing matrix definition, we look at the inverse matrix $\mathbf{W}_{\text{ICE}} = \mathbf{A}_{\text{ICE}}^{-1}$.

Let \mathbf{a} and \mathbf{W}_{ICE} be partitioned, respectively, as $\mathbf{a} = [\gamma; \mathbf{g}]$ and $\mathbf{W}_{\text{ICE}} = [\mathbf{w}^T; \mathbf{B}]$. \mathbf{B} is required to be orthogonal to \mathbf{a} , i.e. $\mathbf{B}\mathbf{a} = \mathbf{0}$, which ensures that $\mathbf{B}\mathbf{x}$ do not contain any contribution of s . A useful selection is $\mathbf{B} = [\mathbf{g}, -\gamma\mathbf{I}_{d-1}]$

where \mathbf{I}_d denotes the $d \times d$ identity matrix. Let \mathbf{w} be partitioned as $\mathbf{w} = [\beta; \mathbf{h}]$. Then,

$$\mathbf{W}_{\text{ICE}} = \begin{pmatrix} \mathbf{w}^T \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} \beta & \mathbf{h}^T \\ \mathbf{g} & -\gamma \mathbf{I}_{d-1} \end{pmatrix}, \quad (6)$$

and from $\mathbf{A}_{\text{ICE}} \cdot \mathbf{W}_{\text{ICE}} = \mathbf{I}_d$ it follows that

$$\mathbf{A}_{\text{ICE}} = [\mathbf{a}, \mathbf{Q}] = \begin{pmatrix} \gamma & \mathbf{h}^T \\ \mathbf{g} & (\mathbf{g}\mathbf{h}^T - \mathbf{I}_{d-1})\gamma^{-1} \end{pmatrix}, \quad (7)$$

where β and γ are linked through

$$\beta\gamma = 1 - \mathbf{h}^T \mathbf{g}. \quad (8)$$

The latter equation can be also written in the form $\mathbf{w}^T \mathbf{a} = 1$, which corresponds to the *distortionless response* constraint [16]. The role of \mathbf{a} , as follows from (4), is the *mixing vector* related to s , while \mathbf{w} is the *separating vector* as $s = \mathbf{w}^T \mathbf{x}$. For the background signal \mathbf{z} , it holds that $\mathbf{z} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{y} = \mathbf{B}\mathbf{A}_2 \mathbf{u}_2$.

Similarly to the indeterminacies in ICA, the scales of s and of \mathbf{a} are ambiguous in the sense that they can be replaced, respectively, by αs and $\alpha^{-1} \mathbf{a}$ where $\alpha \neq 0$. The scaling ambiguity can be avoided by fixing β or γ . Next, the role of $s = u_1$ can be interchanged with u_i , for any $i = 2, \dots, d$. This is the permutation problem [13].

In this paper, we assume that an initial guess of \mathbf{a} or of \mathbf{w} is given, which determines the SOI. The initial value is typically deviated by an estimation error, which increases the probability that the given algorithm finally extracts a different source than the SOI. In experiments (Sect. 4), we therefore conduct a sensitivity analysis, which compares the size of the attraction area of different BSE algorithms.

2.2 Statistical Model

The main principle of ICE is the same as that of ICA. We take the assumption that s and \mathbf{z} are *independent*, so the goal is to find \mathbf{a} and \mathbf{w} such that $\mathbf{w}^T \mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are independent (or as independent as possible).

Let the pdf of s and of \mathbf{z} be, respectively, $p_s(s)$ and $p_{\mathbf{z}}(\mathbf{z})$. The joint pdf of the mixed signals $\mathbf{x} = \mathbf{A}_{\text{ICE}} \mathbf{v}$ is

$$p_{\mathbf{x}}(\mathbf{x}) = p_s(\mathbf{w}^T \mathbf{x}) \cdot p_{\mathbf{z}}(\mathbf{B}\mathbf{x}) \cdot |\det \mathbf{W}_{\text{ICE}}| \quad (9)$$

where it can be shown that

$$\det \mathbf{W}_{\text{ICE}} = (-1)^{d-1} \gamma^{d-2} = (-1)^{d-1} \beta^{-(d-2)} (1 - \mathbf{h}^T \mathbf{g})^{d-2}. \quad (10)$$

Since the background signals remain unmixed after ICE (up to special cases such as $d = 2$), we proposed in [10, 11] to model the unknown $p_{\mathbf{z}}$ as Gaussian with zero mean and covariance $\mathbf{C}_{\mathbf{z}}$. In this paper, we generalize the background model to arbitrary (non-)Gaussian pdf. Thus, the unknown densities $p_s(s)$ and

$p_{\mathbf{z}}(\mathbf{z})$ are replaced, respectively, by model densities $f(s)$ and $q(\mathbf{z})$. The quasi-loglikelihood function for N i.i.d. signal samples, according to (9), takes the form

$$\mathcal{L}(\mathbf{a}, \mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \left\{ \log f(\mathbf{w}^T \mathbf{x}(n)) + \log q(\mathbf{B}\mathbf{x}(n)) \right\} + (d-2) \log |\gamma|. \quad (11)$$

Orthogonal Constraint. The first term on the right-hand side of (11) depends purely on \mathbf{w} , while the second and the third terms depend purely on \mathbf{a} . The only link between \mathbf{a} and \mathbf{w} thus resides in (8). Therefore, the likelihood function can have spurious maxima where \mathbf{a} and \mathbf{w} do not correspond to the same source.

To make the interconnection between \mathbf{a} and \mathbf{w} tighter, the orthogonal constraint (OG) can be imposed [2]. Let \mathbf{W}_{ICE} be a current ICE de-mixing matrix estimate having the structure of (6), and $\widehat{\mathbf{V}} = [\widehat{\mathbf{s}}; \widehat{\mathbf{Z}}] = \mathbf{W}_{\text{ICE}} \mathbf{X}$ be the estimated matrix of de-mixed signal samples. The OG reads

$$\frac{1}{N} \widehat{\mathbf{s}} \cdot \widehat{\mathbf{Z}}^T = \frac{1}{N} \mathbf{w}^T \mathbf{X} \mathbf{X}^T \mathbf{B}^T = \mathbf{w}^T \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{B}^T = \mathbf{0}, \quad (12)$$

where $\widehat{\mathbf{C}}_{\mathbf{x}} = \mathbf{X} \mathbf{X}^T / N$ is the sample-based estimate of $\mathbf{C}_{\mathbf{x}} = \mathbf{E}[\mathbf{x} \mathbf{x}^T]$. The reader can verify that the OG together with (8) introduce the following links between \mathbf{a} and \mathbf{w} :

$$\mathbf{a} = \frac{\widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}}{\mathbf{w}^T \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}}, \quad (13a)$$

$$\mathbf{w} = \frac{\widehat{\mathbf{C}}_{\mathbf{x}}^{-1} \mathbf{a}}{\mathbf{a}^T \widehat{\mathbf{C}}_{\mathbf{x}}^{-1} \mathbf{a}}. \quad (13b)$$

In this paper, we will consider the former coupling, that is, \mathbf{w} will be the free variable while \mathbf{a} will be treated as dependent.

3 Gradient-Based Algorithm

3.1 Gradient of the Contrast Function

The gradient of \mathcal{L} with respect to \mathbf{w} under the coupling (13a), is

$$\left. \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \right|_{\text{w.r.t. (13a)}} = -\frac{1}{N} \mathbf{X} \widehat{\boldsymbol{\phi}}^T + \frac{1}{N} \frac{\widehat{\mathbf{C}}_{\mathbf{x}}}{\mathbf{w}^T \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}} \left(\begin{array}{c} \text{tr}(\mathbf{E} \mathbf{X} \widehat{\boldsymbol{\Psi}}) + (d-2) N \gamma^{-1} \\ -\widehat{\boldsymbol{\Psi}} \mathbf{X}^T \mathbf{e}_1 \end{array} \right) + 2\mathbf{a} \left(\frac{1}{N} \text{tr}(\widehat{\boldsymbol{\Psi}}^T \widehat{\mathbf{Z}}) - (d-2) \right), \quad (14)$$

where $\text{tr}(\cdot)$ denotes the trace, $\mathbf{E} = [\mathbf{0}, \mathbf{I}_{d-1}]$, \mathbf{e}_1 denotes the first column of \mathbf{I}_d , $\widehat{\boldsymbol{\phi}} = \phi(\widehat{\mathbf{s}})$, and $\widehat{\boldsymbol{\Psi}} = \psi(\widehat{\mathbf{Z}})$, where

$$\phi(\xi) = -\frac{\partial \log f(\xi)}{\partial \xi} \quad \text{and} \quad \psi_i(\mathbf{z}) = -\frac{\partial \log q(\mathbf{z})}{\partial z_i}, \quad \psi(\mathbf{z}) = [\psi_1(\mathbf{z}), \dots, \psi_{d-1}(\mathbf{z})]^T, \quad (15)$$

are the score function of the model pdfs $f(\cdot)$ and $q(\cdot)$, respectively, which are applied element/column-wise in case of the vector/matrix argument. We skip details of the lengthy computation of (14) here due to the lack of space.

By exploring this gradient when $N \rightarrow +\infty$ and when \mathbf{w} is the ideal separating vector, that is, when $\mathbf{w}^T \mathbf{x} = s$ and $\mathbf{B}\mathbf{x} = \mathbf{z}$, an important fact can be shown: The ideal separating vector is a stationary point of the contrast function (the gradient is zero) only if ϕ and ψ satisfy $\mathbb{E}[s\phi(s)] = 1$ and $\mathbb{E}[\mathbf{z}\psi(\mathbf{z})^T] = \mathbf{I}_{d-1}$, respectively. Both conditions are automatically satisfied when ϕ and ψ are the true score functions of the respective variables. However, since these are not known in the blind scenario, we introduce the following normalizing conditions: For any estimates of \mathbf{a} and \mathbf{w} , let

$$\widehat{\mathbf{s}}\widehat{\phi}^T = N \quad \text{and} \quad \widehat{\mathbf{Z}}\widehat{\psi}^T = N\mathbf{I}_{d-1}. \quad (16)$$

With these conditions and after few computations, (14) simplifies to

$$\left. \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \right|_{\text{w.r.t. (13a)}} = \mathbf{a} - \frac{1}{N} \mathbf{X}\widehat{\phi}^T + \frac{1}{\mathbf{w}^T \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}} \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{B}^T \mathbf{p}, \quad (17)$$

where $\mathbf{p} = \widehat{\psi}\widehat{\mathbf{s}}^T/N$.

A practical way to select ϕ and ψ meeting the conditions in (16) is by taking some appropriate prototype functions ϕ_1 and ψ_1 instead. Then, the normalization can be done through defining

$$\widehat{\phi} = N(\widehat{\mathbf{s}}\widehat{\phi}_1^T)^{-1}\widehat{\phi}_1 \quad \text{and} \quad \widehat{\psi} = \mathbf{R}^{-1}\widehat{\psi}_1, \quad (18)$$

where $\mathbf{R} = \widehat{\mathbf{Z}}\widehat{\psi}_1^T/N$, $\widehat{\phi}_1 = \phi_1(\widehat{\mathbf{s}})$, and $\widehat{\psi}_1 = \psi_1(\widehat{\mathbf{Z}})$.

A special case that is worth to mention at this point is when the background signals \mathbf{z} are Gaussian, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{z}})$. The covariance $\mathbf{C}_{\mathbf{z}}$ is an unknown nuisance parameter, which must be replaced by the sample-based covariance of $\widehat{\mathbf{Z}}$, that is, by $\widehat{\mathbf{C}}_{\mathbf{z}} = \widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T/N$. It means that the model density $q(\cdot)$ corresponds to $\mathcal{N}(\mathbf{0}, \widehat{\mathbf{C}}_{\mathbf{z}})$, whose score function is $\psi(\mathbf{z}) = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}$. Then, $\widehat{\psi} = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\widehat{\mathbf{Z}}$, $\mathbf{R} = \mathbf{I}_{d-1}$, and $\mathbf{p} = \widehat{\psi}\widehat{\mathbf{s}}^T/N = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\widehat{\mathbf{Z}}\widehat{\mathbf{s}}^T/N = \mathbf{0}$ due to the OG (12). Consequently, the third term on the right-hand side of (17) is zero, and the gradient simplifies to

$$\left. \frac{\partial \mathcal{L}}{\partial \mathbf{w}} \right|_{\text{w.r.t. (13a)}} = \mathbf{a} - \frac{1}{N} \mathbf{X}\widehat{\phi}^T. \quad (19)$$

This result coincides with those derived in [10, 12] under the Gaussian assumption.

The third term on the right-hand side of (17) can be seen as a correction term due to the non-Gaussianity of \mathbf{z} , as \mathbf{p} consists of higher-order correlations between $\widehat{\mathbf{s}}$ and $\widehat{\mathbf{Z}}$, unless ψ is purely linear.

3.2 Proposed Algorithm

We propose a gradient-based algorithm whose steps are described in Algorithm 1. In every step, the OG is imposed through (13a), the normalization steps given

by (18) are done, and the method updates \mathbf{w} in the direction of the steepest ascent of \mathcal{L} . This is repeated until the norm of the gradient is smaller than tol ; μ is the step length parameter; \mathbf{w}_{ini} is the initial guess. We call this method OGICENGB.

Algorithm 1. OGICENGB: separating vector estimation based on orthogonally constrained gradient-ascent algorithm

Input: \mathbf{X} , \mathbf{w}_{ini} , μ , tol , $\phi(\cdot)$, $\psi(\cdot)$
Output: \mathbf{a} , \mathbf{w}

```

1  $\widehat{\mathbf{C}}_{\mathbf{x}} = \mathbf{X}\mathbf{X}^T/N$ ;
2  $\mathbf{w} = \mathbf{w}_{\text{ini}}$ ;
3 repeat
4    $\lambda_{\mathbf{w}} \leftarrow (\mathbf{w}^T \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w})^{-1}$ ;
5    $\mathbf{a} \leftarrow \lambda_{\mathbf{w}} \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{w}$ ;                                /* OG constraint (13a) */
6    $\mathbf{B} = [\mathbf{a}_{2:d}, -\mathbf{a}_1 \mathbf{I}_{d-1}]$ ;                            /* by (6) */
7    $\widehat{\mathbf{s}} \leftarrow \mathbf{w}^T \mathbf{X}$ ;                                /* current SOI estimate */
8    $\widehat{\mathbf{Z}} \leftarrow \mathbf{B}\mathbf{X}$ ;                                /* current background estimate */
9    $\nu \leftarrow \widehat{\mathbf{s}} \phi(\widehat{\mathbf{s}})^T / N$ ;                        /* normalizing constant from (18) */
10   $\mathbf{T} \leftarrow \mathbf{X} \psi(\widehat{\mathbf{Z}})^T / N$ ;                    /* auxiliary matrix due to (18) */
11   $\mathbf{p} = (\mathbf{B}\mathbf{T})^{-1} \mathbf{T}^T \mathbf{w}$ ;                        /* by the definition of  $\mathbf{p}$  */
12   $\Delta \leftarrow \mathbf{a} - \nu^{-1} \mathbf{X} \phi(\widehat{\mathbf{s}})^T / N + \lambda_{\mathbf{w}}^{-1} \widehat{\mathbf{C}}_{\mathbf{x}} \mathbf{B}^T \mathbf{p}$ ; /* by (17) */
13   $\mathbf{w} \leftarrow \mathbf{w} + \mu \Delta$ ;                            /* gradient ascent */
14 until  $\|\Delta\| < \text{tol}$ ;
```

4 Simulations

We compare OGICENGB with its special variant OGICE (assuming Gaussian background) [10, 12], with One-unit FastICA (FICA) from [6], and with the Natural Gradient algorithm (NG) for ICA [1]. In one trial, an instantaneous mixture of $d = 10$ signals is generated according to (1), and the SOI is extracted and evaluated in terms of Signal-to-Interference Ratio (SIR). The SOI u_1 as well as u_2, \dots, u_d are drawn from the Laplacean distribution. The scales of the components are selected so that $\text{SIR}_{\text{in}} = (d-1) \text{E}[|u_1|^2] (\sum_{i=2}^d \text{E}[|u_i|^2])^{-1}$ is 10 dB. The elements of mixing matrices are drawn uniformly from $[1, 2]$, which ensures approximately equal SIR across all channels. The improvement of SIR (SIR_{imp}) is defined as the ratio between the average SIR on channels and the output SIR of the extracted source; the extraction is rated as *successful* if $\text{SIR}_{\text{imp}} > 0$ dB. The percentage of successful trials will be referred to as *success rate*.

The algorithms are initialized by $\mathbf{a}_{\text{ini}} = \mathbf{a} + \mathbf{e}_{\text{ini}}$, where \mathbf{a} is the true mixing vector, and \mathbf{e}_{ini} is a random vector with Gaussian entries such that $\|\mathbf{e}_{\text{ini}}\|^2 = \epsilon^2$. NG is initialized by a de-mixing matrix yielding background subspace that is

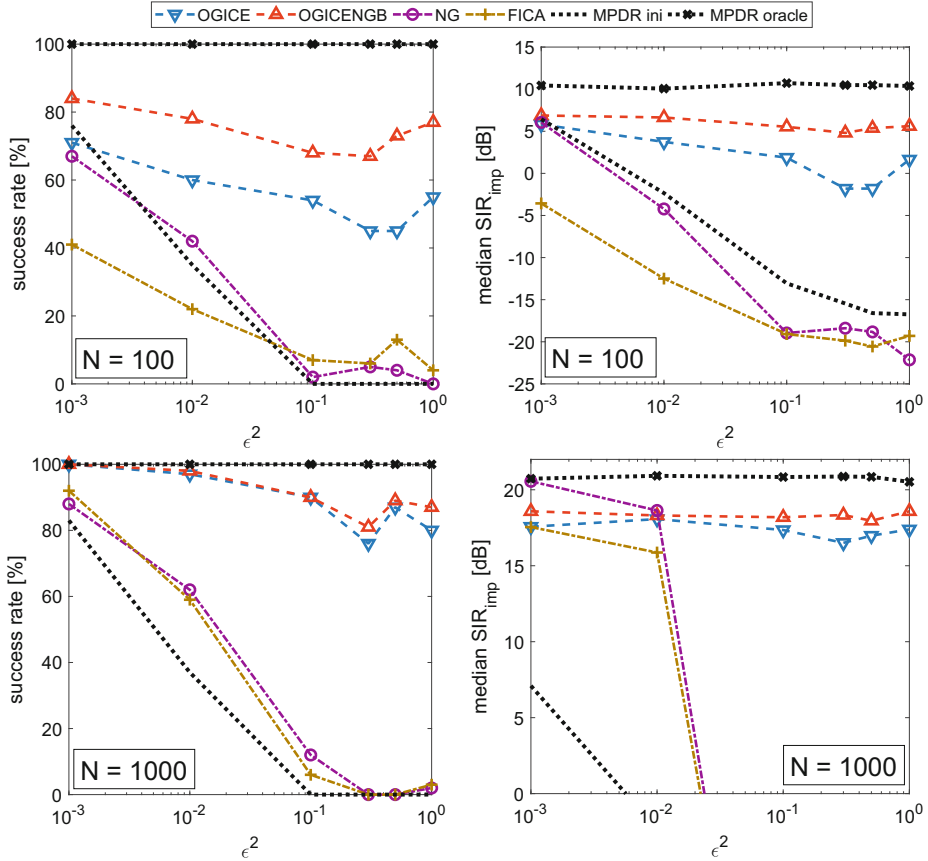


Fig. 1. Success rate and median SIR improvement as functions of ϵ^2 achieved by the compared algorithms in 100 trials for $N = 100$ (row 1) and $N = 1000$ (row 2).

orthogonal to the initial SOI estimate². To compare, the SOI estimates using (13b) with the true mixing vector (MPDR oracle) and with $\mathbf{a} = \mathbf{a}_{ini}$ (MPDR ini) are evaluated, also.

In algorithms, we choose $\phi(s) = \tanh(s)$, which is a smooth approximation of $\text{sign}(x)$ (the true score function for the Laplacean pdf). For choosing ψ in OGICENGB, we adopt the idea from [9] for modeling dependent variables using the multivariate super-Gaussian distribution with covariance \mathbf{C}_z . Thus, the model pdf and the corresponding score function are, respectively,

$$q(\mathbf{z}) \propto \exp \left\{ -\sqrt{\mathbf{z}^T \mathbf{C}_z^{-1} \mathbf{z}} \right\} \quad \text{and} \quad \psi(\mathbf{z}) = \mathbf{C}_z^{-1} \mathbf{z} / \sqrt{\mathbf{z}^T \mathbf{C}_z^{-1} \mathbf{z}}. \quad (20)$$

² Note that the separated sources by NG are *not* reordered after the separation, because the BSE problem is assumed to be resolved correctly only if the SOI appears in the assumed output channel.

Based on this, our final choice of ψ is $\psi(\mathbf{z}) = \widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}/\sqrt{\mathbf{z}^T\widehat{\mathbf{C}}_{\mathbf{z}}^{-1}\mathbf{z}}$ where $\widehat{\mathbf{C}}_{\mathbf{z}}$ is the sample-based estimate of $\mathbf{C}_{\mathbf{z}}$, namely, $\widehat{\mathbf{C}}_{\mathbf{z}} = \widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T/N = \mathbf{B}\widehat{\mathbf{C}}_{\mathbf{x}}\mathbf{B}^T$. The problem of choosing more appropriate nonlinearities, especially ψ , is beyond the scope of this paper.

For all algorithms, the maximum number of iterations is 50000; the stopping criterion is $\text{tol} = 10^{-4}$ for OGICE and OGICENGB, 10^{-3} for NG and 10^{-6} for FICA. The step length μ was set to 0.1; 0.02 in NG; these values were selected to ensure good performance of the methods.

Figure 1 shows the success rate and median SIR_{imp} achieved in 100 trials when the number of samples is, respectively, critical ($N = 100$) and moderate ($N = 1000$). A performance bound is given by MPDR oracle, which yields 100% success rate and 10 dB (resp. 22 dB) of median SIR_{imp} for every ϵ^2 .

For $N = 100$ (row 1 in Fig. 1), NG and FICA fail to improve the initial median SIR given by MPDR ini. By contrast, OGICE and OGICENGB show higher success rate and median SIR_{imp} than MPDR ini when $\epsilon^2 > 0.001$. OGICENGB yields significant improvements compared to OGICE, which points to its ability to exploit the non-Gaussianity of background.

The median SIR_{imp} for $N = 1000$ (row 2, column 2 in Fig. 1) shows that the accuracy of NG and FICA is superior provided that they are initialized in a very close vicinity of the SOI ($\epsilon^2 \leq 0.01$). Here, OGICE achieves similar SIR_{imp} to that of FICA, OGICENGB gives slightly higher SIR_{imp} than OGICE and FICA, and NG outperforms the other methods. This is in a good agreement with the theory as NG exploits the nonGaussianity of background through separating all sources, while OGICENGB performs only a partial separation. For $\epsilon^2 > 0.1$, the median SIR_{imp} of NG and FICA drops below -20 dB, which means that these algorithms mostly converge to a different source (in more than 50% of trials).

The ICE methods show superior global convergence (success rate), which is almost independent of ϵ^2 . Other simulations not shown here due to lack of space confirm that the global convergence of these algorithms is related to the fact that the SOI is significantly dominant in the mixture. The practical use of this interesting property will be subject of further investigations.

5 Conclusions and Future Works

We have shown that OGICENGB can achieve higher separation accuracy than OGICE and One-unit FastICA that assume Gaussian background. The algorithm shows excellent global convergence similarly to OGICE when the SOI is dominant, also in the scenario with a small number of samples ($N = 100$).

Open issues are the choice of a more suitable nonlinearity $\psi(\cdot)$, which might improve the accuracy of OGICENGB, and a faster optimization strategy like that of FastICA, which could considerably increase the convergence speed. Finally, the idea of this paper can be extended to the extraction of a vector component from a set of dependent instantaneous mixtures as an analogy to Independent Vector Analysis; see [9, 10].

References

1. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. In: *Proceedings of Neural Information Processing Systems*, pp. 757–763 (1996)
2. Cardoso, J.F.: On the performance of orthogonal source separation algorithms. In: *Proceedings of European Signal Processing Conference*, pp. 776–779 (1994)
3. Comon, P.: Independent component analysis, a new concept? *Sign. Process.* **36**, 287–314 (1994)
4. Huber, P.J.: Projection pursuit. *Ann. Statist.* **13**(2), 435–475 (1985)
5. Hyvärinen, A.: One-unit contrast functions for independent component analysis: a statistical analysis. In: *Proceedings of the 1997 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing VII*, pp. 388–397 (1997)
6. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Comput.* **9**(7), 1483–1492 (1997)
7. Javidi, S., Mandic, D.P., Cichocki, A.: Complex blind source extraction from noisy mixtures using second-order statistics. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **57**(7), 1404–1416 (2010)
8. Kautský, V., Koldovský, Z., Tichavský, P.: Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-gaussian independent component extraction. In: *2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 94–97 (2017)
9. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **15**(1), 70–79 (2007)
10. Koldovský, Z., Tichavský, P.: Gradient algorithms for complex non-gaussian independent component/vector extraction (2018). [ArXiv1803.10108](https://arxiv.org/abs/1803.10108)
11. Koldovský, Z., Tichavský, P., Kautský, V.: Orthogonally constrained independent component extraction: Blind MPDR beamforming. In: *Proceedings of European Signal Processing Conference*, pp. 1195–1199 (2017)
12. Pham DTA: Blind partial separation of instantaneous mixtures of sources. In: *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 868–875. Springer, Heidelberg (2006)
13. 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. Speech Audio Process.* **12**(5), 530–538 (2004)
14. Tichavský, P., Koldovský, Z., Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. Sign. Process.* **54**(4), 1189–1203 (2006)
15. Tichavský, P., Koldovský, Z., Oja, E.: Corrections to “Performance analysis of the FastICA algorithm and Cramér Rao bounds for linear independent component analysis”. *IEEE Trans. Sign. Process.* **56**(4), 1715–1716 (2008)
16. Van Trees, H.L.: *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley, New York (2002)



A New Link Between Joint Blind Source Separation Using Second Order Statistics and the Canonical Polyadic Decomposition

Dana Lahat^(✉) and Christian Jutten

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
dana@lahat.org.il

Abstract. In this paper, we discuss the joint blind source separation (JBSS) of real-valued Gaussian stationary sources with uncorrelated samples from a new perspective. We show that the second-order statistics of the observations can be reformulated as a coupled decomposition of several tensors. The canonical polyadic decomposition (CPD) of each such tensor, if unique, results in the identification of one or two mixing matrices. The proposed new formulation implies that standard algorithms for joint diagonalization and CPD may be used to estimate the mixing matrices, although only in a sub-optimal manner. We discuss the uniqueness and identifiability of this new approach. We demonstrate how the proposed approach can bring new insights on the uniqueness of JBSS in the presence of underdetermined mixtures.

Keywords: Joint blind source separation
Independent vector analysis · Tensor
Canonical polyadic decomposition · Uniqueness · Identifiability

1 Introduction

In this paper, we present a new type of link between joint blind source separation (JBSS) [1,2] and the canonical polyadic decomposition (CPD) [3], in the special case that each of the sources, in each mixture, is a real-valued Gaussian random process with independent and identically distributed (i.i.d.) samples. To the best of our knowledge, until now, this link has been shown only when the data had some additional type of diversity, e.g., nonstationarity or higher-order statistics (HOS) [4–7]. Our model assumptions, as well as previous related results, are described in Sect. 2. The new algebraic formulation is introduced in Sect. 3. In Sect. 4 we discuss the uniqueness of the proposed new formulation. In Sect. 5, we suggest to use this new formulation as an alternative to

D. Lahat—This work is supported by the project CHESSE, 2012-ERC-AdG-320684. GIPSA-Lab is a partner of the LabEx PERSYVAL-Lab (ANR-11-LABX-0025).

existing JBSS algorithms on the one hand, and to the closed-form solution via generalized eigenvalue decomposition (GEVD) on the other hand. In Sect. 6 we demonstrate how the proposed approach leads to new insights and new results on the identifiability of JBSS in underdetermined cases, beyond existing results in the literature.

In this paper, scalars, column vectors, matrices, and tensors, are denoted a , \mathbf{a} , \mathbf{A} , and \mathcal{A} , respectively. The r th entry of \mathbf{a} , and the r th column of \mathbf{A} , are denoted a_r and \mathbf{a}_r , respectively. \cdot^\top denotes transpose. $\mathbf{A}^{-[k]}$, $\mathbf{X}^{-[k,l]}$, and $\mathbf{A}^{-[k]\top}$ stand for $(\mathbf{A}^{[k]})^{-1}$, $(\mathbf{X}^{[k,l]})^{-1}$, and $(\mathbf{A}^{-[k]})^\top$, respectively. The outer product is denoted as \circ , where $\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^\top$, and $\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}$ is a third-order array (tensor) whose (i, j, k) th element is $a_i b_j c_k$. $\text{Diag}\{\mathbf{a}\}$ is a diagonal matrix with the values of \mathbf{a} on its main diagonal. $E\{\cdot\}$ denotes expectation. $k_{\mathbf{A}}$ denotes the *Kruskal rank* of matrix \mathbf{A} , which is equal to the largest integer $k_{\mathbf{A}}$ such that every subset of $k_{\mathbf{A}}$ columns of \mathbf{A} is linearly independent [8].

2 Problem Formulation

In this paper, we consider the JBSS problem in $K \geq 2$ mixtures,

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]}\mathbf{s}^{[k]} \quad , \quad k = 1, \dots, K \tag{1}$$

where the random vector $\mathbf{x}^{[k]} \in \mathbb{R}^{I^{[k]} \times 1}$ represents the observations at $I^{[k]}$ sensors at the k th mixture. Within each mixture, the R elements of the random vector $\mathbf{s}^{[k]} \triangleq [s_1^{[k]}, \dots, s_R^{[k]}] \in \mathbb{R}^{R \times 1}$ are statistically independent. Each random variable $s_r^{[k]}$ generates a real-valued Gaussian random process with i.i.d. samples. The K mixing matrices $\mathbf{A}^{[k]} \triangleq [\mathbf{a}_1^{[k]} \mid \dots \mid \mathbf{a}_R^{[k]}] \in \mathbb{R}^{I^{[k]} \times R}$ are assumed to be different from each other.

The k th mixture (sometimes referred to as “dataset”) can be written as a sum of contributions from $R \geq 2$ different sources,

$$\mathbf{x}^{[k]} = \sum_{r=1}^R \mathbf{a}_r^{[k]} s_r^{[k]} = \sum_{r=1}^R \mathbf{x}_r^{[k]} . \tag{2}$$

It is clear that $\mathbf{x}_r^{[k]}$ remains invariant if the pair $(\mathbf{a}_r^{[k]}, s_r^{[k]})$ is replaced with $(z^{-1}\mathbf{a}_r^{[k]}, z s_r^{[k]})$, for any $z \neq 0$. Therefore, in the absence of additional information, only $\text{span}(\mathbf{a}_r^{[k]})$ and $\mathbf{x}_r^{[k]}$ may be uniquely identified. Furthermore, the order of the summands in (2) is immaterial. When the model is subject only to these trivial ambiguities, we say that it is essentially unique.

In this paper, we focus on JBSS using second-order statistics (SOS). In this model, the cross-correlation $s_{r\rho}^{[k,l]}$ between any two sources $s_r^{[k]}$ and $s_\rho^{[l]}$ satisfies

$$s_{r\rho}^{[k,l]} \triangleq E\{s_r^{[k]} s_\rho^{[l]\top}\} = \begin{cases} s_{rr}^{[k,l]} & r = \rho \\ 0 & r \neq \rho \end{cases} , \quad \begin{matrix} k, l = 1, \dots, K \\ r, \rho = 1, \dots, R \end{matrix} \tag{3}$$

We assume that each mixture contains R non-zero sources, hence, $s_{rr}^{[k,k]} \neq 0 \forall k$. Furthermore, due to the arbitrary scaling between each source and the column of the mixing matrix associated with it, we can always choose, without loss of generality (w.l.o.g.), $s_{rr}^{[k,k]} \neq 1 \forall k$. For $k \neq l$, each of $s_{rr}^{[k,l]}$ can be zero or non-zero; $s_{rr}^{[k,l]} \neq 0$ can be interpreted as a statistical link (correlation) between the r th source in the k th and l th datasets. The cross-correlation $\mathbf{S}^{[k,l]}$ between $\mathbf{s}^{[k]}$ and $\mathbf{s}^{[l]}$ can thus be written as

$$\mathbf{S}^{[k,l]} \triangleq E\{\mathbf{s}^{[k]}\mathbf{s}^{[l]\top}\} = \text{Diag}\{s_{11}^{[k,l]}, \dots, s_{RR}^{[k,l]}\} \in \mathbb{R}^{R \times R} \quad \forall k, l. \quad (4)$$

In this paper, we assume that all the SOS exist and are finite-valued. Given our assumptions, the sufficient statistics for the estimation of $\text{span}(\mathbf{a}_r^{[k]})$ and $\mathbf{x}_r^{[k]}$ are the cross-correlation matrices of the observations:

$$\mathbf{X}^{[k,l]} \triangleq E\{\mathbf{x}^{[k]}\mathbf{x}^{[l]\top}\} = \mathbf{A}^{[k]}\mathbf{S}^{[k,l]}\mathbf{A}^{[l]\top} = \sum_{r=1}^R s_{rr}^{[k,l]}\mathbf{a}_r^{[k]}\mathbf{a}_r^{[l]\top} \quad \forall k, l \quad (5)$$

where the second equality on the right-hand side (RHS) of (5) is due to (1), and the last equality is due to (2) and (4). In the statistical JBSS formulation (1), each dataset has its own set of parameters, and the link (coupling) between datasets is probabilistic, using additional parameters that represent cross-correlations; in the algebraic formulation in (5), the link between two cross-correlation matrices $\mathbf{X}^{[k,l]}$ and $\mathbf{X}^{[k,l']}$ is deterministic, via a shared mixing matrix $\mathbf{A}^{[k]}$. These two types of links are sometimes referred to as “soft” versus “hard”, see [9] and references therein. Equation (5) implies that the statistically-motivated JBSS can be written algebraically, as a coupled decomposition of the ensemble of matrices $\{\mathbf{X}^{[k,l]}\}_{k,l=1}^K$ [2].

2.1 JBSS via GEVD: A Closed-Form Solution for Two Datasets

Given our assumptions, when $K = 2$ and $\mathbf{A}^{[k]}$ are nonsingular (hence, $I^{[k]} = R$) for $k = 1, 2$, the estimates of $\mathbf{A}^{[k]}$ can always be obtained algebraically, using the GEVD [10, Chapter 12.2, Equation (53)] (see also [11, Sec. 4.3]):

$$\mathbf{X}^{[2,1]}\mathbf{X}^{-[1,1]}\mathbf{X}^{[1,2]}\mathbf{V}^{[1]} = \mathbf{X}^{[2,2]}\mathbf{V}^{[1]}\mathbf{\Lambda} \quad (6a)$$

$$\mathbf{X}^{[1,2]}\mathbf{X}^{-[2,2]}\mathbf{X}^{[2,1]}\mathbf{V}^{[2]} = \mathbf{X}^{[1,1]}\mathbf{V}^{[2]}\mathbf{\Lambda} \quad (6b)$$

where $\mathbf{\Lambda} = \text{Diag}\{\lambda_1, \dots, \lambda_R\}$, and the r th column of $\mathbf{V}^{[k]} \in \mathbb{R}^{R \times R}$ is the generalized eigenvector associated with the generalized eigenvalue λ_r . Therefore, when $K = 2$, JBSS is identifiable if and only if (iff) this GEVD is unique, that is, if its generalized eigenvalues are distinct. The resulting estimates of $\mathbf{A}^{[k]}$, $\mathbf{A}^{[1]\text{GEVD}} \triangleq \mathbf{V}^{[1]\top}$ and $\mathbf{A}^{[2]\text{GEVD}} \triangleq \mathbf{V}^{[2]\top}$, exactly diagonalize $\{\mathbf{X}^{[k,l]}\}_{k,l=1}^K$:

$$(\mathbf{A}^{[k]\text{GEVD}})^{-1}\mathbf{X}^{[k,l]}(\mathbf{A}^{[l]\text{GEVD}})^{-\top} \in \text{Diag} \quad \text{for any } k, l = 1, 2 \quad (7)$$

This solution always exists, and this exact diagonalization can always be achieved, regardless of any perturbation of the observations with respect to

Equation (13) is a decomposition in sum of rank-1 terms of $\mathcal{X}^{[k,m]}$. When R is minimal, (13) is the CPD of $\mathcal{X}^{[k,m]}$, whose three factor matrices are $\mathbf{A}^{[k]}$, $\mathbf{A}^{[m]}$, and $\mathbf{C}^{[k,m]}$.

We now discuss some degeneracies in this representation. Since $\mathbf{S}^{[k,k]}\mathbf{S}^{-[k,k]}\mathbf{S}^{[k,m]} = \mathbf{S}^{[k,m]} = \mathbf{S}^{[k,m]}\mathbf{S}^{-[m,m]}\mathbf{S}^{[m,m]} \forall k, m$, we have

$$\mathbf{C}^{[k,k,m]} = \mathbf{C}^{[k,m,m]} = \mathbf{S}^{[k,m]} \quad \forall k, m \quad (14)$$

and

$$\mathbf{X}^{[k,k,m]} = \mathbf{X}^{[k,m,m]} = \mathbf{A}^{[k]}\mathbf{S}^{[k,m]}\mathbf{A}^{[m]\top} = \mathbf{X}^{[k,m]} \quad \forall k, m \quad (15)$$

where the rightmost equality in (15) follows from (5). Therefore, for fixed $k \neq m$, each of the sequences $\{\mathbf{C}^{[k,l,m]}\}_{l=1}^K$ and $\{\mathbf{X}^{[k,l,m]}\}_{l=1}^K$ contains (at most) $(K-1)$ distinct matrices, whereas for $k = m$, all the matrices in $\{\mathbf{C}^{[k,l,k]}\}_{l=1}^K$ and $\{\mathbf{X}^{[k,l,k]}\}_{l=1}^K$ may be distinct. In order to avoid this degeneracy, from this point and on, we implicitly assume that all the tensors $\mathcal{X}^{[k,m]}$ are constructed such that they do not contain redundant frontal slices, i.e., they do not contain both $\mathbf{X}^{[k,k,m]}$ and $\mathbf{X}^{[k,m,m]}$ for fixed k and m , if $k \neq m$. Therefore, the third “depth” dimension of the tensors $\mathcal{X}^{[k,k]}$ and $\mathcal{X}^{[k,m]}|_{k \neq m}$ is not, in general, the same. As an example, if all $\mathbf{A}^{[k]}$ are nonsingular, then the largest tensors that we can construct for $k \neq m$ are $\mathcal{X}^{[k,m]} \in \mathbb{R}^{R \times R \times (K-1)}$, and for $k = m$, $\mathcal{X}^{[k,k]} \in \mathbb{R}^{R \times R \times K}$. Another point to keep in mind is that due to symmetry, $\mathcal{X}^{[m,k]}$ does not contribute any information beyond $\mathcal{X}^{[k,m]}$.

The bottom line is that we have restated the JBSS problem that we defined in Sect. 2 as an ensemble of tensors $\{\mathcal{X}^{[k,m]}\}_{k,m=1}^K$ that admit a CPD (13). The tensors in this ensemble are coupled because each of them shares, deterministically, a factor matrix $\mathbf{A}^{[k]}$ and/or $\mathbf{A}^{[m]}$ with one or several other CPDs. Hence, we say that the ensemble $\{\mathcal{X}^{[k,m]}\}_{k,m=1}^K$ admits a coupled CPD.

It is worth noting that if $\mathbf{A}^{[k]}$ is nonsingular for some k , the CPD of $\mathcal{X}^{[k,k]}$ amounts to joint diagonalization (JD), and if both $\mathbf{A}^{[k]}$ and $\mathbf{A}^{[m]}$ are nonsingular for some $m \neq k$, then $\mathcal{X}^{[k,m]}$ is an asymmetric two-sided tensor diagonalization. If $\mathbf{A}^{[k]}$ is nonsingular $\forall k$, this coupled CPD amounts to generalized joint diagonalization (GJD) [4]. In general, these are simpler problems.

3.1 Previous Links Between JBSS and Coupled CPD

In fact, the coupled CPD model that we have just described, in Sect. 3, and its link with JBSS, are not new, as we now explain. Let us ignore for a moment the latent structure of the entries of $\mathbf{C}^{[k,m]}$ due to (9) and (10). Then, a representation of the sufficient statistics of a JBSS model in terms of tensors $\{\mathcal{X}^{[k,m]}\}_{k,m=1}^K$ that admit a coupled CPD as in (13) has already been introduced (e.g., [4, 5], [7, Section VI]). However, until now, a three-way structure has been considered only when the sources had an additional type of diversity, such as statistical dependence (correlation) among samples, or nonstationarity; in other words, only when the i.i.d. assumption was violated. In these cases, the distinct frontal slices of

each tensor $\mathcal{X}^{[k,m]}$ represented, for example, different correlation matrices taken at different time lags. In these cases, the third “depth” dimension of the tensors reflected the additional diversity in the data. In this paper, we show for the first time that a coupled CPD formulation is possible even if none of these additional types of diversity is present.

3.2 Discussion

Equation (15) implies that each tensor $\mathcal{X}^{[k,m]}$ has one frontal slice that is identical to the (k, m) th matrix in (5). The difference from (5) is that now, for fixed k and m , we have more than one equation that involves $\mathbf{A}^{[k]}$ and $\mathbf{A}^{[m]}$. Therefore, the new coupled CPD formulation subsumes the coupled matrix factorization in (5). In previous work (e.g., [11]), it was shown that the JBSS problem in Sect. 2 can be solved optimally, in the maximum likelihood (ML) sense, i.e., in terms of the minimal mean square error (MMSE), using the simpler coupled matrix factorization in (5). The proposed coupled CPD formulation uses exactly the same information, and thus cannot achieve a better MMSE. Similarly, in terms of uniqueness, recall that the coupled factorization in (5) uses all the sufficient statistics, and thus, its uniqueness is tantamount to JBSS identifiability [12–14]. Consequently, the alternative formulation of the same statistics in terms of a coupled CPD cannot achieve stronger uniqueness properties. It is thus natural to ask what we can obtain from the more complicated coupled CPD formulation in Sect. 3. The rest of this paper is dedicated to this question.

4 Uniqueness

The identifiability of the JBSS model in Sect. 2 was characterized in [12–14], for nonsingular $\mathbf{A}^{[k]} \forall k$. The model was shown to be identifiable except for very special cases that were fully characterized in [12–14], and depended only on the values of $\{\mathbf{S}^{[k,l]}\}_{k,l=1}^K$. Our goal in this section is to show how these non-identifiable scenarios are reflected in the tensorized framework of Sect. 3. In this paper, we focus only on one of these non-identifiable cases; conclusions for the other cases can be obtained similarly.

In [12–14], it was shown that our JBSS model is not identifiable if there exists a pair (i, j) , $i \neq j$, of sources, whose statistics satisfy

$$s_{jj}^{[k,l]} = \varphi^{[k]} \varphi^{[l]} s_{ii}^{[k,l]} \quad \forall k, l \quad (16)$$

where $\varphi^{[k]} \neq 0 \forall k$. Equation (16) implies that, for each k , the subspaces associated with the i th and j th columns of $\mathbf{A}^{[k]}$ cannot be distinguished. Substituting (16) in (10), we obtain

$$c_j^{[k,l,m]} = s_{jj}^{[k,l]} s_{jj}^{-[l,l]} s_{jj}^{[l,m]} = \varphi^{[k]} \varphi^{[m]} s_{ii}^{[k,l]} s_{ii}^{-[l,l]} s_{ii}^{[l,m]} = \varphi^{[k]} \varphi^{[m]} c_i^{[k,l,m]} \quad \forall k, l, m \quad (17)$$

which implies

$$\mathbf{c}_j^{[k,m]} = \varphi^{[k]} \varphi^{[m]} \mathbf{c}_i^{[k,m]} \quad \forall k, m \quad (18)$$

The other non-identifiable cases are associated with pairs (i, j) of source correlations $\{s_{ii}^{[k,l]}\}_{k,l=1}^K$ and $\{s_{jj}^{[k,l]}\}_{k,l=1}^K$ that contain zeros. Calculations similar to those in (17) show that in these non-identifiable cases, some of $\mathbf{C}^{[k,m]}$ contain two zero columns ($\mathbf{c}_i^{[k,m]} = \mathbf{0} = \mathbf{c}_j^{[k,m]}$), and the remaining $\mathbf{C}^{[k,m]}$ have pairs of proportional columns that contain zeros in some of their entries, in specific locations. Due to lack of space, we omit the details.

Let us assume for a moment that $s_{rr}^{[k,l]} \neq 0 \forall k, l, r$. Hence, the model is not identifiable, and the coupled CPD is not unique, iff (16) holds for some pair (i, j) , $i \neq j$. Equation (18) implies that, in this case, the i th and j th columns of $\mathbf{C}^{[k,m]}$ are proportional $\forall k, m$. Hence, $k_{\mathbf{C}^{[k,m]}} = 1 \forall k, m$, and none of the tensors $\mathcal{X}^{[k,m]}$ has a unique CPD [8]. However, it is important to note that the notation “ $k_{\mathbf{C}^{[k,m]}} = 1 \forall k, m$ ” does not provide information about the indices of the proportional columns. Hence, it does not necessarily imply that the coupled CPD is not unique: if, for some pair of (k, m) , the proportional columns are not in the same indices (i, j) as in the other CPDs, then (16) does not hold, and the JBSS is identifiable. In this case, the coupled matrix factorization (5), as well as the coupled CPD associated with it, are unique. This result is of potential interest because it is the first time that the uniqueness of the coupled CPD is stated explicitly for $\mathbf{C}^{[k,m]}$ that do not have full column rank $\forall k, m$ (the uniqueness analysis of the coupled CPD in [15] assumes that $\mathbf{C}^{[k,m]}$ has full column rank $\forall k, m$, and that the tensors do not have any latent structure). Similar conclusions can be obtained from observing the structure of $\mathbf{C}^{[k,m]}$ in the other non-identifiable cases [12–14], as we have mentioned earlier in this section.

5 Estimating Mixing Matrices from a Single CPD

In this section, we focus our attention on a single CPD (or JD, or a two-sided tensor diagonalization) in (13), within the context of the JBSS model in Sect. 2. It follows from Sect. 4 that it is possible to have $k_{\mathbf{C}^{[k,m]}} \geq 2$. In this case, the CPD of $\mathcal{X}^{[k,m]}$ may be unique. The uniqueness of the CPD is guaranteed, for example, if it satisfies [8]

$$k_{\mathbf{A}^{[k]}} + k_{\mathbf{A}^{[m]}} + k_{\mathbf{C}^{[k,m]}} \geq 2R + 2. \quad (19)$$

Equation (19) implies that, in certain cases, a single CPD may be unique even if $\mathbf{A}^{[k]}$ and/or $\mathbf{A}^{[m]}$, as well as $\mathbf{C}^{[k,m]}$, do not have full column rank. If the CPD of $\mathcal{X}^{[k,m]}$ is unique, then we can extract both $\mathbf{A}^{[k]}$ and $\mathbf{A}^{[m]}$ (if $k \neq m$) or just $\mathbf{A}^{[k]}$ (if $k = m$) from it.¹

¹ In practice, due to finite sample size and noise, (13) is just an approximation. Questions related to uniqueness and estimation in the presence of perturbations from the exact model are beyond the scope of this paper.

Using a single CPD in (13) to estimate one or two mixing matrices can be regarded as an intermediate stage between GEVD and coupled decomposition of the whole ensemble. It allows to compute a single mixing matrix, or two mixing matrices, using any standard JD or CPD algorithm, from a subset of the available cross-correlations. The result of this computation may be used for initialization, or validation, instead of GEVD. An advantage w.r.t. GEVD is that JD and CPD can take into account more than two frontal slices, providing a more accurate initialization (for example). Note also that the GEVD solution in (6) is applicable only to data whose mixing matrices have full column rank, whereas this restriction is relaxed when using a CPD. A drawback of this approach is that we may lose the inherent ability of JBSS to fix a single permutation for all the rank-1 terms in all datasets [1].

6 Application to Underdetermined JBSS

In this section, we demonstrate how our new formulation of the coupled CPD can bring new insights about JBSS with underdetermined mixtures, when $\mathbf{C}^{[k,m]}$ does not have full column rank for at least one pair of (k, m) . To the best of our knowledge, this case has not yet been addressed in the literature.

Consider a JBSS setup as in Sect. 2. In this example, we assume that $s_{rr}^{[k,l]} \neq 0 \forall k, l, r$. Assume that $K - 1$ mixtures, indexed, w.l.o.g., $k = 1, \dots, K - 1$, with nonsingular mixing matrices, satisfy (16) (with $K - 1$ instead of K), and thus are not identifiable. The K th mixture is underdetermined, with a mixing matrix $\mathbf{A}^{[K]}$ that has more columns than rows, i.e. $I^{[K]} < R$. We assume that the cross-correlations of the sources, when taking into account all datasets $k = 1, \dots, K$, do not satisfy (16). Our goal is to demonstrate that this model may be identifiable.

We suggest to solve this problem by constructing a tensor $\mathcal{X}^{[K,m]}$, whose CPD will yield a unique estimate of $\mathbf{A}^{[K]}$ and $\mathbf{A}^{[m]}$. Let us begin by looking at $\mathbf{C}^{[K,m]}$. As an example, let $K = 4$, and $m = 1$. Then,

$$\mathbf{C}^{[4,1]} = \begin{bmatrix} s_{ii}^{[4,1]} s_{ii}^{-[1,1]} s_{ii}^{[1,1]} & s_{jj}^{[4,1]} s_{jj}^{-[1,1]} s_{jj}^{[1,1]} \\ \dots s_{ii}^{[4,2]} s_{ii}^{-[2,2]} s_{ii}^{[2,1]} \dots & s_{jj}^{[4,2]} s_{jj}^{-[2,2]} s_{jj}^{[2,1]} \dots \\ s_{ii}^{[4,3]} s_{ii}^{-[3,3]} s_{ii}^{[3,1]} & s_{jj}^{[4,3]} s_{jj}^{-[3,3]} s_{jj}^{[3,1]} \end{bmatrix} \in \mathbb{R}^{3 \times R} \quad (20a)$$

$$= \begin{bmatrix} s_{ii}^{[4,1]} s_{ii}^{-[1,1]} s_{ii}^{[1,1]} & s_{jj}^{[4,1]} \\ \dots s_{ii}^{[4,2]} s_{ii}^{-[2,2]} s_{ii}^{[2,1]} \dots & \varphi^{-[2]} \varphi^{[1]} s_{jj}^{[4,2]} s_{ii}^{-[2,2]} s_{ii}^{[2,1]} \dots \\ s_{ii}^{[4,3]} s_{ii}^{-[3,3]} s_{ii}^{[3,1]} & \varphi^{-[3]} \varphi^{[1]} s_{jj}^{[4,3]} s_{ii}^{-[3,3]} s_{ii}^{[3,1]} \end{bmatrix} \quad (20b)$$

Equation (20) depicts explicitly the i th and j th columns of $\mathbf{C}^{[4,1]}$. The transition to (20b) is due to (16). In this scenario, $\mathbf{C}^{[K,m]}$ has size $(K - 1) \times R$; recall that $\mathbf{A}^{[K]}$ is not invertible, and thus, cannot take part in (8). Equation (20) shows

that if, as we assume, the cross-correlations $s_{rr}^{[4,m]}$ are such that the ensemble $\{s_{rr}^{[k,l]}\}_{k,l=1}^K$ does not satisfy (16), then there is no linear dependence between $\mathbf{c}_i^{[4,1]}$ and $\mathbf{c}_j^{[4,1]}$. Consequently, $\mathbf{C}^{[4,1]}$ has full rank (although not necessarily full column rank).

We now turn to the uniqueness of $\mathcal{X}^{[K,m]}$, when $m < K$. By (19), the CPD of $\mathcal{X}^{[K,m]}$ is unique if

$$\underbrace{I^{[K]}}_{k_{\mathbf{A}^{[K]}}} + \underbrace{R}_{k_{\mathbf{A}^{[m]}}} + \underbrace{\min(R, K-1)}_{k_{\mathbf{C}^{[K,m]}}} \geq 2R + 2 \quad (21)$$

It follows from (21) that the CPD of $\mathcal{X}^{[K,m]}$ is unique, for example, when $R = 3$, $I^{[K]} = 2$, and $K = 4$, or when $R = 4$, $I^{[k]} = 3$, and $K = 4$.

As soon as $\mathbf{A}^{[K]}$ and $\mathbf{A}^{[m]}$ have been identified, for some fixed m , we can identify all the remaining mixtures $\mathbf{A}^{[k]}$, $k < K$, $k \neq m$, using the fact that now $\mathbf{A}^{[m]}$ is known and invertible, and the diagonal matrix $\mathbf{S}^{[k,m]}$ is nonsingular [15]:

$$\mathbf{X}^{[k,m]} = \mathbf{A}^{[k]} \mathbf{S}^{[k,m]} \mathbf{A}^{[m]\top} \Rightarrow \mathbf{X}^{[k,m]} \mathbf{A}^{-[m]\top} = \mathbf{A}^{[k]} \mathbf{S}^{[k,m]} \quad (22)$$

In this identifiable setup, the tensors $\mathcal{X}^{[k,m]}$ with $k, m < K$ have nonsingular factors $\mathbf{A}^{[k]}$ and $\mathbf{A}^{[m]}$ and, as explained in Sect. 4, a third factor matrix $\mathbf{C}^{[k,m]}$ with $k_{\mathbf{C}^{[k,m]}} = 1$. The tensors $\mathcal{X}^{[K,m]}$ involve one underdetermined factor $\mathbf{A}^{[K]}$, a nonsingular $\mathbf{A}^{[m]}$, and a third factor matrix $\mathbf{C}^{[K,m]}$ that has full rank, but may have more columns than rows, as explained earlier in Sect. 6. Our results show that this model is identifiable, and hence, the overall coupled CPD must be unique, too. This result has been obtained using only one CPD in the ensemble. It is likely that the overall coupled CPD has an even stronger uniqueness.

7 Conclusion

In this paper, we have shown, for the first time, that JBSS of $K \geq 3$ mixtures can be associated with the CPD even in the simplest case where each source in each mixture is a real-valued Gaussian stationary random process with uncorrelated samples. Apart from the theoretical interest in showing another type of link between the statistically-motivated JBSS and an algebraic tensor-based model, we proposed several practical uses to this new formulation. We have shown that this new formulation can bring new insights and new stronger uniqueness results on coupled CPD and on JBSS. In a broader perspective, we provided another evidence for the richness of coupled decompositions.

In Sect. 3.2, we mentioned that the new formulation cannot improve on the MMSE. However, it remains to be seen if the proposed formulation, of coupled CPD, can achieve a smaller estimation error when a norm that does not achieve the ML, e.g., the Frobenius norm, is used in the optimization, and, if so, is the improvement justified w.r.t. the higher computational cost.

The same type of analysis that we presented in this paper for decomposition in sum of rank-1 elements can be extended to terms of any low rank [7], and

to complex-valued data. Finally, if the JBSS data have additional diversity, e.g., sample nonstationarity, this information can be added to each tensor $\mathcal{X}^{[k,m]}$ as additional frontal slices, and thus further enhance the estimation.

References

1. 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). https://doi.org/10.1007/11679363_21
2. Li, Y.O., Adalı, T., Wang, W., Calhoun, V.D.: Joint blind source separation by multiset canonical correlation analysis. *IEEE Trans. Signal Process.* **57**(10), 3918–3929 (2009)
3. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.* **6**(1), 164–189 (1927)
4. Li, X.L., Adalı, T., Anderson, M.: Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Process.* **91**(10), 2314–2322 (2011)
5. Congedo, M., Phlypo, R., Chatel-Goldman, J.: Orthogonal and non-orthogonal joint blind source separation in the least-squares sense. In: Proceedings of the EUSIPCO, Bucharest, pp. 1885–1889, August 2012
6. Lahat, D., Jutten, C.: Joint analysis of multiple datasets by cross-cumulant tensor (block) diagonalization. In: Proceedings of the SAM, Rio de Janeiro, July 2016
7. Lahat, D., Jutten, C.: Joint independent subspace analysis using second-order statistics. *IEEE Trans. Signal Process.* **64**(18), 4891–4904 (2016)
8. Kruskal, J.B.: Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Proc. London Math. Soc.* **18**(2), 95–138 (1977)
9. Lahat, D., Adalı, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges and prospects. *Proc. IEEE* **103**(9), 1449–1477 (2015)
10. Anderson, T.W.: An introduction to multivariate statistical analysis. Wiley, New York (1958)
11. Vía, J., Anderson, M., Li, X.L., Adalı, T.: A maximum likelihood approach for independent vector analysis of Gaussian data sets. In: Proceedings of the MLSP, Beijing, September 2011
12. Vía, J., Anderson, M., Li, X.L., Adalı, T.: Joint blind source separation from second-order statistics: Necessary and sufficient identifiability conditions. In: Proceedings of the ICASSP, Prague, pp. 2520–2523, May 2011
13. Anderson, M., Fu, G.S., Phlypo, R., Adalı, T.: Independent vector analysis: identification conditions and performance bounds. *IEEE Trans. Signal Process.* **62**(17), 4399–4410 (2014)
14. Lahat, D., Jutten, C.: An alternative proof for the identifiability of independent vector analysis using second order statistics. In: Proceedings of the ICASSP, Shanghai, March 2016
15. Gong, X.F., Lin, Q.H., Cong, F.Y., De Lathauwer, L.: Double coupled canonical polyadic decomposition for joint blind source separation (2017). [arXiv:1612.09466](https://arxiv.org/abs/1612.09466) [stat.ML]

Nonlinear Mixtures



A Blind Source Separation Method Based on Output Nonlinear Correlation for Bilinear Mixtures

Andréa Guerrero^(✉), Yannick Deville, and Shahram Hosseini

Université de Toulouse, UPS, CNRS, CNES,
IRAP (Institut de Recherche en Astrophysique et Planétologie),
14 av. Edouard Belin, 31400 Toulouse, France
{andrea.guerrero,yannick.deville,shahram.hosseini}@irap.omp.eu

Abstract. In this paper, a blind source separation method for bilinear mixtures of two source signals is presented, that relies on nonlinear correlation between separating system outputs. An estimate of each source is created by linearly combining observed mixtures and maximizing a cost function based on the correlation between the element-wise product of the estimated sources and the corresponding quadratic term. A proof of the method separability, i.e. of the uniqueness of the solution to the cost function maximization problem, is also given. The algorithm used in this work is also presented. Its effectiveness is demonstrated through tests with artificial mixtures created with real Earth observation spectra. The proposed method is shown to yield much better performance than a state-of-the-art method.

Keywords: Blind source separation methods · Bilinear mixtures
Nonlinear correlation · Hyperspectral imaging

1 Introduction

Blind source separation (BSS) consists of restoring source signals contained in observed mixed data. In this paper, the target application field is Earth observation. In theoretical BSS investigations, the observed data are usually linear combinations of the sources, but various applications involve nonlinear mixtures. In particular, Linear-Quadratic (LQ) memoryless mixtures, which include the bilinear mixing studied here, appear in the *show-through* effect [3], in chemistry applications [8] and also in Earth observation [1]. The bilinear case is considered as a difficult problem because nonlinearity complicates BSS.

BSS methods have been largely studied in the past: see [9, 11–13, 15] for example. Generally, they rely on source properties to resolve the problem. Independent Component Analysis (ICA) methods exploit statistical independence of sources, by using non-Gaussianity, non-stationarity or time correlation [5]. Sparse Component Analysis (SCA) [6] methods are based on source sparsity, whereas Non-negative Matrix Factorization (NMF) [7, 12], which became popular these last

years, only requires non-negativity of sources and mixing coefficients, although this yields major indeterminacies. For a survey of the BSS methods which have been proposed in the literature for LQ and bilinear mixtures, see [14].

In this work, the (over)determined bilinear mixing model for two sources is considered, and only the linear independence of source vectors and some of their element-wise products is required to build a new powerful method, without relying on any other information. In Sect. 2, the mixing model studied in this paper is presented. Then, the proposed method and the principle on which it is based are explained in Sect. 3. We also describe our algorithm and the tools used to execute it. In Sect. 4, the uniqueness of the solution is proved. Then, an evaluation of the performance of the method is provided through tests and comparison with a state-of-the-art method in Sect. 5. Lastly, a conclusion about the effectiveness of our method and perspectives of our work are given.

2 Bilinear Mixing Model

In Earth observation, especially in urban scenes, observed data are often multi-spectral or hyperspectral images where every pixel spectrum may be a mixture of several pure material spectra. Because of the reflection of a sunbeam on multiple materials like asphalt, building walls and ground, the observed data are then produced by a bilinear mixing model [1]:

$$x(n) = \tilde{A}\tilde{s}(n) \quad (1)$$

$$= As_a(n) + Bs_b(n) \quad (2)$$

where x is the observation vector defined in (3), \tilde{s} the extended source vector defined in (4) and \tilde{A} the mixing matrix. We consider real-valued signals that depend on a discrete variable n corresponding to the wavelengths. The considered image consists of P pixels, so x contains P observations and reads

$$x(n) = [x_1(n) \cdots x_P(n)]^T \quad (3)$$

whereas

$$\begin{aligned} \tilde{s}(n) &= [s_a(n)^T \ s_b(n)^T]^T \\ &= [s_1(n) \cdots s_N(n) \ s_1(n)s_2(n) \cdots s_{N-1}(n)s_N(n)]^T. \end{aligned} \quad (4)$$

s_a contains the N actual sources $s_i(n)$ and s_b contains the $K = N(N-1)/2$ quadratic cross terms $s_i(n)s_j(n)$, $i < j$. We here consider an (over)determined case where $P \geq N + K$. In the same way, we define

$$\begin{aligned} \tilde{A} &= [AB] \quad (5) \\ &= \begin{bmatrix} a_{11} & \cdots & a_{1N} & b_{1,1,2} & \cdots & b_{1,N-1,N} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{P1} & \cdots & a_{PN} & b_{P,1,2} & \cdots & b_{P,N-1,N} \end{bmatrix}. \end{aligned}$$

Equation (5) contains two types of terms: A corresponding to mixing coefficients for the N actual sources in s_a and B which contains the mixing coefficients for the quadratic terms in s_b . So the mixed spectra contained in the observations read

$$x_i(n) = \sum_{j=1}^N a_{ij} s_j(n) + \sum_{k=1}^{N-1} \sum_{l=k+1}^N b_{i,k,l} s_k(n) s_l(n) \quad (6)$$

$$i \in \{1, \dots, P\}.$$

From now on, we focus on the case of $N = 2$ actual sources to simplify calculations, so $K = 1$. To address the determined case, we choose $P = N + K = 3$ observations. The mixing model (2) then becomes

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & b_{1,1,2} \\ a_{21} & a_{22} & b_{2,1,2} \\ a_{31} & a_{32} & b_{3,1,2} \end{bmatrix} \begin{bmatrix} s_1(n) \\ s_2(n) \\ s_1(n)s_2(n) \end{bmatrix}. \quad (7)$$

3 Blind Source Separation Based on Output Nonlinear Correlation

The proposed separating system is defined as $y(n) = Cx(n)$ so that

$$\begin{bmatrix} y_1(n) \\ y_2(n) \\ y_3(n) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \\ x_3(n) \end{bmatrix} \quad (8)$$

where C represents the separating matrix and y represents the estimated extended sources. Each estimated signal reads

$$y_i(n) = c_{i1}x_1(n) + c_{i2}x_2(n) + c_{i3}x_3(n) \quad (9)$$

$$i \in \{1, 2, 3\}.$$

3.1 Principle of Proposed Method

The proposed method, named BOCSS for Bilinear Output Correlation-based Source Separation, linearly combines the observations to provide the estimated sources y_i as described in (9), considering the quadratic term as an additional source. Thus, if the outputs y_i are well estimated up to indeterminacies, y_1 and y_2 are proportional to s_1 and s_2 (not necessarily equal to them, due to the scale indeterminacy of BSS) in an arbitrary order (due to the permutation indeterminacy), whereas y_3 is proportional to $s_1 \odot s_2$, with \odot the element-wise product of source vectors. Therefore, y_3 is proportional to $y_1 \odot y_2$. BOCSS then consists of adapting C so as to ensure the latter proportionality condition. To this end, it maximizes the cost function J defined as

$$J = \rho^2(y_1 \odot y_2, y_3). \quad (10)$$

ρ is the correlation coefficient defined as

$$\rho(\alpha, \beta) = \frac{\text{cov}(\alpha, \beta)}{\sigma(\alpha)\sigma(\beta)} \quad (11)$$

where cov is the covariance and σ is the standard deviation of the considered signals. So when separation succeeds, (10) reaches the value 1, equivalent to perfect correlation. We prove the uniqueness of the solution in Sect. 4.

3.2 Algorithm

The above method leads to the following algorithm. To optimize J , nine parameters must be adapted in the first variant of BOCSS, namely the nine coefficients of C . We use the MATLAB® *fminsearch* function to this end.

Algorithm 1. Separation method

```

for all  $t = 1$  to  $T$  do
  Initialize supposed mixing matrix  $\tilde{A}_t^{init}$ 
   $C_t = \text{inv}(\tilde{A}_t^{init})$ 
  Adapt  $C_t$  to maximize  $J$  in (10)
  if  $\text{inv}(C_t) \geq 0$  then
    Save  $C_t$  in  $C_{global}$ 
  end if
end for
Cluster  $C_{global}$ 

```

T is the number of tests. Several tests, corresponding to several initializations \tilde{A}_t^{init} , are needed because the optimization result depends on the matrix initialization (see Sect. 5 for details). We then cluster the 3-entry vectors corresponding to the first and second rows of all matrices C_t stored in C_{global} , and we keep the medians of the two obtained clusters as the first two rows of C , then used to derive $y_1(n)$ and $y_2(n)$ with (8). We use the k -medians method [10] to this end.

In each test, we check the relevance of the obtained matrix C_t based on the mixing matrix non-negativity: it has been demonstrated in [1] that mixing coefficients in Earth observation with an LQ or bilinear model are always non-negative:

$$\begin{cases} 0 \leq a_{ij} \leq 1 & i \in \{1, \dots, P\}, j \in \{1, \dots, N\} \\ 0 \leq b_{ikl} \leq 0.5 & k \in \{1, \dots, N-1\}, l \in \{k+1, \dots, N\} \end{cases} \quad (12)$$

It means the mixing matrix \tilde{A} has to be non-negative. So we verify the non-negativity of the inverse of the separating matrix, $\text{inv}(C_t)$, in each test after optimization, and only keep those satisfying this property in the C_{global} set.

Two variants of BOCSS are proposed: with and without a constraint on the separating matrix diagonal. This constraint consists of keeping its diagonal coefficients constant during the optimization, so only 6 coefficients are adapted instead of 9 and this eliminates the scale indeterminacy.

4 Separability

In this section, we show the uniqueness of the solution to the maximization of the cost function (10). To this end, we express the output of the separating system as

$$\begin{aligned} y(n) &= C\tilde{A}\tilde{s}(n) \\ &= G\tilde{s}(n). \end{aligned} \quad (13)$$

In (13), G contains the mixing step A and the separating step C . y is directly connected to s through the coefficients g_{ij} , $i \in \{1, \dots, N + K\}$ and $j \in \{1, \dots, N + K\}$ with $N + K = 3$ extended sources. This yields

$$y_i(n) = g_{i1}s_1(n) + g_{i2}s_2(n) + g_{i3}s_1(n)s_2(n). \quad (14)$$

Our criterion (10) is based on the comparison between the element-wise product of y_1 and y_2 , and y_3 . Using (14), with $i = 1$ and 2 yields

$$\begin{aligned} y_1 \odot y_2 &= g_{11}g_{21}s_1 \odot s_1 + (g_{11}g_{22} + g_{12}g_{21})s_1 \odot s_2 \\ &\quad + g_{12}g_{22}s_2 \odot s_2 + (g_{11}g_{23} + g_{13}g_{21})s_1 \odot s_1 \odot s_2 \\ &\quad + (g_{12}g_{23} + g_{13}g_{22})s_1 \odot s_2 \odot s_2 \\ &\quad + g_{13}g_{23}s_1 \odot s_1 \odot s_2 \odot s_2. \end{aligned} \quad (15)$$

We hereafter consider the case when the eight vectors s_1 , s_2 , $s_1 \odot s_1$, $s_1 \odot s_2$, $s_2 \odot s_2$, $s_1 \odot s_1 \odot s_2$, $s_1 \odot s_2 \odot s_2$, $s_1 \odot s_1 \odot s_2 \odot s_2$ are linearly independent (see Sect. 5 for more explanations about this property), and y_1 , y_2 and y_3 non-zero. Then, (15) and y_3 (see (14)) show that the collinearity of $y_1 \odot y_2$ and y_3 requested by our method is achieved if and only if both vectors are collinear to $s_1 \odot s_2$. For y_3 , using (14), this yields $g_{31} = g_{32} = 0$. Besides, (15) then becomes

$$y_1 \odot y_2 = (g_{11}g_{22} + g_{12}g_{21})s_1 \odot s_2 \quad (16)$$

and this case corresponds to the following constraints on g_{ij} :

$$\begin{cases} g_{11}g_{21} = 0 \\ g_{12}g_{22} = 0 \\ (g_{11}g_{23} + g_{13}g_{21}) = 0 \\ (g_{12}g_{23} + g_{13}g_{22}) = 0 \\ g_{13}g_{23} = 0 \end{cases} \quad (17)$$

Moreover, (16) with non-zero vectors y_1 and y_2 yields

$$\begin{cases} g_{11}g_{22} \neq 0 \\ \text{or} \\ g_{12}g_{21} \neq 0 \end{cases} \quad (18)$$

The two equations in (18) are mutually exclusive because if the two terms weren't null then the first condition in (17) would not be met, which would yield a contradiction.

Then we need to study each possible case to see if the uniqueness condition is met. The first possible case based on (16) and (18) is

$$g_{11} = 0 \text{ then } g_{12} \neq 0 \text{ and } g_{21} \neq 0. \tag{19}$$

Then Eq. (16) becomes

$$y_1 \odot y_2 = g_{12}g_{21}s_1 \odot s_2. \tag{20}$$

Equations (17), (18) and (19) yield

$$\begin{cases} g_{12} \neq 0 \text{ then } g_{22} = 0 \\ (g_{11}g_{23} + g_{13}g_{21}) = g_{13}g_{21} = 0 \text{ then } g_{13} = 0. \\ (g_{12}g_{23} + g_{13}g_{22}) = g_{12}g_{23} = 0 \text{ then } g_{23} = 0 \end{cases} \tag{21}$$

With these constraints on G , each output of interest is proportional to one source since (14) with $i = 1$ and 2 yields

$$\begin{aligned} y_1 &= g_{12}s_2 \\ y_2 &= g_{21}s_1. \end{aligned} \tag{22}$$

For this first studied case, matrix G thus becomes

$$G_1 = \begin{bmatrix} 0 & g_{12} & 0 \\ g_{21} & 0 & 0 \\ 0 & 0 & g_{33} \end{bmatrix}. \tag{23}$$

The only other possible case is $g_{11} \neq 0$. Then (17) yields $g_{21} = 0$. Therefore, due to (18), $g_{22} \neq 0$. Hence, (17) yields $g_{12} = 0$. With the same type of calculations as above, G is shown to then become

$$G_2 = \begin{bmatrix} g_{11} & 0 & 0 \\ 0 & g_{22} & 0 \\ 0 & 0 & g_{33} \end{bmatrix}. \tag{24}$$

As a conclusion, if the eight vectors $s_1, s_2, s_1 \odot s_1, s_1 \odot s_2, s_2 \odot s_2, s_1 \odot s_1 \odot s_2, s_1 \odot s_2 \odot s_2, s_1 \odot s_1 \odot s_2 \odot s_2$ are linearly independent, the cost function J reaches its global maximum if and only if y_1 and y_2 are proportional to the sources in an arbitrary order, and y_3 is proportional to their product.

5 Test Results

To analyze the performance of our method, we choose a criterion called NRMSE (Normalized Root Mean Square Error) defined in [4]:

$$NRMSE = \frac{\sqrt{\min_{i \neq j \in \{1,2\}} (F_{ij})}}{\sqrt{\|s_1^2\| + \|s_2^2\|}} \tag{25}$$

where F_{ij} represents

$$\min_{\epsilon_1=\pm 1} \left(\|s_1 + \epsilon_1 \frac{\|s_1\|}{\|y_i\|} y_i\|^2 \right) + \min_{\epsilon_2=\pm 1} \left(\|s_2 + \epsilon_2 \frac{\|s_2\|}{\|y_j\|} y_j\|^2 \right). \quad (26)$$

We test our method on artificial mixtures built from two real sources extracted from the USGS hyperspectral database [16]. We create 2 artificial sources: each of their samples is derived as the average of 200 (or 20 depending on the test) adjacent samples of a USGS spectrum. This yields sources with only 10 (or 100) samples, which corresponds to a difficult case since few separating methods are able to separate sources containing few samples. We verify the linear independence of the eight vectors $s_1, s_2, s_1 \odot s_1, s_1 \odot s_2, s_2 \odot s_2, s_1 \odot s_1 \odot s_2, s_1 \odot s_2 \odot s_2, s_1 \odot s_1 \odot s_2 \odot s_2$ needed for the separability: we create a matrix including these eight vectors and we calculate the matrix rank. We obtain a rank of 8 which shows the linear independence.

We create the mixing matrix \tilde{A} according to the model in [2]: the mixing coefficients a_i are uniformly drawn in $[0, 1]$ and then divided by their sum, and the b_{ikl} coefficients are randomly chosen with a uniform distribution over $[0, 0.3]$. In the first tests, the separating matrix C_t is initialized to the inverse of the actual matrix \tilde{A} plus uniform “noise”, over a range set to 10, 20, 50 or 100 % of the considered entry of \tilde{A} . This aims at analyzing the robustness of the proposed method to the initialization of C_t . Additional tests are performed with all entries of \tilde{A}_t^{init} uniformly drawn over $[0, 1]$. All the implemented tests use $N = 2$ sources and $P = 3$ mixtures. Tests were performed for $T = 1000$ and $T = 10000$, in order to analyze the trade-off that the proposed method achieves between clustering efficiency and computational complexity. The two variants of BOCSS are tested here, i.e. with and without the diagonal constraint.

For a random initialization, i.e. no assumption on the mixing coefficients except the range $[0, 1]$, the method leads to only approximately 1% error. Using a MATLAB code on a computer with an Intel Core i7 CPU with a frequency of 2.6 GHz and a RAM of 15.6 GB, the CPU time for the test without constraint is approximately 35 min for $T = 10000$ and around 30 min with the constraint,

Table 1. Test results for spectra with **10 samples**

T = 1000 tests without diagonal constraint					
Noise added to C_t^{init}	10%	20%	50%	100%	Random initialization
NRMSE (%)	0.08	0.6	0.32	0.46	1.44
with diagonal constraint					
NRMSE (%)	0.022	0.1	0.33	0.11	0.67
T = 10000 tests without diagonal constraint					
NRMSE (%)	0.08	0.09	0.25	0.4	1.15
with diagonal constraint					
NRMSE (%)	0.016	0.11	0.34	0.48	4.7

Table 2. Test results for spectra with **100 samples**

T = 1000 tests without diagonal constraint					
Noise added to C_t^{init}	10%	20%	50%	100%	Random initialization
NRMSE (%)	0.02	0.14	0.31	0.42	1.44
with diagonal constraint					
NRMSE (%)	0.02	0.1	0.25	0.33	7.93
T = 10000 tests without diagonal constraint					
NRMSE (%)	0.01	0.13	0.28	0.45	0.64
with diagonal constraint					
NRMSE (%)	0.0058	0.14	0.33	0.37	0.65

which is not surprising since only 6 coefficients are optimized in the latter case. Tables 1 and 2 show that for $T = 10000$ tests and a random initialization, the performance is globally better with 100-sample sources than with 10-sample sources. Besides, the number of tests has no major impact on the method performance since the NRMSE remains quite low whatever T .

We compare the BOCSS method with a different kind of approach, namely the NMF Gradient-Newton LQ (NMF-Grd-Newton-LQ) method in [2] restricted to the bilinear case, which is based on the NMF principle, i.e. on the non-negativity of the data like in Earth observation. The maximization of the cost function is done with the gradient descent with a Newton adaptive step. The same configuration is kept: we test for 10 and 100-sample sources and for the same type of initialization except that this method operates with the mixing matrix A directly and not with the separating matrix C , and the method needs a spectrum initialization. We choose the first spectrum initialization proposed in [2], i.e. constant spectra equal to 0.5.

Table 3 shows that the NMF-Grd-Newton-LQ method yields errors around 9%, whereas the BOCSS method leads to much better results: its error is often nearly 10 times lower. However the NMF-Grd-LQ method converges in a few minutes.

Table 3. Tests with the NMF-Grd-Newton-LQ method

Tests for 10-sample sources					
Noise added to A^{init}	10%	20%	50%	100%	Random initialization
NRMSE (%)	9.06	9.06	9.06	9.07	9.1
Tests for 100-sample sources					
NRMSE (%)	9.09	9.09	9.09	9.09	9.13

6 Conclusion

The proposed BOCSS method based on output nonlinear correlation is really interesting e.g. for the Earth observation field, because it doesn't require information on sources like sparsity, statistical independence or non-negativity. It only requires linear independence between source vectors and some of their element-wise products, as shown in Sect. 4. This work is based on bilinear mixtures and possibly mixing matrix non-negativity property. Comparison with work reported in [2] showed the effectiveness of our method with a major gain on the NRMSE. Uniqueness of the solution is also shown in this paper. We could explore other application fields in future work like Chemistry for example. It would be possible to remove the non-negativity property of the mixing coefficients and study the case with more than two sources to generalize BOCSS. Further investigations on the algorithm will be done to improve the practical results and avoid algorithms to get trapped in cost function local maxima.

References

1. Meganem, I., Déliot, P., Briottet, X., Deville, Y., Hosseini, S.: Linear-quadratic mixing model for reflectances in urban environments. *IEEE Trans. Geosci. Remote Sens.* **52**, 544–558 (2014)
2. Meganem, I., Deville, Y., Hosseini, S., Déliot, P., Briottet, X.: Linear-quadratic blind source separation using NMF to unmix urban hyperspectral images. *IEEE Trans. Sign. Process.* **62**, 1822–1833 (2014)
3. Duarte, L.T., Jutten, C., Moussaoui, S.: Bayesian source separation of linear and linear-quadratic mixtures using truncated priors. *J. Sign. Process. Syst.* **65**, 311–323 (2011)
4. Deville, Y.: Matrix factorization for bilinear blind source separation: methods, separability and conditioning. In: *Proceedings of the 23rd European Signal Processing Conference, Nice, France, pp. 1945–1949* (2015)
5. Cardoso, J.F.: The three easy routes to independent component analysis, contrasts and geometry. In: *Proceedings of the ICA 2001 workshop, San Diego, pp. 1–6* (2001)
6. Gribonval, R., Lesage, S.: A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges. In: *Proceedings of the 14th European Symposium on Artificial Neural Networks, Bruges, Belgium, pp. 323–330* (2006)
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: *Neural Information Processing Systems, pp. 556–562* (2001)
8. Ando, R.A., Jutten, C., Rivet, B., Attux R., Duarte, L.T.: Nonlinear blind source separation for chemical sensor arrays based on a polynomial representation. In: *24th European Signal Processing Conference (EUSIPCO), pp. 2146–2150* (2016)
9. Comon, P., Jutten, C.: *Handbook of Blind Source Separation. Independent Component Analysis and Applications.* Academic Press, Oxford (2010)
10. Theodoridis, S., Koutroumbas, K.: *Pattern Recognition.* Academic Press, Oxford (2009)
11. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis.* Wiley, New York (2001)

12. 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, Chichester (2009)
13. Makino, S., Lee, T.W., Sawada, H.: Blind speech separation. Springer, Dordrecht (2007)
14. Deville, Y., Duarte, L.T.: An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 155–167. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_18
15. Deville, Y.: Blind source separation and blind mixture identification methods. In: Wiley Encyclopedia of Electrical and Electronics Engineering. pp. 1–33 (2016). J. Webster
16. Kokaly, R.F., Clark, R.N., Swayze, G.A., Livo, K.E., Hoefen, T.M., Pearson, N.C., Wise, R.A., Benzel, W.M., Lowers, H.A., Driscoll, R.L., Klein, A.J.: USGS Spectral Library Version 7: U.S. Geological Survey Data Series 1035 (2017). <https://doi.org/10.3133/ds1035>



Using Taylor Series Expansions and Second-Order Statistics for Blind Source Separation in Post-Nonlinear Mixtures

Denis G. Fantinato¹(✉), Leonardo T. Duarte², Yannick Deville³,
Christian Jutten⁴, Romis Attux², and Aline Neves¹

¹ Federal University of ABC, Santo André, SP, Brazil
{denis.fantinato,aline.neves}@ufabc.edu.br

² University of Campinas, Campinas, SP, Brazil

leonardo.duarte@fca.unicamp.br, attux@dca.fee.unicamp.br

³ Université de Toulouse, UPS, CNRS, CNES, IRAP, Toulouse, France
yannick.deville@irap.omp.eu

⁴ GIPSA-Lab, Grenoble INP, CNRS, Grenoble, France
christian.jutten@gipsa-lab.grenoble-inp.fr

Abstract. In the context of Post-Nonlinear (PNL) mixtures, source separation based on Second-Order Statistics (SOS) is a challenging topic due to the inherent difficulties when dealing with nonlinear transformations. Under the assumption that sources are temporally colored, the existing SOS-inspired methods require the use of Higher-Order Statistics (HOS) as a complement in certain stages of PNL demixing. However, a recent study has shown that the sole use of SOS is sufficient for separation if certain constraints on the separation system are obeyed. In this paper, we propose the use of a PNL separating model based on constrained Taylor series expansions which is able to satisfy the requirements that allow a successful SOS-based source separation. The simulation results corroborate the proposal effectiveness.

Keywords: Blind source separation · Post-Nonlinear mixtures
Second-Order Statistics

1 Introduction

The Blind Source Separation (BSS) problem is concerned with retrieving a set of unknown source signals from samples that are mixtures of them [1, 2]. Under the assumption that the sources are statistically mutually independent, a number of separation methods were proposed considering that the mixing process is linear and instantaneous, counting with a wide range of applications. Some of these methods use the ubiquitous tool known as Independent Component Analysis (ICA) [1], whose very essence is the recovery of statistical independence through the use of the Higher-Order Statistics (HOS) of the output signals. Other methods alternatively consider the use of only Second-Order Statistics (SOS) of the

output signals, which is a reliable approach when the sources are temporally correlated [1]. This latter approach contributed to the development of techniques like SOBI [3], TDSEP and AMUSE [1], which are computationally simpler in comparison with ICA.

There are certain cases, however, in which the linear mixing assumption is not sufficient, such as in hyperspectral imaging [4] and in chemical sensor arrays [5], which demands source separation methods that take into account the nonlinear mixing process. The main issue is that, from a general nonlinear standpoint, neither the ICA framework nor the SOS-based methods are sufficient for performing source separation [1]. In view of this limitation, the studies on this topic focused on a set of constrained nonlinear models in which the ICA methods are still valid [6], e.g., the Post-Nonlinear (PNL) models [7, 8].

Indeed, ICA methods can be applied to solve the PNL mixing problem [7], but, for the SOS-based methods, this statement does not hold, even considering that sources are temporally colored [6]. This motivated the use of the second-order framework in a partial manner, i.e., by combining it to HOS: one approach is, for instance, to solve the nonlinear part in a first step through a HOS-based method and then apply an SOS-based method to the remaining linear BSS problem [8, 9]; or, alternatively, some HOS source priors can be used as additional information to aid the SOS-based methods [10]. Nevertheless, a recent study pointed out that the sole use of SOS can be sufficient for separation of PNL mixtures if certain constraints on the separation system are obeyed, which includes the existence of a linear mixture term during adaptation [11]. Based on this approach, we propose in this paper the use of a set of constrained Taylor series-based expansions to compose the PNL separating system, which will be able to satisfy the requirements for the sole use of an SOS-based method. As we intend to show, the constrained separation model fully preserves its nonlinear flexibility, allowing its use in a wide set of applications. Given the complexity of the nonlinear framework, a metaheuristic known as Differential Evolution (DE) [12] will be used to aid coefficient adaptation.

This work is organized as follows. Section 2 presents a brief background on the PNL mixing problem and on the use of SOS for source separation. In Sect. 3, we propose a PNL separating system based on a set of constrained Taylor series-based expansions, whose parameters can be adapted via SOS-based criteria. Some performance analyses are presented in Sect. 4. Finally, Sect. 5 concludes the work.

2 Background on Post-Nonlinear Mixtures and Second-Order Statistics

Within the BSS problem, the PNL model arises as a natural extension of the standard linear instantaneous mixture process to a nonlinear one, in which ICA methods – or, more generally, statistical independence – are able to retrieve the original sources [1, 7]. Basically, in the PNL mixture process, N mutually independent sources – denoted by the vector $\mathbf{s}(n) = [s_1(n) \dots s_N(n)]^T$ – are

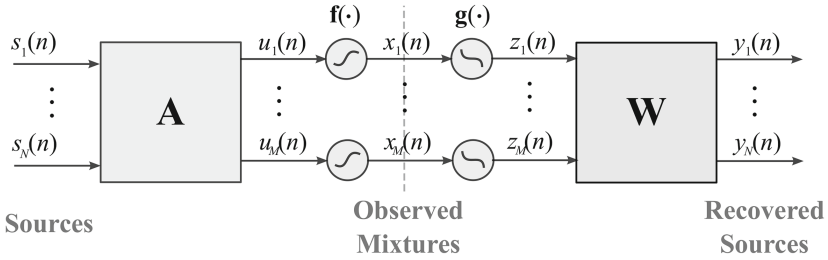


Fig. 1. Mixing and separating systems in the PNL model.

mixed by a linear combination followed by a set of univariate nonlinear functions, resulting in M observed mixtures: $\mathbf{x}(n) = \mathbf{f}(\mathbf{u}(n)) = \mathbf{f}(\mathbf{A}\mathbf{s}(n))$, being \mathbf{A} an $M \times N$ matrix and $\mathbf{f}(\cdot) = [f_1(\cdot) \dots f_M(\cdot)]^T$ a set of M component-wise functions, as illustrated in Fig. 1. As usual in BSS, the aim is to retrieve the sources $\mathbf{s}(n)$ from the observed mixtures $\mathbf{x}(n)$, without prior information about \mathbf{A} or $\mathbf{f}(\cdot)$. In order to do so, a mirrored version of the mixing system is used as a separation system [7], whose output is given by $\mathbf{y}(n) = \mathbf{W}\mathbf{z}(n) = \mathbf{W}\mathbf{g}(\mathbf{x}(n))$, where \mathbf{W} is an $N \times M$ matrix and $\mathbf{g}(\cdot)$ is a set of M component-wise functions, ideally the inverse of $\mathbf{f}(\cdot)$. In this work, the analysis will be restrained to the determined case, i.e., when $M = N$.

To perform blind separation, the PNL separating system can be adjusted using the ICA framework, being sufficient conditions [13]: (i) the mixing matrix \mathbf{A} is invertible and effectively mixes the sources; (ii) $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are monotonic functions; (iii) at most one source is Gaussian, and (iv) the joint distribution of the sources is differentiable and its derivative is continuous on its support.

One disadvantage of ICA methods is their dependence on HOS, which may result in certain difficulties regarding estimation accuracy and/or computational complexity. In that sense, a possible approach to overcome these difficulties is to resort to the temporal statistical information of the sources (by assuming temporally correlated sources) and their SOS, which are simpler to estimate [1]. However, the sole use of SOS is insufficient for performing separation of PNL mixtures, except for cases in which certain properties and/or additional *a priori* information can be exploited, as shown next.

2.1 Second-Order Statistics in PNL Mixtures

Different approaches can be used along with SOS for separation of PNL mixtures in the context of temporally correlated sources. We highlight three of them, in which the following features are exploited: the Gaussianization effect [9], the *a priori* knowledge of one of the source distributions [10] and the residual nonlinear term [11].

The first approach is based on the fact that, after the linear mixing part, the random variables tend to become ‘more’ Gaussian. Hence, ideally, $\mathbf{u}(n)$ and $\mathbf{z}(n)$ (after a successful nonlinear compensation) present distributions relatively

close to a Gaussian. In that sense, the separation task can be performed in two stages: first $\mathbf{g}(\cdot)$ is adjusted so that $\mathbf{z}(n)$ is as Gaussian as possible and, then, \mathbf{W} is adjusted according to an SOS-based method [9], such as WASOBI and TDSEP [1, 2]. This strategy reveals to be more effective when the sources are Gaussian distributed or when the number N of sources is large (in the spirit of the central limit theorem); otherwise a performance loss is observed [10]. We will refer to this method as the ‘Gaussianization-based approach’.

In the second approach, the distribution of one of the sources is assumed to be known in aid of the separation task [10] – e.g., it is known that one of the sources is Gaussian. In this case, the contrast function is a weighted composition of a distribution matching part (so that one of the recovered sources matches the known distribution type) plus the SOBI criterion, which diagonalizes the correlation matrices. Thus, $\mathbf{g}(\cdot)$ and \mathbf{W} are jointly adapted. The performance of the approach will depend on the quality of the distribution type estimation. This method will here be called the ‘One Source Matching-based approach’.

In these two approaches, the use of HOS is required either in the Gaussianization process or in the distribution matching part. However, a recent result has shown that separation of PNL mixtures is possible using only SOS if certain conditions regarding the nonlinear functions are met, allowing the removal of nonlinear residual errors [11]. The essence of this third approach – which will be of great interest for us in this work – can be posed as follows. Suppose that, for a given $\mathbf{f}(\cdot)$, $\mathbf{g}(\cdot)$ is chosen so that the PNL system output yields:

$$\mathbf{y}(n) = \mathbf{W}\mathbf{g} \circ \mathbf{f}(\mathbf{A}\mathbf{s}(n)) = \mathbf{W}\mathbf{A}\mathbf{s}(n) + \mathbf{W}\tilde{\mathbf{f}}(\mathbf{A}\mathbf{s}(n)), \quad (1)$$

which means that the composite nonlinear function $\mathbf{g} \circ \mathbf{f}(\cdot)$ results in a linear term plus a nonlinear residual term, denoted by $\tilde{\mathbf{f}}(\cdot)$ – we assume that scale and/or permutation factors are encompassed by matrix \mathbf{W} . In such a case, the time-lagged autocorrelation matrix of $\mathbf{y}(n)$, $\mathbf{R}_Y(m) = E[\mathbf{y}(n)\mathbf{y}^T(n-m)]$, where m is some lag constant, $m \in 1, \dots, d$, can be expressed as

$$\begin{aligned} \mathbf{R}_Y(m) &= \mathbf{W}\mathbf{A}E[\mathbf{s}(n)\mathbf{s}^T(n-m)]\mathbf{A}^T\mathbf{W}^T + \mathbf{W}\mathbf{A}E\left[\mathbf{s}(n)\left(\tilde{\mathbf{f}}(\mathbf{A}\mathbf{s}(n-m))\right)^T\right]\mathbf{W}^T \\ &+ \mathbf{W}E\left[\tilde{\mathbf{f}}(\mathbf{A}\mathbf{s}(n))\mathbf{s}^T(n-m)\right]\mathbf{A}^T\mathbf{W}^T \\ &+ \mathbf{W}E\left[\tilde{\mathbf{f}}(\mathbf{A}\mathbf{s}(n))\left(\tilde{\mathbf{f}}(\mathbf{A}\mathbf{s}(n-m))\right)^T\right]\mathbf{W}^T, \end{aligned} \quad (2)$$

in which either the first term or the remaining terms can be made diagonal for all considered lags [11]. If the first term (associated with the linear part) is made diagonal, then, since separation is not achieved in (1), the remaining terms present non-null off-diagonal elements for some delays [11]. Then, for $\mathbf{R}_Y(m)$ to be diagonal for all $m \in 0, \dots, d$ the nonlinear residual error must be null, ensuring a successful source separation. However, if the first term in (1) does not exist due to a different choice of $\mathbf{g}(\cdot)$, then $\mathbf{R}_Y(m)$ can be made diagonal even when the nonlinear residual error is not suppressed [11]. In this

sense, when performing source separation, $\mathbf{g}(\cdot)$ must be chosen so that the first term prevails. Once this condition is satisfied, the refinement of $\mathbf{g}(\cdot)$ (to cancel the nonlinear residual error) and the adjustment of the coefficients \mathbf{W} can be performed through a matrix diagonalization-based criterion (for $\mathbf{R}_Y(m)$), such as the SOBI criterion. Nevertheless, the authors propose an alternative criterion called *Second-Order Mutual Information - Quadratic version*, or simply, SOMIq, which seems to be more robust against local minima convergence for metaheuristic optimization [11].

In the SOMIq criterion, the temporal statistical information of the sources is jointly exploited in an extended correlation matrix (similarly to [14]). Basically, column vectors are composed of signals at time instant n concatenated with their d delayed versions in the following form:

$$\begin{aligned} \underline{\mathbf{y}}(n) &= [y_1(n) \dots y_1(n-d) \ y_2(n) \dots y_2(n-d) \dots y_N(n) \dots y_N(n-d)]^T \\ &= \left[\underline{\mathbf{y}}_1^T(n) \ \underline{\mathbf{y}}_2^T(n) \ \dots \ \underline{\mathbf{y}}_N^T(n) \right]^T, \end{aligned} \quad (3)$$

where d is the largest considered lag and $\underline{\mathbf{y}}_i(n) = [y_i(n) \dots y_i(n-d)]^T$, for $i = \{1, \dots, N\}$. From these vectors, we obtain the correlation matrices:

$$\mathbf{R}_{\underline{\mathbf{Y}}} = E \left[\underline{\mathbf{y}}(n) \underline{\mathbf{y}}^T(n) \right]; \quad \mathbf{R}_{\underline{\mathbf{Y}}_i} = E \left[\underline{\mathbf{y}}_i(n) \underline{\mathbf{y}}_i^T(n) \right]. \quad (4)$$

By combining the mutual information measure and assuming Gaussian distributed sources, it is possible to write the SOMIq criterion:

$$J_{SOMIq} = \min_{\mathbf{g}(\cdot), \mathbf{W}} \left(\prod_{i=1}^N |\mathbf{R}_{\underline{\mathbf{Y}}_i}| - |\mathbf{R}_{\underline{\mathbf{Y}}}| \right)^2, \quad (5)$$

where $|\cdot|$ is the determinant operator.

Since $\mathbf{f}(\cdot)$ is not known a priori, the main difficulty of this third approach is how to choose $\mathbf{g}(\cdot)$ so that (1) is satisfied even during adaptation. In certain cases, this task may be less difficult because $\mathbf{f}(\cdot)$ can be partially known – for instance, as occurs in chemical sensor array data, whose mixtures can be described according to the Nicolsky-Eisenman model [5]. However, as we intend to show in this paper, a wide set of nonlinear functions $\mathbf{f}(\cdot)$ can be addressed if a constrained model for $\mathbf{g}(\cdot)$ is assumed: a Taylor series expansions-based parametric model able to satisfy (1) and to preserve the nonlinear flexibility.

3 Proposal: Constrained Taylor Series Expansions

Taylor series expansion is an efficient mathematical tool that allows the representation of a function as an infinite power series. In nonlinear BSS, the expansion is particularly useful when truncated Taylor series expansions are considered for nonlinear function approximation, which, besides contributing with performance improvement, can also reveal insightful theoretical aspects [15].

Mathematically, the Taylor series of a continuous and (infinitely) differentiable function $f(u)$ around zero is the power series:

$$f(u) = \sum_{k=0}^{\infty} f^{(k)}(0) \frac{(u)^k}{k!}, \quad (6)$$

where $(\cdot)!$ denotes the factorial operator and $f^{(k)}(\cdot)$ denotes the k th derivative of $f(\cdot)$. This means that any differentiable nonlinear function $f(\cdot)$ can be decomposed into a sum of polynomial terms.

As previously mentioned, our objective is to specify $\mathbf{g}(\cdot)$ so that Eq. (1) holds – allowing us to use separation criteria solely based on SOS [11] – and the first-order term in the Taylor series expansion will be the key for achieving this goal, as shown in the following. Suppose that the PNL components of $\mathbf{f}(\cdot)$ can be represented by Taylor series expansions, as shown in Eq. (6), and that the compensating nonlinear function for the i th mixture, $g_i(\cdot)$, is a t th-order truncated Taylor series-based expansion around zero, whose terms are given by

$$g_i(x_i(n)) = g_{i,1}x_i(n) + g_{i,2}x_i^2(n) + \cdots + g_{i,t}x_i^t(n), \quad (7)$$

where $g_{i,j}$ is the j th coefficient of the i th nonlinear function. Note that the zero-order coefficient $g_{i,0}$ is disregarded, since we are assuming that $g_i(\cdot)$ and $f_i(\cdot)$ pass through the origin. Based on these assumptions, the composite $\mathbf{g} \circ \mathbf{f}(\cdot)$ gives

$$\begin{aligned} g_i(x_i(n)) &= g_i(f_i(u_i(n))) = g_{i,1}f_i(u_i(n)) + g_{i,2}f_i^2(u_i(n)) + \cdots \\ &= g_{i,1}f'_i(0)u_i(n) + \tilde{f}_i(u_i(n)), \end{aligned} \quad (8)$$

where $\tilde{f}_i(u_i(n))$ gathers the other remaining terms. Thus, by constraining $g_{i,1}$ to be non-null and assuming that $f'_i(0) \neq 0$, we are able to split the PNL system output into a linear and a nonlinear term, satisfying (1) (recall that $\mathbf{u}(n) = \mathbf{A}\mathbf{s}(n)$ and that the scale factor can be compensated by \mathbf{W}). Therefore, $\mathbf{g}(\cdot)$ must be of sufficient order but constrained with $g_{i,1}$ equal to a non-null constant, for $i = 1, \dots, M$. Note that the other coefficients of $g_i(\cdot)$ associated with nonlinear terms are not constrained, preserving the nonlinear flexibility.

Interestingly, this Taylor series-based approach can be applied to a wide set of nonlinear functions since $f_i(\cdot)$ and $g_i(\cdot)$ are required to be both monotonic functions in PNL mixtures [13], thus presenting odd-order terms, usually including the first one. Exceptions exist, however, when $f_i(\cdot)$ is a polynomial whose Taylor series expansion does not encompass the first-order term – in this case, another approach must be followed for $g_i(\cdot)$, but this will not be considered in the present paper. In the following, we consider some simulation analysis when applying the constraint $g_{i,1} = 1$ and using the SOMIq criterion to adapt the other coefficients.

4 Simulation Results

In order to analyze the efficiency of the proposed approach, the compensating nonlinear function $\mathbf{g}(\cdot)$ is chosen to be as given by Eq. (7) – truncated at the 7th-order – considering: (i) the constrained case, where $g_{i,1} = 1$ remains fixed and the

other coefficients $g_{i,j}$ are allowed to vary freely, for $i = 1, \dots, M$ and $j = 2, \dots, 7$; and (ii) the unconstrained case. Along with $\mathbf{g}(\cdot)$, the linear separating matrix \mathbf{W} is adapted according to the SOMIq criterion, given by Eq. (5). Our objective is to evaluate the performance of these constrained and unconstrained cases and to compare them with the performance of the other two aforementioned SOS-based approaches – i.e., the Gaussianization-based [9] and the One Source Matching-based approach [10]. For the Gaussianization approach, in a first stage, $\mathbf{g}(\cdot)$ – unconstrained – will be adapted so that each i th output $z_i(n)$ has a null kurtosis (meaning that $z_i(n)$ is Gaussian distributed) and, in a second stage, \mathbf{W} will be adapted according to the SOBI criterion [3]. For the One Source Matching-based approach, it will be assumed that at least one of the sources is Gaussian distributed, being $\mathbf{g}(\cdot)$ – unconstrained – and \mathbf{W} jointly adapted so that $y_1(n)$ presents null kurtosis and that the time-lagged correlation matrices $\mathbf{R}_Y(m)$, for $m = 0, \dots, d$, be diagonal via the SOBI criterion.

The coefficient adjustment will be performed with the aid of the metaheuristic known as Differential Evolution (DE), a population-based technique that efficiently exploits the search space using the information contained in the population of solutions instead of the usual random operators [12]. The DE parameters are: the population size N_P , the crossover constant CR , the adaptation step F and the number of iterations N_{it} . After training, the performance of the best individual in the population shall be evaluated for a test set of 700000 samples and measured in terms of Signal-to-Interference Ratio (SIR), defined as $\text{SIR} = 10 \log (E[y_i^2(n)]/E[(s_i(n) - y_i(n))^2])$, after sign and variance correction.

Two scenarios are considered: in the first one, we are mainly interested in the constrained Taylor series-based approach effectiveness and in its performance; in the second one, we investigate how the number of samples impacts on the performance. In both scenarios, we consider two sources and two mixtures. In scenario 1, one of the sources is a temporally correlated Gaussian and the other one is a sequence with trapezoidal distribution between $[-1, 1]$, as shown in Fig. 3(a), the mixing system is given by $\mathbf{A} = [0.450 \ -0.551; -0.683 \ 0.317]$, $f_1(u_1(n)) = \arcsin(u_1(n))$ and $f_2(u_2(n)) = \text{arctanh}(u_2(n))$, with $-1 \leq u(n) \leq +1$. The separating system is composed of two 7th-order polynomials with coefficients $g_{1,1}, g_{1,2}, \dots, g_{1,7}, g_{2,1}, g_{2,2}, \dots, g_{2,7}$ – according to Eq. (7) – (recall that $g_{1,1} = g_{2,1} = 1$ for the constrained case) and a 2-by-2 separating matrix \mathbf{W} , with four coefficients. 250000 samples of the mixtures are available for statistics estimation. In scenario 2, the sources are two temporally correlated Gaussians. The elements of the mixing system are $\mathbf{A} = [0.667 \ 0.333; 0.445 \ -0.555]$, $f_1(u_1(n)) = \arcsin(u_1(n))$ and $f_2(u_2(n)) = \arcsin(u_2(n))$, with $-1 \leq u(n) \leq +1$. We adopt again a 2-by-2 separating matrix \mathbf{W} and two 7th-order polynomial for $\mathbf{g}(\cdot)$, but, since the nonlinear mixing functions are equivalent ($f_1(\cdot) = f_2(\cdot) = \arcsin(\cdot)$), we assume that $g_{1,j} = g_{2,j}$, for $j = 1, \dots, 7$, thus, reducing the search space. The number of mixture samples will vary from 500 up to 500000 in this scenario. The considered number of delays for SOBI and SOMIq is $d = 8$ in both scenarios.

To obtain a high global convergence rate, the DE parameters were empirically chosen to be $N_P = 500$, $F = 0.5$, $CR = 0.9$ and $N_{it} = 7000$ for scenario 1, and

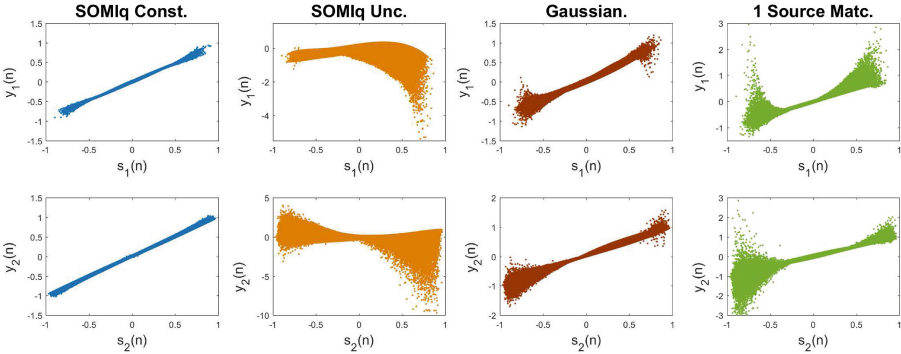


Fig. 2. Scatter plots of the sources and of the outputs for each method in scenario 1.

$N_P = 300$, $F = 0.5$, $CR = 0.9$ and $N_{it} = 2000$ for scenario 2 (smaller search space). For the Gaussianization approach, two DE runs are necessary, one for the nonlinear and the other for the linear stage.

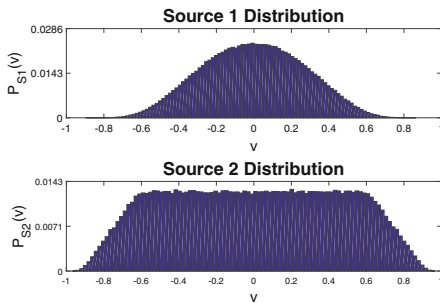
Figure 2 shows $y(n)$ versus $s(n)$ plots – where a diagonal line means that a perfect separation was achieved – considering the outputs of the constrained and unconstrained SOMIq, of the Gaussianization-based and of the One Source Matching-based approach, from the left to the right. It is possible to note that the constrained SOMIq output tends to a diagonal line, being able to recover the sources (with small noise), while the unconstrained case completely fails, indicating that the constrained Taylor Series-based approach can be an efficient method for separation. The other two methods are also able to separate the sources with certain noise. Remark that they require the use of HOS and demand certain knowledge of the sources.

For 20 Monte Carlo simulations, the average SIR performance for scenario 1 is as shown in Table 1, where it is possible to note that the constrained SOMIq approach is able to achieve the best performance, probably due to its simpler SOS-based construction. The unconstrained SOMIq case exposes the fact that the sole use of the SOS is insufficient for nonlinear separation in this case. The Gaussianization-based and the One Source Matching-based approaches present a good performance, but the two-stage adaptation of the Gaussianization-based approach reduces the search space and contributes to a higher global convergence rate. Although no monotonicity constraint over $\mathbf{g}(\cdot)$ was applied, the solutions found for the constrained SOMIq case resulted in monotonic $\mathbf{g}(\cdot)$.

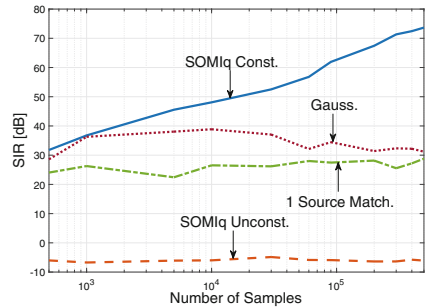
Table 1. Performance in terms of SIR [dB] for scenario 1

Sources	Constrains.SOMIq	Unconstr.SOMIq	Gaussianization approach	One source matching
1	47.04	-5.58	37.63	30.86
2	55.43	-3.92	41.75	32.63

For scenario 2, Fig. 3(b) displays the average SIR performance of the considered methods for 20 Monte Carlo simulations as a function of the number of samples (in logarithmic scale). The SIR performance of the constrained SOMIq is improved as the number of samples increases, being able to achieve almost 50 dB with 10000 samples and 70 dB with 300000. On the other hand, the unconstrained SOMIq fails independently of the number of samples, which confirms the necessity of holding the condition given by (1). A number of samples larger than 1000 seems to cause minor effects on the performance of the Gaussianization-based and the One Source Matching-based approaches, which vary around 35 dB and 27 dB, respectively. For all number of samples considered, the constrained SOMIq obtained the best average SIR performance.



(a) Source distributions (Scenario 1).



(b) SIR performance as a function of the number of samples (Scenario 2).

Fig. 3. Source distributions for scenario 1 and SIR performance for scenario 2.

5 Conclusions

In this work, we have proposed a constrained Taylor series approach for SOS-based source separation in the PNL model. The method is based on the idea that any nonlinear residual error can be suppressed by SOS-based separation methods if a linear mixture term always exists in the separation process. Interestingly, by defining the separating nonlinear function as a constrained Taylor series-based expansion (whose first-order term is fixed), the linear mixture term is kept and source separation can be successfully performed with the aid of an SOS-based criterion named SOMIq. This method assumes that the sources are temporally colored, but differently from the other SOS-based methods for separation of PNL mixtures, no other prior information is necessary nor the use of HOS. The result is a simple and robust nonlinear separation method. Along with the use of the DE metaheuristic for coefficient adaptation, the simulations indicated that the proposed method is able to outperform the concurrent SOS-based methods in the chosen scenarios, although a reasonable amount of data might be necessary to provide reliable SOS estimates. For future works, we consider the proposition

of a gradient-based algorithm and the analysis in scenarios with a higher number of sources.

Acknowledgements. This work was partly supported by FAPESP (2017/11488-5), CNPq (305621/2015-7) and ERC project 2012-ERC-AdG-320684 CHESS.

References

1. Comon, P., Jutten, C.: *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Cambridge (2010)
2. Romano, J.M.T., Attux, R., Cavalcante, C.C., Suyama, R.: *Unsupervised Signal Processing: Channel Equalization and Source Separation*. CRC Press, Boca Raton (2010)
3. Belouchrani, A., Abed-Meraim, K., Cardoso, J.F., Moulines, E.: A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* **45**(2), 434–444 (1997)
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: 19th IEEE European Signal Processing Conference, pp. 1859–1863 (2011)
5. Duarte, L.T., Jutten, C., Moussaoui, S.: A Bayesian nonlinear source separation method for smart ion-selective electrode arrays. *IEEE Sens. J.* **9**(12), 1763–1771 (2009)
6. Hosseini, S., Jutten, C.: On the separability of nonlinear mixtures of temporally correlated sources. *IEEE Signal Process. Lett.* **10**(2), 43–46 (2003)
7. Taleb, A., Jutten, C.: Source separation in post-nonlinear mixtures. *IEEE Trans. Signal Process.* **47**(10), 2807–2820 (1999)
8. Deville, Y., Duarte, L.T.: An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) *LVA/ICA 2015*. LNCS, vol. 9237, pp. 155–167. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_18
9. Ziehe, A., Kawanabe, M., Harmeling, S., Müller, K.: Blind Separation of post-nonlinear mixtures using Gaussianizing transformations and temporal decorrelation. *J. Mach. Learn. Res.* **4**, 1319–1338 (2003)
10. Fantinato, D.G., Duarte, L.T., Zanini, P., Rivet, B., Attux, R., Jutten, C.: A joint second-order statistics and density matching-based approach for separation of post-nonlinear mixtures. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) *LVA/ICA 2017*. LNCS, vol. 10169, pp. 499–508. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_47
11. Fantinato, D., Duarte, L.T., Deville, Y., Attux, R., Jutten, C., Neves, A.: A Second-Order Statistics Method for Blind Source Separation in Post-Nonlinear Mixtures. *Signal Processing Elsevier* (2018, Submitted)
12. Price, K., Storn, R., Lampinen, J.: *Differential Evolution: A Practical Approach to Global Optimization*. Springer, Heidelberg (2005). <https://doi.org/10.1007/3-540-31306-0>
13. Achard, S., Jutten, C.: Identifiability of post-nonlinear mixtures. *IEEE Sig. Process. Lett.* **12**(5), 423–426 (2005)

14. Buchner, H., Aichner, R., Kellermann, W.: A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Process.* **13**(1), 120–134 (2005)
15. Drumetz, L., Ehsandoust, B., Chanussot, J., Rivet, B., Babaie-Zadeh, M., Jutten, C.: Relationships between nonlinear and space-variant linear models in hyperspectral image unmixing. *IEEE Signal Process. Lett.* **24**(10), 1567–1571 (2017)



New Classes of Blind Quantum Source Separation and Process Tomography Methods Based on Spin Component Measurements Along Two Directions

Yannick Deville¹(✉) and Alain Deville²

¹ IRAP (Institut de Recherche en Astrophysique et Planétologie), Université de Toulouse, UPS, CNRS, CNES, 14 avenue Edouard Belin, 31400 Toulouse, France

yannick.deville@irap.omp.eu

² IM2NP UMR 7334, Aix-Marseille Université, CNRS, 13397 Marseille, France

alain.deville@univ-amu.fr

Abstract. We here present major extensions of the fields of blind quantum source separation (BQSS) and blind quantum process tomography (BQPT) for the Heisenberg Hamiltonian. They are based on a new type of spin component measurements performed directly for the available quantum states, which yields new nonlinear mixing models, with extended source signals and mixing parameters. The first two types of proposed BQSS and/or BQPT methods are based on quantum-source independent component analysis. They therefore require typically one thousand quantum states to estimate the mixing parameters, and some of them yield closed-form solutions. We then define a complementary, inversion-based, BQSS/BQPT method which requires only one quantum state, but which is based on solving nonlinear equations numerically.

Keywords: Blind quantum source separation
Blind system identification · Blind quantum process tomography
Nonlinear mixtures · Independent component analysis
Heisenberg Hamiltonian identification

1 Prior Work and Problem Statement

The general framework of latent variable analysis includes two closely related problems, namely system identification and system inversion, especially in their blind version (unknown input values for the considered system) [1]. For classical, i.e. non-quantum, signals and systems, these problems have been extensively studied, not only in the single-input single-output case, but also in multiple-input multiple-output configurations, where blind system inversion and blind source separation (BSS) are almost the same problem, but indeterminacies (i.e. some residual transforms in restored signals) are accepted in BSS: see e.g. [3].

Let us now consider quantum “signals” (i.e. states) and systems. Then, among the above problems, the one which was first studied is non-blind system

identification, called “quantum process tomography” by the Quantum Information Processing community: see e.g. [2, 11–14]. Besides, we introduced the field of “quantum source separation” (QSS) and especially its blind version (BQSS) in the 2007 edition of this (LVA/)ICA conference: see [4] and subsequent references below. Moreover, we recently developed connections between BQSS and system identification, thus introducing the field of “blind quantum process tomography” (BQPT) at the 2015 LVA/ICA conference, and focusing on Heisenberg Hamiltonian identification: see especially [7, 9].

More precisely, the first class of BQSS and BQPT methods that we developed has the following features (see especially [5, 6]): (i) these methods first derive classical-form data from the directly available quantum states, by means of measurements of the components of associated spin operators along a single direction, (ii) this only allows them to estimate part of the “source and mixing parameters” defined below, (iii) to estimate the considered mixing parameter, they process the above classical-form data with quantum-source independent component analysis (QSICA) algorithms [6], so that these identification methods typically require 10^7 quantum state preparations (10^3 different states, each prepared 10^4 times, as explained hereafter in Sect. 3).

We then proposed a second, quite different, class of BQSS and BQPT methods which have the following features (see especially [10]): (i) their standard versions process data in quantum form, except that they estimate all unknown mixing parameters by using measurements of the components of output spin operators along two directions (a single direction was shown not to be sufficient for these methods, e.g. in [10]), (ii) they estimate all source parameters, (iii) to estimate the mixing parameters, they process the available quantum states with methods based on disentangling a few output quantum states of the separating system (instead of using QSICA), so that these identification methods typically require 6×10^4 state preparations (3 different states, each prepared 10^4 times, and with measurements along 2 directions) when using advanced algorithms [8].

Our second class of methods thus has various attractive features, as compared with the first class. One may wonder whether this results from using quantum-form data, spin components along *two* directions or disentanglement-based adaptation. The present paper therefore aims at answering this question. To this end, after summarizing the required information about the considered physical configuration (see Sect. 2) and about our first class of methods (Sect. 3), we extend that approach to measurements of spin components along two directions, thus deriving a new, nonlinear, mixing model, a new class of BQSS methods (Sect. 4) and a new class of BQPT methods (Sect. 5), which have complementary features. We then propose an inversion-based joint BQSS/BQPT method which requires far fewer quantum state preparations (Sect. 6).

2 Quantum States and Heisenberg Coupling Model

Computations in the field of Quantum Information Processing use quantum bits, also called qubits, instead of classical bits [12]. A qubit with index i considered

at a given time t_0 has a quantum state. If this state is pure, it belongs to a two-dimensional space \mathcal{E}_i and may be expressed as

$$|\psi_i(t_0)\rangle = \alpha_i|+\rangle + \beta_i|-\rangle \quad (1)$$

in the basis of \mathcal{E}_i defined by the two orthonormal vectors that we hereafter denote as $|+\rangle$ and $|-\rangle$, whereas α_i and β_i are two complex-valued coefficients constrained to be such that the state $|\psi_i(t_0)\rangle$ is normalized (that is, $|\alpha_i|^2 + |\beta_i|^2 = 1$).

In the BQSS and BQPT configurations studied in this paper, we first consider a system composed of two qubits, called “qubit 1” and “qubit 2” hereafter, at a given time t_0 . This system has a quantum state. If this state is pure, it belongs to the four-dimensional space \mathcal{E} defined as the tensor product (denoted \otimes) of the spaces \mathcal{E}_1 and \mathcal{E}_2 respectively associated with qubits 1 and 2, i.e. $\mathcal{E} = \mathcal{E}_1 \otimes \mathcal{E}_2$. We hereafter denote as \mathcal{B}_+ the basis of \mathcal{E} composed of the four orthonormal vectors $|++\rangle, |+-\rangle, |-+\rangle, |--\rangle$, where e.g. $|+-\rangle$ is an abbreviation for $|+\rangle \otimes |-\rangle$, with $|+\rangle$ corresponding to qubit 1 and $|-\rangle$ corresponding to qubit 2. In particular, we study the case when the two qubits are independently initialized, with states defined by (1) respectively with $i = 1$ and $i = 2$. The state of the two-qubit system at the initial time t_0 then reads

$$\begin{aligned} |\psi(t_0)\rangle &= |\psi_1(t_0)\rangle \otimes |\psi_2(t_0)\rangle \quad (2) \\ &= \alpha_1\alpha_2|++\rangle + \alpha_1\beta_2|+-\rangle + \beta_1\alpha_2|-+\rangle + \beta_1\beta_2|--\rangle. \quad (3) \end{aligned}$$

Besides, we consider the case when the two qubits correspond to two distinguishable [10] spins 1/2 which have undesired coupling after they have been initialized according to (2). The considered coupling is based on the Heisenberg model with a cylindrical-symmetry axis presently collinear to the applied magnetic field, which has a magnitude B . This common axis is chosen as the “quantization axis”, called Oz . Moreover, we assume an isotropic \bar{g} tensor, with principal value g . The time interval when these spins are considered is supposed to be short enough for their coupling with their environment to be negligible. In these conditions, the temporal evolution of the quantum state of the system composed of these two spins is governed by the following effective [10] Hamiltonian:

$$H = Gs_{1z}B + Gs_{2z}B - 2J_{xy}(s_{1x}s_{2x} + s_{1y}s_{2y}) - 2J_zs_{1z}s_{2z} \quad (4)$$

where:

- $G = g\mu_e$, where μ_e is the Bohr magneton, i.e. $\mu_e = e\hbar/2m_e = 0.927 \times 10^{-23} JT^{-1}$ and \hbar is the reduced Planck constant.
- s_{ix} , s_{iy} and s_{iz} , with $i \in \{1, 2\}$, are the three components of the vector operator \vec{s}_i associated with spin i in a cartesian frame. Besides, the above-mentioned vectors $|+\rangle$ and $|-\rangle$ corresponding to spin i are eigenvectors of s_{iz} , for the eigenvalues 1/2 and $-1/2$ respectively.
- J_{xy} and J_z are the principal values of the exchange tensor.

Among the above parameters, the value of g may be experimentally determined, and B can be measured. The values of J_{xy} and J_z are here assumed to be unknown.

The final quantum state $|\psi(t)\rangle$ of the above two-spin system at an arbitrary time $t > t_0$, which results from the above Hamiltonian, was derived in [5], and its expression is provided in Sect. 4.1. We hereafter consider the results of measurement procedures applied to that state $|\psi(t)\rangle$.

3 BQSS and BQPT with Monodirectional Measurements: Achievements and Limitations

The first class of BQSS methods that we especially detailed in [6] consists of (i) first deriving classical-form data from the above-defined final quantum state $|\psi(t)\rangle$, by measuring the components of the considered two spins along direction Oz , and then (ii) processing the above classical-form data with original statistical methods which have relationships with classical ICA. More precisely, the result of each measurement of the above couple of spin components has four possible values, whose probabilities are denoted as p_{1zz} to p_{4zz} hereafter. Besides, we use the polar representation of the qubit parameters α_i and β_i , which reads

$$\alpha_i = r_i e^{i\theta_i} \quad \beta_i = q_i e^{i\phi_i} \quad i \in \{1, 2\} \quad (5)$$

where i is the imaginary unit, and with $0 \leq r_i \leq 1$ and

$$q_i = \sqrt{1 - r_i^2} \quad (6)$$

because $|\psi(t_0)\rangle$ is normalized. The above probabilities then read [6]

$$p_{1zz} = r_1^2 r_2^2 \quad (7)$$

$$p_{2zz} = r_1^2(1 - r_2^2)(1 - v^2) + (1 - r_1^2)r_2^2 v^2 - 2r_1 r_2 \sqrt{1 - r_1^2} \sqrt{1 - r_2^2} \sqrt{1 - v^2} v \sin \Delta_I \quad (8)$$

$$p_{4zz} = (1 - r_1^2)(1 - r_2^2) \quad (9)$$

with

$$\Delta_I = (\phi_2 - \theta_2) - (\phi_1 - \theta_1) \quad (10)$$

$$\Delta_E = -\frac{J_{xy}(t - t_0)}{\hbar} \quad (11)$$

$$v = \text{sgn}(\cos \Delta_E) \sin \Delta_E. \quad (12)$$

Probability p_{3zz} is redundant with the above ones, because

$$p_{1zz} + p_{2zz} + p_{3zz} + p_{4zz} = 1. \quad (13)$$

p_{3zz} is therefore not considered.

In BSS terms, the mixing model (7)–(9) involves three “observed signals”, namely p_{1zz} , p_{2zz} and p_{4zz} . Besides, they depend on only *three* unknown “source signals”, namely r_1 , r_2 and Δ_I , which is an attractive feature because this allows

the mixing transform (7)–(9) applied to these source signals to be invertible over a bounded domain of source values if the fixed value of v is such that $0 < v^2 < 1$ (see details in [5,6]). It should be noted that this mixing model is nonlinear.

When developing our first class of BQSS methods, we restricted ourselves to the above view of source and observed signals. We here have to further analyze them as follows. Equations (3) and (5), which define the unknown initial state $|\psi(t_0)\rangle$, involve 8 polar parameters. However, only 4 of them should be considered (thus defining 4 classical-form “potential source signals”), namely $r_1, r_2, (\phi_1 - \theta_1)$ and $(\phi_2 - \theta_2)$, for the following two reasons. First, the parameters q_i are redundant with r_i , due to (6). Second, a global phase factor in any quantum state $|\psi_i(t_0)\rangle$ has no physical consequence, so that using (5) and rewriting (1) as

$$|\psi_i(t_0)\rangle = e^{i\theta_i} \left(r_i|+\rangle + q_i e^{i(\phi_i - \theta_i)}|-\rangle \right) \quad (14)$$

shows that the state $|\psi_i(t_0)\rangle$ only involves a single relevant phase parameter, that is, the phase difference $(\phi_i - \theta_i)$. This means that, by using only a single type of spin measurements, our first class of BQSS methods is able to estimate only *part of* the (relevant) parameters of the unknown initial quantum state $|\psi(t_0)\rangle$: it cannot separately estimate each of the phase differences $(\phi_1 - \theta_1)$ and $(\phi_2 - \theta_2)$, but only Δ_I , which is the difference between them, as shown by (10). One of the goals of this paper is to determine whether $(\phi_1 - \theta_1)$ and $(\phi_2 - \theta_2)$ can be separately estimated by also performing other types of spin component measurements for (other preparations of) the state $|\psi(t)\rangle$.

The above mixing model (7)–(9) involves a single, unknown, mixing parameter, namely v . In [6], we presented various statistical QSICA methods which allow one to estimate v from a set of values of the triplet $(p_{1zz}, p_{2zz}, p_{4zz})$ of observed signals (this set typically consists of 10^3 values of this triplet, and each value corresponds to one quantum state $|\psi(t_0)\rangle$ and is estimated from typically 10^4 preparations of $|\psi(t_0)\rangle$). Several of these QSICA methods yield closed-form expressions for the estimates of v . As shown by (11)–(12), estimating v yields an estimate of only one of the unknown physical parameters of the considered Hamiltonian, namely J_{xy} (and up to some indeterminacies). The corresponding class of BQPT methods, briefly introduced in [7], therefore only achieves *partial* Hamiltonian estimation, since it cannot estimate J_z . In this paper, we therefore also aim at investigating whether J_z can be estimated with this type of methods, by performing additional types of spin component measurements for (other preparations of) the state $|\psi(t)\rangle$.

4 New BQSS Methods, with Bidirectional Measurements

4.1 New Mixing Model

We here consider (additional preparations of) the above-defined two-spin final state $|\psi(t)\rangle$. Since we here aim at investigating measurements of the component of each spin along the Ox axis, we have to express that state $|\psi(t)\rangle$ in the basis

\mathcal{B}_{+x} of the above-defined space \mathcal{E} composed of the four orthonormal vectors $|+x+x\rangle, |+x-x\rangle, |-x+x\rangle, |-x-x\rangle$ where e.g. $|+x-x\rangle$ is an abbreviation for $|+x\rangle \otimes |-x\rangle$, with $|+x\rangle$ corresponding to spin 1 and $|-x\rangle$ corresponding to spin 2. In these expressions, the vectors $|+x\rangle$ and $|-x\rangle$ corresponding to spin i , with $i \in \{1, 2\}$, are eigenvectors of s_{ix} , for the eigenvalues $1/2$ and $-1/2$ respectively.

To derive the required expression of $|\psi(t)\rangle$ in basis \mathcal{B}_{+x} , we start from its expression in the basis composed of the eigenvectors of the matrix representing the Hamiltonian H in the \mathcal{B}_+ basis. These eigenvectors read [5]

$$|1, 1\rangle = |++\rangle, \quad |1, 0\rangle = \frac{|+-\rangle + |-+\rangle}{\sqrt{2}} \quad (15)$$

$$|0, 0\rangle = \frac{|+-\rangle - |-+\rangle}{\sqrt{2}}, \quad |1, -1\rangle = |--\rangle. \quad (16)$$

In [5], we showed that $|\psi(t)\rangle$ may be expressed with respect to them as

$$\begin{aligned} |\psi(t)\rangle &= \alpha_1\alpha_2 e^{-i\omega_{1,1}(t-t_0)}|1, 1\rangle + \frac{\alpha_1\beta_2 + \beta_1\alpha_2}{\sqrt{2}} e^{-i\omega_{1,0}(t-t_0)}|1, 0\rangle \\ &+ \frac{\alpha_1\beta_2 - \beta_1\alpha_2}{\sqrt{2}} e^{-i\omega_{0,0}(t-t_0)}|0, 0\rangle + \beta_1\beta_2 e^{-i\omega_{1,-1}(t-t_0)}|1, -1\rangle \end{aligned} \quad (17)$$

where all $\omega_{k,l}$ are angular frequencies defined in [5]. Moreover, a well-known result of quantum physics is the link between the eigenvectors of s_{iz} and s_{ix} :

$$|+\rangle = \frac{|+x\rangle + |-x\rangle}{\sqrt{2}}, \quad |-\rangle = \frac{|+x\rangle - |-x\rangle}{\sqrt{2}}. \quad (18)$$

Combining all above equations yields the required expression of $|\psi(t)\rangle$:

$$|\psi(t)\rangle = \sum_{j=1}^4 c_{jx}(t-t_0)|b_{jx}\rangle \quad (19)$$

where $|b_{jx}\rangle$ stands for the above-defined vectors composing basis \mathcal{B}_{+x} (in the same order as above), and the corresponding coefficients read

$$\begin{aligned} c_{1x}(t-t_0) &= \frac{1}{2} e^{-i\omega_{1,1}(t-t_0)}(T_1 + T_4), & c_{2x}(t-t_0) &= \frac{1}{2} e^{-i\omega_{1,1}(t-t_0)}(T_2 - T_3) \\ c_{3x}(t-t_0) &= \frac{1}{2} e^{-i\omega_{1,1}(t-t_0)}(T_2 + T_3), & c_{4x}(t-t_0) &= \frac{1}{2} e^{-i\omega_{1,1}(t-t_0)}(T_1 - T_4) \end{aligned}$$

where

$$\begin{aligned} T_1 &= \alpha_1\alpha_2 + \beta_1\beta_2 e^{i(\omega_{1,1}-\omega_{1,-1})(t-t_0)}, & T_4 &= (\alpha_1\beta_2 + \beta_1\alpha_2) e^{i(\omega_{1,1}-\omega_{1,0})(t-t_0)} \\ T_2 &= \alpha_1\alpha_2 - \beta_1\beta_2 e^{i(\omega_{1,1}-\omega_{1,-1})(t-t_0)}, & T_3 &= (\alpha_1\beta_2 - \beta_1\alpha_2) e^{i(\omega_{1,1}-\omega_{0,0})(t-t_0)}. \end{aligned}$$

When measuring the couple of spin components (s_{1x}, s_{2x}) , the obtained couple of values is equal to one of its four possible values, that is $(+\frac{1}{2}, +\frac{1}{2})$, $(+\frac{1}{2}, -\frac{1}{2})$,

$(-\frac{1}{2}, +\frac{1}{2})$ or $(-\frac{1}{2}, -\frac{1}{2})$ in normalized units. The probabilities of these four values are respectively denoted as p_{1xx} , p_{2xx} , p_{3xx} and p_{4xx} hereafter. They are equal to the squared moduli of the coefficients $c_{jx}(t-t_0)$ of the corresponding vectors $|\pm x \pm x\rangle$ in the state expression (19), that is

$$\begin{aligned} p_{1xx} &= |c_{1x}(t-t_0)|^2 = \frac{1}{4}|T_1 + T_4|^2, & p_{2xx} &= |c_{2x}(t-t_0)|^2 = \frac{1}{4}|T_2 - T_3|^2 \\ p_{3xx} &= |c_{3x}(t-t_0)|^2 = \frac{1}{4}|T_2 + T_3|^2, & p_{4xx} &= |c_{4x}(t-t_0)|^2 = \frac{1}{4}|T_1 - T_4|^2. \end{aligned}$$

Expressing T_1 to T_4 with respect to the polar parameters of $|\psi(t_0)\rangle$, defined in (5), yields the final expressions of p_{1xx} to p_{4xx} . Here again, only three of these probabilities or of their combinations need to be considered, because their sum is equal to one. Together with p_{1zz} , p_{2zz} , p_{4zz} , this yields 6 non-redundant classical-form “observed signals” in BSS terms. These signals again potentially depend on all 4 above-defined relevant parameters of the unknown initial state $|\psi(t_0)\rangle$, which are then the 4 “source signals” of this new mixing model. This model is therefore potentially invertible, at least over bounded intervals of the source signals. This extended mixing model is nonlinear and original. It involves 2 unknown mixing parameters (through all four $\omega_{k,l}$ [5]), namely J_{xy} and J_z .

As explained e.g. in [4–6], for any given state $|\psi(t_0)\rangle$, the observed signals available in practice for our previous BQSS methods are not the exact probabilities p_{1zz} to p_{4zz} , but their estimates obtained with our Repeated Write/Read (RWR) procedure, which consists of (i) repeatedly preparing $|\psi(t_0)\rangle$, deriving $|\psi(t)\rangle$, performing measurements, and (ii) then deriving the sample frequencies of all possible measured values. The same principle here applies to p_{1xx} to p_{4xx} .

4.2 BQSS Methods

Here focusing on the BQSS task, as opposed to BQPT, the methods of Sect. 3 already allow us to extract the signals r_1 , r_2 and Δ_I , and what we still have to do is to develop a method for then separately extracting $(\phi_1 - \theta_1)$ and $(\phi_2 - \theta_2)$, by also using p_{1xx} to p_{4xx} . Although each of these probabilities alone has a complicated expression, an analysis of their structure and calculations showed us that a particularly attractive BQSS method may be developed by considering their following combination:

$$p_{1xx} + p_{4xx} = \frac{1}{2} (|T_1|^2 + |T_4|^2) \quad (20)$$

$$\begin{aligned} &= \frac{1}{2} + r_1 r_2 \sqrt{1 - r_1^2} \sqrt{1 - r_2^2} [\cos \Delta_I \\ &\quad + \cos((\phi_1 - \theta_1) + (\phi_2 - \theta_2) - \Delta\Phi_{1,-1})] \end{aligned} \quad (21)$$

where the expressions of $\omega_{1,1}$ and $\omega_{1,-1}$ in [5] show that

$$\Delta\Phi_{1,-1} = (\omega_{1,-1} - \omega_{1,1})(t - t_0) = -\frac{2GB(t - t_0)}{\hbar} \quad (22)$$

and its value is therefore known in the considered configuration. The quantity $p_{1xx} + p_{4xx}$ is therefore attractive for BQSS because (unlike p_{1xx} and p_{4xx}) it does not depend on the unknown mixing parameters J_{xy} and J_z ! This allows us to build various complete BQSS methods which operate as follows:

- First, the “adaptation phase” essentially aims at estimating the required mixing parameters, so as to fix the transform performed by the corresponding separating system [10]. To this end, we consider a set (typically 10^3 : see above) of (repeatedly prepared) states $|\psi(t_0)\rangle$, which yield the associated available states $|\psi(t)\rangle$, that we use to estimate the mixing parameter v (related to J_{xy}) of (8). To this end, we apply any of the methods based on measurements of Oz spin components which were considered in Sect. 3.
- Then, in the “inversion phase” [10], for each (repeatedly prepared) available state $|\psi(t)\rangle$, we aim at restoring the associated four source signals which define the corresponding state $|\psi(t_0)\rangle$. To this end, we first use the separating structure corresponding to the methods of Sect. 3, e.g. described in [6]. This yields estimates of r_1 , r_2 and Δ_I (the indeterminacies may be removed as explained in [6]). We then use Ox measurements and invert (21)–(22) as

$$(\phi_1 - \theta_1) + (\phi_2 - \theta_2) = \pm \arccos \left(\frac{p_{1xx} + p_{4xx} - \frac{1}{2}}{r_1 r_2 \sqrt{1 - r_1^2} \sqrt{1 - r_2^2}} - \cos \Delta_I \right) - \frac{2GB(t - t_0)}{\hbar} + 2k\pi \quad (23)$$

where one may reduce the \pm and $2k\pi$ indeterminacies by using bounded intervals for the considered quantities, as in [6]. Computing the (half) difference and sum of (23) and (10) eventually yields $(\phi_1 - \theta_1)$ and $(\phi_2 - \theta_2)$.

It should be noted that extending BQSS methods to measurements along the Ox axis thus does not increase “complexity” (in terms of the number of quantum state preparations during the adaptation phase), as compared to our previous methods restricted to measurements along the Oz axis, especially because we succeed in achieving this extended BQSS without having to estimate J_z in addition. The counterpart of not estimating J_z is of course that this class of BQSS methods do not achieve *complete* BQPT/Hamiltonian identification. Other methods are therefore proposed hereafter to this end.

5 New BQPT Methods, with Bidirectional Measurements

We here again only use the available data defined in Sect. 4.1. Additional calculations yield the following expressions:

$$p_{1xx} - p_{4xx} = \Re(T_1 T_4^*) = R_{14} \cos \Delta\Phi_{1,0} - I_{14} \sin \Delta\Phi_{1,0} \quad (24)$$

where $\Re(\cdot)$ stands for real part, $*$ stands for complex conjugate, and where (again using [5]) to transform $\omega_{1,1}$ and $\omega_{1,0}$)

$$\begin{aligned} R_{14} &= r_1^2 r_2 \sqrt{1 - r_2^2} \cos(\phi_2 - \theta_2) + r_2^2 r_1 \sqrt{1 - r_1^2} \cos(\phi_1 - \theta_1) \\ &\quad + (1 - r_1^2) r_2 \sqrt{1 - r_2^2} \cos(\phi_2 - \theta_2 - \Delta\Phi_{1,-1}) \\ &\quad + (1 - r_2^2) r_1 \sqrt{1 - r_1^2} \cos(\phi_1 - \theta_1 - \Delta\Phi_{1,-1}) \end{aligned} \quad (25)$$

$$\begin{aligned} I_{14} &= -r_1^2 r_2 \sqrt{1 - r_2^2} \sin(\phi_2 - \theta_2) - r_2^2 r_1 \sqrt{1 - r_1^2} \sin(\phi_1 - \theta_1) \\ &\quad + (1 - r_1^2) r_2 \sqrt{1 - r_2^2} \sin(\phi_2 - \theta_2 - \Delta\Phi_{1,-1}) \\ &\quad + (1 - r_2^2) r_1 \sqrt{1 - r_1^2} \sin(\phi_1 - \theta_1 - \Delta\Phi_{1,-1}) \end{aligned} \quad (26)$$

$$\Delta\Phi_{1,0} = (\omega_{1,0} - \omega_{1,1})(t - t_0) = \frac{(t - t_0)}{\hbar} (-J_{xy} + J_z - GB). \quad (27)$$

This allows us to propose the class of BQPT methods operating as follows. First, we again use the adaptation phase of one of the methods of Sect. 3 to estimate v and hence J_{xy} (up to indeterminacies). We then only need a *single* state $|\psi(t)\rangle$ and we use the method of the inversion phase of Sect. 4.2 to estimate all four above-defined source signals. Using (22), Eq. (24) then has a single unknown, that is $\Delta\Phi_{1,0}$, and it has the standard form $a \cos x + b \sin x = c$, so that it is easily solved. This yields $\Delta\Phi_{1,0}$ (up to some indeterminacies) and hence J_z thanks to (27).

6 A New Inversion-Based Joint BQSS/BQPT Method

We here use a third combination of p_{1xx} to p_{4xx} , not redundant with (20), (24):

$$p_{2xx} - p_{3xx} = -\Re(T_2 T_3^*) = -(R_{23} \cos \Delta\Phi_{0,0} - I_{23} \sin \Delta\Phi_{0,0}) \quad (28)$$

with (again using [5]) to transform $\omega_{1,1}$ and $\omega_{0,0}$)

$$\Delta\Phi_{0,0} = (\omega_{0,0} - \omega_{1,1})(t - t_0) = \frac{(t - t_0)}{\hbar} (J_{xy} + J_z - GB) \quad (29)$$

and R_{23} , I_{23} , respectively derived from (25) and (26) by changing the signs of their second and fourth (i.e. last) terms.

With the same approach as in the previous sections, one would consider (7)–(9), (21), (24), and (28) as a nonlinear mixing model (involving the above-defined four source signals and two mixing parameters), one would first use statistical QSICA methods for estimating the mixing parameters, and one would then use these estimated parameters to restore (new values of) source signals. However, introducing all above equations opens the way to the following, completely different, approach. One may consider (7)–(9), (21), (24), and (28) as a set of *six* nonlinear equations which, fortunately, involve only *six* unknowns, namely r_1 ,

r_2 , $(\phi_1 - \theta_1)$, $(\phi_2 - \theta_2)$, J_{xy} and J_z . One may analytically solve part of these equations: see [5] for (7) and (9). The others may be solved by using a nonlinear numerical optimization algorithm from the literature. This is especially attractive because it requires only *one* (repeatedly prepared) state $|\psi(t)\rangle$, whereas the above-defined QSICA methods typically need 10^3 (repeatedly prepared) such states. On the contrary, some of the complete BQSS/BQPT methods defined in the previous sections yield closed-form estimates for all above six source and mixing quantities, whereas the numerical inversion-based method proposed here might yield spurious global minima and convergence to local minima, which should be further investigated.

7 Conclusion

In this paper, we introduced major extensions of the fields of blind quantum source separation (BQSS) and blind quantum process tomography (BQPT): we first considered a new type of spin component measurements for the directly available quantum states (that correspond to the observed mixed signals processed in *classical* BSS), which led us to introduce new nonlinear mixing models, for which we proposed three types of BQSS and/or BQPT methods. Our future works will especially consist of (i) analyzing the theoretical properties of the inversion-based method defined in Sect. 6 and (ii) developing software tools to simulate qubits and evaluate the performance of the proposed methods.

References

1. Abed-Meraim, K., Qiu, W., Hua, Y.: Blind system identification. Proc. IEEE **85**, 1310–1322 (1997)
2. Branderhorst, M. P. A., Nunn, J., Walmsley, I. A., Kosut, R. L.: Simplified quantum process tomography (2009). <https://arxiv.org/abs/0910.4609> version 2
3. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation. Independent Component Analysis and Applications. Academic Press, Oxford (2010)
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). https://doi.org/10.1007/978-3-540-74494-8_88
5. Deville, Y., Deville, A.: Classical-processing and quantum-processing signal separation methods for qubit uncoupling. Quant. Inf. Proc. **11**, 1311–1347 (2012)
6. Deville, Y., Deville, A.: Quantum-source independent component analysis and related statistical blind qubit uncoupling methods. In: Naik, G.R., Wang, W. (eds.) Blind Source Separation. SCT, pp. 3–37. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-55016-4_1
7. Deville, Y., Deville, A.: From Blind quantum source separation to blind quantum process tomography. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 184–192. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_21

8. Deville, Y., Deville, A.: Fast disentanglement-based blind quantum source separation and process tomography: a closed-form solution using a feedback classical adapting structure. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 438–448. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_41
9. Deville, Y., Deville, A.: The blind version of quantum process tomography: operating with unknown input values. In: Proceedings of the 20th World Congress of the International Federation of Automatic Control (IFAC 2017), pp. 12228–12234 (2017)
10. Deville, Y., Deville, A.: Blind quantum source separation: quantum-processing qubit uncoupling systems based on disentanglement. *Digit. Sig. Proc.* **67**, 30–51 (2017)
11. Merkel, S. T., Gambetta, J. M., Smolin, J. A., Poletto, S., Córcoles, A. D., Johnson, B. R., Ryan, C. A., Steffen, M.: Self-consistent quantum process tomography. *Phys. Rev. A* **87**, 062119–1–062119–9 (2013)
12. Nielsen, M.A., Chuang, I.L.: *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge (2000)
13. Shukla, A., Mahesh, T. S.: Single-scan quantum process tomography. *Physical Review A* **90**, 052301–1 to 052301–6 (2014)
14. Takahashi, M., Bartlett, S. D., Doherty, A. C.: Tomography of a spin qubit in a double quantum dot. *Physical Review A* 022120–1 to 022120–9 (2013)

Audio Data and Methods



Blind Signal Separation by Synchronized Joint Diagonalization

Hiroshi Sawada^(✉)

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
sawada.hiroshi@lab.ntt.co.jp

Abstract. Joint Diagonalization (JD) is a well-known method for blind signal separation (BSS) by exploiting the nonstationarity of signals. In this paper, we propose *Synchronized Joint Diagonalization* (SJD) that solves multiple JD problems simultaneously and tries to synchronize the activity of the same signal along the time axis over the multiple JD problems. SJD attains not only signal separation by the mechanism of JD but also permutation alignment by the synchronization when applied to frequency-domain BSS. Although the formulation of SJD starts from the minimization of multi-channel Itakura-Saito divergences between a covariance matrix and a diagonal matrix, the simplified cost function with the finest time blocks becomes similar to that of Independent Vector/Component Analysis (IVA/ICA). We discuss the relationship among SJD and existing techniques. Experimental results on speech separation are shown to demonstrate the behavior of these methods.

1 Introduction

As a blind signal separation (BSS) method that exploits the nonstationarity of signals, there have been proposed many Joint Diagonalization (JD) methods [1–7] applied to the covariance matrices of multiple time blocks. These are basically separation methods for instantaneous mixtures. When applied to convolutive mixtures with delays and reverberations, these methods need to be extended or followed by some post-processing. Typically, the convolutive mixtures are transformed into time-frequency domain, and multiple JD problems that are associated with multiple frequency bins are solved [3] followed by permutation alignment [8]. Such a way to solve a convolutive BSS problem is called frequency-domain BSS.

In this paper, we propose a method called *Synchronized Joint Diagonalization* (SJD). The method solves multiple JD problems with synchronizing the diagonal elements of the same source along the time axis. We model the nonstationarity of a source signal with parameters that depend only on signal identity and a time block but not on a frequency bin. The modeling is typically effective for speech separation as will be demonstrated later. We employ multichannel Itakura-Saito (IS) divergence [9] between a covariance matrix of a separated signal and a diagonal matrix that is modeled with above mentioned parameters. We propose

an algorithm that minimizes the IS divergence sum over all the time blocks and all the frequency bins. Consequently, SJD is expected to produce solutions for multiple JD problems with their permutations aligned among frequency bins.

JD and SJD methods utilize the nonstationarity of a signal by taking multiple time blocks into account. Under some conditions later described, we observe a relationship between other BSS methods based on other principles. More specifically, we discuss (1) the relationship between JD and Independent Component Analysis (ICA) [10, 11] and (2) the relationship between SJD and Independent Vector Analysis (IVA) [12–16] that is expected to perform permutation alignment as well.

This paper is organized as follows. Section 2 formulates frequency-domain BSS and explains JD. Section 3 proposes SJD and discusses the relationships among other BSS methods. Section 4 reports experimental results.

2 Preliminary

2.1 Formulation of Frequency-Domain BSS

Suppose we have M observed signals and their time-frequency representations $x_{ijm} \in \mathbb{C}$ by applying a short-time Fourier transform (STFT), with $i = 1, \dots, I$ and $j = 1, \dots, J$ being frequency bins and time frames, respectively. The M observed signals form a complex vector as $\mathbf{x}_{ij} = [x_{ij1}, \dots, x_{ijM}]^T \in \mathbb{C}^M$, which is assumed to be a linear mixture of N independent complex source signals $s_{ijn} = [\mathbf{s}_{ij}]_n$, $n = 1, \dots, N$ as

$$\mathbf{x}_{ij} = \mathbf{A}_i \mathbf{s}_{ij}. \quad (1)$$

The $M \times N$ complex mixing matrix \mathbf{A}_i is assumed to be frequency-bin dependent, but time-frame invariant. The purpose of frequency-domain BSS is to estimate $N \times M$ separation matrices \mathbf{W}_i for all frequency bins $i = 1, \dots, I$ and to obtain separated signals $y_{ijn} = [\mathbf{y}_{ij}]_n$, $n = 1, \dots, N$ that should be close to the original source signals by

$$\mathbf{y}_{ij} = \mathbf{W}_i \mathbf{x}_{ij}. \quad (2)$$

2.2 Joint Diagonalization (JD) of Covariance Matrices

This subsection explains a BSS method based on JD of covariance matrices [1–7]. The process is frequency-bin i wise. The total J time frames are partitioned $\bigcup_{b=1}^B \mathcal{J}_b = \{1, \dots, J\}$ into B time blocks \mathcal{J}_b , $b = 1, \dots, B$. For each time block \mathcal{J}_b , we calculate the covariance matrix (assuming zero mean) of observed signals

$$\mathbf{X}_{ib} = \frac{1}{|\mathcal{J}_b|} \sum_{j \in \mathcal{J}_b} \mathbf{x}_{ij} \mathbf{x}_{ij}^H \quad (3)$$

and the covariance matrix (assuming zero mean, too) of separated signals

$$\mathbf{Y}_{ib} = \frac{1}{|\mathcal{J}_b|} \sum_{j \in \mathcal{J}_b} \mathbf{y}_{ij} \mathbf{y}_{ij}^H = \mathbf{W}_i \mathbf{X}_{ib} \mathbf{W}_i^H. \quad (4)$$

The goal here is to estimate a separation matrix \mathbf{W}_i that diagonalizes (or makes the off-diagonal elements $[\mathbf{Y}_{ib}]_{nm}$, $n \neq m$ close to zero) all the B covariance matrices \mathbf{Y}_{ib} , $b = 1, \dots, B$, jointly.

If $B = 2$, exact diagonalization (making off-diagonal elements exactly zero) is possible. A separation matrix \mathbf{W}_i is obtained by solving generalized eigenvalue decomposition $\mathbf{X}_{i1} \mathbf{W}_i^H = \mathbf{X}_{i2} \mathbf{W}_i^H \mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is a diagonal matrix. Figure 1 shows an example when $B = 2$.

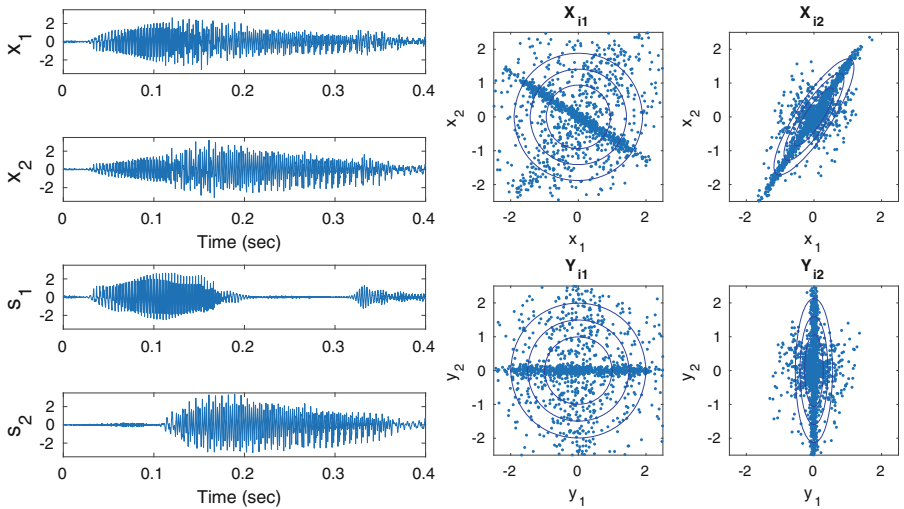


Fig. 1. An example of JD when the number B of time blocks is two. The left hand side shows two-channel observed mixtures (upper) and two source signals (lower). Let the first time block \mathcal{J}_1 be from 0 to 0.2 (sec) and the second block \mathcal{J}_2 be from 0.2 to 0.4 (sec). The right hand side shows covariance matrices. The upper right corresponds to those $\mathbf{X}_{i1} \mathbf{X}_{i2}$ of observed mixtures, which are not jointly diagonalized. The lower right corresponds to those $\mathbf{Y}_{i1} \mathbf{Y}_{i2}$ of separated signals, which are jointly diagonalized.

If $B \geq 3$, exact diagonalization is not possible in general. Therefore, we typically define a distance or divergence such as

$$d(\mathbf{Y}_{ib}) = \|\mathbf{Y}_{ib} - \text{diag}(\mathbf{Y}_{ib})\|_F \quad (5)$$

or

$$d(\mathbf{Y}_{ib}) = \log \det \text{diag}(\mathbf{Y}_{ib}) - \log \det \mathbf{Y}_{ib}, \quad (6)$$

and minimize the sum over the whole time blocks

$$\mathcal{C}(\mathbf{W}_i) = \sum_{b=1}^B d(\mathbf{Y}_{ib}). \quad (7)$$

The operation $\text{diag}(\mathbf{Y})$ keeps the diagonal elements of \mathbf{Y} unchanged and makes the off-diagonal elements zero. $\|\mathbf{Y}\|_F = \sqrt{\sum_{n=1}^N \sum_{m=1}^M |y_{nm}|^2}$ is a Frobenius norm. For a covariance matrix \mathbf{Y} , which is Hermitian, $\det \text{diag}(\mathbf{Y}) \geq \det \mathbf{Y}$ satisfies. Equations (5) and (6) are employed in [2, 3, 6, 7] and [1, 4, 5], respectively. Therein, various minimization algorithms have been proposed.

3 Proposed Method

3.1 Synchronized Joint Diagonalization (SJD)

In this paper, we generalize Eq. (6) and consider the multichannel Itakura-Saito divergence [9] or log-determinant divergence

$$d_{IS}(\mathbf{Y}_{ib}, \hat{\mathbf{Y}}_{ib}) = \text{tr}(\mathbf{Y}_{ib} \hat{\mathbf{Y}}_{ib}^{-1}) - \log \det \mathbf{Y}_{ib} \hat{\mathbf{Y}}_{ib}^{-1} - N \quad (8)$$

between a covariance matrix \mathbf{Y}_{ib} and a diagonal matrix $\hat{\mathbf{Y}}_{ib}$. If $\hat{\mathbf{Y}}_{ib} = \text{diag}(\mathbf{Y}_{ib})$, Eq. (8) reduces to Eq. (6).

In the formulation of SJD, we model the diagonal matrix $\hat{\mathbf{Y}}_{ib}$ as

$$[\hat{\mathbf{Y}}_{ib}]_{nm} = \begin{cases} v_{bn} & \text{if } n = m \\ 0 & \text{if } n \neq m \end{cases} \quad (9)$$

The model parameter v_{bn} depends on a time block b and a signal n , and does not depend on frequency i . The intention of introducing this parameter is to make signal n activity on each block b synchronized over all frequency bins i . For the purpose of regularization, we assume a prior distribution for v_{bn} using an inverse-gamma distribution

$$v_{bn} \sim \mathcal{IG}(v_{bn} | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} v_{bn}^{-\alpha-1} \exp\left(-\frac{\beta}{v_{bn}}\right). \quad (10)$$

For notational convenience, let \mathbf{V} be a $B \times N$ matrix and $v_{bn} = [\mathbf{V}]_{bn}$.

SJD minimizes the sum of the total divergence and the negative log-likelihood. The total divergence is over not only the whole time blocks but also all frequency bins.

$$\mathcal{C}(\{\mathbf{W}_i\}_{i=1}^I, \mathbf{V}) = \sum_{i=1}^I \sum_{b=1}^B d_{IS}(\mathbf{Y}_{ib}, \hat{\mathbf{Y}}_{ib}) - \sum_{b=1}^B \sum_{n=1}^N \log \mathcal{IG}(v_{bn} | \alpha, \beta). \quad (11)$$

Substituting Eqs. from (8) to (10) into Eq. (11) and eliminating constant terms, we have a simplified cost function to be minimized

$$\mathcal{C} = \sum_{b=1}^B \sum_{n=1}^N \left[\frac{\sum_{i=1}^I [\mathbf{Y}_{ib}]_{nn} + \beta}{v_{bn}} + (I + \alpha + 1) \log v_{bn} \right] - 2B \sum_{i=1}^I \log |\det \mathbf{W}_i|. \quad (12)$$

3.2 Optimization Algorithm

We minimize Eq. (12) by alternatively updating $\{\mathbf{W}_i\}_{i=1}^I$ and \mathbf{V} .

Regarding \mathbf{V} , we have an update rule for each element

$$v_{bn} = \frac{\sum_{i=1}^I [\mathbf{Y}_{ib}]_{nn} + \beta}{I + \alpha + 1} \quad (13)$$

as a solution of the partial derivative

$$\frac{\partial \mathcal{C}}{\partial v_{bn}} = -\frac{\sum_{i=1}^I [\mathbf{Y}_{ib}]_{nn} + \beta}{v_{bn}^2} + \frac{I + \alpha + 1}{v_{bn}} \quad (14)$$

of \mathcal{C} with respect to the element v_{bn} being zero.

Regarding a frequency bin-wise separation matrix

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{w}_{i1}^H \\ \vdots \\ \mathbf{w}_{iN}^H \end{bmatrix}, \quad (15)$$

we update it by the following procedure [15] derived by an auxiliary function method. First, we calculate a weighted mean

$$\mathbf{U}_{in} = \frac{1}{B} \sum_{b=1}^B \frac{1}{v_{bn}} \mathbf{x}_{ib} \quad (16)$$

of the observation covariance matrices for all the signal $n = 1, \dots, N$. Then, we solve the Hybrid Exact-Approximate Diagonalization (HEAD) problem [17] for these N matrices \mathbf{U}_{in} , and update \mathbf{W}_i as the HEAD solution. An efficient way [15] to solve the HEAD problem is to calculate

$$\mathbf{w}_{in} = (\mathbf{W}_i \mathbf{U}_{in})^{-1} \mathbf{e}_n \quad (17)$$

for each n , where \mathbf{e}_n is a vector whose n -th element is one and all the others are zero, and

$$\mathbf{w}_{in} = \frac{\mathbf{w}_{in}}{\sqrt{\mathbf{w}_{in}^H \mathbf{U}_{in} \mathbf{w}_{in}}}. \quad (18)$$

to accommodate a HEAD condition $\mathbf{w}_{in}^H \mathbf{U}_{in} \mathbf{w}_{in} = 1$.

As described in [18], normalization on the scale of \mathbf{W}_i makes the algorithm stable. Thus, we apply $\mathbf{W}_i \leftarrow \mathbf{W}_i / \sqrt[2]{|\mathbf{W}_i|}$ after each iteration.

3.3 Example

Figure 2 shows examples in which the model parameters \mathbf{V} are estimated for speech mixtures. When the number of time blocks is large $B = 162$, the dynamics of speech is finely estimated. When the number of blocks is relatively small $B = 20$, the averaged power is estimated for each time block.

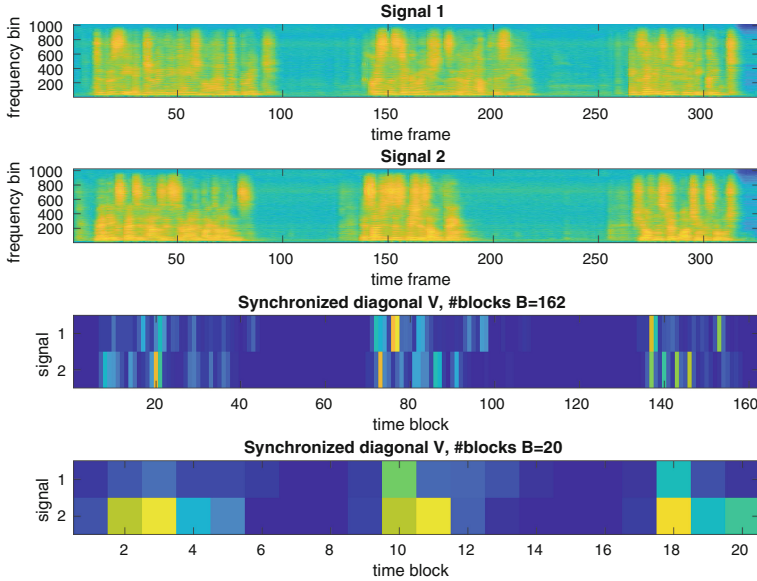


Fig. 2. Two speech signals whose spectrograms are shown in the top two rows are assumed to be mixed. The model parameters \mathbf{V} estimated by the optimization algorithm are shown in the bottom two rows with different number B of time blocks.

3.4 Relationship to IVA

Time blocks \mathcal{J}_b can be specified arbitrary. Now we consider a condition.

Condition 1. *The number B of time blocks equals to the number J of time frames and the block size $|\mathcal{J}_b|$ is one, i.e., $\mathcal{J}_j = \{j\}$ for $j = 1, \dots, J$.*

Then, the summation in the covariance matrix calculation (4) disappears $\mathbf{Y}_{ij} = \mathbf{y}_{ij}\mathbf{y}_{ij}^H$, and the cost function (12) becomes

$$C = \sum_{j=1}^J \sum_{n=1}^N \left[\frac{\sum_{i=1}^I |y_{ijn}|^2 + \beta}{v_{jn}} + (I + \alpha + 1) \log v_{jn} \right] - 2J \sum_{i=1}^I \log |\det \mathbf{W}_i| \quad (19)$$

by considering $[\mathbf{Y}_{ij}]_{nn} = |y_{ijn}|^2$.

Now we see some relationship between (19) and Independent Vector Analysis (IVA) [12–16]. In IVA, we consider a vector $\tilde{\mathbf{y}}_{jn} = [y_{1jn}, \dots, y_{Ijn}]^T \in \mathbb{C}^I$ that represents all the complex values over all frequency bins of a separated signal n at a time frame j . And we minimize the negative log-likelihood as a cost function

$$C_{IVA} = - \sum_{j=1}^J \sum_{n=1}^N \log p(\tilde{\mathbf{y}}_{jn}) - 2J \sum_{i=1}^I \log |\det \mathbf{W}_i|. \quad (20)$$

The probability density function of the vector is typically assumed to be a super-Gaussian distribution [12–15] with scale parameter γ

$$p(\tilde{\mathbf{y}}_{jn}) \propto \exp\left(-\frac{\|\tilde{\mathbf{y}}_{jn}\|}{\gamma}\right) = \exp\left(-\frac{\sqrt{\sum_{i=1}^I |y_{ijn}|^2}}{\gamma}\right) \quad (21)$$

or assumed to be a Gaussian distribution with time-varying variance σ_{jn}^2 [16]

$$p(\tilde{\mathbf{y}}_{jn}) \propto \exp\left(-\frac{\|\tilde{\mathbf{y}}_{jn}\|^2}{\sigma_{jn}^2}\right) = \exp\left(-\frac{\sum_{i=1}^I |y_{ijn}|^2}{\sigma_{jn}^2}\right). \quad (22)$$

If we compare (19) and (20) together with (21) or (22), the difference regarding the separation matrix \mathbf{W}_i and the separated signals y_{ijn} is only in $\frac{\sum_{i=1}^I |y_{ijn}|^2 + \beta}{v_{jn}}$ and $\frac{\sqrt{\sum_{i=1}^I |y_{ijn}|^2}}{\gamma}$ or $\frac{\sum_{i=1}^I |y_{ijn}|^2}{\sigma_{jn}^2}$. Therefore, the optimization algorithm shown from (16) to (18) has been derived by the auxiliary function technique [15, 16] with the above mentioned difference in mind.

3.5 Relationship to JD and ICA

Ordinary JD can be seen as a special case of SJD where the number of frequency bins is one $I = 1$. This means that we obtain a new algorithm for JD by assuming $I = 1$ in Subsect. 3.2.

When $I = 1$, the SJD cost function (12) and parameter update (13) become

$$\mathcal{C} = \sum_{b=1}^B \sum_{n=1}^N \left[\frac{[\mathbf{Y}_{ib}]_{nn} + \beta}{v_{bn}} + (\alpha + 2) \log v_{bn} \right] - 2B \log |\det \mathbf{W}_i|, \quad (23)$$

$$v_{bn} = \frac{[\mathbf{Y}_{ib}]_{nn} + \beta}{\alpha + 2}, \quad (24)$$

respectively. By substituting (24) into (23), we have

$$\mathcal{C} = (\alpha + 2) \sum_{b=1}^B \sum_{n=1}^N \log ([\mathbf{Y}_{ib}]_{nn} + \beta) - 2B \log |\det \mathbf{W}_i| \quad (25)$$

with constant terms omitted.

Now we discuss the relationship to Independent Component Analysis (ICA) [10, 11]. Again we assume Condition 1. Then, (25) becomes

$$\mathcal{C} = \sum_{j=1}^J \left[(\alpha + 2) \sum_{n=1}^N \log (|y_{ijn}|^2 + \beta) - 2 \log |\det \mathbf{W}_i| \right] \quad (26)$$

This cost function (26) has a relationship to complex-valued FastICA [11] with $G(|y_{ijn}|^2) = \log(|y_{ijn}|^2 + \beta)$ contrast function. If we consider the fact that FastICA assumes unitary separation matrix $|\det \mathbf{W}_i| = 1$, we see that these two methods optimize the same function.

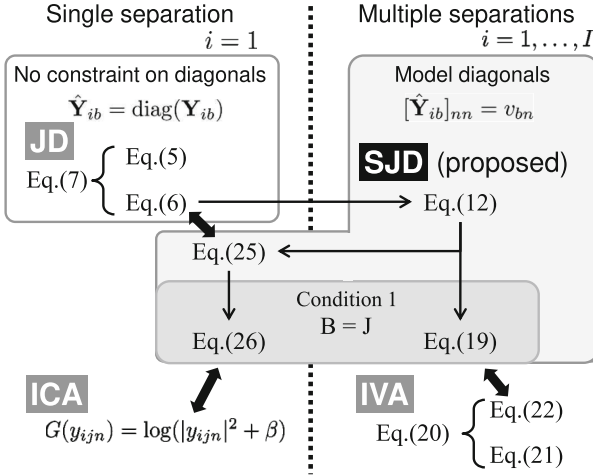


Fig. 3. Summary of the relationship among SJD and other related existing techniques

3.6 Summary of the Relationship

Figure 3 summarizes the relationship discussed so far. We have extended JD with the divergence (6) to SJD by modeling diagonal elements with parameters v_{bn} . Then, we considered Condition 1 where $B = J$, and observed the relationship between SJD and IVA. On another path, we went back to single separation (JD) from SJD by setting $I = 1$, and then observed the relationship to ICA by assuming Condition 1 again.

We consider that Eq. (12) is the most general form. Section 3.2 shows an algorithm to optimize it. As shown in Fig. 3 with the arrows, Eqs. (19), (25) and (26) can be derived as special cases of Eq. (12). This means that we have optimization algorithms for all the four equations. In the next Section, we compare the results of BSS with these equations.

4 Experiments

To examine the behavior of SJD and its variants, we performed experiments for blind speech separation. We measured impulse responses from two loudspeakers to two microphones in a room whose reverberation time was $RT_{60} = 200$ ms. Then, we made mixtures by convolving the impulse responses and 10-second speech signals. The sampling frequency was 16 kHz. The frame width and shift of STFT were 128 ms and 32 ms, respectively. We set $\alpha = \beta = 0.5$ for (10). The separation performance was evaluated in terms of Signal-to-Distortion Ratio (SDR) [19]. The algorithm was coded with Matlab and run on an Intel Xeon E3-1290.

Figure 4 shows the separation performance with varying B . The left plot shows the results of SJD (12) and IVA (19). The center and right plots show the

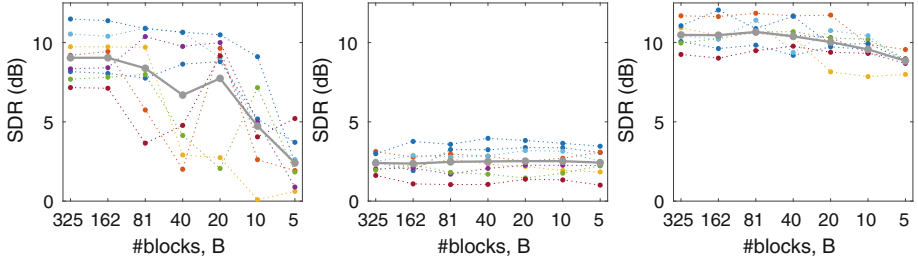


Fig. 4. Left: separation performances by SJD with varying the number B of time blocks. A dotted line corresponds to a mixture of two speeches. The solid grey line represents the average over eight mixtures. Center: separation performances by multiple JDs. Right: separation performances by multiple JDs followed by a permutation alignment method [20].

results of JD (25) and ICA (26). In each plot, the left most values (when $B = 325$) correspond to IVA or ICA. In the center case, there was no care regarding permutation ambiguities among frequency bins. In the right case, permutation ambiguities were aligned by a post-processing method [20].

We observe the followings. SJD with a large B (325, 162, 81) generally contributed to separation performance. This was because permutations were effectively aligned by precisely tracking the activity of source speeches (see Fig. 2 for an example). JD alone did not attain source separation. However, when followed by the alignment post-processing, JD attained excellent separations. In these JD cases, the number B of time blocks did not clearly affect separation performance. This means that every BSS problems at every frequency bin is effectively solved by JD regardless of B . In total, even though SJD/IVA is designed to align permutations among frequency bins, there was still a small gap regarding the alignment performance between SJD/IVA and the post-processing [20]. Future work includes improvement of SJD/IVA on permutation alignment capability.

The optimization algorithms sufficiently converged with 50 iterations in all cases. The execution time was around 12 s with $B = J = 325$ and was around 7.4 s with $B = 81$, respectively, for a 10-second mixture.

5 Conclusion

As a BSS method to exploit the nonstationarity of source signals, we extended JD to SJD by introducing model parameters $v_{bn} = [\mathbf{V}]_{bn}$ that represent frequency-independent source n activities. The SJD cost function (12) can efficiently be optimized by the algorithm shown in Sect. 3.2, and can be considered as a general cost function for BSS as summarized in Fig. 3. The findings described in this paper help us to organize and understand various existing BSS methods including ICA and IVA that exploit the nongaussianity of source signals.

References

1. Matsuoka, K., Ohya, M., Kawamoto, M.: A neural net for blind separation of nonstationary signals. *Neural Netw.* **8**(3), 411–419 (1995)
2. Wax, M., Sheinvald, J.: A least-squares approach to joint diagonalization. *IEEE Sign. Process. Lett.* **4**(2), 52–53 (1997)
3. Parra, L., Spence, C.: Convolutional blind separation of non-stationary sources. *IEEE Trans. Speech Audio Process.* **8**(3), 320–327 (2000)
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., Cardoso, J.F.: Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Trans. Sign. Process.* **49**(9), 1837–1848 (2001)
6. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Trans. Sign. Process.* **50**(7), 1545–1553 (2002)
7. 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. *J. Mach. Learn. Res.* **5**, 777–800 (2004)
8. 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. Speech Audio Process.* **12**(5), 530–538 (2004)
9. Sawada, H., Kameoka, H., Araki, S., Ueda, N.: Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Trans. Audio Speech Lang. Process.* **21**(5), 971–982 (2013)
10. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*. Wiley, New York (2001)
11. Bingham, E., Hyvärinen, A.: A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Syst.* **10**(1), 1–8 (2000)
12. Hiroe, A.: Solution of permutation problem in frequency domain ICA, using multivariate probability density functions. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds.) *ICA 2006*. LNCS, vol. 3889, pp. 601–608. Springer, Heidelberg (2006). https://doi.org/10.1007/11679363_75
13. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **15**(1), 70–79 (2007)
14. Lee, I., Kim, T., Lee, T.W.: Fast fixed-point independent vector analysis algorithms for convolutional blind source separation. *Sign. Process.* **87**(8), 1859–1871 (2007)
15. Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, pp. 189–192 (2011)
16. Ono, N.: Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions. In: *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Asia-Pacific, IEEE, pp. 1–4 (2012)
17. Yeredor, A.: On hybrid exact-approximate joint diagonalization. In: *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, pp. 312–315 (2009)
18. Kitamura, D., Ono, N., Sawada, H., Kameoka, H., Saruwatari, H.: Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(9), 1626–1641 (2016)

19. Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., Duong, N.: The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Sign. Process.* **92**(8), 1928–1936 (2012)
20. Sawada, H., Araki, S., Makino, S.: Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS. In: *Proceedings of the ISCAS 2007*, pp. 3247–3250 (2007)



Exploiting Structures of Temporal Causality for Robust Speaker Localization in Reverberant Environments

Christopher Schymura¹(✉), Peng Guo¹, Yanir Maymon², Boaz Rafaely², and Dorothea Kolossa¹

¹ Department of Electrical Engineering and Information Technology,
Ruhr-Universität Bochum, Bochum 44801, Germany
christopher.schymura@rub.de

² Department of Electrical and Computer Engineering,
Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Abstract. This paper introduces a framework for robust speaker localization in reverberant environments based on a causal analysis of the temporal relationship between direct sound and corresponding reflections. It extends previously proposed localization approaches for spherical microphone arrays based on a direct-path dominance test. So far, these methods are applied in the time-frequency domain without considering the temporal context of direction-of-arrival measurements. In this work, a causal analysis of the temporal structure of subsequent directions-of-arrival estimates based on the Granger causality test is proposed. The cause-effect relationship between estimated directions is modeled via a causal graph, which is used to distinguish the direction of the direct sound from corresponding reflections. An experimental evaluation in simulated acoustic environments shows that the proposed approach yields an improvement in localization performance especially in highly reverberant conditions.

Keywords: Speaker localization · Spherical microphone arrays
Vector autoregressive models · Multivariate granger causality test

1 Introduction

Acoustic speaker localization (ASL) is a widely and actively investigated topic in digital signal processing. It has important practical applications in speech enhancement [1], teleconferencing and smart rooms [2,3], robot audition [4,5] and many other fields. ASL aims at estimating the position or direction of arrival (DoA) of speakers from audio signals captured by an array of microphones. Popular methods for performing this task include approaches based on beamforming [6,7], the multiple signal classification (MUSIC) algorithm [8,9] and the degenerate unmixing and estimation technique (DUET) [10].

Recently, several novel approaches to ASL using spherical microphone arrays have been proposed, which allow a precise estimation of speaker DoAs in three-dimensional space [11, 12]. The method introduced in [11] utilized a direct-path dominance (DPD) test to cope with reverberation effects by performing ASL exclusively on time-frequency (TF) bins that are dominated by the direct sound. This approach was extended in [12] towards a computationally less demanding framework based on pseudo-intensity vectors. A central aspect of DPD-test-based localization is the clustering of TF bins into components that correspond to different DoAs. This is typically achieved using probabilistic clustering methods like Gaussian mixture models (GMMs) [13]. Prior to clustering, DPD-test-based DoA estimation processes each TF bin individually to decide if a particular element in the acoustic spectrogram is dominated by the direct sound. The GMM components obtained during the clustering step can thus be interpreted as DoAs that correspond to the true source locations. However, mis-detections can occur depending on the available amount of data and the clustering parameters, which might yield GMM components corrupted by strong reflections. In [11] a heuristic approach was proposed, where the DoA that is assumed to stem from the true source location was chosen as the most dominant GMM component. Here, the GMM components were estimated based on DoA observations obtained by locally averaging TF regions without considering temporal structure. An open question in this regard is, if considering the temporal context of the received direct sound and corresponding reflections can be beneficial for selecting the correct component.

This paper introduces a clustering-based DoA estimation framework for spherical microphone arrays which exploits the temporal causal structure of sound propagation. In a reverberant environment, the sound waves received at the microphone array are typically composed of the direct sound, followed by early-reflections and subsequent late reverberation [14, Chap. 4]. This physically justified structure allows us to consider models that incorporate the temporal context of the received acoustic signals explicitly. The underlying assumption is that the DoA of the direct sound that arrives after a period of silence is visible at the acoustic sensors before subsequent DoAs of acoustic reflections. Hence, a temporal cause-effect relationship between direct sound and corresponding reflections can be formulated. A popular framework to determine causal effects between time-series is the Granger causality test (GCT) [15], which has received wide acclaim in econometric time-series analysis [15, 16]. This statistical test is applied here in the context of ASL to distinguish the direct sound from corresponding reflections DoAs based on their causal relationships.

2 Localization Framework

The localization framework that is used in this study is largely based on the method introduced in [11] and will be briefly reviewed in this section. Furthermore, the GMM-based clustering approach from [17] and the generation of time series signals, which serves as a basis for the causal analysis step proposed in Sect. 3, will also be summarized below.

2.1 Spherical Array Processing

A rigid spherical microphone array with radius r and Q acoustic sensors is considered in this study. The sound field around the array is assumed to be composed of L plane waves. Let $p(k, r, \theta_q, \phi_q)$ denote the sound pressure at the q -th microphone, where k is the wave number and the angles θ_q and ϕ_q represent the position of the microphone in spherical coordinates $\mathbf{m}_q = [\theta_q, \phi_q]^T$. By following the approach introduced in [11] and denoting $\boldsymbol{\Omega} \in \{\mathbf{m}_1, \dots, \mathbf{m}_Q\}$, the sound pressure at the surface of the array can be formulated as

$$\mathbf{p}(k, r, \boldsymbol{\Omega}) = \mathbf{Y}(\boldsymbol{\Omega})\mathbf{B}(k, r)\mathbf{Y}^H(\boldsymbol{\Psi})\mathbf{s}(k) + \mathbf{n}(k), \quad (1)$$

with $\mathbf{p}(k, r, \boldsymbol{\Omega}) = [p(k, r, \mathbf{m}_1), \dots, p(k, r, \mathbf{m}_Q)]^T$, where $\mathbf{Y}(\boldsymbol{\Omega}) \in \mathbb{C}^{Q \times (N+1)^2}$ and $\mathbf{Y}(\boldsymbol{\Psi}) \in \mathbb{C}^{L \times (N+1)^2}$ are spherical harmonics matrices with corresponding spherical harmonics functions $Y_n^m(\mathbf{m}_q)$ and $Y_n^m(\mathbf{v}_l)$ of order n and degree m , respectively. $\mathbf{B}(k, r) \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$ is a diagonal matrix modeling the scattering of a plane wave from the rigid sphere via radial functions [18] and N is the order of the spherical array. $\boldsymbol{\Psi} \in \{\mathbf{v}_1, \dots, \mathbf{v}_L\}$ represents the source arrival directions, $\mathbf{s}(k) = [s_1(k), \dots, s_L(k)]^T$ are the acoustic source signals and $\mathbf{n}(k) = [n_1(k), \dots, n_Q(k)]^T$ is a vector modeling additive noise.

A plane wave decomposition (PWD) [18] can be applied to Eq. (1) by left-multiplying with $\mathbf{B}^{-1}(k, r)[\mathbf{Y}(\boldsymbol{\Omega})]^\dagger$, which yields

$$\mathbf{a}(k) = \mathbf{Y}^H(\boldsymbol{\Psi})\mathbf{s}(k) + \tilde{\mathbf{n}}(k), \quad (2)$$

where $\mathbf{a}(k) = [a_{00}(k), \dots, a_{NN}(k)]^T \in \mathbb{C}^{(N+1)^2}$ represents a vector containing the complex spherical harmonics coefficients of the plane wave density function and $\tilde{\mathbf{n}}(k) = \mathbf{B}^{-1}(k, r)[\mathbf{Y}(\boldsymbol{\Omega})]^\dagger \mathbf{n}(k)$. Equation (2) is then transformed into the shorttime Fourier transform (STFT) domain, which leads to

$$\mathbf{a}_{\tau, \nu} = \mathbf{Y}^H(\boldsymbol{\Psi})\mathbf{s}_{\tau, \nu} + \tilde{\mathbf{n}}_{\tau, \nu}, \quad (3)$$

where τ and ν denote the time- and frequency indices, respectively.

2.2 Direction-of-arrival Estimation

Based on Eq. (3), a spatial-spectrum matrix $\mathbf{R}_{\tau, \nu} = \langle \mathbf{a}_{\tau, \nu} \mathbf{a}_{\tau, \nu}^H \rangle$ is computed at each TF bin by averaging over a specified number of neighboring time- and frequency bins. Subsequently, the DPD test [11] is applied to distinguish TF bins that are dominated by the direct sound from bins that are largely influenced by coherent reflections. By following the approach introduced in [11], a singularvalue decomposition (SVD) of the spatial-spectrum matrix is computed, followed by applying the DPD-test to each $\mathbf{R}_{\tau, \nu}$. This allows the construction of a set of TF bins that pass the DPD test

$$\mathcal{D} = \left\{ (\tau, \nu) : \frac{\sigma_1(\mathbf{R}_{\tau, \nu})}{\sigma_2(\mathbf{R}_{\tau, \nu})} \geq \xi \right\}, \quad (4)$$

where σ_1 and σ_2 are the two largest singular values obtained by the SVD and ξ denotes a threshold parameter, which is typically chosen sufficiently larger than one [13]. In the following step, DoA estimation is performed within all TF bins that passed the DPD test $(\tau, \nu) \in \mathcal{D}$ via the MUSIC estimation framework [9]. This yields a set of DoA vectors $\hat{\mathbf{v}}_{\tau,\nu} = [\hat{\phi}_{\tau,\nu} \hat{\theta}_{\tau,\nu}]^T$ for all considered TF bins, which will be subsequently used to generate DoA time series signals.

2.3 Clustering

A dataset of estimated azimuth and elevation angle pairs $\mathbf{V} = \{\hat{\mathbf{v}}_{\tau,\nu}\} \forall \tau, \nu \in \mathcal{D}$ is used to cluster all DoA candidates based on a GMM with the probability density function (PDF)

$$p(\mathbf{V} | \lambda) = \sum_{i=1}^K \pi_i \mathcal{N}(\mathbf{V} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (5)$$

with $\sum_{i=1}^K \pi_i = 1$, where K denotes the number of mixture components, π_i correspond to the mixture weights and $\mathcal{N}(\mathbf{V} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ describes a multivariate normal distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The model parameters are summarized within the set $\lambda = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^K$.

In recently proposed approaches for DoA estimation using the DPD test, the corresponding source DoA was chosen according to Eq. (5) as the mixture component with the largest mixture weight and lowest variance [11, 13, 17]. This rather heuristic approach nevertheless yielded good estimation results in the respective works, however, it did not take into account the temporal context of the individual DoA estimates in the TF plane.

3 Causal Analysis

Causal analysis based on the GCT requires an appropriate time-domain representation of the acoustic signals. In this work, time-series are generated as binary activity patterns of the estimated GMM clusters in the TF domain $\mathbf{x}_{1:T,\nu}^{(i)} = [x_{\tau,\nu}^{(i)}]_{\tau=1,\dots,T}$, with

$$x_{\tau,\nu}^{(i)} = \begin{cases} 1 & \text{if } p(\hat{\mathbf{v}}_{\tau,\nu} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) > p(\hat{\mathbf{v}}_{\tau,\nu} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \forall j = 1 \dots, K, j \neq i \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

The activity pattern in Eq. (6) models the TF bins in which the corresponding i -th cluster is dominant over all other clusters. As each cluster is associated with a certain range of DoAs, the binary activity pattern can be interpreted as a TF representation of dominant incoming sound directions. The remaining question is, which of the clusters corresponds to the true source DoA. A possible solution to this problem based on the GCT will be introduced in this section.

3.1 Granger Causality Test

The GCT [15] was initially proposed as an econometric analysis tool to test for temporal causality between time series. A commonly used framework to perform the GCT utilizes vector autoregressive (VAR) models of time-series data to be analyzed. Herein, the notion of Granger causality follows a predictive interpretation: Let \mathbf{X} and \mathbf{Y} denote jointly distributed, multivariate stochastic process variables. \mathbf{Y} is assumed to Granger-cause \mathbf{X} , denoted as $\mathcal{F}_{\mathbf{Y} \rightarrow \mathbf{X}}$, if the degree to which the past of \mathbf{Y} helps to predict \mathbf{X} is significantly higher than the degree to which \mathbf{X} can already be predicted by only considering its own past [19]. In the context of VAR processes, this can be modeled as

$$\begin{bmatrix} \mathbf{x}_\tau \\ \mathbf{y}_\tau \end{bmatrix} = \sum_{\mu=1}^m \begin{bmatrix} \mathbf{A}_{xx,\mu} & \mathbf{A}_{xy,\mu} \\ \mathbf{A}_{yx,\mu} & \mathbf{A}_{yy,\mu} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{\tau-\mu} \\ \mathbf{y}_{\tau-\mu} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{x,\tau} \\ \boldsymbol{\epsilon}_{y,\tau} \end{bmatrix} \tag{7}$$

where \mathbf{x}_τ and \mathbf{y}_τ are realizations of the stochastic process variables \mathbf{X} and \mathbf{Y} , m denotes the order of the VAR process, μ is the respective time lag and the matrices $\mathbf{A}_{xx,\mu}$, $\mathbf{A}_{xy,\mu}$, $\mathbf{A}_{yx,\mu}$ and $\mathbf{A}_{yy,\mu}$ represent the regression coefficients of the model. The residuals are assumed to be zero-mean, Gaussian distributed random variables $\boldsymbol{\epsilon}_{x,\tau} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x)$ and $\boldsymbol{\epsilon}_{y,\tau} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_y)$ with covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$. Considering only the sub-expression of Eq. (7) related to \mathbf{x}_τ

$$\mathbf{x}_\tau = \sum_{\mu=1}^m \mathbf{A}_{xx,\mu} \mathbf{x}_{\tau-\mu} + \sum_{\tau=1}^m \mathbf{A}_{xy,\mu} \mathbf{y}_{\tau-\mu} + \boldsymbol{\epsilon}_{x,\tau}, \tag{8}$$

it is clear that the regression coefficients $\mathbf{A}_{xy,\mu}$ represent the dependence of \mathbf{X} on the past of \mathbf{Y} . Setting $\mathbf{A}_{xy,\mu} = \mathbf{0} \forall \mu$ leads to a reduced VAR process model

$$\mathbf{x}_\tau = \sum_{\mu=1}^m \mathbf{A}'_{xx,\mu} \mathbf{x}_{\tau-\mu} + \boldsymbol{\epsilon}'_{x,\tau}, \tag{9}$$

which completely omits this dependence. Here, $\mathbf{A}'_{xx,\mu}$ denotes the regression coefficients of the reduced model and $\boldsymbol{\epsilon}'_{x,t} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}'_x)$ are the corresponding residuals with covariance matrix $\boldsymbol{\Sigma}'_x$. A test for Granger causality $\mathcal{F}_{\mathbf{Y} \rightarrow \mathbf{X}}$ can now be conducted by comparing the prediction performances of the full (7) and the reduced (9) VAR models. In this study, a GCT statistic based on the log-likelihood ratio

$$\mathcal{F}_{\mathbf{Y} \rightarrow \mathbf{X}} \equiv \log \left\{ \frac{|\boldsymbol{\Sigma}'_x|}{|\boldsymbol{\Sigma}_x|} \right\} \tag{10}$$

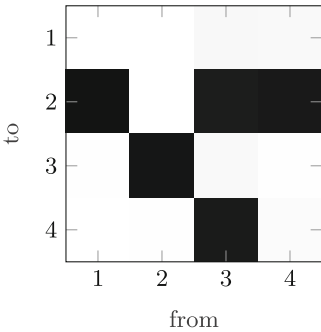
as described in [20] is utilized, which is evaluated using a statistical test based on the F -test statistic [21]. As shown in [19], this test statistic can be used to evaluate the null hypothesis $H_0 : \mathbf{A}_{xy,\mu} = \mathbf{0} \forall \mu$ representing zero causality between the full and the reduced VAR model.

It has to be noted that the GCT as applied here is based on a linear Gaussian assumption, whereas the generated time series in Eq. (6) are binary signals.

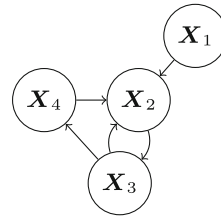
Even though the Gaussian assumption is clearly violated in this case, the applied framework has shown that it is able to cope with the specific type of signals used in this work. As the current study shall serve as a proof of concept, the conceptually simple linear Gaussian model is adopted here. However, advanced GCT measures for signals with different statistical properties exist [22] and provide interesting research directions for future work.

3.2 Causal Graph Formulation

The notion of Granger causality introduced in Sect. 3.1 can be utilized to construct a causal graph [23, Chap. 5], representing causal relationships between the DoA time series introduced in Eq. (6). A causal graph is represented by a set of nodes $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ representing multivariate random variables and a set of edges \mathcal{E} , which allows a directed, pair-wise connection between nodes if they exhibit a causal relationship. In this study, an edge from \mathbf{X}_i to \mathbf{X}_j is added to the causal graph if Granger causality $\mathcal{F}_{\mathbf{X}_i \rightarrow \mathbf{X}_j}$ between the two variables exists, i.e. they pass the GCT based on Eq. (10) given a specified significance level α . An example of a causal analysis is depicted in Figs. 1a and b.



(a) Exemplary Granger matrix. Cases where the null hypothesis of the GCT between a pair of time-series was rejected are denoted as a black square.



(b) Causal graph corresponding to the Granger matrix in Fig. 1a.

Fig. 1. Exemplary causal analysis of time-series data that was generated by a speech source in a simulated acoustic setup with room dimensions $8\text{ m} \times 5\text{ m} \times 3\text{ m}$ and $T_{60} = 1\text{ s}$. The true source DoA with respect to the microphone array is $\phi = 99.4^\circ$ and $\theta = 24.5^\circ$.

The application of this analysis framework to the DoA estimation problem discussed in Sect. 2 is straightforward: First, full and reduced VAR models according to Eqs. (7) and (9) are fitted to each pair of time series acquired through Eq. (6) using ordinary least squares (OLS) estimation for VAR processes [24]. The model order of the VARs is determined via model order selection based on the Akaike information criterion (AIC) [25]. A pair-wise GCT is

then conducted using the estimated VAR models according to Eq. (10). This yields a causal graph which is subsequently analyzed to determine its root node, corresponding to the time-series that is supposed to initially have caused all other time-series. This root node is then selected as the DoA of the true source position.

3.3 Root Node Selection

The causal model introduced in this study enables the analysis of causes and effects between DoA time series that have been generated by the localization framework introduced in Sect. 2. The purpose of this analysis is to distinguish the DoA representing the true source position from other DoAs that correspond to interfering reflections. From the viewpoint of Granger causality, the DoA of the direct sound is directly related to a root node in the causal graph, as all subsequent reflections imply a temporal causal relation with respect to the direct sound. Hence, a root cause analysis (RCA) [26] can be applied to the causal graph to determine the root node.

A root node in a causal graph corresponds to a node that has an indegree of zero [27, Chap. 9]. Referring to the example in Fig. 1b, the node corresponding to the variable \mathbf{X}_1 would represent a root node. However, in many practical cases, the causal graph either might not contain a node that strictly fulfills this condition at all, or there is more than one node with indegree zero. To solve this issue, RCA partitions the causal graph into subgraphs of strongly connected nodes. A subgraph is called strongly connected, if every node can be reached through a path from every other node [27, Chap. 12]. In Fig. 1b, the variables \mathbf{X}_2 , \mathbf{X}_3 and \mathbf{X}_4 form a strongly connected subgraph. There exist several algorithms based on depth-first search, which can find strongly connected components in arbitrary graphs in linear time. In this work, Tarjan’s algorithm [28] is used for the partitioning of the causal graph. Subsequently, the strongly connected subgraph with the largest outdegree is selected as the root subgraph and the node with the largest outdegree within that subgraph is chosen as the direct sound component.

4 Evaluation

The proposed framework is evaluated in a simulated acoustic environment. As a baseline, the DoA selection approach used in [13] was chosen for comparison.

4.1 Experimental Setup

Monte Carlo simulations with a single speaker in a room of size $8\text{ m} \times 5\text{ m} \times 3\text{ m}$ were conducted. The acoustic simulation was performed using the image-source method [29] with a spherical microphone array of radius $r = 4.2\text{ cm}$ composed of 32 microphones. A collection of speech sounds from the “sound event detection in synthetic audio” task of the Detection and Classification of Acoustic Scenes

and Events (DCASE) challenge 2016¹ was used throughout all experiments. 100 Monte Carlo runs each were conducted for reverberation times between 0.5 s and 2.5 s to investigate the performance of the proposed framework also in very challenging acoustic conditions. The length of each audio signal was fixed to 10 s. Diffuse background noise with a signal-to-noise ratio of 40 dB was added to the acoustic signal in all simulations. The GCT was performed with a significance level of $\alpha < 0.05$ in all experiments. The GMM parameters were estimated with a fixed number of $K = 10$ components. Localization root mean square error (RMSE) was selected as the evaluation metric for all experiments.

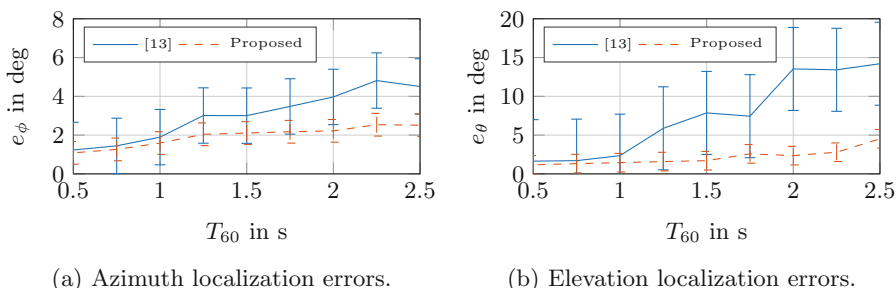


Fig. 2. Averaged localization RMSEs with corresponding error standard deviations obtained during the conducted Monte Carlo simulations.

4.2 Results and Discussion

The results depicted in Fig. 2 show that the proposed framework consistently outperforms the baseline method from [13]. The improvement is especially notable in acoustic environments with high reverberation time. This result indicates that the explicit consideration of temporal causality is beneficial for DoA component selection, especially if the influence of acoustic reflections becomes more significant with respect to the direct sound. To further investigate this effect, Table 1 shows the average coincidence rate of GMM indices selected with the baseline method and the proposed framework based on the GCT. It indicates that with increasing reverberation time, the selected GMM components that are assumed to stem from the true source position also increasingly differ between both methods. This can be explained by the fact that the baseline method solely focuses on the GMM component with largest mixture weight and lowest variance, which becomes a less informative selection criterion if the influence of strong acoustic reflections increases. In comparison, the proposed approach bases its decision on causal relationships between the different DoAs and selects the root component of the causal graph, which might not be the most dominant one in the case of strong reflections. The results indicate that explicit consideration of Granger causality between DoAs is beneficial in the context of ASL.

¹ <http://www.cs.tut.fi/sgn/arg/dccase2016/>.

Table 1. Coincidence rate, representing the degree to which the same DoA components of the GMM were selected by the proposed method and the baseline.

T_{60} in s	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
Coincidence rate	0.97	0.93	0.91	0.81	0.79	0.79	0.72	0.65	0.65

5 Conclusion

This study has proposed a speaker localization framework based on the principle of temporal causality between the direct sound DoA and corresponding reflections. A DoA selection criterion using the Granger causality test was introduced, which considers the DoA of the direct sound as a root node in a causal graph. Experimental evaluation has shown that the proposed method outperforms previously introduced selection criteria, exclusively based on the strength and variance of estimated individual DoAs without considering their temporal relationship. Future work might focus on extending the proposed framework towards different application domains like acoustic localization with arbitrary array geometries or robot audition. Furthermore, an extension to the multi-source case, as well as the incorporation of a statistical model that is more accurate than the Gaussian assumption provide interesting directions for further research.

References

1. Drews, M.: Speaker localization and its application to time delay estimators for multi-microphone speech enhancement systems. In: European Signal Processing Conference (1996)
2. Busso, C., Hernanz, S., Chu, C.W., Kwon, S., Lee, S., Georgiou, P.G., Cohen, I., Narayanan, S.: Smart room: participant and speaker localization and identification. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (2005)
3. Chen, X., Shi, Y., Jiang, W.: Speaker tracking and identifying based on indoor localization system and microphone array. In: International Conference on Advanced Information Networking and Applications (2007)
4. Evers, C., Moore, A.H., Naylor, P.A.: Acoustic simultaneous localization and mapping (A-SLAM) of a moving microphone array and its surrounding speakers. In: International Conference on Acoustics, Speech and Signal Processing (2016)
5. Schymura, C., Grajales, J.D.R., Kolossa, D.: Monte Carlo exploration for active binaural localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2017)
6. Zhang, C., Florencio, D., Ba, D.E., Zhang, Z.: Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. *IEEE Trans. Multimedia* **10**(3), 538–548 (2008)
7. Zohourian, M., Enzner, G., Martin, R.: On the use of beamforming approaches for binaural speaker localization. In: ITG Symposium on Speech Communication (2016)

8. Schmidt, R.: Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propag.* **34**(3), 276–280 (1986)
9. Ishi, C.T., Chatot, O., Ishiguro, H., Hagita, N.: Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. In: *International Conference on Intelligent Robots and Systems* (2009)
10. Rickard, S., Dietrich, F.: DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In: *IEEE Workshop on Statistical Signal and Array Processing* (2000)
11. Nadiri, O., Rafaely, B.: Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(10), 1494–1505 (2014)
12. Moore, A.H., Evers, C., Naylor, P.A.: Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(1), 178–192 (2017)
13. Rafaely, B., Kolossa, D., Maymon, Y.: Towards acoustically robust localization of speakers in a reverberant environment. In: *Hands-free Speech Communications and Microphone Arrays* (2017)
14. Kuttruff, H.: *Room Acoustics*. Taylor & Francis, Boca Raton (2009)
15. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**(3), 424–438 (1969)
16. Granger, C.W.J.: Time series analysis, cointegration, and applications. *Am. Econ. Rev.* **94**(3), 421–425 (2004)
17. Rafaely, B., Kolossa, D.: Speaker localization in reverberant rooms based on direct path dominance test statistics. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2017)
18. Rafaely, B.: *Fundamentals of spherical array processing*. STSP, vol. 8. Springer, Heidelberg (2015). <https://doi.org/10.1007/978-3-662-45664-4>
19. Barnett, L., Seth, A.K.: The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *J. Neurosci. Methods* **223**, 50–68 (2014)
20. Barrett, A., Barnett, L., Seth, A.K.: Multivariate granger causality and generalized variance. *Phys. Rev. E* **81**(4), 041907 (2010)
21. Box, G.E.P.: Non-normality and tests on variances. *Biometrika* **40**(3–4), 318–335 (1953)
22. Kim, S., Putrino, D., Ghosh, S., Brown, E.N.: A granger causality measure for point process models of ensemble neural spiking activity. *PLOS Comput. Biol.* **7**(3), 1–13 (2011)
23. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York (2000)
24. *Vector autoregressive models for multivariate time series. Modeling Financial Time Series with S-PLUS®*. Springer, New York (2006)
25. de Waele, S., Broersen, P.M.T.: Order selection for vector autoregressive models. *IEEE Trans. Sign. Process.* **51**(2), 427–433 (2003)
26. Alaeddini, A., Dogan, I.: Using Bayesian networks for root cause analysis in statistical process control. *Expert Syst. Appl.* **38**(9), 11230–11243 (2011)
27. Bondy, A., Murty, U.S.R.: *Graph theory*. Graduate Texts in Mathematics. Springer, London (2011)
28. Tarjan, R.: Depth-first search and linear graph algorithms. In: *Annual Symposium on Switching and Automata Theory* (1971)
29. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **65**(4), 943–950 (1979)



Relative Transfer Function Estimation from Speech Keywords

Ryan M. Corey^(✉) and Andrew C. Singer

University of Illinois at Urbana-Champaign, Urbana, IL, USA
corey1@illinois.edu

Abstract. Far-field speech capture systems rely on microphone arrays to spatially filter sound, attenuating unwanted interference and noise and enhancing a speech signal of interest. To design effective spatial filters, we must first estimate the acoustic transfer functions between the source and the microphones. It is difficult to estimate these transfer functions if the source signals are unknown. However, in systems that are activated by a particular speech phrase, we can use that phrase as a pilot signal to estimate the relative transfer functions. Here, we propose a method to estimate relative transfer functions from known speech phrases in the presence of background noise and interference using template matching and time-frequency masking. We find that the proposed method can outperform conventional estimation techniques, but its performance depends on the characteristics of the speech phrase.

Keywords: Relative transfer function
Multichannel source separation · Keyword spotting · Microphone array

1 Introduction

In many audio processing applications, such as voice assistants and augmented listening devices, we wish to isolate a single speech signal of interest from background noise and interference. These systems can use microphone arrays to spatially filter audio signals, emphasizing sounds from a target direction and attenuating signals from other directions [1]. Multichannel processing has been shown to improve the performance of speech recognition systems in noisy environments [2]. Arrays can also be used in hearing aids and other listening devices to enhance human hearing [3]. In order to filter out interference, a system must determine which signals are coming from which source. We can differentiate sources using their relative transfer functions (RTF), which describe differences in sound propagation between sources and microphones and are generally different for sources in different locations [4]. In environments with significant reverberation, particularly when devices are placed next to walls or other reflecting surfaces, the RTFs are difficult to predict geometrically and must be estimated from observed data.

Parts of this research were completed at Arm Research. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1144245.

Because the target speech and unwanted interference signals are generally unknown, RTF estimation is a difficult problem. If the desired signal is stronger than the background noise or if the noise statistics can be reliably estimated, then the RTFs can be estimated using subspace techniques [5, 6]. The RTFs can also be estimated using a variety of blind source separation techniques that rely on assumptions about the properties of the signals [7, 8]. It would be more reliable to estimate RTFs using a known pilot signal like those used in communication systems [9], but such signals are typically unavailable. In some applications, however, we do have partial knowledge of the content of the speech. We can therefore use the speech itself as a pilot signal to estimate the RTFs.

In this work, we consider audio capture systems that are activated by a certain speech phrase, known as a keyword. Such keywords are often used to remotely activate voice assistants on mobile phones and other electronic devices. These systems use low-power keyword spotting algorithms to continuously monitor for the speech phrase, then activate the full recognition system once it is detected [10]. Because the content of this speech phrase is known in advance, we can use the keyword to better estimate the RTFs of the speaker. Here, we propose an RTF estimation system that matches a multichannel recording to a prerecorded template of the keyword, uses that template to isolate the keyword in each channel, and estimates the RTFs from those isolated recordings. To demonstrate the source separation utility of the keyword alone, we do not apply any other blind source separation techniques and we use no information about the array geometry. A key question in this study is the impact of the choice of keyword on the performance of the system: how do the length and spectral content of the keyword affect the accuracy of the RTF estimate? We will demonstrate the performance of the system and address this question using a crowdsourced database of speech commands and a microphone array similar to those used in commercial voice-assistant-enabled speakers.

2 Far-Field Audio Capture

A far-field audio capture system is shown in Fig. 1. Sound is captured by an array of M microphones, which we assume to behave linearly but which may have arbitrary locations and frequency responses. The system continuously records from all M microphones while it waits for the keyword. The signals are processed as follows:

1. A keyword spotting algorithm, which we assume to work perfectly, activates the system upon detecting the keyword.
2. Once the system is activated, the keyword is used to estimate the relative transfer functions of the source.
3. The RTFs are used to design a source separation filter that isolates the speech following the keyword and suppresses interference and noise.
4. The separated speech is then reproduced, stored, transmitted, or processed further, depending on the application.

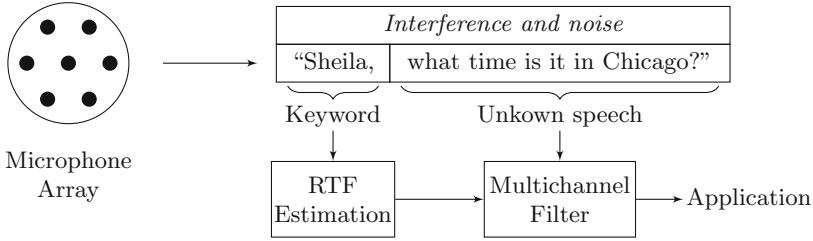


Fig. 1. A far-field audio capture system uses a known speech keyword to design a multichannel filter and separate the unknown speech.

2.1 Microphone Array System

Let $s(n, k)$ be the short-time Fourier transform (STFT) of the signal of interest at the first microphone, where n is the frame index and k is the frequency index. Let $\mathbf{x}(n, k)$ be the M -dimensional STFT vector of mixture signals received by the M microphones. Under the multiplicative transfer function model [8], the mixture is given by

$$\mathbf{x}(n, k) = \mathbf{a}(k)s(n, k) + \mathbf{z}(n, k) \quad (1)$$

$$= \mathbf{c}(n, k) + \mathbf{z}(n, k), \quad (2)$$

where $\mathbf{z}(n, k)$ is the M -dimensional STFT vector of unwanted interference and noise signals received by the microphones, $\mathbf{a}(k)$ is the vector of RTFs, and $\mathbf{c}(n, k) = \mathbf{a}(k)s(n, k)$ is the noise-free vector of source images. Because $s(n, k)$ is defined with respect to the first microphone, $a_1(k) = 1$ for all k . The RTFs depend on the relative positions of the source and microphones, the reverberation characteristics of the space, and the frequency responses and directionalities of the microphones, which may be unknown.

2.2 Source Separation

To isolate the signal of interest, $s(n, k)$, from the mixtures $\mathbf{x}(n, k)$, we use an M -channel spatial filter $\mathbf{w}(k)$, sometimes known as a filter-and-sum beamformer:

$$\hat{s}(n, k) = \mathbf{w}^H(k)\mathbf{x}(n, k). \quad (3)$$

There are many ways to select the coefficients. Here, we restrict our attention to the minimum power distortionless response (MPDR) coefficients [11],

$$\mathbf{w}(k) = \frac{\Sigma_x^{-1}(k)\mathbf{a}(k)}{\mathbf{a}^H(k)\Sigma_x^{-1}(k)\mathbf{a}(k)}, \quad (4)$$

where $\Sigma_x(k) = \mathbb{E}[\mathbf{x}(n, k)\mathbf{x}^H(n, k)]$ is the covariance matrix of the mixture. The MPDR filter minimizes the expected power of $\mathbf{w}^H(k)\mathbf{x}(n, k)$ while ensuring that

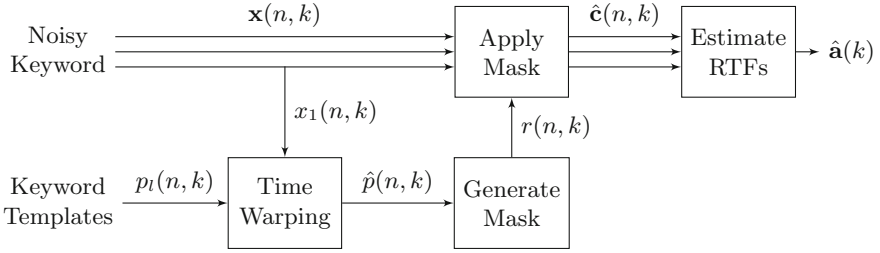


Fig. 2. The relative transfer functions are estimated from the noisy recording using a time-warped template and a time-frequency mask.

$\mathbf{w}^H(k)\mathbf{a}(k)s(n, k) = s(n, k)$. To compute the coefficients, we must first estimate both $\Sigma_x(k)$ and $\mathbf{a}(k)$. In our experiments, the mixture covariance matrix is estimated from the recording itself,

$$\hat{\Sigma}_x(k) = \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n, k)\mathbf{x}^H(n, k). \quad (5)$$

The MPDR filter is known to be sensitive to errors in the estimate of $\mathbf{a}(k)$ [11]. While this is a disadvantage in practice, it is helpful in illustrating the RTF estimation performance of the system.

3 Relative Transfer Function Estimation

If the source and microphone positions or the room acoustics are unknown, then the RTFs must be estimated blindly from the noisy mixture data. Fortunately, in keyword-activated systems, the keyword itself can act as a pilot signal to measure the acoustic channel. Of course, the keyword signal as uttered by the speaker is not known exactly; it must itself be estimated from the noisy mixture.

The proposed method, shown in Fig. 2, combines classic template matching algorithms and modern single-channel source separation methods:

1. Use dynamic time warping to match the recorded keyword to a template keyword from a database.
2. Use the warped template to generate a time-frequency mask consistent with the recorded keyword.
3. Apply the mask to each of the M channels of the mixture to isolate the recorded keyword from interference and noise.
4. Estimate the RTFs from the spatial correlation of the masked data.

To better analyze the performance of keyword-based RTF estimation and to compare different keywords, we do not apply any other blind source separation techniques and we do not use information from the speech following the keyword.

3.1 Template Matching

Template matching is a classic small-vocabulary speech recognition technique [12]. The recorded keyword signal $x_1(n, k)$ is matched to one of L templates $p_l(n, k)$ from a database. Since the sounds within a keyword can be uttered at different speeds, the templates are warped to match the time scale of the recording. Mathematically, we find the best-fitting template and the corresponding time mapping by solving the minimization problem

$$\hat{l}, \hat{t}(n) = \arg \min_{l, t(n)} \sum_n \text{Cost}(x_1(n, 1), \dots, x_1(n, K); p_l(t(n), 1), \dots, p_l(t(n), K)), \quad (6)$$

where $t(n)$ is nondecreasing. In our experiments, the cost function is the Euclidean distance between the Mel frequency cepstral coefficients of each pair of frames. The optimization problem (6) can be solved using dynamic programming [12]. The warped template is given by

$$\hat{p}(n, k) = p_{\hat{l}}(\hat{t}(n), k), \quad \text{for } k = 1, \dots, K. \quad (7)$$

Note that in dynamic time warping, it is customary to warp the time scales of both the recording and the template to find the closest match. Here, we warp the time scale of the template to match that of the recording.

3.2 Time-Frequency Masking

Because speech and other signals are sparse in the time-frequency domain, mixtures of several such sources can be effectively separated by assigning each time-frequency bin to a single source [13]. This process is known as time-frequency masking, and is often used in single-channel source separation. First, a mask is generated by comparing the power in the warped template to a threshold:

$$r(n, k) = \begin{cases} 1, & \text{if } |\hat{p}(n, k)|^2 > \gamma(k) \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The threshold $\gamma(k)$ is a tuning parameter. In our experiments, we set it so that roughly 10% of the mask frames are 1 for each frequency bin k .

To isolate the keyword in the recording from interference and noise, we apply the time-frequency mask to each channel:

$$\hat{c}_m(n, k) = x_m(n, k)r(n, k). \quad (9)$$

If the signals are indeed sparse and if the mask is a good fit, then for nonzero values of $\hat{c}_m(n, k)$, we have $|a_m(k)s(n, k)|^2 \gg |z_m(n, k)|^2$, so that

$$\hat{\mathbf{c}}(n, k) \approx \mathbf{a}(k)s(n, k). \quad (10)$$

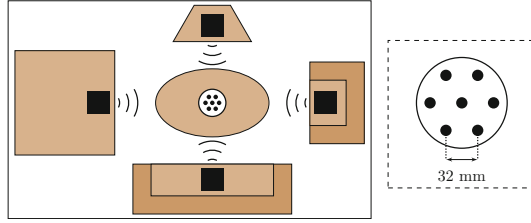


Fig. 3. Experimental setup using a MEMS microphone array in a living room.

3.3 Relative Transfer Functions

Finally, we use the masked signals to estimate the relative transfer functions. We compute the sample covariance matrix of the masked source spatial images:

$$\hat{\Sigma}_c(k) = \frac{1}{N} \sum_{n=1}^N \hat{\mathbf{c}}(n, k) \hat{\mathbf{c}}^H(n, k). \quad (11)$$

If (10) held exactly, then $\hat{\Sigma}_c(k)$ would be a rank-1 matrix proportional to $\mathbf{a}(k) \mathbf{a}^H(k)$. Let $\mathbf{u}(k)$ be the singular vector corresponding to the largest singular value of $\hat{\Sigma}_c(k)$. Then the estimated RTF vector is

$$\hat{\mathbf{a}}(k) = \frac{\mathbf{u}(k)}{u_1(k)}. \quad (12)$$

This is a special case of covariance whitening RTF estimation [6] where the noise is reduced by time-frequency masking rather than whitening. Related classification-based RTF estimation methods incorporate speech presence probabilities [14] and sparsity assumptions [15] to improve the mask.

4 Experiments

To evaluate the performance of the proposed separation method, we present empirical results for RTF estimation and source separation in a cocktail party scenario. The recording device, which is designed for voice assistant applications, is a circular array of $M = 7$ digital MEMS microphones spaced about 32 mm apart, as shown in Fig. 3. The array sits on a coffee table in the center of a living room ($T_{60} \approx 400$ ms) and four signals are emitted from loudspeakers placed on a television stand, sofa, chair, and dining table between one and two meters away. One source is designated the target and the other three are interference.

Impulse responses were measured using sweep signals and used to simulate speech mixtures from prerecorded data. The keywords, examples of which are shown in Fig. 4, are taken from a crowdsourced database of one-second spoken commands [16]. The samples were recorded in widely varying environments with different equipment, reverberation characteristics, and noise levels, so the acoustic simulation is less realistic than it would be with samples recorded in controlled

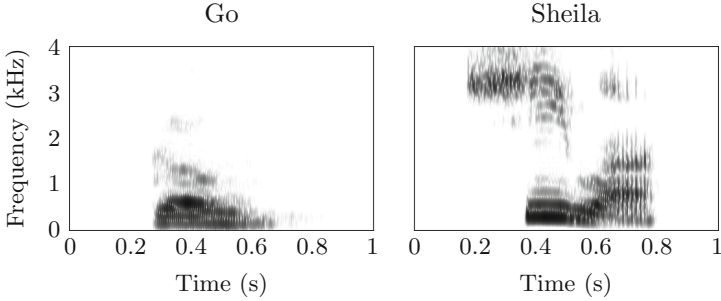


Fig. 4. Spectrograms of two keywords with different RTF estimation performance.

anechoic conditions. Recordings with excessive background noise were removed and the clips were normalized to a constant average power. The experiments below use a set of $L = 500$ templates and a separate test set of 100 utterances for each keyword. For each trial, four ten-second speech clips are selected at random from a subset of the TIMIT database [17]. The mixtures also include a multichannel recording of living room background noise from appliances and ventilation. In these experiments, all signals are sampled at 8 kHz and the STFT uses a length-1024 discrete Fourier transform, a von Hann window of length 1024 samples (128 ms), and a step size of 256 samples (32 ms) .

4.1 Relative Transfer Function Estimation Results

The MPDR beamformer, like many related multichannel filters, reduces noise and interference by projecting the mixture vector onto the RTF vector of the target source. If the estimated RTF vector is not parallel to the source image vector, the source will be distorted and unwanted noise might be amplified. Thus, to measure RTF estimation performance, we use the angle between the true and estimated RTF vectors, averaged across frequency bins:

$$\text{RTF Error} = \frac{1}{K} \sum_{k=0}^{K-1} \arccos \text{Re} \left[\frac{\hat{\mathbf{a}}^H(k) \mathbf{a}(k)}{|\hat{\mathbf{a}}(k)| |\mathbf{a}(k)|} \right]. \quad (13)$$

Figure 5 shows RTF estimation error as a function of the input signal-to-interference-plus-noise ratio (SINR) of the keyword recording. The plots on the left show estimation performance using the ideal binary mask (IBM), which is one when $|s(n, k)|^2 > |z_1(n, k)|^2$ and zero otherwise. The IBM experiment shows the effect of keyword choice on RTF estimation performance if the keyword and noise signals were known perfectly. The plots on the right show the performance of the proposed method with template matching and mask estimation.

It is clear that longer keywords are better than shorter keywords, but there is significant variation even between keywords with the same number of syllables. Keywords that contain sibilants (“yes”, “Sheila”) and thus strong high-frequency

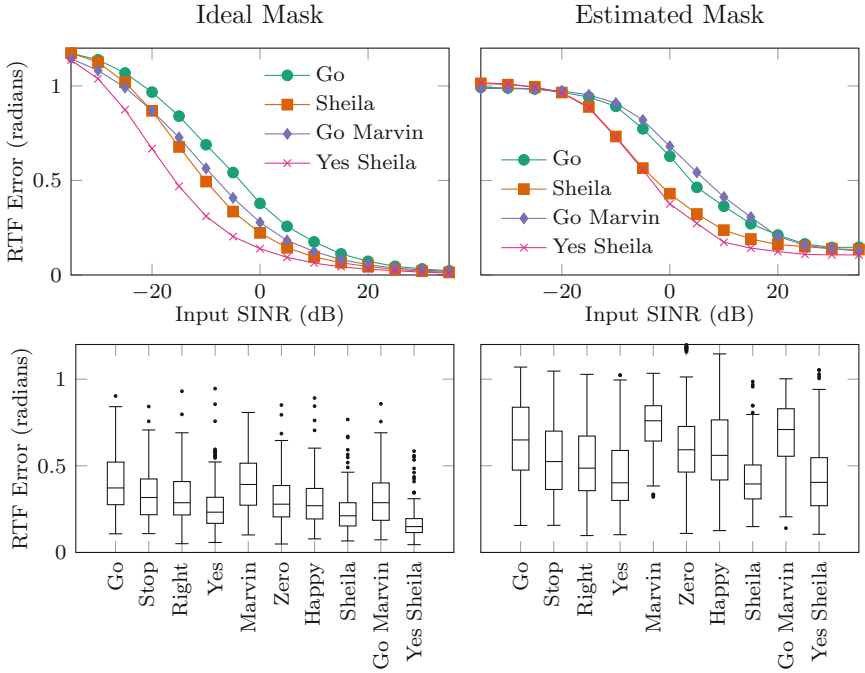


Fig. 5. RTF estimation performance using different keywords. Top: RTF error versus input SINR. Bottom: RTF error at 0 dB input SINR.

content appear to outperform keywords that do not. These keywords are easier to align with templates and cover more of the speech spectrum.

4.2 Source Separation Results

The ultimate goal of the proposed method is to improve source separation performance in a far-field speech capture system. We measure separation performance using the signal-to-error ratio (SER), computed in the time domain:

$$\text{SER} = 10 \log_{10} \frac{\sum_t s^2(t)}{\sum_t (\hat{s}(t) - s(t))^2}. \quad (14)$$

Figure 6 shows the SER for mixtures of four speech sources and background noise at an input SINR of about -4 dB. The plot on the left shows the SER as a function of the keyword input SINR (the input SINR of the unknown speech was not varied). The proposed method provides a roughly 20 dB keyword SINR improvement over the blind RTF estimator, which selects the dominant singular vector of $\hat{\Sigma}_x(k)$ at each frequency. There is a significant gap between the ideal and estimated mask performance, suggesting that there is room for improvement in the template-matching and mask estimation algorithms. The plot on the right

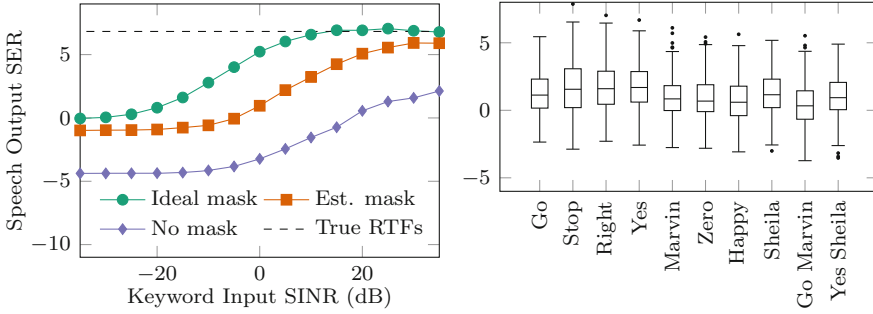


Fig. 6. Source separation performance with four speech sources. Left: Median speech output SER versus keyword input SINR with keyword “Yes Sheila”. Right: Speech output SER at 0 dB keyword input SINR.

shows the output SER when the keyword input SER is 0 dB. The output SERs vary less than the RTF errors for different keywords, and keywords that include sibilants do not have a clear advantage. Since the average spectrum of speech signals is concentrated at low frequencies, high-frequency RTF errors have a smaller impact on the separated speech signal.

5 Conclusions

The experiments show that speech keywords can be used as pilot signals to estimate the RTFs of a source in a noisy mixture. The proposed method is most useful when the interference and noise statistics are not known in advance, so covariance whitening and other model-based RTF estimation methods cannot be applied. In these situations, our experiments suggest that keyword-based RTF estimation can dramatically improve source separation performance.

The accuracy of the RTF estimate appears to depend on the length and the spectral content of the keyword. The most useful keywords have a variety of sounds, making them easy to separate by masking and ensuring that the full speech spectrum is captured by the template. The choice of keyword has a smaller impact on the performance of the separator, suggesting that the method may be useful for some applications even with keywords that are short and spectrally concentrated. In this work, we have used relatively simple algorithms for template matching, mask estimation, and source separation. While these are adequate for this proof of concept, our results suggest that more sophisticated algorithms could improve performance.




Many source separation methods rely on assumptions about the geometry of the array or the statistics of the source signals. However, we can also leverage information about the *content* of the signals. This study has shown that we can effectively separate a speech source from strong interference based only on our knowledge of a single word.

References

1. Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A.: A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(4), 692–730 (2017)
2. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second ‘CHiME’ speech separation and recognition challenge: datasets, tasks and baselines. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 126–130 (2013)
3. Doclo, S., Kellermann, W., Makino, S., Nordholm, S.E.: Multichannel signal enhancement algorithms for assisted listening devices. *IEEE Sign. Process. Mag.* **32**(2), 18–30 (2015)
4. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Sign. Process.* **49**(8), 1614–1626 (2001)
5. Cohen, I.: Relative transfer function identification using speech signals. *IEEE Trans. Speech Audio Process.* **12**(5), 451–459 (2004)
6. Markovich, S., Gannot, S., Cohen, I.: Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Trans. Audio Speech Lang. Process.* **17**(6), 1071–1086 (2009)
7. Makino, S., Lee, T.W., Sawada, H.: *Blind Speech Separation*. vol. 615. Springer, New York (2007)
8. Pedersen, M.S., Larsen, J., Kjems, U., Parra, L.C.: Convolutional blind source separation methods. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (eds.) *Springer Handbook of Speech Processing*. SH, pp. 1065–1094. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9_52
9. Corey, R., Singer, A.: Real-world evaluation of multichannel audio enhancement using acoustic pilot signals. In: *Asilomar Conference on Signals, Systems, and Computers*. (2017)
10. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4087–4091 (2014)
11. Van Trees, H.: *Optimum Array Processing*. Wiley, New York (2004)
12. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics Speech Sign. Process.* **26**(1), 43–49 (1978)
13. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Sign. Process.* **52**(7), 1830–1847 (2004)
14. Araki, S., Okada, M., Higuchi, T., Ogawa, A., Nakatani, T.: Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 385–389 (2016)
15. Koldovský, Z., Málek, J., Gannot, S.: Spatial source subtraction based on incomplete measurements of relative transfer function. *IEEE Trans. Audio Speech Lang. Process.* **23**(8), 1335–1347 (2015)
16. Warden, P.: *Speech commands: a public dataset for single-word speech recognition* (2017). Dataset available from http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz
17. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D.: *TIMIT acoustic-phonetic continuous speech corpus LDC93S1*. Web Download (1993)



On the Number of Signals in Multivariate Time Series

Markus Matilainen^{1,2}, Klaus Nordhausen³, and Joni Virta^{1,4}

¹ University of Turku, Turku, Finland

² Turku PET Centre, Turku University Hospital and University of Turku, Turku, Finland

³ Vienna University of Technology, Vienna, Austria

⁴ Aalto University, Helsinki, Finland
joni.virta@aalto.fi

Abstract. We assume a second-order source separation model where the observed multivariate time series is a linear mixture of latent, temporally uncorrelated time series with some components pure white noise. To avoid the modelling of noise, we extract the non-noise latent components using some standard method, allowing the modelling of the extracted univariate time series individually. An important question is the determination of which of the latent components are of interest in modelling and which can be considered as noise. Bootstrap-based methods have recently been used in determining the latent dimension in various methods of unsupervised and supervised dimension reduction and we propose a set of similar estimation strategies for second-order stationary time series. Simulation studies and a sound wave example are used to show the method's effectiveness.

1 Time Series Modelling via Blind Source Separation

Consider a multivariate time series $\mathbf{x}_t = (x_{1t}, \dots, x_{pt})^\top \in \mathbb{R}^p$, $t \in \{1, \dots, T\}$, commonly encountered in contemporary applications in the form of e.g. climate, financial, EEG, MEG of fMRI-data [1]. Naturally, in each of these cases the series can have dependency both within and between the individual series and it is this richness of structure that sets multivariate time series analysis apart from its univariate counterpart. Needless to say, the added complexity comes with a price: already in the simplest first-order vector autoregressive VAR(1)-model [2], where each time point linearly depends on the values of the previous time only, it takes a total of $2p^2$ parameters to describe the full covariance structure of the model, and with any more sophisticated models the number of parameters inflates even further. The problem with modelling is further amplified when the dimensionality p is large: as multivariate data often contains varying quantities of redundancy and noise some of the model parameters are actually used to model them while in reality we could resort to a simpler model.

A simultaneous solution to both previous problems is given by (linear) *blind source separation* (BSS) [3]. In our time series context, we assume in BSS that

the observed series \mathbf{x}_t is an invertible mixture of some latent series \mathbf{z}_t with a simpler dependency structure, i.e.

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_t, \quad t \in \{1, \dots, T\}, \quad (1)$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location vector and the *mixing matrix* $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ is invertible. Furthermore, \mathbf{z}_t is usually assumed to be weak second-order stationary and its component series temporally uncorrelated,

$$\mathbb{E}((\mathbf{z}_t - E(\mathbf{z}_t))(\mathbf{z}_{t+\tau} - E(\mathbf{z}_{t+\tau}))^\top) = \boldsymbol{\Lambda}_\tau \text{ is diagonal for all lags } \tau \in \mathbb{Z}_+.$$

The assumption on stationarity further allows us to fix $E(\mathbf{z}_t) = \mathbf{0}$, $\text{Cov}(\mathbf{z}_t) = \mathbf{I}_p$ as the two moments are in (1) confounded with $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$, respectively. The BSS model (1) equipped with the previous assumptions is commonly known as the *second order separation* (SOS) model [3].

Measurement error and noise are commonly included in the model (1) additively, as $\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Omega}\mathbf{z}_t + \boldsymbol{\epsilon}_t$ where $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$ is a white noise vector [3] representing the two sources of external variation. However, as in this case all estimates of the signals will always be distorted by some noise, we work in the following with the contrasting idea that the noise is not an external but an internal part of the model. That is, we assume that the latent series can be partitioned as $\mathbf{z}_t = (\mathbf{s}_t^\top, \mathbf{w}_t^\top)^\top$ where $\mathbf{w}_t \in \mathbb{R}^{p-k}$ is white noise and the sources of interest (“signals”) in $\mathbf{s}_t \in \mathbb{R}^k$ contain all the time dependency manifested in \mathbf{x}_t . Similar models (with different definitions of “noise”) have been previously used in the context of both unsupervised and supervised dimension reduction in e.g. [4–7]. Compared to the additive noise model the proposed one makes the modelling and predicting of \mathbf{x}_t particularly simple, the process consisting of four steps: estimate the latent series \mathbf{z}_t using some standard method, identify the $p - k$ white noise series among \mathbf{z}_t and discard them, model the remaining k temporally uncorrelated signal series individually, and finally, back-transform the model to the original scale. This recipe avoids both of the previous problems affecting multivariate time series models: the number of parameters is kept in control as instead of modelling a full p -variate time series we model k univariate time series, and the modelling of noise is averted as we discard it prior to the modelling step.

However, the second of the four steps, the estimation of the dimensionality k , is often heavily overlooked in similar contexts in the literature. BSS as a solution to the modelling problem can be seen to have succeeded only partially if our estimate d of k is inconsistent: on one hand, having $d > k$ means that we model noise in the third step, further biasing any predictions made with the model later, and on the other hand, having $d < k$ means that not all of the signal gets captured by the model and we have voluntarily discarded information. The task is similar to that of selecting the number of principal components in principal component analysis where naïve descriptive tools such as the scree plot or the Kaiser rule [8] are commonly used. [9] approached the estimation via the ladle estimator of [10] but, as far as the authors know, no other work towards this goal in the context of time series has been done in the literature. As our current approach, we propose a semi-parametric, bootstrap-based strategy for estimating k .

2 Two SOS Methods and Test Statistics

To motivate our approach we next go through the steps taken in the two most popular SOS methods, AMUSE (algorithm for multiple signals extraction) [11] and SOBI (second order blind identification) [12]. Also, without loss of generality, we assume that all our series are centered, i.e. $\boldsymbol{\mu} = \mathbf{0}$. AMUSE and SOBI both assume the model (1) and the assumptions following it. We denote the lag- τ autocovariance matrix of the series \mathbf{x}_t by $\boldsymbol{\Sigma}_\tau(\mathbf{x}_t) = \mathbb{E}(\mathbf{x}_t \mathbf{x}_{t+\tau}^\top)$, the choice $\tau = 0$ giving the marginal covariance matrix of the series.

The usual starting point in BSS is whitening the data: we estimate the marginal covariance matrix $\boldsymbol{\Sigma}_0(\mathbf{x}_t)$ and standardize the series using its (unique symmetric) inverse root $\boldsymbol{\Sigma}_0(\mathbf{x}_t)^{-1/2}$. This yields us the standardized series $\mathbf{x}_t^{st} = \boldsymbol{\Sigma}_0(\mathbf{x}_t)^{-1/2} \mathbf{x}_t$ with the property that $\boldsymbol{\Sigma}_0(\mathbf{x}_t^{st}) = \mathbf{I}_p$. Some algebra reveals the importance of the standardization for the BSS model: the standardized series satisfies $\mathbf{x}_t^{st} = \mathbf{U} \mathbf{z}_t$ for some unknown orthogonal matrix $\mathbf{U} \in \mathbb{R}^{p \times p}$ [13, 14].

This insight instantly suggests using the eigendecompositions of the autocovariance matrices to recover the missing matrix \mathbf{U} . Following our assumptions, for any fixed lag $\tau_0 > 0$ we have $\boldsymbol{\Sigma}_{\tau_0}(\mathbf{x}_t^{st}) = \mathbf{U} \boldsymbol{\Lambda}_{\tau_0} \mathbf{U}^\top$ where $\boldsymbol{\Lambda}_{\tau_0}$ is diagonal. The diagonal elements of $\boldsymbol{\Lambda}_{\tau_0}$ contain the marginal τ_0 th autocovariances of the latent series and for the white noise series \mathbf{w}_t they naturally equal 0. Thus, assuming that all the k signal series correspond to distinct, non-zero eigenvalues, the related eigenvectors $\mathbf{U}_1 \in \mathbb{R}^{p \times k}$ can be identified up to sign and order, and finally, we obtain the signal series \mathbf{s}_t via the transformation $\mathbf{x}_t \mapsto \mathbf{U}_1^\top \boldsymbol{\Sigma}_0(\mathbf{x}_t)^{-1/2} \mathbf{x}_t$, yielding the AMUSE-solution with lag τ_0 . In practice the time series of interest are selected from the estimated p latent series by inspecting the diagonal values of the estimated $\hat{\boldsymbol{\Lambda}}_{\tau_0}^2$, where the squaring is used simply for convenience to order the components in a decreasing order of interestingness. The noise components can now be identified as being the last $p - d$ components that have “small enough” eigenvalues, $d = 0, \dots, p$, the key question then being what actually is small enough. An equivalent formulation for the problem can be stated via the running means m_{p-d} of the last $p - d$ squared eigenvalues by asking for which d the estimate \hat{m}_{p-d} is “too large”. This prompts to use \hat{m}_{p-d} as a test statistic for testing the null hypothesis,

$$H_{0,d} : \text{The last } p - d \text{ latent series are white noise.}$$

For a fixed d , if the observed value of \hat{m}_{p-d} exceeds some pre-defined critical value we conclude that the result is too unlikely to have been originated under the null hypothesis and infer that the number of signal components is larger than d . Chaining together tests for several null hypotheses $H_{0,d_1}, H_{0,d_2}, \dots$ then allows us to pinpoint the true value $d = k$. However, obtaining the distribution of our test statistic under the null hypothesis is a highly non-trivial task under the general SOS-model, and we thus resort to the bootstrap [15] to obtain the quantiles, the next section detailing several bootstrapping strategies we can use to accurately replicate the null distribution.

AMUSE already gives us a reasonable starting point for devising a test statistic for the signal dimensionality, but suffers from a clear drawback: the signal

components must all have non-zero τ_0 th autocovariances in order to be distinguished from the noise (to be distinguishable from each other the signal autocovariances also need to be mutually distinct but that is irrelevant with respect to our current problem of separating the *noise subspace* from the *signal subspace* as a whole). In practice this necessitates a careful choosing of the single lag τ_0 , possibly using some expert knowledge on the phenomenon at hand. Such inconvenience is avoided with our preferred SOS-method, SOBI. In SOBI one instead chooses a set of lags, \mathcal{T}_0 , and *jointly diagonalizes* all $|\mathcal{T}_0|$ autocovariance matrices of the standardized series corresponding to the lags (with $|\mathcal{T}_0| = 1$ we revert back to AMUSE). The joint diagonalization is captured by the optimization problem

$$\mathbf{U}^\top = \operatorname{argmax}_{\mathbf{V}^\top \mathbf{V} = \mathbf{I}_p} \sum_{\tau \in \mathcal{T}_0} \|\operatorname{diag}(\mathbf{V}^\top \boldsymbol{\Sigma}_\tau(\mathbf{x}_t^{st}) \mathbf{V})\|_F^2,$$

commonly solved using the Jacobi rotation algorithm [16]. For two latent components to be mutually distinguishable by the joint diagonalization it is sufficient that the corresponding marginal autocovariances differ for some lag in \mathcal{T}_0 [12]. In particular, we can distinguish the noise subspace from the signal subspace if all signal series exhibit autocorrelation for at least one lag in \mathcal{T}_0 (which can be a different lag for different signals), prompting us to choose a relatively large set of lags, $\mathcal{T}_0 = \{1, \dots, 12\}$ being a common choice. Thus, a natural test statistic for the null hypothesis $H_{0,d}$ is again obtained by considering “eigenvalues”, the diagonal elements of the estimated as-diagonal-as-possible matrices $\mathbf{A}_\tau = \operatorname{diag}(\mathbf{U}^\top \boldsymbol{\Sigma}_\tau(\mathbf{x}_t^{st}) \mathbf{U})$. Ordering the sums of the squared elements in decreasing order, the running means \hat{m}_{p-d} of the last $p - d$ components of the sample estimate of $\sum_{\tau \in \mathcal{T}_0} \mathbf{A}_\tau^2$ will be “small” for large enough values of d and their null distributions can be used to find the value of d where \hat{m}_{p-d} is too large to have originated under the null hypothesis, again allowing us to identify the correct dimensionality.

3 Bootstrap Tests for the White Noise Dimension

Bootstrap-based methods have recently been used in determining the noise subspace dimension for principal component analysis (PCA), independent component analysis (ICA) and sliced inverse regression (SIR) in [5] and for non-Gaussian component analysis in [6]. As an alternative testing method both works also discuss tests that are based on limiting behaviors of certain functions of the noise eigenvalues. Such asymptotic procedures are indeed efficient when the sample size is high and could certainly be considered in our context as well, if not for the general difficulty of obtaining limiting results for time series models (see however the limiting behaviour of AMUSE and SOBI for linear processes in [17–21]). As such we leave the development of asymptotic testing procedures to a subsequent work and proceed now with bootstrapping tests.

Assume a time series coming from the model (1), fix a candidate for the signal dimension d and let \hat{m}_{p-d} be the test statistic of the previous section, the mean of the last $p - d$ squared eigenvalues produced by either AMUSE or

SOBI (mean of the sums of the squared “eigenvalues” in the case of SOBI). To test the null hypothesis $H_{0,d}$ we need a way to generate samples from the distribution of the model (1) under the null hypothesis. We will consider four different strategies where we always leave the signal part untouched and take bootstrap samples of the noise part under the current null hypothesis, denoted $z_{i,s}^*$ where $i = d + 1, \dots, p$ denotes the component and $s = 1, \dots, T$ the time point.

Parametric bootstrap: The most widely used assumption about the white noise is that it is Gaussian, making all noise features independently and identically $N(0, 1)$ -distributed. The bootstrap samples are then

$$z_{i,s}^* \sim N(0, 1).$$

Naturally, the parametric bootstrap makes the strongest assumptions, in this case that (i) the noise processes are independent, (ii) within a noise process the time points are serially independent and (iii) the noise is Gaussian. Using next non-parametric bootstrap these assumptions can be relaxed in different ways.

Non-parametric bootstrap I: First we relax the distributional assumption while keeping assumptions (i) and (ii), and assume only that the noise distribution is for all noise components the same but not necessarily Gaussian. Then all $(p - k) \times T$ elements in the noise part are iid samples from the same distribution and we can use the combined sample to estimate the empirical distribution function (ecdf) and to sample $(p - d) \times T$ elements from it. Thus

$$z_{i,s}^* \sim \text{ecdf}\{(\hat{\mathbf{z}}_{d+1}^\top, \dots, \hat{\mathbf{z}}_p^\top)^\top\}, \quad i = d + 1, \dots, p, \quad s = 1, \dots, T,$$

where $\hat{\mathbf{z}}_j$ is the T -vector of the estimated j th latent series and $\text{ecdf}\{\mathbf{x}\}$ denotes the ecdf of the samples in \mathbf{x} .

Non-parametric bootstrap II: Another way to relax the third assumption is to keep assumptions (i) and (ii) but assume that each process has a possibly different standardized distribution. In that case each noise series should be bootstrapped individually and independently from the others. Therefore using this strategy the bootstrap samples are obtained as

$$z_{i,s}^* \sim \text{ecdf}\{\hat{\mathbf{z}}_i^\top\}, \quad i = d + 1, \dots, p, \quad s = 1, \dots, T.$$

Non-parametric bootstrap III: The last approach considered relaxes also the independence between the noise processes and just requires that they are uncorrelated and serially independent. Hence the ecdf is now multivariate and a bootstrap sample of vectors is obtained as

$$\mathbf{z}_{n,s}^* \sim \text{ecdf}\{\hat{\mathbf{z}}_{n,1}, \dots, \hat{\mathbf{z}}_{n,T}\}, \quad s = 1, \dots, T,$$

where $\mathbf{z}_{n,s}^* = (z_{d+1,s}^*, \dots, z_{p,s}^*)^\top$ and $\hat{\mathbf{z}}_{n,t} = (\hat{z}_{d+1,t}, \dots, \hat{z}_{p,t})^\top$.

In Algorithm 1 we describe the entire testing procedure for $H_{0,d}$ using SOBI (where the version for AMUSE is obtained by using only a single lag).

Algorithm 1. Testing $H_{0,d}$

Set proposed dimension d , number of resamples R , observed sample \mathbf{X}_i ;
 Estimate the SOBI-solution for \mathbf{X}_i : $\hat{\mathbf{U}}^\top \hat{\Sigma}_0^{-1/2}, \hat{m}_{p-d}$;
for $i \in \{1, \dots, R\}$ **do**
 $\mathbf{Z}_i^* \leftarrow$ bootstrap the last $p - d$ series of $\hat{\mathbf{Z}}_i = \hat{\mathbf{U}}^\top \hat{\Sigma}_0^{-1/2} \mathbf{X}_i$;
 $\mathbf{X}_i^* \leftarrow \hat{\Sigma}_0^{1/2} \hat{\mathbf{U}} \mathbf{Z}_i^*$;
 Estimate the SOBI-solution for \mathbf{X}_i^* : \hat{m}_{p-d}^* ;
Return the p -value: $[\#(\hat{m}_{p-d}^* \geq \hat{m}_{p-d}) + 1]/(R + 1)$;

The addition of one in both the numerator and denominator of the p -value is a commonly used “correction” to avoid the event of obtaining a zero p -value. For some other guidelines concerning bootstrap hypothesis testing, see [22].

The procedure above tests only for a specific value of the signal/noise dimension. To obtain an estimate for the dimension, the changing point from rejection to acceptance of the sequence of null hypotheses is of interest. For that the tests have to be applied sequentially and different strategies are possible. For example, one can start with the assumption that all components are noise and then increase successively the hypothetical signal dimension until for the first time the null hypothesis cannot be rejected or one can start with the hypothesis of a single noise component and increase the noise dimension until the first time the null hypothesis is rejected. Another possibility is to use some divide-and-conquer strategy. Comparing different estimation strategies is however beyond the scope of this paper and will be explored in a future work. The following simulation study focuses on validating the bootstrap hypothesis tests as suggested above.

4 Simulations

In order to assess the performance of the bootstrap tests, we conducted a simulation study with three different settings using 5-dimensional time series. The first two are taken as ARMA-processes: $z_1 \sim ARMA(2, 1)$ with parameters $\phi_1 = 0.5, \phi_2 = 0.2$ and $\theta_1 = 0.5$ and $z_2 \sim MA(5)$ with the parameter vector $\boldsymbol{\theta} = (-0.4, 0.6, -0.3, 0.1, -0.3)$. The final three series are noise with the following distributions in the different settings: Setting 1: $z_3, z_4, z_5 \sim N(0, 1)$; Setting 2: $(z_3, z_4, z_5) \sim \mathbf{t}_5$; Setting 3: $z_3 \sim N(0, 1), z_4 \sim \mathbf{t}_5$ and $z_5 \sim U(-\sqrt{3}, \sqrt{3})$.

In all settings the signal subspace has the true dimension $k = 2$. Setting 1 is possibly the most natural one, in Setting 2 the noise has a spherical 3-variate \mathbf{t}_5 -distribution which means that there is some dependence among the components and in Setting 3 the noise components are independent but have different marginal distributions. As a mixing matrix we used a random matrix $\boldsymbol{\Omega}$, where the elements of the matrix were drawn randomly from the $N(0, 1)$ -distribution. Next the bootstrap p -values based on $M = 200$ and 500 bootstrap

Table 1. Rejection rates in Setting 1 over 200 bootstrap samples and 2000 repetitions.

n	Bootstrap method	AMUSE			SOBI		
		$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$
200	Parametric	0.998	0.042	0.004	1.000	0.049	0.006
200	Non-parametric I	0.998	0.042	0.006	0.998	0.047	0.008
200	Non-parametric II	0.998	0.048	0.006	1.000	0.046	0.005
200	Non-parametric III	0.999	0.046	0.005	1.000	0.052	0.005
500	Parametric	1.000	0.047	0.008	1.000	0.052	0.010
500	Non-parametric I	1.000	0.043	0.007	1.000	0.047	0.008
500	Non-parametric II	1.000	0.046	0.010	1.000	0.050	0.010
500	Non-parametric III	1.000	0.045	0.010	1.000	0.054	0.008
2000	Parametric	1.000	0.053	0.008	1.000	0.048	0.007
2000	Non-parametric I	1.000	0.042	0.006	1.000	0.057	0.007
2000	Non-parametric II	1.000	0.052	0.006	1.000	0.050	0.008
2000	Non-parametric III	1.000	0.052	0.008	1.000	0.048	0.008
5000	Parametric	1.000	0.052	0.008	1.000	0.053	0.006
5000	Non-parametric I	1.000	0.050	0.009	1.000	0.054	0.006
5000	Non-parametric II	1.000	0.054	0.010	1.000	0.050	0.007
5000	Non-parametric III	1.000	0.052	0.007	1.000	0.050	0.006

samples were calculated and the procedure was repeated 2000 times. We used the time series lengths $T = 200, 500, 2000, 5000$ and AMUSE with lag 1 and SOBI with lags $1, \dots, 12$. Tables 1, 2 and 3 show the proportions of rejections at the α -level 0.05 based on 2000 repetitions for hypotheses $H_{0,1}, H_{0,2}$ (the true value which should be the first test we do not reject) and $H_{0,3}$ for each combination of settings and methods with $M = 200$. The results based on $M = 500$ gave very similar results and were thus omitted from the tables.

Based on the simulation results we conclude that all the tests had quite good power and successfully detected if there were non-noise components among the hypothetical noise part. Interestingly, the parametric bootstrap test seems quite robust – it works also in Settings 2 and 3 where the data were generated using other noise processes. The non-parametric bootstrap test I, which is the closest to the parametric one, seems however to be the least effective of the non-parametric bootstrap tests. As the non-parametric bootstrap test III is valid in all three settings, and the other tests do not gain much in the settings they were designed for, this test might be the best choice in practise. As the differences between AMUSE and SOBI seem minor, we advocate SOBI for practical applications as it is usually preferable over AMUSE and most likely estimates the signals better.

Table 2. Rejection rates in Setting 2 over 200 bootstrap samples and 2000 repetitions.

n	Bootstrap method	AMUSE			SOBI		
		$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$
200	Parametric	1.000	0.046	0.007	1.000	0.052	0.008
200	Non-parametric I	0.998	0.045	0.006	1.000	0.030	0.004
200	Non-parametric II	1.000	0.046	0.006	1.000	0.048	0.010
200	Non-parametric III	1.000	0.044	0.006	0.999	0.051	0.012
500	Parametric	1.000	0.052	0.008	1.000	0.043	0.007
500	Non-parametric I	1.000	0.050	0.005	1.000	0.046	0.006
500	Non-parametric II	1.000	0.046	0.006	1.000	0.044	0.006
500	Non-parametric III	1.000	0.051	0.006	1.000	0.046	0.008
2000	Parametric	1.000	0.042	0.003	1.000	0.047	0.007
2000	Non-parametric I	1.000	0.044	0.005	1.000	0.051	0.006
2000	Non-parametric II	1.000	0.048	0.002	1.000	0.047	0.010
2000	Non-parametric III	1.000	0.044	0.003	1.000	0.045	0.008
5000	Parametric	1.000	0.050	0.008	1.000	0.047	0.009
5000	Non-parametric I	1.000	0.068	0.010	1.000	0.055	0.006
5000	Non-parametric II	1.000	0.049	0.005	1.000	0.048	0.008
5000	Non-parametric III	1.000	0.049	0.007	1.000	0.046	0.006

5 Sound Example

To evaluate the method in practice we used it to estimate the dimension of a set of sound recordings mixed with noise. The signal part \mathbf{s}_t was 3-dimensional with the length $T = 50000$ and was obtained from http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi as in [23]. To this we added 17 channels of $N(0, 1)$ -noise to obtain the latent $\mathbf{z} = (s_1, s_2, s_3, w_4, w_5, \dots, w_{20})^\top$ which was mixed with $\Omega \in \mathbb{R}^{20 \times 20}$ containing iid $N(0, 1)$ variables to obtain the “observed” data $\mathbf{x}_t = \Omega \mathbf{z}_t$.

We considered all four bootstrap strategies using AMUSE with lag 1 and SOBI with lags $1, \dots, 12$. Each combination then produced a string of p -values p_0, \dots, p_{19} corresponding respectively to the null hypotheses $H_{0,0}, \dots, H_{0,19}$. The forwards estimate for d is then the first k for which $H_{0,k}$ is not rejected and the backwards estimate for d is $k + 1$ where k is the last $H_{0,k}$ to be rejected. The resulting estimates are shown in Table 4 and reveal that all combinations correctly identify the true signal dimension. Note that as the forwards and backwards estimates yield the true dimension then also any divide-and-conquer methods are bound to find the true dimension in this case.

Table 3. Rejection rates in Setting 3 over 200 bootstrap samples and 2000 repetitions.

n	Bootstrap method	AMUSE			SOBI		
		$H_{0,1}$	$H_{0,2}$	$H_{0,3}$	$H_{0,1}$	$H_{0,2}$	$H_{0,3}$
200	Parametric	0.999	0.044	0.006	0.998	0.051	0.008
200	Non-parametric I	0.999	0.063	0.008	0.999	0.047	0.009
200	Non-parametric II	0.998	0.044	0.006	0.998	0.047	0.008
200	Non-parametric III	0.998	0.044	0.007	0.999	0.046	0.007
500	Parametric	1.000	0.045	0.008	1.000	0.054	0.006
500	Non-parametric I	1.000	0.052	0.005	1.000	0.047	0.006
500	Non-parametric II	1.000	0.044	0.005	1.000	0.052	0.009
500	Non-parametric III	1.000	0.051	0.006	1.000	0.056	0.006
2000	Parametric	1.000	0.050	0.004	1.000	0.051	0.007
2000	Non-parametric I	1.000	0.050	0.006	1.000	0.060	0.006
2000	Non-parametric II	1.000	0.050	0.006	1.000	0.052	0.009
2000	Non-parametric III	1.000	0.049	0.003	1.000	0.046	0.008
5000	Parametric	1.000	0.060	0.007	1.000	0.046	0.009
5000	Non-parametric I	1.000	0.053	0.008	1.000	0.053	0.009
5000	Non-parametric II	1.000	0.058	0.004	1.000	0.045	0.009
5000	Non-parametric III	1.000	0.058	0.006	1.000	0.046	0.006

Table 4. The estimates for d for each combination of bootstrap strategy and methods in the sound example.

Estimator	BSS	Parametric	Non-par I	Non-par II	Non-par III
Forwards	AMUSE	3	3	3	3
Forwards	SOBI	3	3	3	3
Backwards	AMUSE	3	3	3	3
Backwards	SOBI	3	3	3	3

6 Summary

We proposed four bootstrap tests to test the signal subspace dimension in an SOS framework using AMUSE or SOBI. Simulations showed that the different bootstrap tests work generally well and keep the α -level with good rejection power. To estimate the subspace dimension, the tests would need to be applied sequentially, maybe with different strategies and a possible need for multiple testing adjustments. These issues will be addressed in future work, although an application to sound wave data already yielded some evidence that the sequential estimation works in practice. Note that the suggested tests ignore any possible variation coming from the estimation of the signal as these parts are not bootstrapped. Time series bootstrap strategies as described, for example, in [24]

could then be applied also here for the signal parts. This extension will also be explored in future research.

Acknowledgements. The work of KN was supported by the CRoNoS COST Action IC1408.

References

1. Tang, A.C., Sutherland, M.T., McKinney, C.J.: Validation of SOBI components from high-density EEG. *Neuroimage* **25**(2), 539–553 (2005)
2. Lütkepohl, H.: *New Introduction to Multiple Time Series Analysis*. Springer, Heidelberg (2005)
3. Comon, P., Jutten, C.: *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press (2010)
4. Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., Müller, K.R.: In search of non-Gaussian components of a high-dimensional distribution. *J. Mach. Learn. Res.* **7**, 247–282 (2006)
5. Nordhausen, K., Oja, H., Tyler, D.: Asymptotic and bootstrap tests for subspace dimension (2017). <https://arxiv.org/abs/1611.04908v2>
6. Nordhausen, K., Oja, H., Tyler, D., Virta, J.: Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. *IEEE Sign. Process. Lett.* **24**(6), 887–891 (2017)
7. Matilainen, M., Croux, C., Nordhausen, K., Oja, H.: Supervised dimension reduction for multivariate time series. *Econom. Stat.* **4**, 57–69 (2017)
8. Jolliffe, I.: *Principal Component Analysis*. Springer, New York (2002)
9. Nordhausen, K., Virta, J.: Ladle estimator for time series signal dimension. In: *Proceedings of IEEE Statistical Signal Processing Workshop 2018, IEEE SSP 2018*. (2018, To appear)
10. Luo, W., Li, B.: Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**(4), 875–887 (2016)
11. Tong, L., Soon, V., Huang, Y., Liu, R.: AMUSE: A new blind identification algorithm. In: *Proceedings of IEEE International Symposium on Circuits and Systems, IEEE*, pp. 1784–1787 (1990)
12. Belouchrani, A., Abed Meraim, K., Cardoso, J.F., Moulines, E.: A blind source separation technique based on second order statistics. *IEEE Trans. Sign. Process.* **45**, 434–444 (1997)
13. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. *IEE Proc. F* **140**(6), 362–370 (1993)
14. Miettinen, J., Taskinen, S., Nordhausen, K., Oja, H.: Fourth moments and independent component analysis. *Stat. Sci.* **30**, 372–390 (2015)
15. Efron, B.: Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**(1), 1–26 (1979)
16. Clarkson, D.B.: Remark AS R74: A least squares version of algorithm AS 211: The F-G diagonalization algorithm. *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)* **37**(2), 317–321 (1988)
17. Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Statistical properties of a blind source separation estimator for stationary time series. *Stat. Probab. Lett.* **82**, 1865–1873 (2012)

18. Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S.: Deflation-based separation of uncorrelated stationary time series. *J. Multivar. Anal.* **123**, 214–227 (2014)
19. Illner, K., Miettinen, J., Fuchs, C., Taskinen, S., Nordhausen, K., Oja, H., Theis, F.J.: Model selection using limiting distributions of second-order blind source separation algorithms. *Sign. Process.* **113**, 95–103 (2015)
20. Miettinen, J., Illner, K., Nordhausen, K., Oja, H., Taskinen, S., Theis, F.: Separation of uncorrelated stationary time series using autocovariance matrices. *J. Time Ser. Anal.* **37**(3), 337–354 (2016)
21. Taskinen, S., Miettinen, J., Nordhausen, K.: A more efficient second order blind identification method for separation of uncorrelated stationary time series. *Stat. Probab. Lett.* **116**, 21–26 (2016)
22. Hall, P., Wilson, S.R.: Two guidelines for bootstrap hypothesis testing. *Biometrics*, 757–762 (1991)
23. Miettinen, J., Nordhausen, K., Taskinen, S.: Blind source separation based on joint diagonalization in R: the packages JADE and BSSasympt. *J. Stat. Softw.* **76**(2), 1–31 (2017)
24. Lahiri, S.: *Resampling Methods for Dependent Data*. Springer, New York (2003)



A Generative Model for Natural Sounds Based on Latent Force Modelling

William J. Wilkinson^(✉), Joshua D. Reiss, and Dan Stowell

Queen Mary University of London, London, UK
w.j.wilkinson@qmul.ac.uk

Abstract. Generative models based on subband amplitude envelopes of natural sounds have resulted in convincing synthesis, showing subband amplitude modulation to be a crucial component of auditory perception. Probabilistic latent variable analysis can be particularly insightful, but existing approaches don't incorporate prior knowledge about the physical behaviour of amplitude envelopes, such as exponential decay or feedback. We use latent force modelling, a probabilistic learning paradigm that encodes physical knowledge into Gaussian process regression, to model correlation across spectral subband envelopes. We augment the standard latent force model approach by explicitly modelling dependencies across multiple time steps. Incorporating this prior knowledge strengthens the interpretation of the latent functions as the source that generated the signal. We examine this interpretation via an experiment showing that sounds generated by sampling from our probabilistic model are perceived to be more realistic than those generated by comparative models based on nonnegative matrix factorisation, even in cases where our model is outperformed from a reconstruction error perspective.

Keywords: Latent force model · Gaussian processes
Natural sounds · Generative model

1 Introduction

Computational models for generating audio signals are a means of exploring and understanding our perception of sound. Natural sounds, defined here as everyday non-music, non-speech sounds, are an appealing medium with which to study perception since they exclude cognitive factors such as language and musical interpretation. McDermott [1] used synthesis as a means to demonstrate that the human auditory system utilises time-averaged statistics of subband amplitudes to classify sound textures. In a similar vein, Turner [2] constructed a synthesis model based on probabilistic latent variable analysis of those same subband amplitudes. One main advantage of a latent variable approach is the possibility that the uncovered latent behaviour may represent either (*i*) the primitive source that generated the signal, or (*ii*) the latent information that the human auditory system encodes when it calculates time-averaged statistics.

Latent variable analysis captures correlations across multiple dimensions by modelling the data’s shared dependence on some unobserved (latent) variable or function. It is, by its very nature, ill-posed; we typically aim to simultaneously predict both the latent functions *and* the mapping from this latent space to the observation data. As such, infinitely many potential solutions exist and we cannot guarantee that our prediction will encode the true sound source or our true perceptual representation.

The ill-posed nature of the problem necessitates the use of prior information. It is commonly suggested that nonnegativity, smoothness and sparsity form a suitable set of prior assumptions about real life signals. We argue that, even after imposing such constraints, a simple scalar mapping between the latent space and observation space is insufficient to capture all the complex behaviour that we observe in the subband amplitude envelopes of an audio signal. We construct a latent force model (LFM) [3] to incorporate prior knowledge about how amplitude envelopes behave via a discrete differential equation that models exponential decay [4].

Utilising the state space formulation [5], we augment the standard LFM by explicitly including in the current state information from many discrete time steps. This allows us to capture phenomena such as feedback, damping and to some extent reverberation. In this probabilistic approach the latent functions are modelled with Gaussian processes, which provide uncertainty information about our predictions whilst also guaranteeing that the latent functions are smooth. Nonnegativity is imposed via a nonlinear transformation.

Evaluating latent representations is not straightforward. Objective measures of our ability to reconstruct the observation data don’t inform us about the interpretability of our predictions. We hypothesise that if the latent functions capture physically or perceptually meaningful information, then a generative model based on synthesising latent functions that are statistically similar should generate realistic data when projected back to the observation space.

In this paper we introduce a generative model, applicable to a wide range of natural sounds, based on an extended LFM¹ (Sect. 3). Comparative models based on variants of nonnegative matrix factorisation (NMF) are implemented to perform evaluation-by-synthesis, which shows how listeners often perceive the LFM approach to generate more realistic sounds even in cases where NMF is more efficient from a reconstruction error perspective (Sect. 4).

2 Background

The perceptual similarity of two sounds is not determined by direct comparison of their waveforms, but rather by comparison of their statistics [1]. Hence it is argued that prior information for natural sounds should take a statistical form [2]. We argue in Sect. 3 that these statistical representations can be improved

¹ Matlab source code and example stimuli can be found at c4dm.eecs.qmul.ac.uk/audioengineering/natural_sound_generation.

through the inclusion of assumptions about the physical behaviour of sound, resulting in a hybrid statistical-physical prior.

In order to analyse sound statistics, both McDermott [1] and Turner [2] utilise the subband filtering approach to time-frequency analysis, in which the signal is split into different frequency channels by a bank of band-pass filters. The time-frequency representation is then formed by tracking the amplitude envelopes of each subband. McDermott generates sound textures by designing an objective function which allows the statistics of a synthetic signal to be matched to that of a target signal. Turner utilises probabilistic time-frequency analysis combined with probabilistic latent variable analysis to represent similar features. Turner’s approach has the advantage that once the parameters have been optimised, new amplitude envelopes can be generated by drawing samples from the latent distribution. It should be noted that samples drawn from the model will not exhibit the fast attack and slow decay we observe in audio amplitude envelopes, since the model is temporally symmetric.

NMF is a ubiquitous technique for decomposing time-frequency audio data [6–8], however a common criticism is its inability to take into account temporal information. The most common approach to dealing with this issue is to impose smoothness on the latent functions, the idea being that smoothness is a proxy for local correlation across neighbouring time steps. Temporal NMF (tNMF) imposes smoothness by penalising latent functions which change abruptly [8] or by placing a Gaussian process prior over them [9]. An alternative approach is to use a hidden Markov model to capture the changes in an audio signal’s spectral make up over time [10]. High resolution NMF (HR-NMF) models the temporal evolution of a sound by utilising the assumption that natural signals are a sum of exponentially modulated sinusoids, with each frequency channel being assigned its own decay parameter estimated using expectation-maximisation [11].

2.1 Latent Force Models

To incorporate our prior assumptions into data-driven analysis we use latent force models (LFMs) [3], a probabilistic modelling approach which assumes M observed output functions x_m are produced by some $R < M$ unobserved (latent) functions u_r being passed through a set of differential equations. If the chosen set of differential equations represents some physical behaviour present in the system we are modelling, even if only in a simplistic manner, then such a technique can improve our ability to learn from data [12]. This is achieved by placing a Gaussian process (GP) prior [13] over the R latent functions, calculating the cross-covariances (which involves solving the ODEs), and performing regression.

It was shown by Hartikainen and Särkka [5] that, under certain conditions, an equivalent regression task can be performed by reformulating the model (i.e. the ODE representing our physical knowledge of the system) into state space (SS) form, reformulating the GP as a stochastic differential equation (SDE), and then combining them into a joint SS model:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t)) + Lw(t) . \quad (1)$$

Here $\mathbf{x}(t)$ represents the state vector containing $\{x_m(t)\}_{m=1}^M$ and the states of the SDE $\{u_r(t), \dot{u}_r(t), \dots\}_{r=1}^R$, $w(t)$ is a white noise process, \mathbf{f} is the transition function which is dependent on θ , the set of all ODE parameters and GP/SDE hyperparameters, and L is a vector determining which states are driven by the white noise. The model's discrete form is

$$\mathbf{x}[t_k] = \hat{\mathbf{f}}(\mathbf{x}[t_{k-1}], \Delta t_k) + \mathbf{q}[t_{k-1}], \quad (2)$$

where Δt is the time step size, $\hat{\mathbf{f}}$ is the discretised transition function and $\mathbf{q}[t_{k-1}] \sim N(\mathbf{0}, Q[\Delta t_k])$ is the noise term with process noise matrix Q . The corresponding output measurement model is

$$\mathbf{y}[t_k] = H\mathbf{x}[t_k] + \epsilon[t_k], \quad \epsilon[t_k] \sim N(0, \sigma^2), \quad (3)$$

where measurement matrix H simply selects the outputs from the joint model.

The posterior process $\mathbf{x}[t_k]$, i.e. the solution to (2), is a GP in the linear case such that the filtering distribution $p(\mathbf{x}[t_k] \mid \mathbf{y}[t_1], \dots, \mathbf{y}[t_k])$ is Gaussian. Hence state estimation can be performed via Kalman filtering and smoothing [14].

However, if \mathbf{f} is a nonlinear function, as is the case if we wish to impose nonnegativity on the latent functions, then calculation of the predictive and filtering distributions involves integrating equations which are a combination of Gaussian processes and nonlinear functions. We may approximate the solutions to these integrals numerically using Gaussian cubature rules. This approach is known as the cubature Kalman filter (CKF) [15].

The Kalman update steps provide us with the means to calculate the marginal data likelihood $p(\mathbf{y}[t_{1:T}] \mid \theta)$. Model parameters θ can therefore be estimated from the data by maximising this likelihood using gradient-based methods.

3 Latent Force Models for Audio Signals

To obtain amplitude data in the desired form we pass an audio signal through an equivalent rectangular bandwidth (ERB) filter bank. We then use Gaussian process probabilistic amplitude demodulation (GPPAD) [16] to calculate the subband envelopes and their corresponding carrier signals. GPPAD allows for control over demodulation time-scales via GP lengthscale hyperparameters. We are concerned with slowly varying behaviour correlated across the frequency spectrum, in accordance with the observation that the human auditory system summarises sound statistics over time [1]. Fast-varying behaviour is relegated to the carrier signal and will be modelled as independent filtered noise.

The number of channels in the filter bank and the demodulation lengthscales must be set manually during this first analysis stage. Keeping the number of total model parameters small is a priority (see Sect. 3.1), so we typically set the number of filters to 16, and the lengthscales such that we capture amplitude behaviour occurring over durations of 10 ms and slower.

3.1 Augmented Latent Force Models for Amplitude Envelopes

We use a first order differential equation to model the exponential decay that occurs in audio amplitude envelopes [4]. However this overly simplistic model does not take into account varying decay behaviour due to internal damping, or feedback and other nonstationary effects which occur as a sound is generated and propagates towards a listener.

Since we require nonnegativity of our latent functions, which is imposed via nonlinear transformation, we use the nonlinear LFM whose general form is (2) with nonlinear \hat{f} . For a first order ODE its discrete form is

$$\dot{x}_m[t_k] = -D_m x_m[t_k] + \sum_{r=1}^R S_{mr} g(u_r[t_k]), \quad (4)$$

for $m = 1, \dots, M$ where M is the number of frequency channels. D_m and S_{mr} are the damping and sensitivity parameters respectively and $g(u) = \log(1 + e^u)$ is the positivity-enforcing nonlinear transformation. The model progresses forwards in time with step size Δ_t using Euler’s method: $x_m[t_{k+1}] = x_m[t_k] + \Delta_t \dot{x}_m[t_k]$.

To account for the complex behaviour mentioned above that occurs in real audio signals, we extend this discrete model such that predictions at the current time step t_k can be influenced explicitly by predictions from multiple time steps in the past. As in [4] we augment the model by adding a parameter γ_m which controls the “linearity” of decay. Our final model becomes

$$\dot{x}_m[t_k] = -D_m x_m^{\gamma_m}[t_k] + \sum_{p=1}^P B_{mp} x_m[t_{k-p}] + \sum_{q=0}^P \sum_{r=1}^R S_{mrq} g(u_r[t_{k-q}]). \quad (5)$$

We restrict $\gamma_m \in [0.5, 1]$, and for sounding objects with strong internal damping we expect γ_m to be small, representing an almost linear decay. Parameters B_{mp} are *feedback* coefficients which determine how the current output is affected by output behaviour from p time steps in the past. S_{mrq} are *lag* parameters which determine how sensitive the current output is to input r from q time steps ago.

The lag term is important since modes of vibration in a sounding object tend to be activated at slightly different times due to deformations in the object as it vibrates, and due to the interaction of multiple modes of vibration. It can also capture effects due to reverberation. The feedback terms allow for long and varied decay behaviour that can’t be described by simple exponential decay.

The challenge is to incorporate (5) into our filtering procedure. We do this by augmenting our state vector $\mathbf{x}[t_k]$ and transition model

$$\hat{f}(\mathbf{x}[t_{k-1}], \Delta t_k) = \mathbf{x}[t_k] + \Delta_t \dot{\mathbf{x}}[t_k] \quad (6)$$

with new rows corresponding to the delayed terms. Figure 1 shows how after each time step the current states $X[t_k] = \{x_m[t_k]\}_{m=1}^M$, $U[t_k] = \{u_r[t_k]\}_{r=1}^R$ are “passed down” such that at the next time step they are in the locations corresponding to feedback and lag terms. When performing the Kalman filter

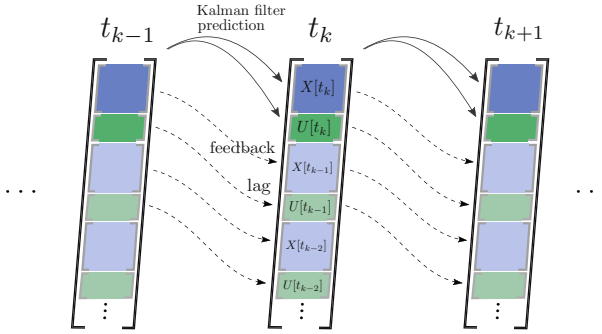


Fig. 1. The augmented LFM stores terms from previous time steps in the state vector. Blue represents output predictions X (amplitudes), green represents latent predictions U . Each step, predictions pass down to feedback and lag state locations. The entire state is used to predict the next step’s outputs and latents via Kalman filtering.

prediction step, augmented states are included since they influence predictions for the current state, however the predictions for these augmented entries are simply exact copies from the previous time step.

Figure 2 shows the latent prediction for a metal impact sound with one latent force, $R = 1$. The mean of the distribution is the minimum least squares error estimate, so we pass it through discrete model (5) to reconstruct the amplitude envelopes. Despite the single latent force, we observe that some of the complex behaviour has been learnt. Additionally, the latent force is both smooth and sparse, and the reconstructed envelopes have a slow decay despite this sparsity.

3.2 Generating Novel Instances of Natural Sounds

A significant benefit of probabilistic approaches such as LFM or tNMF is that, as well as providing us with uncertainty information about our predictions, they provide the means to sample new latent functions from the learnt distribution. By passing these new functions through the model we can generate amplitude envelopes. These envelopes modulate carrier signals produced using a sinusoids-plus-noise approach based on analysis of the original carriers. The subbands are then summed to create a new synthetic audio signal distinct from the original but with similar characteristics.

Sampling from the prior of the learnt distribution generates functions with appropriate smoothness and magnitude, however the desired energy sparsity is not guaranteed. Latent functions are modelled independently, but in practice they tend to co-occur and are activated in similar regions of the signal. We use GPPAD again to demodulate our latent functions with a slowly varying envelope, then fit a GP with a squared exponential covariance function to this envelope [13]. We sample from this high-level envelope and use it to modulate our newly generated latent functions; the results of this product is latent behaviour with sparse energy, as demonstrated in Fig. 3(d).

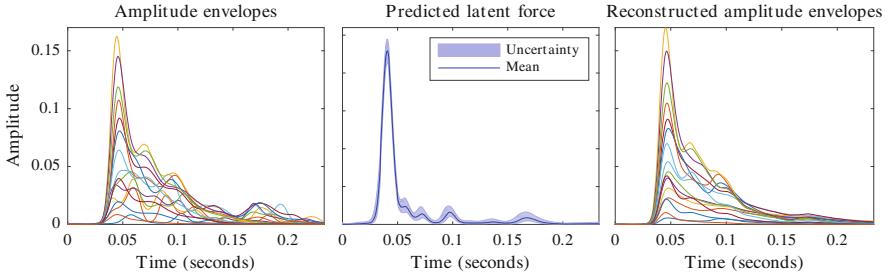


Fig. 2. LFM applied to a metal impact sound, with mean and 95% confidence of the latent distribution shown. The mean is passed through the model (5) to reconstruct the envelopes. Complex behaviour is maintained despite using a single force.

3.3 Optimisation Settings

The set of model parameters $\{D_m, B_{mp}, S_{mrq}, \gamma_m, \lambda_r\}$, with GP lengthscales λ_r , becomes large as R, P increase. To alleviate issues that occur when our parameter space becomes large we sparsify the feedback and sensitivity parameters. For example, if $P = 10$, we may manually fix B_{mp} to zero for $p \in [3, 4, 6, 7, 9]$ such that only half the parameters are included in the optimisation procedure.

Reliability of the optimisation procedure suffers as the number of parameters increases, so in practice all M frequency channels are not optimised together. We select the 6 envelopes contributing the most energy and train the model on the observations from only these channels. The remaining channels are then appended on and optimised whilst keeping the already-trained parameters fixed. This improves reliability but prioritises envelopes of high energy. We also skip prediction steps for periods of the signal that are of very low amplitude, which speeds up the filtering step. Despite these adjustments, optimisation still takes up to 72 h for a 2 s sound sample.

4 Evaluation

To evaluate our method we collated a set of 20 audio recordings, selected as being representative of everyday natural sounds². Music and speech sounds were not included, nor were sounds with significant frequency modulation, since our model doesn't capture this behaviour. We compare against NMF, optimised using alternating least squares, and the GP-based implementation of tNMF [9].

4.1 Reconstruction Error of Original Sound

We analyse our ability to reconstruct the original data by projecting the latent representation back to the output space. For the LFM this means passing the

² From freesound.org and from the Natural Sound Stimulus set: mcdermottlab.mit.edu/svnh/Natural-Sound/Stimuli.html.

mean of the learnt distribution through model (5). Figure 4 shows reconstruction RMS error and cosine distance of LFM and tNMF relative to NMF for the 20 recordings. The smoothness constraint enforced by placing a GP prior over the latent functions negatively impacts the reconstruction. This is demonstrated by the fact that tNMF performs poorly from an RMS error perspective. Despite this, the LFM has much descriptive power, and is sometimes capable of achieving a lower RMS error than the unconstrained NMF. Interestingly however, tNMF consistently outperforms the other two models based on cosine distance.

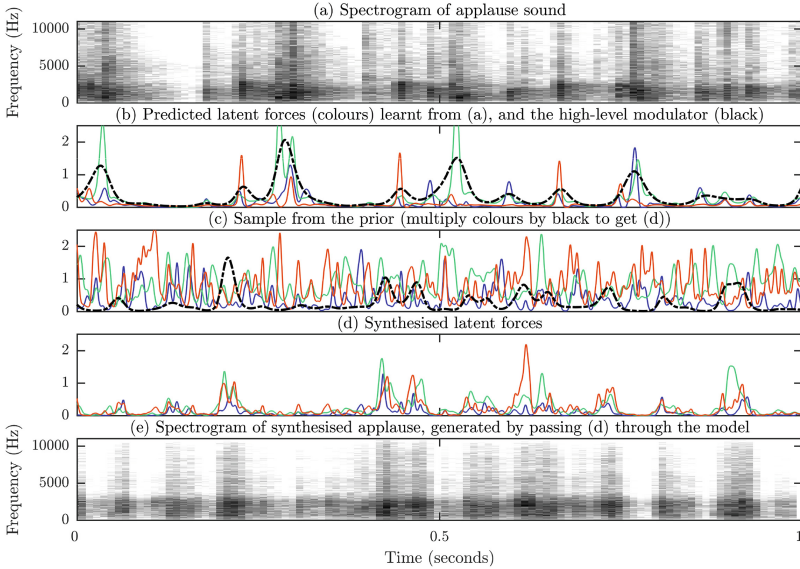


Fig. 3. LFM generative model with 3 latent forces applied to an applause sound. The high-level modulator (black line in (b)) is calculated by demodulating the latent forces.

4.2 Listening Test for Novel Sounds

Objective results suggest that smoothness constraints harm reconstruction of the original signal. However, our aim is to learn realistic latent representations that will be the foundation of a generative model. To test their suitability, we designed an experiment to compare generative models based on LFM, NMF and tNMF. The approach outlined in Sect. 3.2 was used for all model types. Since NMF is non-probabilistic, it does not provide an immediate way in which to sample new data, therefore GPs were fit to the latent functions after analysis.

Our experiment followed a multi-stimulus subjective quality rating paradigm³: 24 participants were shown 20 pages (order randomised), one per

³ The test was run online and implemented with the Web Audio Evaluation Tool: github.com/BrechtDeMan/WebAudioEvaluationTool.

sound example, and asked to listen to the reference recording and then rate 7 generated sounds (2 from each model plus an anchor) based on their credibility as a new sound of the same type as the reference. Ratings were on a scale of 0 to 1, with a score of 1 representing a very realistic sound. Figure 5 shows the mean realism ratings. Whilst variation was large between sound examples, LFM was generally rated as more realistic than the other methods.

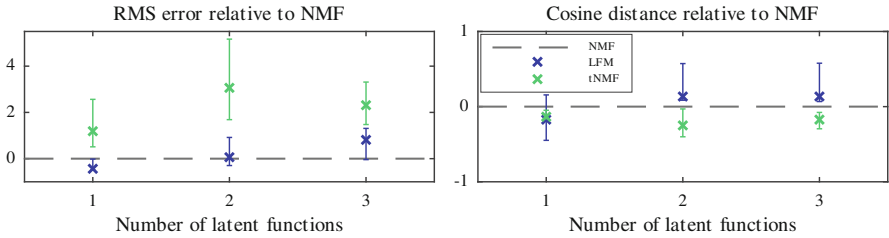


Fig. 4. Reconstruction error of LFM and tNMF plotted relative to NMF. Crosses represent the median, error bars range from first to third quartile.

To test for significance we applied a generalised linear mixed effects model (GLMM), with beta regression, in which *sound example* and *participant* were treated as random effects. Table 1 shows that the mean realism rating was highest for LFM regardless of number of latent functions. The difference was significant at a 5% level except for LFM vs. NMF with 3 latent functions. This suggests that for sounds requiring many latent functions to capture their behaviour, such as textural sounds, LFM may not offer a significant gain over purely statistical approaches. For example, the wind recording in Fig. 5, a textural sound whose envelopes do not exhibit clear exponential decay, was captured best with tNMF.

Table 1. GLMM with three-way comparison applied to listening test results. LFM received higher mean ratings, but confidence decreases with number of latent forces, indicated by increasing *p values*. *Estimate* can be interpreted as the ratio increase in realism rating when choosing model A over model B.

	All sounds		1 latent fn.		2 latent fns.		3 latent fns.	
	Estimate	p value	Estimate	p value	Estimate	p value	Estimate	p value
LFM vs. NMF	0.3839	<1e-04	0.8248	<1e-05	0.3140	0.0448	0.2052	0.2867
LFM vs. tNMF	0.4987	<1e-04	0.7976	<1e-05	0.5134	<0.001	0.3243	0.0285
NMF vs. tNMF	0.1148	0.3750	-0.0272	0.9980	0.1994	0.3218	0.1191	0.7154

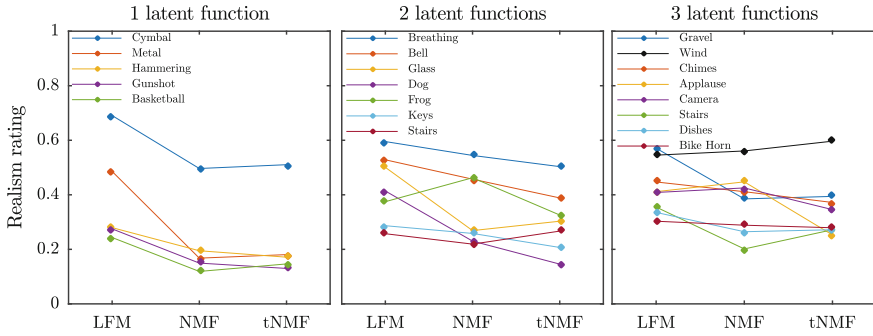


Fig. 5. Mean realism ratings obtained from the listening test.

5 Conclusion

Our results show that in order to extend existing synthesis techniques to a larger class of sounds, it is important to utilise prior knowledge about how natural sound behaves. We achieved this by using latent force modelling to capture exponential decay, and augmented the standard approach to include feedback and delay across many discrete time steps. Doing so allowed us to make smooth, sparse latent predictions that we argue are more representative of the real source that generated a given sound.

This claim is supported by the fact that a generative model based on LFM was consistently rated as more realistic by listeners than alternatives based on variants of NMF, even in cases where it was not superior in reconstruction of the original signal. Resonance, decay and modulations in the subband amplitudes were captured well by our model, which is flexible enough to be applicable to sounds ranging from glass breaking to dogs barking.

The nonlinear ODE representing our physical knowledge contains a large number of parameters, making our approach impractical in some cases, so a more compact model would be of huge benefit. Efficient nonlinear filtering methods or numerical ODE solvers would make the computation time more acceptable. Future work includes amplitude behaviour occurring on multiple time scales at once, and models for frequency modulation and other nonstationary effects would further expand the class of sounds to which such techniques can be applied.

References

1. McDermott, J.H., Simoncelli, E.P.: Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **71**(5), 926–940 (2011)
2. Turner, R.E.: Statistical Models for Natural Sounds. Ph.D. thesis, UCL (2010)
3. Alvarez, M.A., Luengo, D., Lawrence, N.D.: Latent force models. In: International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 12, pp. 9–16 (2009)

4. Wilkinson, W.J., Reiss, J.D., Stowell, D.: Latent force models for sound. In: International Conference on Digital Audio Effects (2017)
5. Hartikainen, J., Särkkä, S.: Sequential inference for latent force models. In: International Conference on Uncertainty in Artificial Intelligence (UAI-11), pp. 311–318 (2011)
6. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the itakura-saito divergence: with application to music analysis. *Neural Comput.* **21**(3), 793–830 (2009)
7. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. In: WASPAA, pp. 177–180 (2003)
8. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process.* **15**(3), 1066–1074 (2007)
9. Turner, R.E., Sahani, M.: Time-frequency analysis as probabilistic inference. *IEEE Trans. Signal Process.* **62**(23), 6171–6183 (2014)
10. 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). https://doi.org/10.1007/978-3-642-15995-4_18
11. Badeau, R.: Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF). In: WASPAA, pp. 253–256 (2011)
12. Alvarez, M.A., Luengo, D., Lawrence, N.D.: Linear latent force models using Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2693–2705 (2013)
13. Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. MIT Press, Cambridge (2006)
14. Särkkä, S.: *Bayesian Filtering and Smoothing*. Cambridge University Press, New York (2013)
15. Hartikainen, J., Seppänen, M., Sarkka, S.: State-space inference for non-linear latent force models with application to satellite orbit prediction. In: 29th International Conference on Machine Learning (ICML), pp. 903–910 (2012)
16. Turner, R.E., Sahani, M.: Demodulation as probabilistic inference. *IEEE Trans. Audio Speech Lang. Process.* **19**(8), 2398–2411 (2011)



Independent Vector Analysis Exploiting Pre-learned Banks of Relative Transfer Functions for Assumed Target's Positions

Jaroslav Čmejla^(✉), Tomáš Kounovský, Jiří Málek, and Zbyněk Koldovský

Acoustic Signal Analysis and Processing Group, Faculty of Mechatronics,
Informatics and Interdisciplinary Studies, Technical University of Liberec,
Studentská 2, 461 17 Liberec, Czech Republic
jaroslav.cmejla@tul.cz

Abstract. On-line frequency-domain blind separation of audio sources performed through Independent Vector Analysis (IVA) suffers from the problem of determining the order of the separated outputs. In this work, we apply a supervised IVA based on pilot components obtained using a bank of Relative Transfer Functions (RTF). The bank is assumed to be available for potential positions of a target speaker within a confined area. In every frame, the most suitable RTF is selected from the bank based on a criterion. The pilot components are obtained as pre-separated target and interference, respectively, through the Minimum-Power Distortionless Beamforming and Null Beamforming. The supervised IVA is tested in a real-world scenario with various levels of up-to-dateness of the bank. We show that the global permutation problem is resolved even when the bank contains only pure delay filters. The Signal-to-Interference Ratio in separated signals is mostly better than that achieved by the pre-separation, unless the bank contains very precise RTFs.

Keywords: Independent vector analysis · Relative transfer function
Source separation · Speech enhancement

1 Introduction

Frequency-Domain Independent Component Analysis (FD-ICA) [18] provides an effective tool for audio signal separation and enhancement. It is an unsupervised method where each frequency band is treated separately as an instantaneous mixture. This causes the permutation problem as the order of separated frequency components is random [17]. Rather than solving the separation in two steps, i.e., by applying FD-ICA and some depermutation method afterwards,

This paper was supported by The Czech Science Foundation through Project No. 17-00902S and partly supported by the Student Grant Scheme 2018 project of the Technical University in Liberec and by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040.

a fast and effective solution is provided through Independent Vector Analysis (IVA). Here, all frequencies are separated jointly; the separated sources should be independent while frequency components corresponding to the same source are forced to be as dependent as possible [5, 8, 15].

The random global order of separated sources is the remaining problem of IVA. Classical solutions impose a constraint on the de-mixing filters obtained through IVA. For example, the filters are constrained to remain close to the pure delay filters where the delays correspond to the expected Directions of Arrival (DOA) of the sources [4]. Unconstrained modifications of IVA have also been proposed exploiting prior or side information. For example, a priori knowledge of temporal power variations of sources is used in [16]. In [9], prior knowledge of target positions was used to initialize the IVA, resulting in faster convergence and global permutation of the targets in the de-mixed signals. This partly solves the global permutation problem, but only when the sources remain in static locations.

A general formulation, referred to as Supervised IVA (S-IVA), has been recently proposed in [13] where higher-order dependencies between so-called pilot components and the separated signals are used. For example, the outputs of a voice activity detector (VAD) and of a video speech detector (VSD) were used as the pilots in [14] to distinguish speakers. In this paper, we propose a cheaper solution relying purely on audio. It is assumed that the position of a targeted speaker is confined to a limited area and that a bank of Relative Transfer Functions (RTFB) for some possible positions of the speaker is available. This bank can be directly used for separation as in [6]. However, it is more realistic to assume that the bank is not that up-to-date due to various changes (variations of acoustic conditions, rotations of the target speaker, new locations, etc.) so that its direct application yields a limited separation accuracy. Therefore, we propose to use the bank for pre-separating the target from the interference and to use these outputs as pilots in the supervised IVA.

The paper is organized as follows. In Sect. 2, S-IVA and a corresponding algorithm are briefly described. In Sect. 3, the concept of the RTFB and its deployment for obtaining pilot components for S-IVA are proposed. Section 4 is devoted to experiments with real-world on-line separation where S-IVA is compared with the original IVA and with the beamforming-based separation relying purely on the RTFB. Section 5 concludes the paper.

2 Blind Separation Using Supervised IVA

2.1 Problem Definition

In this paper, we will constrain to situations where two source signals are recorded by two microphones. Let S_n^k and X_m^k be the Short-Term Fourier Transform (STFT) coefficients of the n th source and the m th microphone, respectively, where k is the frequency bin index. The source and the mixture vector will be denoted, respectively, as $\mathbf{S}^k = [S_1^k, S_2^k]^T$ and $\mathbf{X}^k = [X_1^k, X_2^k]^T$. The mixing model within the k th frequency bin reads

$$\mathbf{X}^k = \mathbf{H}^k \mathbf{S}^k + \mathbf{V}^k, \tag{1}$$

where \mathbf{H}^k is the mixing matrix. The objective of IVA is to jointly estimate the set of de-mixing matrices $\{\mathbf{W}^k\}_{k=1,\dots,K}$; K is the number of frequency bins; see, e.g., [5]. The vector of the n th separated source will be denoted by $\mathbf{Y}_n = [Y_n^1, \dots, Y_n^K]$ where

$$Y_n^k = \sum_{m=1}^2 W_{nm}^k X_m^k, \quad k = 1, \dots, K. \tag{2}$$

Each separated source corresponds to one of the two original sources up to the scaling ambiguity, which we subsequently resolve using Minimal Distortion Principle [11].

2.2 Supervised IVA Using Natural Gradient

The supervised IVA (S-IVA) is based on a joint statistical model of the frequency components corresponding to a source and of additional pilot components, because all these components are assumed to be dependent [13]. For simplicity, we assume only one pilot component, for the n th source, denoted by P_n . As in [8], the multivariate super-Gaussian distribution is used for modeling the joint pdf of \mathbf{Y}_n and of P_n , that is,

$$f(\mathbf{Y}_n, P_n) \propto \exp \left(-\sqrt{\sum_{k=1}^K |Y_n^k|^2 + |P_n|^2} \right). \tag{3}$$

The log-likelihood function for the joint estimation of $\{\mathbf{W}^k\}_k$ is given by

$$\mathcal{L}(\{\mathbf{W}^k\}_k) = \sum_{k=1}^K \log |\det \mathbf{W}^k| + \sum_{n=1}^N \text{E}[\log f(\mathbf{Y}_n, P_n)]. \tag{4}$$

which is maximized using the natural gradient-based learning rules

$$\begin{aligned} \Delta W_{nm}^k &= (I_{nm} - \text{E}[\phi^k(\mathbf{Y}_n, P_n)(Y_m^k)^*])W_{nm}^k, \\ \mathbf{W}_{\text{new}}^k &= \mathbf{W}_{\text{old}}^k + \eta \Delta \mathbf{W}^k, \end{aligned} \tag{5}$$

where η is the step length, I_{nm} is the nm th element of the identity matrix, and $\phi^k = -\partial/\partial Y_m^k(\log f)$, $k = 1, \dots, K$, are the score functions related to (3). In practice, we use ad hoc modifications of the score functions given by

$$\phi^k(\tilde{\mathbf{Y}}_n) = \frac{Y_n^k}{\sqrt{(1 - \beta_n) \sum_{k=1}^K |Y_n^k|^2 + \beta_n |P_n|^2}}. \tag{6}$$

where the hyper-parameter $\beta_n \in (0, 1)$ controls the influence of P_n . In (5), the expectation value is either approximated by the average taken over frames or by the instant value in case of on-line processing.

3 Utilization of the Bank of RTFs

3.1 Bank of Relative Transfer Functions

Given a pair of microphones (we will denote them L and R for left and right), the mixing model (1) can be re-written with respect to one particular (target) source, from here denoted by S , and with respect to the left microphone as

$$\begin{aligned} X_L^k &= S^k + V_L^k \\ X_R^k &= G^k S^k + V_R^k. \end{aligned} \quad (7)$$

S^k denotes the spatial image of the target source on the left microphone, and V_L^k and V_R^k involve the contributions of the other source (in practice also of noise). G^k is the Relative Transfer Function (RTF) related to the microphone pair and to the target source.

Although several methods exist that can estimate G^k from noisy mixtures [2], they can hardly achieve the accuracy of noise-free estimates. These can be obtained when a sufficiently long noise and interference-free interval of recording is available. However, the RTF estimate remains accurate only for the given position of the source. In order to cover the area of the most probable target source occurrence, a bank of RTFs (RTFB) was assumed to be available in [7] such that the RTFs in the bank correspond to several potential target's positions within the confined area. It was assumed that such a bank was prepared in advance during noise-free periods. Then, it can be used in dynamical noisy situations when the target performs movements within the assumed area.

Specifically, in every processing frame, null beamforming using all RTFs can be performed. The RTF corresponding to the null beamformer yielding output with the lowest L_p norm is then selected as the most fitting solution [10]. Since we assume that both target and interference are speech signals, the Null Beamformer using the correct RTF should notice an increased sparsity on its output. Therefore, the value of p is chosen to be $p \leq 1$. Several other methods for selecting the best RTF from the RTFB have also been proposed; see, e.g., [6, 12].

3.2 Pre-separation Using Beamformers

Let us assume for now that G^k is known. We now describe simple approaches for obtaining separated signals of the target and interference. To obtain the target, we can apply a minimum power distortionless beamformer (MPDR) whose output is given by

$$\hat{S}^k = \left(\frac{(\mathbf{C}_x^k)^{-1} \mathbf{u}^k}{(\mathbf{u}^k)^H (\mathbf{C}_x^k)^{-1} \mathbf{u}^k} \right)^H \mathbf{X}^k, \quad (8)$$

where $\mathbf{u}^k = (1, G^k)^T$, and \mathbf{C}_x^k is the covariance matrix of \mathbf{X}^k ; the superscript H denotes the conjugate transpose. In the on-line processing regime, \mathbf{C}_x^k has to be estimated in a recursive way as

$$\mathbf{C}_x^{k,\ell} = \lambda \mathbf{C}_x^{k,\ell-1} + (1 - \lambda) \mathbf{X}^k (\mathbf{X}^k)^H, \quad (9)$$

where ℓ stands for the frame index.

Next, a signal containing only the interference can be obtained through blocking the target signal (null beamforming). Specifically, the reference signal is obtained as

$$Z^k = G^k X_L^k - X_R^k = G^k V_L^k - V_R^k, \quad (10)$$

which involves only V_L^k and V_R^k .

3.3 Pilot Component Definition

The performance of the beamforming approaches highly depends on the accuracy of the RTFs in the RTFB. To achieve optimum separation, the RTFs must be up-to-date with respect to changes of the acoustic environment, the RTFB should cover the entire area of possible target's positions, and the time domain length of the RTFs must be sufficiently long with respect to reverberation. Since these requirements are hardly met in practice, it is better to take into account a limited performance of the beamforming methods.

It is more realistic to assume that the separated signals \hat{S}^k and Z^k are only dominated by the target and interference, respectively. Then, we propose to exploit these signals as pilots within S-IVA, which might finally achieve better separation. Thus, the pilot components are defined as

$$P_1 = \sum_{k=1}^K |\hat{S}^k|, \quad \text{and} \quad P_2 = \sum_{k=1}^K |Z^k|, \quad (11)$$

for the target and the interference output, respectively.

4 Experiments

In this section, we present results of experiments whose goal is to demonstrate the influence of the accuracy of the RTFB on the solution of the global permutation by S-IVA, and to compare the separation accuracies achieved though S-IVA and the beamforming methods from Sect. 3.2.

4.1 Scenario

The experimental setup is illustrated in Fig. 1. Two speakers (simulated by loudspeakers) recorded by two microphones with mutual distance of 18 cm are considered in a room with reverberation time $T_{60} = 700$ ms. The target source is located within a 15×15 cm area that is located approx. 1 m in front of the microphones. The area is covered by a regular grid of 16 positions with inter-grid distance of 5 cm, for which the RTFB is prepared using noise-free training recordings played from these exact positions. For the experiment, a testing recording is obtained when the target loudspeaker is randomly moved within the area.

The interference is represented by another loudspeaker which is moving between 0° through 180° around the microphones at the distance of 1.5 m.

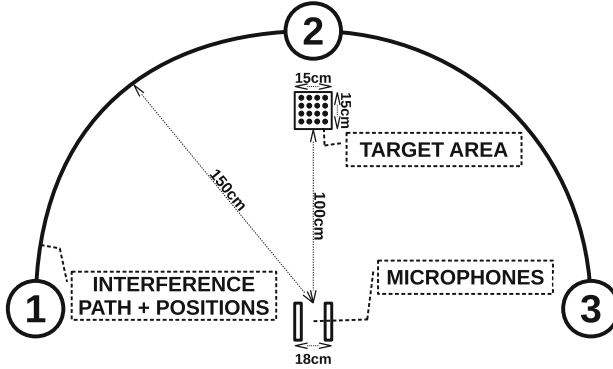


Fig. 1. The illustration of the experimental setup

The interference moves through major positions, denoted in Fig. 1, in the order 1, 2, 3, 2, 1, which is repeated several times. The interference source stopped at each major position for about 20 s.

The utterances played by the loudspeakers were taken from the TIMIT dataset [3]; the length of the entire testing recording is about 8 min; the sample rate is 16 kHz. Target and interference were recorded separately and mixed afterwards at the initial global Signal-to-Interference Ratio (SIR) of 0 dB.

The on-line S-IVA algorithm was used to separate the sources in the STFT domain with the frame length of 4096 samples and 75% frame overlap. The separated signals were reconstructed in the time domain by the overlap-add method. To improve the convergence of S-IVA, the scaled natural gradient modification of (5) described in [1] was used.

For evaluation, the improvement of SIR (iSIR) is computed on each frame and averaged over microphones. This gives us an improvement in SIR for all separated sources. Averages of these results are used as a measure of separation quality.

4.2 Results

The following notation is used for all figures. MPDR denotes the separation provided by the combination of the MPDR and Null beamforming. DOA setting indicates that the RTFs are pure delay filters. S-IVA followed by the specification of the hyper-parameters, β_n , denotes the proportion of piloting by the outputs of the beamformers (11). “S-IVA oracle” corresponds to the S-IVA piloted by original (oracle, separated) signals. IVA indicates the original unsupervised IVA algorithm.

Figure 2 shows the per-frame performance of the above-mentioned methods. It contains results for two different settings of the MPDR: RTFs are set to be delay filters (DOA, top result) and full RTFs having the length of 1000 taps in the time domain (bottom result). The most difficult periods for successful

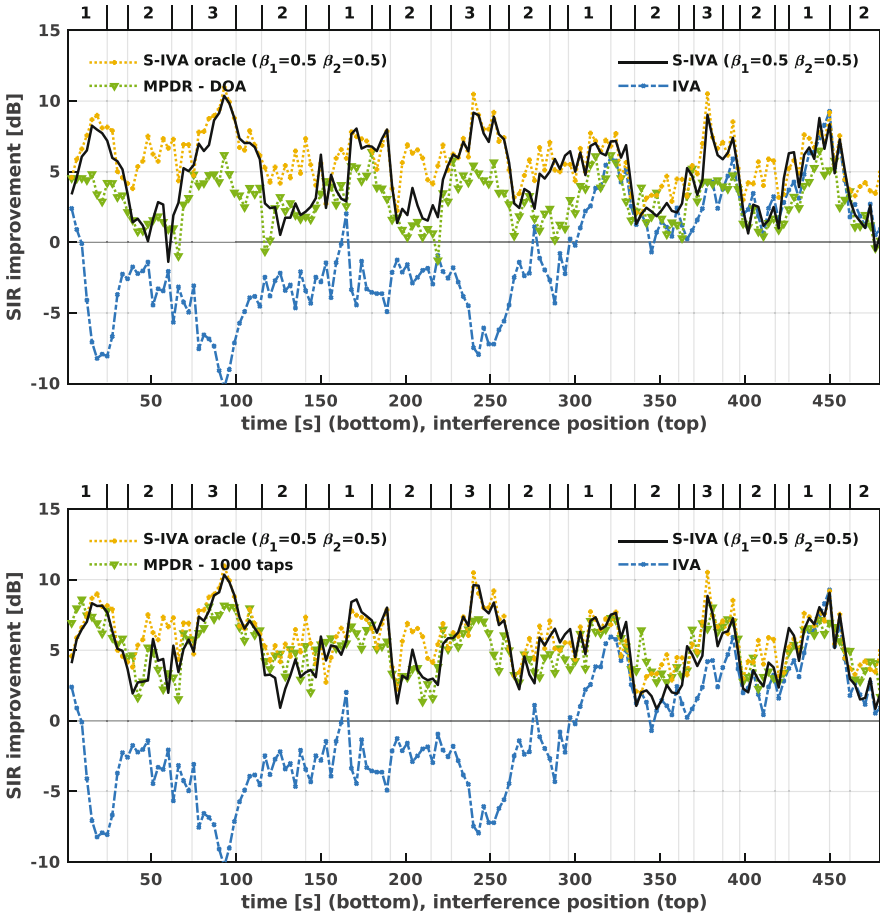


Fig. 2. Results in terms of iSIR as functions of time when the RTFB consists of pure delay filters (DOA) and 1000 taps long RTFs (in the time domain), respectively.

separation are when the interference source stays in position 2, that is, when the angular positions of both sources are the same. It can be observed that for those cases the iSIR of all methods drops down close to 0 dB. In these situations, the original IVA suffers from the global permutation problem, because the order of the separated outputs can be changed with high probability. In our experiment, the IVA performance suffers due to the global permutation (frames with negative iSIR). By contrast, the results show that with S-IVA the problem is solved, even with the DOA pilots. S-IVA piloted by clean signals in average achieves the best results and the result shows limits of the separation provides by S-IVA.

The performance of the beamforming methods is close to that of S-IVA only when the time-domain length of RTFs is 1000 taps or more. So the solution through S-IVA does not seem to bring many advantages compared to the

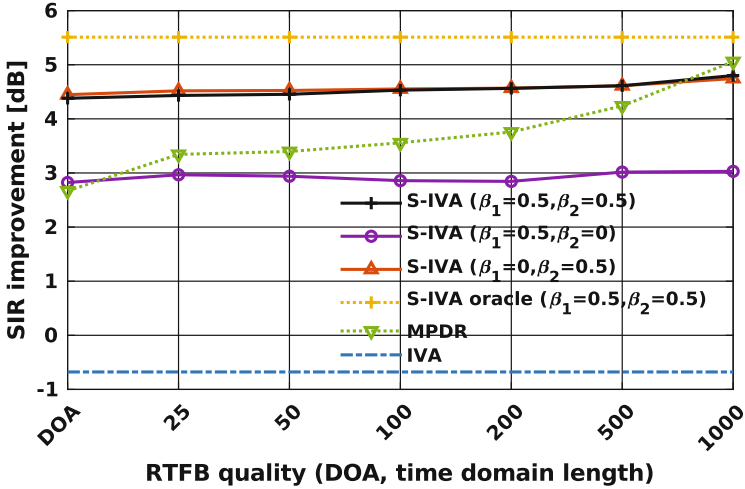


Fig. 3. Average separation performance (iSIR averaged over all frames) for various RTFB settings.

methods from [6, 7, 10] in this configuration. However, S-IVA provides better performance when the RTFs are less accurate.

In the second experiment, we examine different time domain lengths of the RTFs in order to simulate a deteriorated performance of the RTFB. Figure 3 compares average iSIR (over all frames) for all of the above-mentioned methods as functions of the various lengths (including the DOA setting).

The original IVA yields -1 dB of average iSIR due to the global permutation problem. The performance achieved through beamforming steadily grows with the time domain length of RTFs and outperforms S-IVA when the length exceeds 500 taps.

S-IVA is presented with three settings: First, S-IVA piloted by the output of the MPDR beamformer ($\beta_1 = 0.5, \beta_2 = 0$), which is dominated by the target signal. Second, S-IVA piloted only by the output of the null beamformer ($\beta_1 = 0, \beta_2 = 0.5$) that is dominated by the interference signal. Finally, S-IVA piloted by both pilot components ($\beta_1 = 0.5, \beta_2 = 0.5$). The results show that all variants solve the global permutation problem. Nevertheless, S-IVA piloted only by the MDPR beamformer is significantly worse than the other variants. This can be explained by the fact that the separation of the target from the interference is harder than the separation of interference from the target, because the target source is much closer to the microphones. Consequently, the piloting of the global permutation is more efficient when using the output of the null beamformer. Finally, we should mention the fact that the performance of S-IVA is not much influenced by the length of the RTFs.

5 Conclusion

In this work, we have proposed a novel variant of the Supervised IVA where the pilot component is obtained as the output of the MPDR or of the Null beamformer steered by a bank of pre-learned RTFs. We have shown by experiments that this variant of S-IVA is more practical than just using MPDR and Null beamforming taking the most appropriate RTF from the bank, because their performance is highly dependent on the quality of the RTFB. By contrast, we have shown that the performance of S-IVA piloted by outputs of the beamformers is robust against poor accuracy of RTFB, while the global permutation problem is efficiently solved.

In future works, we plan to generalize the proposed method for multiple microphones and sources. A straightforward way is to derive appropriate pilot components for all sources. Alternatively, a practical situation is when only some sources should be extracted from the mixture. Pilot components should be used to supervise the extraction of the sources as independent vector components. The goal is ensure that the blind method extracts the desired signal.

We also plan to compare our method with approaches that impose constraints on de-mixing filters to solve the global permutation problem, such as [4].

Acknowledgements. We are due to Dr. Francesco Nesta from Synaptics for his helpful comments and useful suggestions.

References

1. Douglas, S.C., Gupta, M.: Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2007, vol. 2, pp. II-637–II-640 (2007)
2. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Process.* **49**(8), 1614–1626 (2001)
3. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: Darpa timit acoustic phonetic continuous speech corpus cdrom (1993)
4. Khan, A.H., Taseska, M., Habets, E.A.P.: A geometrically constrained independent vector analysis algorithm for online source extraction. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) *LVA/ICA 2015*. LNCS, vol. 9237, pp. 396–403. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_46
5. Kim, T., Attias, H.T., Lee, S.Y., Lee, T.W.: Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **15**, 70–79 (2007)
6. Koldovský, Z., Málek, J., Tichavský, P., Nesta, F.: Semi-blind noise extraction using partially known position of the target source. *IEEE Trans. Audio Speech Lang. Process.* **21**(10), 2029–2041 (2013)
7. Koldovský, Z., Tichavský, P., Botka, D.: Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters. In: *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, pp. 679–683 (2013)

8. Lee, I., Kim, T., Lee, T.W.: Independent vector analysis for convolutive blind speech separation. In: Makino, S., Sawada, H., Lee, T.W. (eds.) *Blind Speech Separation*. Signals and Communication Technology, pp. 169–192. Springer, Dordrecht (2007). https://doi.org/10.1007/978-1-4020-6479-1_6
9. Liang, Y., Naqvi, S.M., Chambers, J.A.: Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment. *EURASIP J. Adv. Signal Process.* **2012**(1), 183 (2012)
10. Málek, J., Koldovský, Z., Gannot, S., Tichavský, P.: Informed generalized sidelobe canceler utilizing sparsity of speech signals. In: *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE (2013)
11. Matsuoka, K.: Minimal distortion principle for blind source separation. In: *Proceedings of the 41st SICE Annual Conference, SICE 2002*, vol. 4, pp. 2138–2143 (2002)
12. Nesta, F., Fakhry, M.: Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 86–90 (2013)
13. Nesta, F., Koldovský, Z.: Supervised independent vector analysis through pilot dependent components. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 536–540 (2017)
14. Nesta, F., Mosayyebpour, S., Koldovský, Z., Paleček, K.: Audio/video supervised independent vector analysis through multimodal pilot dependent components. In: *Proceedings of European Signal Processing Conference*, pp. 1190–1194 (2017)
15. Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 189–192 (2011)
16. Ono, T., Ono, N., Sagayama, S.: User-guided independent vector analysis with source activity tuning. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2417–2420 (2012)
17. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In: *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 505–510 (2003)
18. Smaragdīs, P.: Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* **22**, 21–34 (1998)



Does k Matter? k -NN Hubness Analysis for Kernel Additive Modelling Vocal Separation

Delia Fano Yela^(✉), Dan Stowell, and Mark Sandler

Queen Mary University of London, Mile End Road, London E1 4NS, UK
d.fanoyela@qmul.ac.uk

Abstract. Kernel Additive Modelling (KAM) is a framework for source separation aiming to explicitly model inherent properties of sound sources to help with their identification and separation. KAM separates a given source by applying robust statistics on the selection of time-frequency bins obtained through a source-specific kernel, typically the k -NN function. Even though the parameter k appears to be key for a successful separation, little discussion on its influence or optimisation can be found in the literature. Here we propose a novel method, based on graph theory statistics, to automatically optimise k in a vocal separation task. We introduce the k -NN hubness as an indicator to find a tailored k at a low computational cost. Subsequently, we evaluate our method in comparison to the common approach to choose k . We further discuss the influence and importance of this parameter with illuminating results.

Keywords: Source separation · Kernel additive modelling
Graph theory · Music processing · Vocal separation

1 Introduction

Source separation is a discipline aiming to isolate different sources from a given observable mixture. Amongst the methods for music source separation in a blind underdetermined scenario (less observable mixtures than sound sources), the major goal becomes to find inherent characteristics of the sources of interest to help with their identification and separation.

In the last decade, a number of computationally inexpensive methods explicitly modelling the target source's properties have gotten some attention [1–7]. These methods can be understood as instances of the wider kernel additive modelling (KAM) framework [8]. The basic idea behind KAM relies on the repetitive nature of music by estimating the target source at a particular point based on points at which the source's output is somehow similar. This is typically applied to time-frequency bins in a spectrogram representation. The function determining the target source similarity between time-frequency bins, while ignoring the

D. Fano Yela—This work was funded by EPSRC grant EP/L019981/1.

entries associated with other sources, is the so-called kernel function. Consequently, if a magnitude of a bin deviates amongst the ones judged to be similar by the target source kernel, one can assume there is another overlaying source and employ order statistics to attenuate its influence.

KAM has been successfully employed for a variety of tasks in source separation, such as vocal separation, speech enhancement, percussive/harmonic separation or interference reduction [4, 8, 9]. In the case of vocal separation, a popular approach is to assume the accompaniment music to be typically more repetitive and dense compared to the vocals, considered to be sparse and varied [2]. Meaning there are more segments in the mix containing the same or similar background music than there is for vocals. The nature of these segments vary amongst methods, such as a single repeating periodic musical pattern [2], the temporal context surrounding every time frame [5] or just a single time frame [1]. In all of these cases, the background music is implicitly assumed to have a higher energy contribution than that of the vocal source.

Amongst these methods, a popular choice for the accompaniment proximity kernel is the k nearest neighbours (k-NN) function, returning the k most similar frames to a given frame. The proximity measure between frames is typically based on the Euclidean distance, and therefore, two frames will be considered to be similar if they share the same centre frequency. Within the k-NN frames selection, if the vocal is indeed sparse it should appear as an outlier and can therefore be separated from the more common source through median filtering across similar bins. Since the breakdown point of the median operator is of 50% of outliers (vocals), one could expect the choice of k to be key for a successful separation. However, there is little or no guidance on how to set this parameter in the literature, nor explanation of its overall influence.

Here we investigate the influence of the parameter k in a vocal separation task and we further propose a novel method for its automatic optimisation, based on consideration of the proximity graph, which is lightweight and needs no prior training. In Sect. 2 will introduce the KAM vocal separation baseline and discuss typical methods to choose the parameter k in the K-NN proximity kernel. We will then propose a novel computationally inexpensive method for k optimisation in Sect. 3 based on graph theory statistics. In Sect. 4 we will further analyse and discuss the impact of this parameter through an experimental evaluation and validate the proposed method in such scenario.

2 Vocal Separation Using k Nearest Neighbours

KAM is a framework capable of combining different approaches to source separation using different assumptions to model sound sources. From the different proximity kernel families described in [8], we will focus on the models for repetitive patterns in a vocal separation task. In particular, we present a subset that can be regarded as an instance of KAM using only one iteration of the kernel backfitting procedure described in [8], which was also used in similar form in the REPET family of methods [2], and later extended to account for different repetitive patterns [1, 7].

These methods take advantage of the repetitive nature of music and define a distinction between a repeating background and a sparse varied foreground. For vocal separation in popular music the background typically corresponds to the music accompaniment and the vocals can be regarded as the sparse foreground. Therefore, one can assume that the musical accompaniment contributes to most of the energy across the frequency spectrum. We follow the method and notation described in [1] serving as the baseline method on which we will investigate the influence and optimisation of its single inherent parameter k .

Formally, we define the magnitude spectrogram of a musical signal as $X \in \mathbb{R}^{M \times N}$, where M is the number of frequency bins and N the number of time frames. For each pair of frames $(j, \ell) \in \{1, \dots, N\} \times \{1, \dots, N\}$, we then compute the squared Euclidean distance between the two corresponding columns in X :

$$D_{j, \ell} = \sum_{m=1}^M (X_{m, j} - X_{m, \ell})^2.$$

The result is a symmetric matrix D , which we can now sort to find the k nearest neighbours to every frame by keeping track of the frame index. Then, for every frame j , we create a matrix $A^j \in \mathbb{R}^{M \times K}$ containing as columns the specific subset of the k most similar frames taken from X . We expect the selected k closest frames to j to share similar musical accompaniment and differ in terms of the vocal part. In other words, the vocal contribution in the k nearest frames to j can be regarded as an outlier and the musical accompaniment as the commonality between them. Consequently, the median filter is the operator of choice in [1] to extract the common background music and separate out the vocal contribution on each frame. The estimated magnitude spectrogram $Y \in \mathbb{R}^{M \times N}$ of the musical accompaniment is:

$$Y_{m, j} := \text{median}(A_{m, 1}^j, \dots, A_{m, K}^j)$$

To extract both magnitude and vocals from the mixture, we use the soft mask $W \in [0, 1]^{M \times N}$ described in [1]. The complex spectrograms for the accompaniment and vocals can then be estimated by applying soft masks W and $(1 - W)$ respectively to the original mixture spectrogram using an element-wise multiplication.

A successful separation between background music and vocals relies largely on the vocals actually being outliers within the selection of the k closest frames. We want to make sure that the k -NN frames have similar background music with no or different vocals. However, there are also frames containing matching background music *and* matching vocals, which will then be very likely to be selected as near neighbours. Those frames are unhelpful for the median filtering but since the breakdown point of the median operator is of 50% of outliers (vocals), the method is robust to the vocal repetitions up to a point. This robustness is closely related to the number of nearest neighbours we choose, i.e. the parameter k .

There seems to be little or no indication on the method to find the optimal parameter k in the literature [1, 5, 7, 8]. In [7] the authors introduce three other parameters to set boundaries for the choice of k . However, no indication was

found on how to actually fix any of those parameters, including k . A recent extension introducing a temporal context R in the proximity kernel [5] performs a parameter sweep to set the new R parameter to the value giving the best mean metric across a dataset.

To our knowledge, there are currently two broad approaches to setting k : perceptual assessment or evaluation metric optimisation. In the first approach one simply listens to the estimates for different k values and adjusts the parameter to the best sounding setting. This is the preferred method to set k when there is a reduced number of songs to be processed. The second approach relies on a metric, typically the Signal to Distortion Ratio (SDR), comparing the estimated sound sources with the ground truth. One will set k to obtain the best metric result. In practice, this means a parameter sweep for different k values, for which no indication was found on how to pick. In addition, the commonly used SDR measure is known to be a proxy for perceptual quality and its precision has been criticised [10]. However, when dealing with large datasets, perceptual assessment of the results can be very time consuming. Therefore, it is more typical to use the second approach to optimise for an overall best performance.

A parameter sweeping approach to find the optimal k value has a number of disadvantages, primarily linked to the optimisation through a performance metric. Firstly, the separation performance metrics usually require to have ground truth separate tracks available, which is not always possible in an application scenario. Further, the commonly used separation performance metrics are computationally expensive [11], limiting the parameter sweep to a reduced number of values in a time constraint situation. In addition, optimising k using an overall performance metric does not assure the best value for all songs in the dataset. Moreover, fixing the k sweep values leaves no room to inform the optimisation with the track's individual properties, such as length.

Ideally we would like to be able to automatically pick k in an unsupervised way for each track separately, taking into account the nature of the song and thus finding a tailored value for k assuring a successful separation. We would also like to do this without having to perform multiple runs of source separation and discarding all but one of them.

3 Properties of the k -NN Graph

For a given music recording, the family of KAM methods we consider depends fundamentally for its behaviour on the set of nearest neighbours selected for each of the N frames. These nearest neighbour relationships can be represented as a directed graph with frames as nodes, and each node having k arcs leading outward to its nearest neighbours. Note that if frame i is a neighbour of frame j , the reverse is not necessarily true. At extreme settings, if $k = 0$ then the graph has no arcs and thus no structure, while if $k = N$ the graph is fully connected and likewise exhibits no structure. What are desirable characteristics for a k -NN graph to be used in KAM?

Unlike many problems defined on a graph, in KAM we do not wish our graph to take on simple structure such as well-separated clusters: instead, we want all

frames to have connections to frames which are similar according to the current source kernel, but dissimilar in terms of the other sources. It is not clear how these structural considerations can best be quantified numerically, though such structure would have some impact on summary statistics considered in graph theory.

Consider a set of frames containing a background musical phrase which is repeated often: we would expect these to form a densely connected component in the graph. The frames also containing sparsely-present and variable vocal energy would be expected to have arcs pointing to that densely connected component but few arcs pointing back out to them. Therefore, the number of incoming arcs (i.e. in-degree) would be unevenly distributed across the nodes, directly as a result of the observed signal properties which one assumes in KAM.

One way to analyse such properties in graph theory is the concept of ‘hubs’, which are nodes with an unusually high in-degree [12]. This has been of particular influence in social network theory as researchers studied effects such as ‘small world’ phenomena, which can have important effects such as the speed at which news or illness spreads through a social network. For a given graph, one can define summary statistics which reflect the general presence of hubs. One referred to as the ‘hubness’ is simply the skewness of the k -occurrence statistics, i.e. the skewness of the distribution of the in-degrees of nodes in the graph. Here, the k -occurrence of a frame corresponds to the number of times that frame is amongst the k nearest neighbours, and the ‘hubness’ is therefore the skewness of the distribution of all frames’ k -occurrence. In a k -NN graph we assign a fixed number of arcs, and so the average in-degree is always k ; however if the graph contains strong hubs then the skewness of the in-degree will be high.

In our vocal separation application in KAM it is clear that a graph with relatively *high* hubness should typically be one which has appropriate structure. We typically have very little *a priori* guidance over what value of k to choose, so it is advantageous that, for each track separately, we can iterate over a selection of possible k , inspect graph statistics such as hubness for the graphs thus produced, and select k which produces the optimal statistics. Therefore, we here propose to select the k producing the maximum hubness of the associated k -NN graph.

However, in a situation where we vary k , the hubness h will vary even in the null case of a randomly-constructed graph. (This can be seen in the extreme cases: for $k = 0$ or $k = N$ the graph is symmetric and the hubness is 0, whereas for other k it can be nonzero.) A standard null model can be generated by selecting k neighbours for each frame purely at random. This is related to the classic Erdős-Rényi random graph except that it is directed rather than undirected [13]. The distribution of k -occurrences in this null model follows a binomial distribution with parameters N and k/N , leading to an expression for the expected hubness as:

$$h_{\text{null}} = (1 - 2k/N) / \sqrt{k(1 - k/N)} \quad (1)$$

We can thus define a normalised hubness statistic as the ‘excess’ hubness, i.e. the raw observed hubness minus the hubness expected under the null model, which should then be less biased than the raw hubness in selecting k .

The above null model is one of the simplest random graphs. In practice, graphs constructed from high-dimensional similarity measures do not behave strictly in that fashion, and it is an ongoing research topic to model how k -NN graphs behave in general [12]. In preliminary work we found that the general scaling of the hubness statistic was out of line (larger) than in the simple null model, and so our empirical normalisation is given as

$$h_{\text{norm}} = \frac{h}{\max(h)} - \frac{h_{\text{null}}}{\max(h_{\text{null}})} \quad (2)$$

where maxima are across the sweep of k settings.

Using the maximum hubness as a metric to choose k has numerous advantages:

1. It does not require any ground truth information
2. k is optimised per track as a pre-processing step before the separation actually takes place
3. It is quick to compute so we can sweep through a lot of different k values, so we can have a finer optimisation
4. The hubness has been demonstrated to have perceptual relevance for song similarity in music recommendation, suggesting that it reflects properties of the nearest neighbour graph that have impact on its applied use. However, it has not been used for frame selection in KAM and so that is to be explored here.

4 Experiments and Discussion

To evaluate the proposed method, we quantitatively compare it against the standard parameter sweep for setting k in KAM for a vocal separation task. We chose to follow the vocal separation method described in [1] with FFT size of 4096 and hop size of 1024 samples, as it represents a baseline instance of the larger KAM framework.

To encourage reproducibility, we use the publicly available Test Demixing Secrets Dataset (DSD100) [14], containing 50 full length songs of diverse genres sampled at 44.1 kHz. Since the kernel implemented relies on musical repetition, we evaluated our proposed method on full length songs to ensure as much sound material as possible for KAM’s source reconstruction. However, the literature only offers some indication on k values for 30 second segments. We therefore use a broad range of fix k values for the traditional parameter sweep, letting $k \in \{0, 25, 50, 100, 200, 400, 800, 1600, 3200\}$, and a finer percent increase sweep for the computational inexpensive proposed method taking the song length into account, letting $k \in \{(0.001, 0.011, 0.021, 0.031, \dots, 0.45) \times N\}$ where N is the total number of time frames in the song.

Following common practice in the field, we employ the Signal to Distortion Ratio (SDR) in the BSS Eval toolbox 3.0 [11] as the quantitative indicator of the separation performance. Therefore, we would expect to observe a positive correlation between SDR and hubness for different k values. Due to the diversity

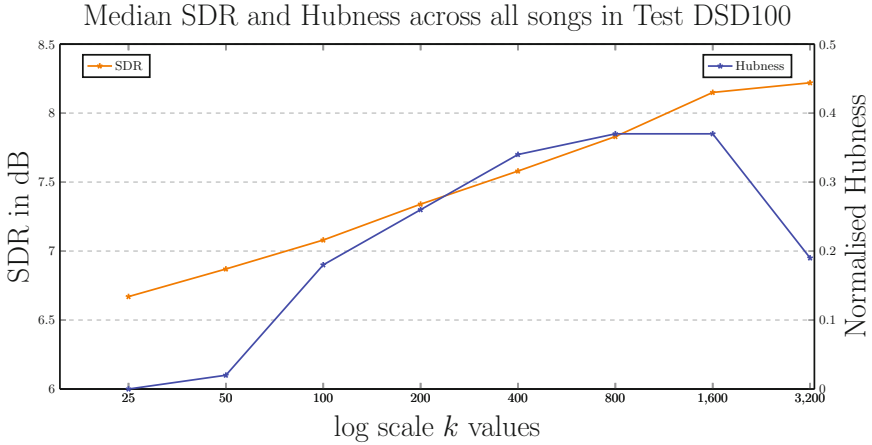


Fig. 1. Median SDR and hubness across all songs in Test DSD100 for different fixed k values

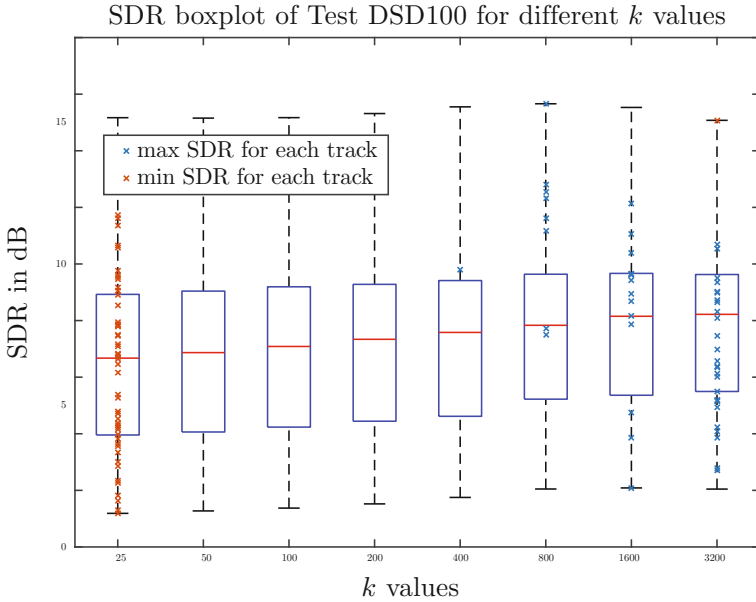


Fig. 2. SDR Boxplot of every song in the Test DSD100 dataset for different k values. The maximum and minimum SDR obtained for each song are marked in blue and orange respectively, showing a general trend of higher separation performance with increasing k value. (Color figure online)

of styles in the dataset, one could also expect an improvement in the overall separation performance (and so SDR) by using a tailored k for each song following the proposed method.

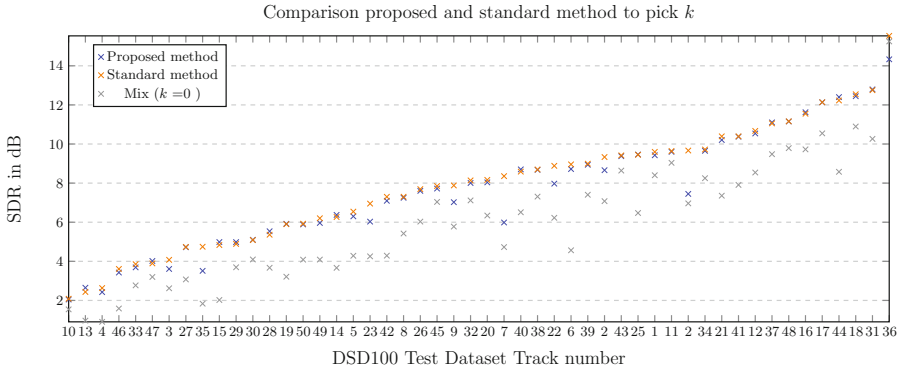


Fig. 4. SDR values for each song in the Test DSD100 dataset sorted in ascending order, using the optimal k issued from the standard and proposed method, in comparison to the SDR of the raw mixture (i.e. $k=0$).

Figure 3 offers a different perspective on the individual song behaviour which should shed some light on the above dilemma. As expected, very repetitive songs such as track 45, 4 or 50, achieve a higher SDR with highest k values. However, it is also the case for unconventional pop songs such as 43 or 17, where the variance in SDR is extremely low (less than 0.05). For such cases the separation may not have been successful, but Fig. 4 shows otherwise as the median SDR is above the mixture’s SDR (equivalent $k = 0$). Further, the overall SDR variance is surprisingly low, with a median of 1.4 dB potential SDR increase by changing k (maximum of 3.57 dB and minimum of 0.17 dB). With such a low potential SDR improvement, one might wonder if k actually matters at all or again, if the SDR is failing to capture the actual separation performance.

The majority of cases where different values of k induce substantial changes in SDR correspond to popular songs with a classic pop musical set-up and repeating musical structures (Fig. 3)—the ideal scenario for the implemented KAM vocal separation as described in [1]. One could therefore infer that a track sensitive to different k values (i.e. higher SDR variance), fulfills KAM requirements for a successful source separation. Track 44 presents an excellent example as it has a high SDR median and high SDR variance (2.72 dB of potential SDR improvement). However, most of the tracks in the dataset fail to present such characteristics, introducing a question regarding the flexibility and adaptability of the implemented KAM for vocal separation.

Songs which fulfill KAM ideal requirements for vocal separation (sensitive to k or highly repetitive) are expected to present higher SDR values than more complex songs. However, Fig. 3 does not present such logic, which makes one further wonder if the choice of separation performance metric is the adequate choice and so perceptual models or listening tests should be adopted for separation methods evaluation.

Nevertheless, Fig. 4 shows the proposed method can be used as substitute to the current technique for fixing k . Both methods present similar results in most cases and although the proposed one presents lower SDR for some songs,

it seems a small trade-off for a considerable decrease in computation time (1000 times faster than the standard method).

References

1. FitzGerald, D.: Vocal separation using nearest neighbours and median filtering. In: Proceedings of the Irish Signals and Systems Conference (ISSC), pp. 1–5 (2012)
2. Rafii, Z., Pardo, B.: Repeating pattern extraction technique (REPET): a simple method for music/voice separation. *IEEE Trans. Audio Speech Lang. Process.* **21**(1), 71–82 (2013)
3. FitzGerald, D.: Harmonic/percussive separation using median filtering. In: Proceedings of the International Conference on Digital Audio Effects (DAFx), Graz, Austria, pp. 246–253 (2010)
4. Fano Yela, D., Ewert, S., FitzGerald, D., Sandler, M.B.: Interference reduction in music recordings combining kernel additive modelling and non-negative matrix factorization. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, USA (2017)
5. Fano Yela, D., Ewert, S., Fitzgerald, D., Sandler, M.: On the importance of temporal context in proximity kernels: A vocal separation case study. In: Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio, June 2017
6. Fano Yela, D., Ewert, S., O’Hanlon, K., Sandler, M.B.: Shift-invariant kernel additive modelling for audio source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, Canada (2018)
7. Rafii, Z., Pardo, B.: Music/voice separation using the similarity matrix. In: ISMIR, pp. 583–588 (2012)
8. Liutkus, A., FitzGerald, D., Rafii, Z., Pardo, B., Daudet, L.: Kernel additive models for source separation. *IEEE Trans. Sig. Process.* **62**(16), 4298–4310 (2014)
9. Rafii, Z., Pardo, B.: Online repet-sim for real-time speech enhancement. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 848–852. IEEE (2013)
10. Cano, E., FitzGerald, D., Brandenburg, K.: Evaluation of quality of sound source separation algorithms: human perception vs quantitative metrics. In: Proceedings of the European Signal Processing Conference (EUSIPCO) (2016)
11. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
12. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: the emergence and influence of hubs. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 865–872. ACM (2009)
13. Erdős, P., Rényi, A.: On random graphs, i. *Pub. Math. (Debrecen)* **6**, 290–297 (1959)
14. Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontcave, J.: The 2016 signal separation evaluation campaign. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) *LVA/ICA 2017*. LNCS, vol. 10169, pp. 323–332. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_31

**Signal Separation Evaluation Campaign
(SiSEC 2018)**



The 2018 Signal Separation Evaluation Campaign

Fabian-Robert Stöter¹, Antoine Liutkus^{1(✉)}, and Nobutaka Ito²

¹ Inria and LIRMM, University of Montpellier, Montpellier, France
antoine.liutkus@inria.fr

² NTT Communication Science Laboratories, NTT Corporation, Tokyo, Japan

Abstract. This paper reports the organization and results for the 2018 community-based Signal Separation Evaluation Campaign (SiSEC 2018). This year's edition was focused on audio and pursued the effort towards scaling up and making it easier to prototype audio separation software in an era of machine-learning based systems. For this purpose, we prepared a new music separation database: MUSDB18, featuring close to 10 h of audio. Additionally, open-source software was released to automatically load, process and report performance on MUSDB18. Furthermore, a new official Python version for the *BSS Eval* toolbox was released, along with reference implementations for three oracle separation methods: ideal binary mask, ideal ratio mask, and multichannel Wiener filter. We finally report the results obtained by the participants.

1 Introduction

Source separation is a signal processing problem that consists in recovering individual superimposed *sources* from a *mixture*. Since 2008, the role of the Signal Separation Evaluation Campaign (SiSEC) has been to compare performance of separation systems on a voluntary and community-based basis, by defining tasks, datasets and metrics to evaluate methods [1, 14, 18, 19, 29, 30, 34]. Although source separation may find applications in several domains, the focus of SiSEC has always mostly been on audio source separation.

This year, we decided to drop the legacy speech separation and denoising tasks UND and BGN, because they are now the core focus of very large and successful other campaigns such as CHiME [2, 3, 31]. Instead, most of our efforts were spent on music separation, where the SiSEC MUS task is playing an important role, both in terms of datasets and participation. However, we also maintained the ASY task of asynchronous separation, due to its originality and adequation with the objectives of SiSEC.

While the primary objective of SiSEC is to regularly report on the progress made by the community through standardized evaluations, its secondary objective is also to provide useful resources for research in source separation, even outside the scope of the campaign itself. This explains why the SiSEC data has always been made public, to be used for related publications.

Since 2015, the scope of the SiSEC MUS data was significantly widened, so that it could serve not only for evaluation, but also for the design of music separation system. This important shift is motivated by the recent development of systems based on deep learning, which now define the state-of-the-art and require important amounts of learning data. This led to the proposal of the MSD [19] and the DSD100 [14] datasets in the previous editions.

This year’s SiSEC presents several contributions. First, the computation of oracle performance goes further than the usual Ideal Binary Mask (IBM) to also include Ideal Ratio Mask (IRM) and Multichannel Wiener Filters (MWF). Second, we released the MUSDB18, that comprises almost 10 h of music with separated stems. Third, we released a new version 4 for the BSS Eval toolbox, that handles time-invariant distortion filters, significantly speeding up computations¹.

2 Oracle Performance for Audio Separation

We write I as the number of channels of the audio mixture: $I = 2$ for stereo. We write x for the 3-dimensional complex array obtained by stacking the Short-Time Frequency Transforms (STFT) of all channels. Its dimensions are $F \times T \times I$, where F, T stand for the number of frequency bands and time frames, respectively. Its values at Time-Frequency (TF) bin (f, t) are written $x(f, t) \in \mathbb{C}^I$, with entries $x_i(f, t)$. The mixture is the sum of the sources *images*: $x(f, t) = \sum_j y_j(f, t)$, which are also multichannel.

A filtering method \mathbf{m} usually computes estimates $\hat{y}_j^{\mathbf{m}}$ for the source images linearly from x :

$$\hat{y}_j^{\mathbf{m}}(f, t | \theta_{\mathbf{m}}) = M_j^{\mathbf{m}}(f, t | \theta_{\mathbf{m}}) x(f, t), \quad (1)$$

where $\theta_{\mathbf{m}}$ are some parameters specific to \mathbf{m} and $M_j(f, t | \theta_{\mathbf{m}})$ is a $I \times I$ complex matrix called a TF *mask*, computed using $\theta_{\mathbf{m}}$ in a way specific to method \mathbf{m} . Once given the filtering strategy \mathbf{m} , the objective of a source separation system is to analyze the mixture to obtain parameters $\theta_{\mathbf{m}}$ that yield good separation performance.

For evaluation purposes, it is useful to know how good a filtering strategy can be, i.e. to have some upper bound on its performance, which is what an *oracle* is [33]:

$$\theta_{\mathbf{m}}^* = \operatorname{argmin}_{\theta_{\mathbf{m}}} \sum_{f, t, j} \|y_j(f, t) - \hat{y}_j^{\mathbf{m}}(f, t | \theta_{\mathbf{m}})\|, \quad (2)$$

where $\|\cdot\|$ is any norm deemed appropriate. In this SiSEC, we covered the three most commonly used filtering strategies, and assessed performance of their respective oracles:

¹ siseq.inria.fr.

1. The **Ideal Binary Mask** (*IBM*, [35]) is arguably the simplest filtering method. It processes all (f, t, i) of the mixture independently and simply assigns each of them to one source only: $M_{ij}^{\text{IBM}}(f, t) \in \{0, 1\}$. The IBM1 method is defined as $M_{ij} = 1$ iff source j has a magnitude $|y_{ij}(f, t)|$ that is at least half the sum of all sources magnitudes. IBM2 is defined similarly with the sources power spectrograms $|y_{ij}(f, t)|^2$.
2. The **Ideal Ratio Mask** (*IRM*), also called the α -Wiener filter [12], relaxes the binary nature of the IBM. It processes all (f, t, i) through multiplication by $M_{ij}^{\text{IRM}} \in [0, 1]$ defined as:

$$M_{ij}^{\text{IRM}}(f, t) = \frac{v_{ij}(f, t)}{\sum_{j'} v_{ij'}(f, t)}, \quad (3)$$

where $v_{ij}(f, t) = |y_{ij}(f, t)|^\alpha$ is the fractional power spectrogram of the source image y_{ij} . Particular cases include the *IRM2* Wiener filter for $\alpha = 2$ and the *IRM1* magnitude ratio mask for $\alpha = 1$.

3. The **Multichannel Wiener Filter** (*MWF*, [6]) exploits multichannel information, while IBM and IRM do not. $M_j^{\text{MWF}}(f, t)$ is a $I \times I$ complex matrix given by:

$$M_j^{\text{MWF}}(f, t) = C_j(f, t) C_x^{-1}(f, t), \quad (4)$$

where $C_j(f, t)$ is the $I \times I$ covariance matrix for source j at TF bin (f, t) and $C_x = \sum_j C_j$. In the classical local Gaussian model [6], the further parameterization $C_j(f, t) = v_j(f, t) R_j(f)$ is picked, with R_j being the $I \times I$ *spatial covariance matrix*, encoding the average correlations between channels at frequency bin f , and $v_j(f, t) \geq 0$ encoding the power spectral density at (f, t) . The optimal values for these parameters are easily computed from the true sources y_j [13].

These five oracle systems IBM1, IBM2, IRM1, IRM2, MWF have been implemented in Python and released in an open-source license².

3 Data and Metrics

3.1 The MUSDB18 Dataset

For the organization of the present SiSEC, the MUSDB18 corpus was released [21], comprising tracks from MedleyDB [4], DSD100 [14, 19], and other material. It contains 150 full-length tracks, totaling approximately 10 h of audio.

- All items are full-length tracks, enabling the handling of long-term musical structures, and the evaluation of quality over silent regions for sources.
- All signals are stereo and mixed using professional digital audio workstations, thus representative of real application scenarios.

² github.com/sigsep/sigsep-mus-oracle.

- All signals are split into 4 predefined categories: bass, drums, vocals, and other. This promotes automation of the algorithms.
- Many musical genres are represented: jazz, electro, metal, etc.
- It is split into a training (100 tracks, 6.5 h) and a test set (50 tracks, 3.5 h), for the design of data-driven methods.

The dataset is freely available online, along with Python development tools³.

3.2 BSS Eval Version 4

The BSS Eval metrics, as implemented in the MATLAB toolboxes [7, 32] are widely used in the separation literature. They assess separation quality through 3 criteria: Source to Distortion, to Artefact, to Interference ratios (SDR, SAR, SIR) and additionally with the Image to Spatial distortion (ISR) for the BSS Eval v3 toolbox [32].

One particularity of BSS Eval is to compute the metrics after optimally matching the estimates to the true sources through linear *distortion filters*. This provides some robustness to linear mismatches. This matching is the reason for most of the computation cost of BSS Eval, especially considering it is done for each evaluation window.

In this SiSEC, we decided to drop the assumption that distortion filters could be varying over time, but considered instead they are fixed for the whole length of the track. First, this significantly reduces the computational cost because matching is done only once for the whole signal. Second, this introduces more dynamics in the evaluation, because time-varying matching filters over-estimate performance, as we show later. Third, this makes matching more stable, because sources are never silent throughout the whole recording, while they often were for short windows.

This new 4th version for the BSS Eval toolbox was implemented in Python⁴, and is fully compatible with earlier MATLAB-based versions up to a tolerance of 10⁻¹² dB in case time-varying filters are selected.

4 Separation Results

4.1 Oracle Performance with BSS Eval v4

To the best of our knowledge, the results presented in Fig. 2 are the first fair comparison between the different and widely used oracle systems presented in Sect. 2. On this figure, we can see boxplots of the BSS Eval scores obtained by IBM1, IBM2, IRM1, IRM2 and MWF on the 4 sources considered in MUSDB18. The scores were computed on 1 second windows, taken on the whole test-set.

The most striking fact we see on this Fig. 2 is that IBM is *not* achieving the best scores on any metric except ISR. Most particularly, we notice that IBM

³ <https://sigsep.github.io/musdb>.

⁴ `pip install museval`.

systematically induces a small loss in performance of a few dBs on SDR and SIR compared to soft masks for most sources, and to a significant loss for SAR, that can get as bad as around 5 dB for the accompaniment source. This is in line with the presence of strong *musical noise* produced by IBM whenever the source to separate is *dense* and cannot be assumed stronger in magnitude or energy than all others whenever it is active. This also happens for the bass, which is usually weaker than all other sources at high frequencies, yielding significant distortion with IBM. Furthermore, we suspect the strong scores obtained by IBM in vocals and bass ISR to mostly be due to the zeroing of large amounts of frequency bands in those estimates. Indeed, zero estimates lead the projection filters of BSS eval to totally cancel those frequencies in the reference also, artificially boosting ISR performance.

Now, comparing soft masks, it appears that IRM2 and MWF produce the best overall performance as compared to IRM1. However, this result is expected: BSS Eval scores are *in fine* relative to squared-error criteria, which are precisely optimised with those filters. Previous perceptual studies showed that IRM1 may be preferred in some cases [12]. This may be reflected in the slightly better performance that IRM1 obtains for SAR. Finally, although IRM2 seems slightly better than MWF for most metrics, we highlight that it also comes with twice as many parameters: power spectral densities for left and right channels, instead of just one for MWF, shared across channels.

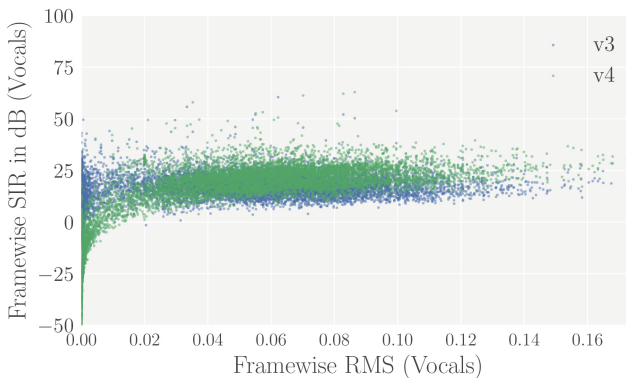


Fig. 1. Vocals SIR score vs vocals energy for BSS eval v3 and v4.

Concerning the discrepancies between BSS Eval v3 and v4 (time-invariant distortion filters), we observe several differences. First, computations were 8 times faster for v4 than for v3, which allowed using small 1 s frames and thus get an estimate of the performance along time at a reasonable computing cost. Second, computing distortion filters only once for the whole duration of the signal brings an interesting side-effect, that can be visualized on Fig. 1. The new v4 brings a much higher dynamics for the scores: we clearly see that lower

energy for the true source brings lower performance. However, the marginal distributions for the scores over the whole dataset were not statistically different between v3 and v4, which validates the use of fewer distortion filters to optimize computing time and get to similar conclusions.

4.2 Comparison of Systems Submitted to SiSEC-MUS 2018

This year’s participation has been the strongest ever observed for SiSEC, with 30 systems submitted in total. Due to space constraints, we cannot detail all the methods here, but refer the interested reader to the corresponding papers. We may distinguish three broad groups of methods, that are:

Model-based. These methods exploit prior knowledge about the spectrograms of the sources to separate and do not use the MUSDB18 training data for their design. They are: MELO as described in [24], as well as all the method implemented in NUSSL [15]: 2DFT [25], RPCA [9], REP1 [22], REP2 [20], HPSS [8].

No additional data. These methods are data-driven and exploit only the training data for MUSDB18 to learn the models. They are: RGT1-2 [23], STL, HEL1 [10], MDL1 [17], MDLT [16], JY1-3 [11], WK [36], UHL1 [27], UHL2 [28], TAK1 [26].

With additional data. These methods are also data-driven, and exploit additional training data on top of the MUSDB18 training set. They are: UHL3 [28], TAK2-3 [26], TAU [26,28].

As may be seen, the vast majority of methods submitted this year to SiSEC MUS are based on deep learning, reflecting a shift in the community’s methodology. The MIX method additionally serves as a negative anchor, that corresponds to using the mixture as an estimate for all sources.

In the first set of results depicted on Fig. 2, we display boxplots of the BSSeval scores for the evaluation. For each track, the median value of the score was taken and used for the boxplots. Inspecting these results, we immediately see that data-driven methods clearly outperform model-based approaches by a large margin. This fact is noticeable for most targets and metrics.

In the second set of results displayed on Fig. 3, we computed the track-wise median SDR score for all methods on the vocals (top) and accompaniment (bottom) targets. The striking fact we notice there is that methods exploiting additional training data (UHL3, TA*) do perform comparably to the oracles for approximately half of the tracks. After inspection, it turns out that room for improvement mostly lies in tracks featuring significant amounts of distortion in either the vocals or the accompaniment. We may also notice on these plots that tracks where accompaniment separation is easy often come with a challenging estimation of vocals. After inspection, this is the case when vocals are rarely active. Consequently, correctly detecting vocals presence seems a good asset for separation methods.

Our third round of analysis concerns the pair-wise post-hoc Conover-Inman test, displayed on Fig. 4, to assess which methods perform significantly better

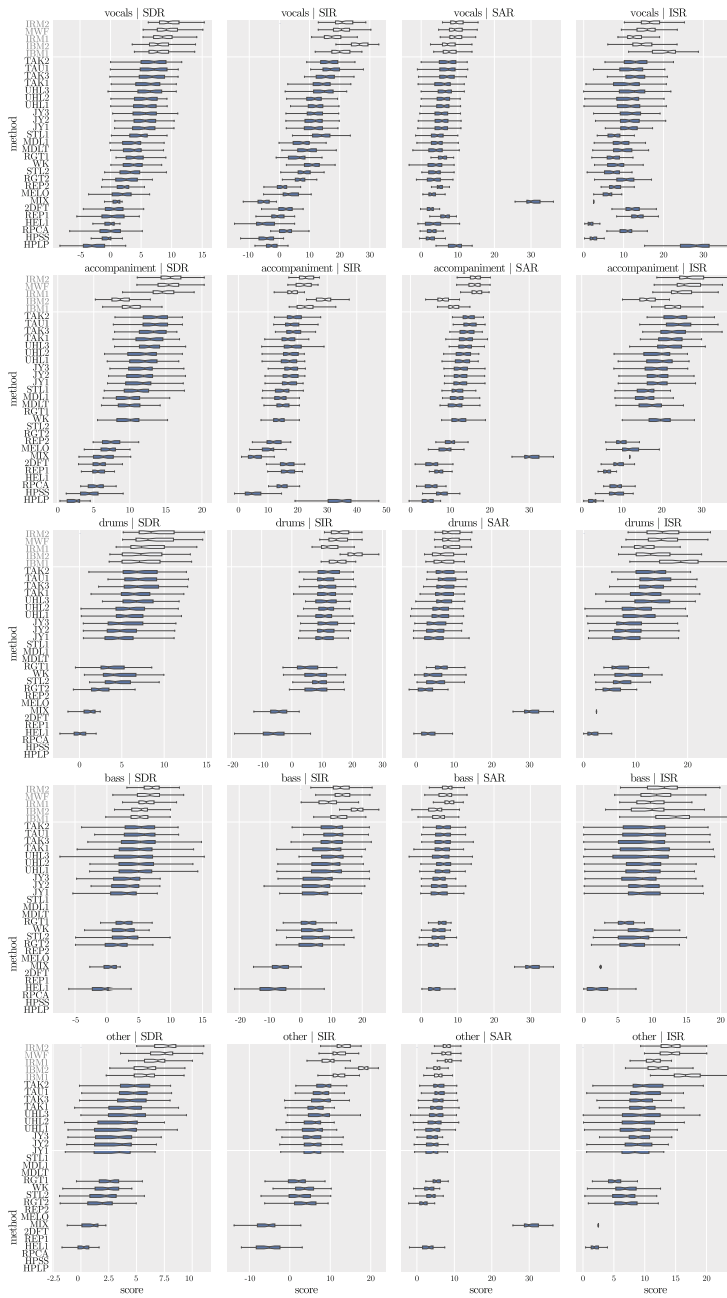


Fig. 2. Details of results for all metrics, targets and methods.

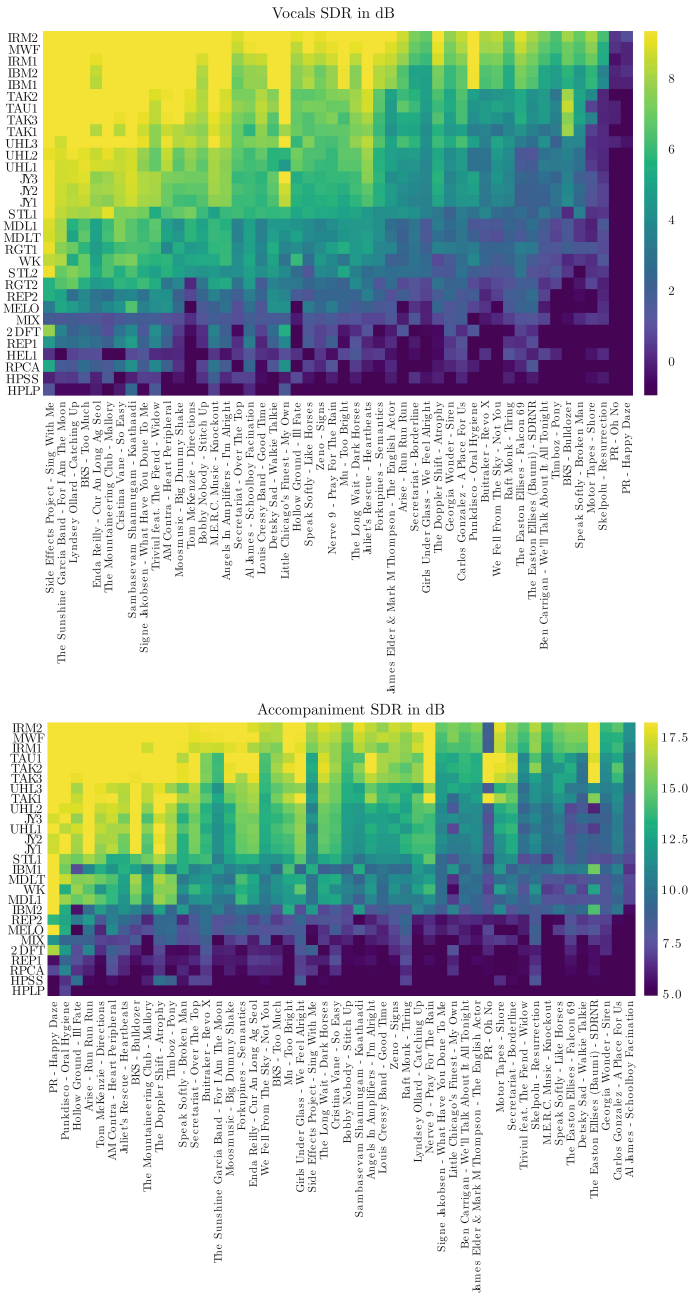


Fig. 3. Vocals (top) and accompaniment (below) SDR for all tracks and methods.

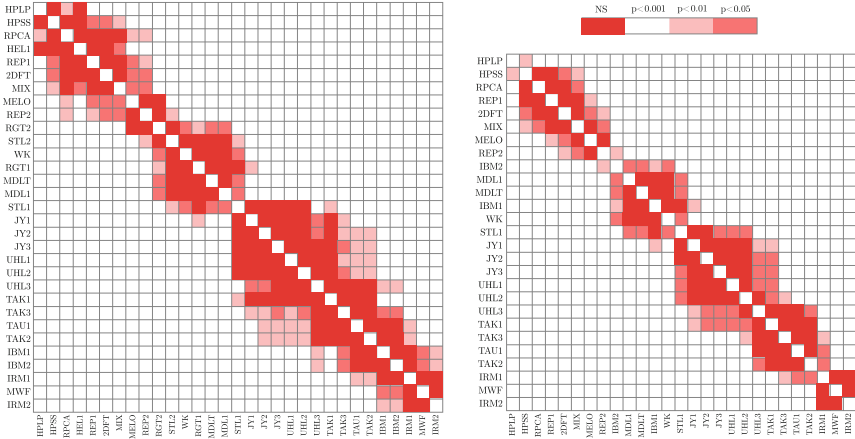


Fig. 4. Pair-wise statistical significance of the differences between separation quality. Left: vocals SDR. Right: accompaniment SDR.

than others, for both vocals and accompaniment separation. In this plot, an obvious fact is that DNN-based methods exploiting additional training data perform best. Remarkably, they do not perform significantly differently than the oracles for accompaniment, suggesting that the automatic karaoke problem can now be considered solved to a large extent, given sufficient amounts of training data. On the contrary, vocals separation shows room for improvement.

Concerning model-based methods, we notice they perform worse, but that among them, MELO stands above for vocal separation, while it is comparable to others for accompaniment. For DNN approaches not using additional training data, we notice different behaviours for vocals and accompaniment separation. We may summarize the results by mentioning that RGT1-2, STL and MDL1 do not behave as well as MDLT, STL1, JY1-3, WK and UHL1-2, which all behave comparably. It is noteworthy that TAK1 and UHL2 compare well with methods exploiting additional data for vocals separation.

This evaluation highlights a methodological question that should be investigated in future campaigns, which is the relative importance of the system architecture and the amount of training data. It indeed appears that very different architectures do behave comparably and that the gap in performance now rather comes from additional training data, as exemplified by the difference between UHL2 and UHL3. This confirms the importance of using standard training and test datasets such as MUSDB18 for evaluation, and we believe that obtaining good performance with reduced training data remains an interesting and challenging machine learning problem.

4.3 Comparison of Systems Submitted to SiSEC-ASY 2018

As shown in Table 1, there was one submission to the task “Asynchronous recordings of speech mixtures” by Corey *et al.* [5]. This method does not resample the microphone signals in order to separate them. Rather, it uses a separate time-varying two-channel Wiener filter for each synchronous pair of microphones. The remaining asynchronous microphone pairs are used to compute a speech presence probability for each source in each time-frequency bin. The speech presence information from the remote microphone pairs allows the reference recorder to separate more than two speech signals using a two-channel filter.

Table 1. Result for the task “Asynchronous recordings of speech mixtures”. Result by Miyabe *et al.* in SiSEC2015 is also shown as a reference.

Systems	Criteria	3src			4src		
		Realmix	Sumrefs	Mix	Realmix	Sumrefs	Mix
Corey [5]	SDR	-4.0	-4.0	-4.1	3.1	2.9	1.7
	ISR	-0.1	-0.1	-0.1	7.0	6.7	5.8
	SIR	-2.2	-1.7	-1.9	5.4	5.0	2.4
	SAR	-13.2	-13.1	-12.4	7.9	7.8	6.1
Miyabe	SDR	6.9	6.8	10.6	4.0	3.8	3.3
	ISR	11.2	11.1	15.1	8.8	8.5	7.3
	SIR	11.0	10.9	14.9	6.7	6.4	6.0
	SAR	11.7	11.6	15.5	7.8	7.6	7.4

5 Conclusion

We reported our work on the organization of SiSEC 2018, that comprised the development of a new Python version 4 for BSS Eval to assess performance, that is fully compatible with earlier MATLAB versions and additionally allows for time-invariant distortion filters, significantly reducing computational load. Furthermore, we presented the new MUSDB18 dataset, that gathers 150 music tracks with isolated stems, totaling almost 10 h of music. Finally, we also provide open-source implementations of 3 popular oracle methods to provide various upper bounds for performance.

Then, we reported the impact of choosing time-invariant distortion filters for BSS Eval over time-varying ones and quickly summarized the discrepancies in the performance of the proposed oracles methods with BSS Eval v3 and v4.

Finally, we provided an overall presentation of the scores obtained by the participants to this year’s edition. More detailed analysis and sound excerpts can be accessed online on the SiSEC webpage.

References

1. Araki, S., Nesta, F., Vincent, E., Koldovský, Z., Nolte, G., Ziehe, A., Benichoux, A.: The 2011 signal separation evaluation campaign (SiSEC2011): - audio source separation -. In: Theis, F., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) LVA/ICA 2012. LNCS, vol. 7191, pp. 414–422. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28551-6_51
2. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The third chimespeech separation and recognition challenge: dataset, task and baselines. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 504–511. IEEE (2015)
3. Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P.: The pascal chime speech separation and recognition challenge. *Comput. Speech Lang.* **27**(3), 621–633 (2013)
4. Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.P.: MedleyDB: a multitrack dataset for annotation-intensive mir research. In: 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, October 2014
5. Corey, R.M., Singer, A.C.: Underdetermined methods for multichannel audio enhancement with partial preservation of background sources. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 26–30 (2017)
6. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1830–1840 (2010)
7. Févotte, C., Gribonval, R., Vincent, E.: Bss_eval toolbox user guide-revision 2.0 (2005)
8. Fitzgerald, D.: Harmonic/percussive separation using median filtering (2010)
9. Huang, P.-S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M.: Singing-voice separation from monaural recordings using robust principal component analysis. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 57–60. IEEE (2012)
10. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Singing-voice separation from monaural recordings using deep recurrent neural networks. In: ISMIR, pp. 477–482 (2014)
11. Liu, J.-Y., Yang, Y.-H.: JY Music Source Separation submission for SiSEC, Research Center for IT Innovation, Academia Sinica, Taiwan (2018). <https://github.com/ciaua/MusicSourceSeparation>
12. Liutkus, A., Badeau, R.: Generalized Wiener filtering with fractional power spectrograms. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, QLD, Australia, April 2015
13. Liutkus, A., Badeau, R., Richard, G.: Low bitrate informed source separation of realistic mixtures. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 66–70. IEEE (2013)
14. Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontcave, J.: The 2016 signal separation evaluation campaign. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 323–332. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_31
15. Manilow, E., Seetharaman, P., Pishdadian, F., Pardo, B.: NUSSL: the north-western university source separation library (2018). <https://github.com/interactiveaudiolab/nussl>

16. Mimitakis, S.I., Drossos, K., Santos, J., Schuller, G., Virtanen, T., Bengio, Y.: Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask (2017)
17. Mimitakis, S.I., Drossos, K., Virtanen, T., Schuller, G.: A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation (2017)
18. Ono, N., Koldovský, Z., Miyabe, S., Ito, N.: The 2013 signal separation evaluation campaign. In: 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), September 2013
19. Ono, N., Rafii, Z., Kitamura, D., Ito, N., Liutkus, A.: The 2015 signal separation evaluation campaign. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 387–395. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_45
20. Rafii, Z., Liutkus, A., Pardo, B.: REPET for background/foreground separation in audio. In: Naik, G.R., Wang, W. (eds.) Blind Source Separation. SCT, pp. 395–411. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-55016-4_14
21. Rafii, Z., Liutkus, A., Stter, F.-R., Mimitakis, S.I., Bittner, R.: The MUSDB18 corpus for music separation, December 2017
22. Rafii, Z., Pardo, B.: Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE Trans. Audio Speech Lang. Process.* **21**(1), 73–84 (2013)
23. Roma, G., Green, O., Tremblay, P.-A.: Improving single-network single-channel separation of musical audio with convolutional layers. In: International Conference on Latent Variable Analysis and Signal Separation (2018)
24. Salamon, J., Gómez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Trans. Audio Speech Lang. Process.* **20**(6), 1759–1770 (2012)
25. Seetharaman, P., Pishdadian, F., Pardo, B.: Music/voice separation using the 2d fourier transform. In: 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 36–40. IEEE (2017)
26. Takahashi, N., Mitsufuji, Y.: Multi-scale multi-band densenets for audio source separation. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 21–25. IEEE (2017)
27. Uhlich, S., Giron, F., Mitsufuji, Y.: Deep neural network based instrument extraction from music. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2135–2139. IEEE (2015)
28. Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., Mitsufuji, Y.: Improving music source separation based on deep neural networks through data augmentation and network blending. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 261–265. IEEE (2017)
29. Vincent, E., Araki, S., Bofill, P.: The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 734–741. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00599-2_92
30. Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., Duong, N.Q.K.: The signal separation evaluation campaign (2007–2010): achievements and remaining challenges. *Signal Process.* **92**(8), 1928–1936 (2012)

31. Vincent, E., Barker, J., Watanabe, S., Roux, J.L., Nesta, F., Matassoni, M.: The second chimespeech separation and recognition challenge: datasets, tasks and baselines. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 126–130. IEEE (2013)
32. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
33. Vincent, E., Gribonval, R., Plumbley, M.D.: Oracle estimators for the benchmarking of source separation algorithms. *Signal Process.* **87**(8), 1933–1950 (2007)
34. 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). https://doi.org/10.1007/978-3-540-74494-8_69
35. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (ed.) *Speech Separation by Humans and Machines*, pp. 181–197. Springer, Boston (2005). https://doi.org/10.1007/0-387-22794-6_12
36. Weninger, F., Hershey, J.R., Roux, J.L., Schuller, B.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 577–581. IEEE (2014)



Improving Single-Network Single-Channel Separation of Musical Audio with Convolutional Layers

Gerard Roma^(✉), Owen Green, and Pierre Alexandre Tremblay

Centre for Research in New Music, University of Huddersfield, Huddersfield, UK
{g.roma,o.green,p.a.tremblay}@hud.ac.uk

Abstract. Most convolutional neural network architectures explored so far for musical audio separation follow an autoencoder structure, where the mixture is considered to be a corrupted version of the original source. On the other hand, many approaches based on deep neural networks make use of several networks with different objectives for estimating the sources. In this paper we propose a discriminative approach based on traditional convolutional neural network architectures for image classification and speech recognition. Our results show that this architecture performs similarly to current state of the art approaches for separating singing voice, and that the addition of convolutional layers allows improving separation results with respect to using only fully-connected layers.

Keywords: Audio source separation · Convolutional neural networks

1 Introduction

The concept of *musical audio* is commonly used to refer to polyphonic mixtures of musical instrument recordings and/or electronic sounds that have been produced and mastered for distribution. Separating such mixtures into source streams has many interesting applications, such as remixing and upmixing [1]. In a production context, being able to adjust an already assembled mix is useful in several situations. A mastering engineer may be able to make adjustments to vocal levels in the order of 1–2 dB without requesting a new mix or vocal stem from their client; or alter the perspectival position of a stream by selectively applying reverb or delay; or use an isolated stream as a key signal for a compressor. In remixing, more radical treatments might use extracted streams as sources for processing, although this is contingent on the degree and type of artefacts introduced by separation. In these contexts, it is important that the raw, extracted streams sum exactly to the original mixture and would pass what sound engineers commonly call a null sum test: inverting the phase of one mixture and adding this to the other should result in silence. In this way, the producer is always working from an uncompromised, original mix and—accepting

that extracted streams will have artefacts—can judge the outcome of adjustments relative to this neutral starting point. As mastering engineer Bob Olhsson notes, “audio processing is the art of balancing subjective enhancement against objective degradation” [2].

The success of deep learning architectures in other domains has sparked interest in applying them to separating musical audio. Here, we are interested in current machine learning systems for musical audio separation to the extent that they can provide *useful approximations* for audio processing applications. Most deep learning approaches rely on having training examples, which requires consistent instrumentation labels. For this reason, many models are limited to separating just singing voice or other specific streams, which is not so useful in a general production context. We should also note how much work these labels are made to do in terms of the territory they cover: for instance “vocals” could span Stevie Wonder, Björk, and T-Pain (without even considering more extreme examples). As such, aiming for a ‘perfect’ decomposition of sources is unrealistic, given the need for labelled data and the complexities of production processes, typically including non-linear effects. A compromise solution is provided by the Demixing Secrets Dataset (DSD100) used in the MUS task of the SiSEC evaluation campaign [3], where each track is consistently seen as a mix of vocals, bass and drums, while other instruments are grouped into a “other” category. On this basis, we take our separation to be yielding *streams* rather than sources that we can think of as being *vocal-like*, *bass-like* and so forth, and take as a priority that the sum of the streams matches exactly the original mixture. For this reason, we adopt the well established framework of time-frequency masking [4].

The SiSEC campaign is also a good measure of the state of the art. In the last iteration the best performance was obtained by systems using Deep Neural Networks (DNN) [5,6], either feed-forward or recurrent. Most recent DNN approaches have been based on two-step algorithms. For example, the system in [6] is formulated as a variant of the Expectation Maximization (EM) algorithm where one DNN tries to separate the sources, while a second one tries to enhance the result. A similar approach was presented in [7]. A recent system proposed in [8] also uses two networks (a “Masker” and a “Denoiser”).

The rapid adoption of Convolutional Neural Networks (CNN) in domains such as computer vision (including image segmentation) has fostered expectations with respect to audio source separation. However, applications of CNNs to musical audio separation have only surfaced recently. Most approaches follow an autoencoder structure, where the network tries to produce a de-noised version of the input. In this case, the mixture is seen as a signal where noise has been added to the target source. These networks follow a U-shaped structure where the input is a slice of the spectrogram, and the output is a slice with the same size. This means that after several downsampling layers (via convolution or max pooling) there is a series of upsampling (often called “deconvolution” [9]) layers that recover the original dimensions. For example, the system in [10] is composed of an encoding step and a decoding step (using deconvolution layers) connected by a fully-connected layer. The approach in [11,12] uses a similar system with

upsampling layers. These systems have both been evaluated with the DSD100 dataset. Another similar system was proposed in [13] specifically for singing-voice separation. Although it was not evaluated with the same dataset, it seems to provide improvements with the iKala [14] dataset used for singing voice extraction in the MIREX¹ evaluation challenge. However, it relies on a large private training dataset, based on artist distribution of instrumental tracks, so it is not clear whether it would extend to other instruments.

In this paper we investigate a different approach to CNNs for musical audio separation, based on the classic models for image classification [15]. As opposed to autoencoder-like approaches, our model can be seen as “discriminative”, in the sense that the problem is modelled as a classification of time-frequency bins. This implies adding some fully-connected layers after the convolutional layers. With this method, we hope to combine the discriminative power of fully-connected networks with the possibility of learning features from a wider temporal context provided by convolutional layers. As we are interested in the potential for real-time implementation, we limit our temporal scope to texture windows of around 200 ms. While this still requires a relatively long latency, it is much shorter than in recent CNN-based systems based on processing spectrograms of several seconds [13]. Each texture window is used to predict a filter for a given spectral frame.

The rest of the paper is organized as follows. In the next section we describe the proposed approach based on CNNs with two variants. In Sect. 3, we assess the potential of this model in experiments with the DSD100 dataset. Finally, in Sect. 4 we reflect on future possibilities for this work.

2 Proposed Approach

2.1 Problem Formulation

Most recent work on audio source separation is based time-frequency representations, typically the Short-Time Fourier Transform (STFT), and relies on the assumption that the transform X of a mixture signal x at time index n and frequency index k results from the sum of i component streams [16]:

$$X_{n,k} = \sum_i S_{n,k}^i \quad (1)$$

As seen in the previous section, in musical audio this may not really need to correspond to the original acoustic sources, but it is assumed that such decomposition would result in useful component signals (also, this assumes that a constant overlap-add window is used for an STFT). Audio source separation attempts to recover an estimate of each stream S_i , typically with the hope that the original sources do not overlap in X . Hence, the most common way of extracting the stream is by applying a time-frequency mask to X , so that

$$\hat{S}_{n,k}^i = M_{n,k}^i X_{n,k}. \quad (2)$$

¹ http://www.music-ir.org/mirex/wiki/MIREX_HOME.

The mask M^i can be either binary or soft. Ideal binary masks, specified to be 1 when the target source dominates a given time-frequency bin and 0 otherwise, are routinely used in the STFT domain as an upper bound of automatic music separation, which shows that, even for such broad band signals, the components do not overlap too much in this representation. However, in general for better-sounding estimates, a soft mask where $M_{n,k}^i$ ranges between 0 and 1 is preferred.

2.2 Mask Estimation

CNNs have become the standard method for image classification and object recognition in images. Conventional CNNs include both convolutional and fully-connected layers. In this setting they are typically used to obtain progressively smaller feature maps, which works for the mentioned tasks, where the output is just a class label, or a label and a set of coordinates. Here, this combination of convolutional and fully-connected layers is useful, given the importance of the temporal context for estimating M_i at a given point. Using DNNs with only fully-connected layers is problematic for this because in order to use multiple frames as input, large numbers of parameters are required. We define the n th input of the network as the sequence of $2c + 1$ magnitude frames:

$$\hat{X}_n = [|x_{n-c}| \dots |x_n| \dots |x_{n+c}|], \quad (3)$$

With respect to the objective, one option is to see the separation problem as a classification of time-frequency bins [4]. In this case, the spectrogram can be encoded as:

$$Y_{n,k} = \arg \max_i (S_{n,k}^i). \quad (4)$$

Here, Y is an integer matrix that contains the index of the source with the largest magnitude at each time-frequency bin. We would seek to estimate \hat{Y} so that the mask $M_{n,k}^i$ for each time-frequency bin can be obtained as

$$M_{n,k}^i = \begin{cases} 1 & \text{if } \hat{Y}_{n,k} = i \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

This setting allows us to use the popular softmax function, which we can compute for each time-frequency bin. Our model predicts a single frame, which means that the sequence of outputs produced by the last fully-connected layer of the network, O , is a $N \times K \times I$ tensor which can be seen as the non-normalized probability of source i at frequency k for frame n . The softmax function then produces normalized probabilities:

$$P_{n,k}^i = \frac{e^{O_{n,k,i}}}{\sum_j e^{O_{n,k,j}}} \quad (6)$$

The goal of the network is then to minimize the negative log likelihood for the correct class, averaged across frequency bins:

$$l_{nll} = \frac{1}{K} \sum_{\substack{k, \\ i=Y_{n,k}}} -\log(P^i_{n,k}). \quad (7)$$

Such setting can be used to obtain binary masks that are guaranteed to split the spectrogram evenly, so the estimates would pass the null sum test. However, binary masks typically introduce audible artifacts. Alternatively, an ideal soft mask is computed as

$$M^i_{n,k} = \frac{|S^i_{n,k}|}{\sum_i |S^i_{n,k}|}. \quad (8)$$

In this case, the estimate can be obtained by using a sigmoid function at the output of the network, which is also a $N \times K \times I$ tensor:

$$P^i_{n,k} = \frac{e^{O_{n,k,i}}}{1 + e^{O_{n,k,i}}}. \quad (9)$$

Then the mean square error loss can be used:

$$l_{mse} = \frac{1}{I} \sum_i \left(\frac{1}{K} \sum_k (M^i_{n,k} - P^i_{n,k})^2 \right). \quad (10)$$

This is equivalent to estimating a soft mask separately for each source, but with all masks being computed simultaneously by the same network. Hence, while the target soft masks $M^i_{n,k}$ are normalized to sum to one, the output of the network is not guaranteed to do so. In order to preserve this quality, the estimate masks need to be normalized again.

2.3 Network Architecture

As mentioned in the previous sections, the proposed architecture consists in combining convolutional layers with a fully-connected output. This architecture has been applied to classification and recognition tasks in speech [17] and musical audio [18], but, to the best of our knowledge, not to musical audio separation (note that [10] included fully-connected layers, but not at the output, which has a different interpretation). This architecture can be used to optimize both targets described on Sect. 2.2.

Figure 1 describes the basic architecture used in our experiments. Dimensionality reduction is achieved via max-pooling layers, as convolutional layers are padded to result on outputs of the same size. Convolutional layers are connected via rectified linear unit (ReLU) functions, while fully-connected layers are connected with sigmoid functions. After the last fully-connected layer, either a softmax function or a sigmoid activation function can be used.

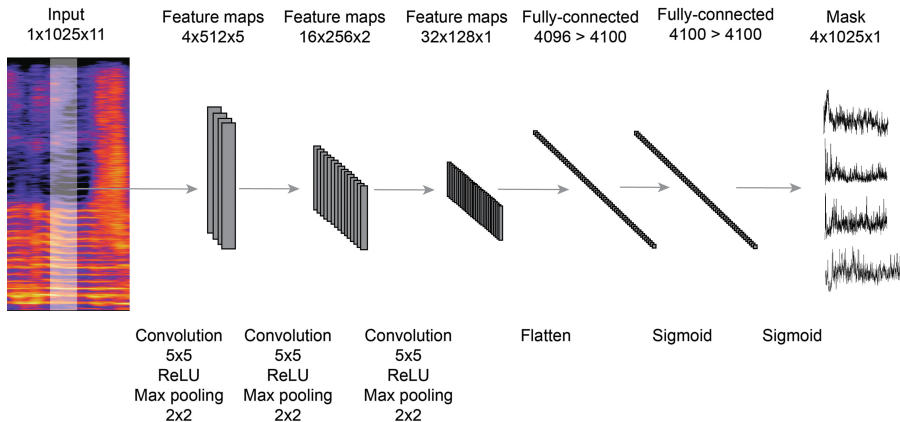


Fig. 1. Base CNN architecture

3 Evaluation

In order to assess the proposed model, we compared three different models on the task of separating musical audio using the DSD100 dataset, roughly following the experimental setting of the SiSEC campaign, which allows us to compare the results to state of the art approaches. Our goal was specifically to assess the addition of convolutional layers to a DNN network. The baseline DNN model (*dnn*) was devised by removing the convolutional layers and extending the input layer to accommodate 5 input frames. The second model (*cnn1*) included the convolutional layers. Both were trained to optimize l_{mse} with a sigmoid function and soft masks. The third model (*cnn2*) was trained to optimize l_{nll} with a softmax output function. For each model, we extracted estimates for vocals, bass, drums and other, the instrument categories of the dataset.

3.1 Experiment Setup

The development set, composed of 50 songs, was used for training, while the test set (also 50 songs) was used for testing. The dataset was managed using the `dsd-tools` package². All songs were mixed to mono by averaging both channels, and downsampled to 22050 Hz for processing. The estimates were upsampled again for evaluation, while the reference tracks were also downmixed. Each track was analyzed using a STFT with a window of 2048 samples (~ 100 ms) and hops of 256 samples (~ 10 ms). For each frame, we grouped a sequence of 11 context frames (~ 200 ms) and obtained the magnitude spectrogram of the mixture, and both the classification target and soft masks described in Sect. 2.2. The networks were trained using the Adaptive Moment Estimation (ADAM) variant of Stochastic Gradient Descent [19].

² <https://github.com/faroit/dsdtools>.

After shuffling the training set, a validation set of 20% of the data was used to determine the number training epochs. A threshold of 5 epochs was used to stop the training process if the loss had not decreased for the validation set during that time. We used batch normalization [20] for each convolutional layer. In our experience, using large enough batches this made normalizing the data unnecessary. For the *dnn* model, training data was normalized to zero mean and unit variance. We extracted the SDR, SIR and SAR measures typically used for source separation [21]. Results for each measure and target stream were compared using a Wilcoxon signed-rank test with Bonferroni correction. Implementation was based on the Pytorch³ python library. The source code can be obtained from <https://github.com/flucoma/LVA-ICA-2018>.

3.2 Results and Discussion

The results of the experiment are shown in Fig. 2. All pair-wise comparisons were found to be significant ($p < 0.01$), with a few exceptions.

The system worked particularly well for separating vocals, with a median SDR just above 4dB. This is similar to state-of-the art results employing multiple network setups, such as [6, 8], but using a single network. Also, the result for vocals is higher than previously published methods based on CNNs [10, 12]. In addition, our system extracts estimates for multiple sources in one pass. Results for other instruments are not as good. This may be due to the fact that early versions of the system were evaluated for extracting of vocals. In early experiments, separating the four streams improved the result for vocals, as opposed to separation of vocals vs accompaniment. The difference may also be due to the breadth of material that non-vocal categories encompass. Balancing the performance between the different instruments would probably require a compromise in terms of window size and overlap factor.

With respect to the different models, we were mainly interested in comparing *cnn1* with the other two, since *dnn* and *cnn2* have a different architecture as well as a different loss function. It should be noted that *dnn* did not have access to the same temporal context, but since this was extended to 5 frames, the number of parameters was higher than for the CNN models (38M vs 34M parameters). Hence, the results show that for a very similar architecture, convolutional layers allow increasing the temporal context seen by the network, resulting in better performance with a small addition of trainable parameters.

Finally, it could be expected that, since it produces a binary mask, *cnn2* would result in better SIR and lower SAR. This model still gives a similar overall result (SDR) and can be possibly adapted to work in remixing applications where artifacts would be diminished by the presence of all sources.

³ <http://pytorch.org/>.

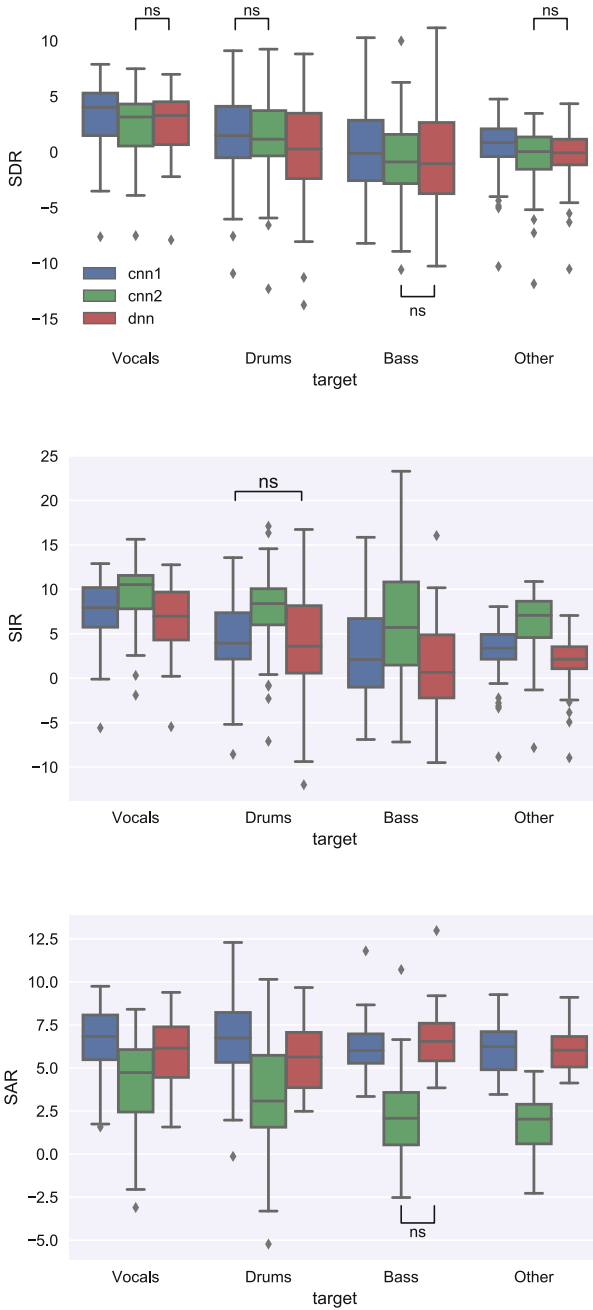


Fig. 2. Results of separation task with the Test set of the DSD100 dataset. All pairwise differences within each measure and target are statistically significant ($p < 0.01$) except where noted (“ns”).

4 Conclusions

In this paper we have studied the application of “traditional” CNN architectures to separation of musical audio. Since the use of DNNs is already well established for this task, this work can be seen as incremental, showing that the addition of convolutional layers can improve the results of DNN architectures by allowing access to a longer temporal context. Another advantage of these layers is that it is easy to add additional features. We hope to study this further to keep advancing this model. Our results show that this architecture allows achieving state-of-the-art separation for vocals using a single-network algorithm. We plan to investigate how to improve the results for other instruments. Some examples of the output of our system can be found in the companion page <http://www.flucoma.org/LVA-ICA-2018/>.

Acknowledgement. This research was part of the Fluid Corpus Manipulation project (FluCoMa), which has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 725899).

References

1. Roma, G., Grais, E.M., Simpson, A.J., Plumbley, M.D.: Music remixing and upmixing using source separation. In: Proceedings of the 2nd AES Workshop on Intelligent Music Production (2016)
2. Katz, B.: *Mastering Audio: The Art and the Science*, 3rd edn. Focal Press, Waltham (2014)
3. Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontcave, J.: The 2016 signal separation evaluation campaign. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 323–332. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_31
4. Wang, Y., Narayanan, A., Wang, D.: On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(12), 1849–1858 (2014)
5. Uhlich, S., Giron, F., Mitsufuji, Y.: Deep neural network based instrument extraction from music. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2135–2139, April 2015
6. Nugraha, A., Liutkus, A., Vincent, E.: Multichannel music separation with deep neural networks. In: Proceedings of the 24th European Signal Processing Conference (EUSIPCO), August 2016, pp. 1748–1752 (2016)
7. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Discriminative enhancement for single channel audio source separation using deep neural networks. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 236–246. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_23
8. Mimilakis, S.I., Drossos, K., Santos, J.F., Schuller, G., Virtanen, T., Bengio, Y.: Monaural Singing Voice Separation with Skip-Filtering Connections and Recurrent Inference of Time-Frequency Mask. [arXiv:1711.01437](https://arxiv.org/abs/1711.01437) [cs, eess], November 2017

9. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2528–2535. IEEE (2010)
10. Chandna, P., Miron, M., Janer, J., Gómez, E.: Monoaural audio source separation using deep convolutional neural networks. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 258–266. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_25
11. Grais, E.M., Plumbley, M.D.: Single channel audio source separation using convolutional denoising autoencoders. In: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1265–1269, November 2017
12. Grais, E.M., Wierstorf, H., Ward, D., Plumbley, M.D.: Multi-Resolution Fully Convolutional Neural Networks for Monoaural Audio Source Separation. [arXiv:1710.11473](https://arxiv.org/abs/1710.11473) [cs, eess], October 2017
13. Jansson, A., Humphrey, E.J., Montecchio, N., Bittner, R., Kumar, A., Weyde, T.: Singing voice separation with deep U-Net convolutional networks. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 323–332 (2017)
14. Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Su, L., Yang, Y.H., Jang, R.: Vocal activity informed singing voice separation with the iKala dataset. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 718–722. IEEE (2015)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Vincent, E., Bertin, N., Gribonval, R., Bimbot, F.: From blind to guided audio source separation: how models and side information can improve the separation of sound. *IEEE Sig. Process. Mag.* **31**(3), 107–115 (2014)
17. Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(10), 1533–1545 (2014)
18. Schlüter, J., Grill, T.: Exploring data augmentation for improved singing voice detection with neural networks. In: ISMIR, pp. 121–126 (2015)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. [arXiv preprint arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. [arXiv preprint arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
21. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)

Deep Learning and Data-driven Methods



Training Strategies for Deep Latent Models and Applications to Speech Presence Probability Estimation

Shlomo E. Chazan, Sharon Gannot^(✉), and Jacob Goldberger

Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel
{Shlomi.Chazan, Sharon.Gannot, Jacob.Goldberger}@biu.ac.il

Abstract. In this study we address models with latent variable in the context of neural networks. We analyze a neural network architecture, mixture of deep experts (MoDE), that models latent variables using the mixture of expert paradigm. Learning the parameters of latent variable models is usually done by the expectation-maximization (EM) algorithm. However, it is well known that back-propagation gradient-based algorithms are the preferred strategy for training neural networks. We show that in the case of neural networks with latent variables, the back-propagation algorithm is actually a recursive variant of the EM that is more suitable for training neural networks. To demonstrate the viability of the proposed MoDE network it is applied to the task of speech presence probability estimation, widely applicable to many speech processing problem, e.g. speaker diarization and separation, speech enhancement and noise reduction. Experimental results show the benefits of the proposed architecture over standard fully-connected networks with the same number of parameters.

Keywords: DNNs · Mixture of experts · Expectation-maximization

1 Introduction

The mixture of experts (MoE) model, introduced by Jacobs et al. [8, 9], provides an important paradigm for combining latent variables in discriminative models. The objective of this framework is to describe the behavior of a certain phenomenon under the assumption that there are separate processes involved in the generation of the data under analysis. The MoE model is comprised of several expert models and a gate model. Each of the experts provides a decision and the gate is a latent variable that selects the relevant expert based on the input data. Most MoE implementations are based on experts that are implemented by shallow models such as linear regression or logistic regression. In spite of the huge success of deep learning, there are only a few studies that have explicitly utilized and analyzed MoEs as an architecture component of a neural network [7, 11].

Neural networks deal with the problem of making probabilistic inference based on a given input data. In many cases, when formulating this problem, it is

natural to consider latent variables that control the network flow from the input features to the network output and affect the final decision.

In this study, we utilize the MoE framework to define a neural network with latent random variables. We propose a mixture of deep experts (MoDE) network architecture where both the experts and the gating are implemented by neural networks. The unobserved gating decision is a latent random variable which is marginalized by the neural network in the process of obtaining the final decision. A common technique for Maximum-Likelihood (ML) estimation of the model parameters in the presence of latent variables is the expectation-maximization (EM) algorithm [6]. The EM algorithm alternates between estimating the unobserved variables given the current model parameters and refitting the model given the estimated, complete data. In spite of the tremendous success of the EM algorithm in parameter estimation tasks, it does not scale well when the parametric model is corresponding to a neural network since the EM framework requires training a neural network in each iteration. For real-world, large-scale networks, even a single training iteration is a non-trivial challenge. Instead, the back-propagation (BP), gradient-based, algorithm is the standard method for training neural networks. The main contribution of this study is the establishment of the link between the BP algorithm and an on-line variant of the EM algorithm for the training of neural networks with latent variables [2, 12].

As an example of the applicability of the proposed MoDE modeling and training scheme, we apply it to the task of speech presence probability (SPP), widely used in speech processing tasks, e.g. speaker diarization and separation as well as speech enhancement and noise reduction. The speech signal comprises several different acoustic states such as the phoneme identity or the coarser distinction between voiced and unvoiced phonemes. Each such state induces a different relationship between the speech signal and the SPP that could be utilized to infer the SPP. In our modeling, each expert is responsible for a specific acoustic state and the gating network is responsible to inferring the speech state at each time frame. Unlike our previous method [3], in the current approach, there is no need for phoneme-labeled data, since the gating (DNN) is capable of splitting the input space in an unsupervised manner.

2 A Mixture of Deep Experts

In this section we first review the MoE framework and then use it to define neural networks with latent variables.

The MoE model is a discriminative latent variable model that produces a decision y given an input feature set x . We first sample an expert using a gating function and then apply the expert to produce the output label. The index of the selected expert is an intermediate hidden random variable denoted by z . Formally, the MoE conditional distribution can be written as follows:

$$p(y|x;\theta) = \sum_{i=1}^m p(z=i|x;\theta_g)p(y|z=i, x;\theta_i) \quad (1)$$

such that x is the feature, y is the classification result, z is a hidden random variable that selects the expert that is applied and m is the number of experts. The model parameter-set θ is composed of the parameter sets of the gating function θ_g and the parameter sets $\theta_1, \dots, \theta_m$ of the m experts. A simple example is a mixture of m linear regressions. In this case, each expert is a linear regression model with parameters $\theta_i = \{a_i, b_i, \sigma_i^2\}$. We first sample a hidden r.v. z from a discrete distribution. Then, if $z = i$, y is sampled from a normal distribution according to the rule: $y|(x, z=i) \sim N(y; a_i x + b_i, \sigma_i^2)$.

We next address the problem of learning the MoE parameters (i.e. the parameters of the experts and the gating function) given a training dataset $(x_1, y_1), \dots, (x_N, y_N)$, where N is the size of the database. The likelihood function of the MoE model parameters is:

$$L(\theta) = L(\theta_g, \theta_1, \dots, \theta_m) = \sum_{t=1}^N \log p(y_t | x_t; \theta). \quad (2)$$

Since the selected expert used to produce y_t from the feature set x_t (i.e. the value of the r.v. z_t) is hidden, it is natural to apply the EM algorithm to find the maximum-likelihood parameters [9]. The EM auxiliary function is:

$$Q(\theta, \tilde{\theta}) = \sum_{t=1}^N E_{p(z_t | x_t, y_t; \tilde{\theta})}(\log p(y_t, z_t | x_t; \theta)) \quad (3)$$

such that $\tilde{\theta}$ is the current parameter estimate. In the E-step we apply Bayes' rule to estimate the value of the selected expert based on the current parameter estimate:

$$w_{ti} = p(z_t = i | x_t, y_t; \tilde{\theta}) = \frac{p(y_t | x_t, z_t = i; \tilde{\theta}_i) p(z_t = i | x_t; \tilde{\theta}_g)}{p(y_t | x_t; \tilde{\theta})} \quad (4)$$

$t = 1, \dots, N, \quad i = 1, \dots, m.$

The M-step decouples the parameter estimation of the different components of the MoE model. We can optimize each of the experts and the gating function separately since in each case there is a separate set of parameters. The updated parameters of the gating function are obtained by maximizing the weighted likelihood function:

$$L_g(\theta_g) = \sum_{t=1}^N \sum_{i=1}^m w_{ti} \log p(z_t = i | x_t; \theta_g) \quad (5)$$

and the updated parameters of the experts are obtained by maximizing the functions:

$$L_i(\theta_i) = \sum_{t=1}^N w_{ti} \log p(y_t | x_t, z_t = i; \theta_i), \quad i = 1, \dots, m. \quad (6)$$

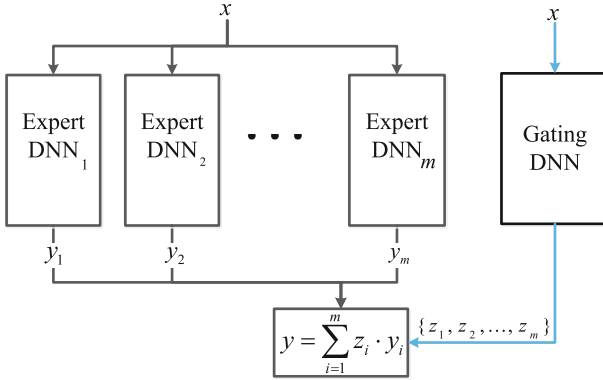


Fig. 1. A Mixture of Deep Experts (MoDE) architecture.

In this work we address the situation where both the experts and the gating functions are implemented by DNNs. This brings the DNNs modeling power to latent variables models. We denote this model mixture of deep experts (MoDE). The MoDE model is expressed as follows:

$$p_{\text{NN}}(y|x; \theta) = \sum_{i=1}^m p_{\text{NN}}(z = i|x; \theta_g) p_{\text{NN}}(y|z = i, x; \theta_i) \tag{7}$$

where p_{NN} is the DNN estimation. In this model, θ_g is the parameter-set of the DNN that implements the gating function and θ_i is the parameter-set of the DNN that implements the i -th expert. The MoDE architecture is illustrated in Fig. 1.

To apply the EM algorithm described above to MoDE, we need to train in the M-step both the gating and the experts neural networks using the objective functions defined by (5) and (6), respectively. We thus iterate between computing the posterior distribution of the latent variables at the E-step and training the DNNs of the experts and the gating at the M-step. This approach, however, requires training a neural network in each iteration of the EM algorithm while for real-world, large-scale networks, even a single training iteration is a non-trivial challenge. Another problem is that the EM algorithm is a greedy optimization procedure that is notorious for getting stuck in local optima. This is a major drawback in optimizing the likelihood functions of MoDEs that are highly non-concave. Finally, the main justification of the EM algorithm is a guarantee that the likelihood function is increased at each iteration. Here, in the M-step we train a neural network and there is no guarantee that we find the global optimum or even improve the likelihood compared to the previous EM iteration.

When estimating the parameters of a DNN with latent variables, we can either focus on the latent variable aspect and use the EM algorithm or focus on the DNN aspect and apply gradient methods via the back-propagation algorithm which simultaneously trains all the sub-networks by directly maximizing the likelihood function:

$$L(\theta) = \sum_{t=1}^N \log \left(\sum_{i=1}^m p_{\text{NN}}(z_t = i | x_t; \theta_g) \cdot p_{\text{NN}}(y_t | z_t = i, x_t; \theta_i) \right). \quad (8)$$

In this architecture, the experts and the gating networks are components of a single network and are simultaneously trained with the same objective function. It can be easily verified that the back-propagation equation for the parameter set of the i -th expert is:

$$\frac{\partial L}{\partial \theta_i} = \sum_{t=1}^N w_{ti} \cdot \frac{\partial}{\partial \theta_i} \log p_{\text{NN}}(y_t | z_t = i, x_t; \theta_i) \quad (9)$$

such that w_{ti} is the posterior distribution of the gating random variable:

$$w_{ti} = p_{\text{NN}}(z_t = i | x_t, y_t; \theta) = \frac{p_{\text{NN}}(y_t | x_t, z_t = i; \theta_i) p_{\text{NN}}(z_t = i | x_t; \theta_g)}{p_{\text{NN}}(y_t | x_t; \theta)}. \quad (10)$$

In a similar way, the back-propagation equation for the parameter set of the gating DNN is:

$$\frac{\partial L}{\partial \theta_g} = \sum_{t=1}^N \sum_{i=1}^m w_{ti} \cdot \frac{\partial}{\partial \theta_g} \log p_{\text{NN}}(z_t = i | x_t; \theta_g). \quad (11)$$

The two algorithms (EM and BP) for training neural networks with latent variables are very similar. Expression (10) coincides with one term of auxiliary function constituting the E-step of the EM algorithm defined in (4). The back-propagation partial derivative w.r.t θ_i (9) is identical to the partial derivatives of the function $L_i(\theta_i)$ (6) that is optimized by the M-step and the partial derivative w.r.t. θ_g coincides with the partial derivative of $L_g(\theta_g)$ (5). We next establish the exact connection between the two training strategies.

There are on-line variants of the EM for latent data models with independent observations. One of the dominant approaches to on-line EM-like estimation is the method proposed by Titterton [12], which consists in using a stochastic approximation algorithm, where the parameters are updated after the acquisition of each new observation. It is a Newton-type algorithm that uses the gradient of the incomplete-data likelihood¹ weighted by the expectation of the Hessian, namely the complete-data Fisher information matrix. A simplified variant of this method is a first-order gradient approach where the complete data Fisher information matrix is replaced by a scalar learning rate parameter. Applying this approximate EM procedure is identical to the standard back-propagation training procedure. In the case of BP, we can use a subset of the training data (mini-batch) or even a single training example to compute the gradient. For the case of using a single example the updating equations of the back-propagation and the simplified Titterton's scheme are identical:

$$\theta_g \leftarrow \theta_g + \epsilon \sum_{i=1}^m w_{ti} \cdot \frac{\partial}{\partial \theta_g} \log p_{\text{NN}}(z_t = i | x_t; \theta_g) \quad (12)$$

¹ The gradient of the incomplete-data likelihood can be calculated by the expectation of the complete-data likelihood by the Fisher identity.

$$\theta_i \leftarrow \theta_i + \epsilon w_{ti} \cdot \frac{\partial}{\partial \theta_i} \log p_{\text{NN}}(y_t | z_t = i, x_t; \theta_i), \quad i = 1, \dots, m \quad (13)$$

such that t is the index of the current example, w_{ti} is defined in (10) and ϵ is the learning rate.

A related study [10], analyzed the relations between the EM algorithm and gradient-based methods for standard generative models such as mixture of Gaussians and hidden Markov models. Here, we provide an analysis for discriminative latent variable models that are implemented by a neural network. We showed that the standard back-propagation training algorithm is essentially an on-line variant of the EM algorithm.

3 Deep Mixture of Experts for SPP Estimation

In this section we apply the MoDE principle to a speech presence probability (SPP) estimation task and describe the network specifics and training procedure.

Let $s(n)$ denote a sample of speech signal at time n . Let $x(n) = s(n) + v(n)$ denote the observed, single microphone, noisy signal where additive noise $v(n)$ was added to the clean speech.

The short-time Fourier transform (STFT) with a frame of length L of $x(n)$ is denoted by $X(t, k)$, where t is the frame index and $k = 0, 1, \dots, L - 1$ denotes the frequency band index. Define the log-spectrum of the noisy signal at a single time frame by $\mathbf{x}(t)$, such that the k -th component is $x_k = \text{Log}|X(t, k)|$ where $k = 0, \dots, L/2$.

The hidden speech state $z(t)$ corresponds to a building block of a speech signal. Unlike our supervised approach in [3] in which each expert is specializing in a specific phoneme, here the network splits the role of each expert in an unsupervised manner.

All m experts in the proposed algorithm are implemented by DNNs with the same structure. The input to each DNN is the noisy log-spectrum frame together with 8 context frames (4 frames from the past and 4 from the future). The network consists of 3 fully-connected hidden layers with 500 (ReLU) neurons each. The targets to the network are the associated binary masks defined by:

$$B(t, k) = \begin{cases} 1 & x_{t,k} > \text{Tr}(k) \\ 0 & \text{o.w} \end{cases} \quad (14)$$

where $\text{Tr}(k)$ is a threshold over the log-spectrum of the clean speech signal for the k -th frequency band. We stress that the threshold is applied here to the ground truth of the clean signal and not to a noisy version thereof. Two values for the threshold are set, the first for the low frequencies band and the second for the high frequencies band. We set the thresholds such that in the low frequencies the harmonics of the clean speech signals will be accurate (high $\text{Tr}(k)$), and in the high frequencies the unvoiced patterns will be preserved (lower $\text{Tr}(k)$).

The output layer that provides the soft SPP decisions is composed of $L/2+1$ sigmoid neurons, one for each frequency band. The SPP decision of the i -th expert for the t -th frame and k -th frequency bin is denoted as $\rho_i(t, k)$.

The architecture of the all the expert DNNs and the gating DNN are identical. Each DNN comprises of 3 fully connected hidden layers with 500 ReLU neurons each. Note, that the output layer of all m experts is a sigmoid function, while the output layer of the gating DNN is a softmax function that produces the gating distribution for the m experts. The gating (p.d.f.) is therefore:

$$p_i(t) = p(z(t) = i | \mathbf{x}(t); \theta_g). \quad (15)$$

The averaged SPP is obtained by a weighted average of the deep experts' decisions:

$$\rho(t, k) = p(B(t, k) = 1 | \mathbf{x}, \theta) = \sum_{i=1}^m p_i(t) \cdot \rho_i(t, k). \quad (16)$$

The proposed MoDE algorithm for SPP estimation is summarized in Algorithm 1.

Algorithm 1. MoDE speech presence probability estimation.

Input :

- Noisy speech log-spectral vector at time t , x_t .
- MoDE model parameters $\theta = \{\theta_1, \dots, \theta_m, \theta_g\}$.

Output: Speech presence probability (SPP) $\rho(t, k)$

- **Experts' DNNs:** Compute SPP decision $\rho_i(t, k)$ for each expert $i \in \{1, \dots, m\}$ and for each time-frequency bin (t, k) .
 - **Gating DNN:** $p_i(t) = p(z_t = i | \mathbf{x}_t; \theta_g)$
 - **Average Experts' decisions:** $\rho(t, k) = \sum_{i=1}^m p_i(t) \cdot \rho_i(t, k)$
-

4 Experimental Results

In the training phase clean signals drawn from the train set of the TIMIT corpus (462-speaker train set) were contaminated with the Speech-like and Babble noises with 2 SNRs, 5 dB and 10 dB.

To test the proposed MoDE algorithm we contaminated different speech signals with several types of noise from the NOISEX-92 database [13], namely *Speech-like*, *Babble*, *Room* and *Factory*. The noise was added to the clean signal drawn from the test set of the TIMIT database (24-speaker core test set), with 5 levels of (SNR) at -5 dB, 0 dB, 5 dB, 10 dB and 15 dB chosen to represent various real-life scenarios.

The number of experts in the experiment section was set to $m = 10$, and thus it is denoted here MoDE-10. We compared the proposed algorithm to the classic model-based (OMLSA) algorithm [5] with the (IMCRA) noise estimator [4]

which is a state-of-the-art algorithm for single microphone speech enhancement. Note, that although the OMLSA is a speech enhancement algorithm, here its SPP estimation is tested. The default parameters of the OMLSA were set according to [1].

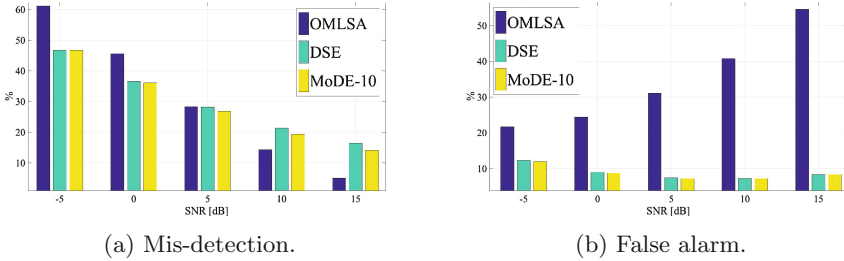


Fig. 2. Objective measurements of the hard decision of the compared SPPs.

Additionally, we compared the proposed MoDE algorithm to another DNN which has a fully-connected architecture and can be viewed as a single-expert network. We denote this network the (DSE). The DSE architecture is a single DNN with 3 fully connected hidden layers with ReLU neurons. The output layer was set to be a sigmoid to estimate the SPP. For a fair comparison the number of parameters of DSE is identical to the number of the total parameters in the proposed MoDE-10. The DSE and the MoDE were both trained with the same database.

To evaluate the performance of the SPP estimation algorithm, hard-decision is applied to (16):

$$\hat{B}(t, k) = \begin{cases} 1 & \rho(t, k) > 0.5 \\ 0 & \text{otherwise} \end{cases}. \quad (17)$$

A binary decision was also applied in other algorithms. Figure 2 depicts the averaged *mis-detection* and *false-alarm* percentage in the four tested noise types. It is clear that the OMLSA algorithm does not perform well, since it has high values of mis-detection and false-alarm. It is also evident that the MoDE-10 algorithm outperforms the DSE in most SNRs in both mis-detection and false-alarm measures.

Figure 3 depicts the results for all algorithms in a scenario with Babble noise type with SNR=5 dB. The true binary mask is also shown for comparison in Fig. 3c. It is evident that the proposed MoDE-10 algorithm outperforms the competing OMLSA and DSE algorithms and produce results that are most similar to the true binary mask \mathbf{B} . As expected, the OMLSA false alarm values here are high while the DNN-based algorithms are much better. The MoDE-10 is more accurate than the DSE algorithm. We encircled in Fig. 3e and f areas where the MoDE-10 detection is better than the its DSE counterpart. It is evident that MoDE-10 reconstructs the harmonic structure in a much better way.

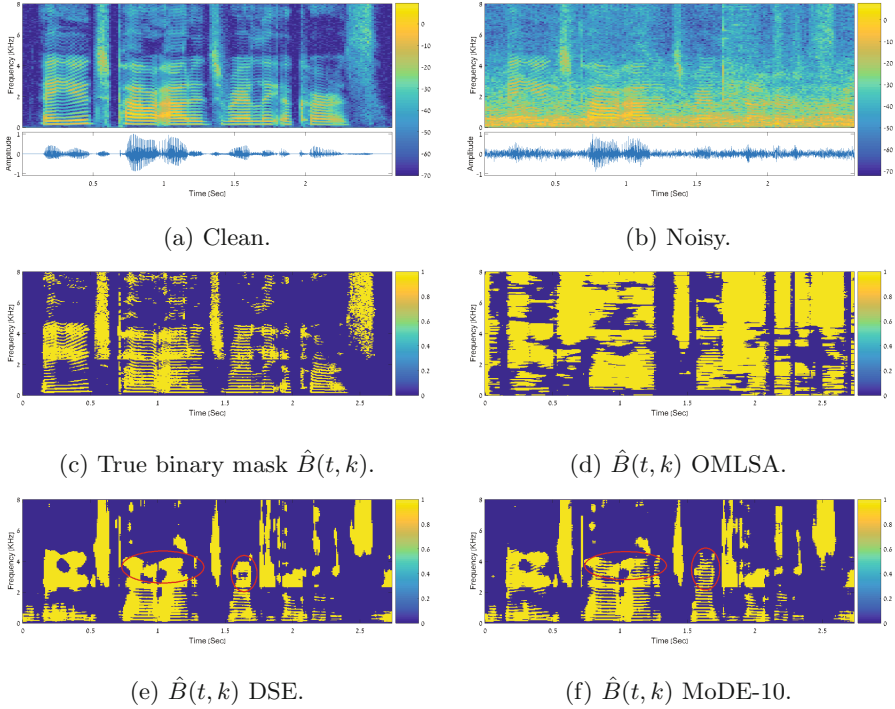


Fig. 3. Performance of the evaluated algorithms in a scenario with Babble noise with SNR= 5 dB.

5 Conclusions

In this study we addressed deep latent variable models. We proposed the MoDE network architecture which implements latent variables in a neural networks setup.

In this study, we showed that in the case of neural network with latent variables, BP is actually an on-line version of the EM algorithm. We demonstrated the benefits of using latent variable in neural network addressing the problem of speech presence estimation.

References

1. Matlab software for speech enhancement based on optimally modified lsa (OMLSA) speech estimator and improved minima controlled recursive averaging (IMCRA) noise estimation approach for robust speech enhancement. <http://webee.technion.ac.il/people/IsraelCohen/>
2. Cappé, O., Moulines, E.: On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B (Stat. Method.)* **71**(3), 593–613 (2009)

3. Chazan, S.E., Gannot, S., Goldberger, J.: A phoneme-based pre-training approach for deep neural network with application to speech enhancement. In: 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5, September 2016
4. Cohen, I., Berdugo, B.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Sig. Process. Lett.* **9**(1), 12–15 (2002)
5. Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. *Sig. Process.* **81**(11), 2403–2418 (2001)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**(1), 1–38 (1977). <http://www.jstor.org/stable/2984875>
7. Eigen, D., Ranzato, M., Sutskever, I.: Learning factored representations in a deep mixture of experts. In: International Conference on Learning Representations (ICLR), Workshop (2014)
8. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. *Neural Comput.* **3**(1), 79–87 (1991)
9. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.* **6**(2), 181–214 (1994)
10. Salakhutdinov, R., Roweis, S., Ghahramani, Z.: Optimization with EM and expectation-conjugate-gradient. In: International Conference on Machine Learning (ICML) (2003)
11. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR) (2017)
12. Titterton, D.M.: Recursive parameter estimation using incomplete data. *J. R. Stat. Soc. Ser. B* **46**, 257–267 (1984)
13. Varga, A., Steeneken, H.J.: Assessment for automatic speech recognition: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **12**(3), 247–251 (1993)



Jointly Detecting and Separating Singing Voice: A Multi-Task Approach

Daniel Stoller¹(✉) , Sebastian Ewert² , and Simon Dixon¹(✉) 

¹ Queen Mary University of London, London, UK
{d.stoller,s.e.dixon}@qmul.ac.uk

² Spotify, London, UK
sewert@spotify.com

Abstract. A main challenge in applying deep learning to music processing is the availability of training data. One potential solution is Multi-task Learning, in which the model also learns to solve related auxiliary tasks on additional datasets to exploit their correlation. While intuitive in principle, it can be challenging to identify related tasks and construct the model to optimally share information between tasks. In this paper, we explore vocal activity detection as an additional task to stabilise and improve the performance of vocal separation. Further, we identify problematic biases specific to each dataset that could limit the generalisation capability of separation and detection models, to which our proposed approach is robust. Experiments show improved performance in separation as well as vocal detection compared to single-task baselines. However, we find that the commonly used Signal-to-Distortion Ratio (SDR) metrics did not capture the improvement on non-vocal sections, indicating the need for improved evaluation methodologies.

Keywords: Singing voice separation · Vocal activity detection
Multi-task learning

1 Introduction and Related Work

Separating the singing voice from the accompaniment in music recordings is a challenging task, with the acoustical properties of the instruments involved and their interactions in a recording being highly complex. Most current approaches train deep neural networks on multi-track recordings in a supervised fashion to estimate the individual sources from a given mixture input [9, 15]. While this approach often leads to considerable improvements over previous methods, it requires suitable input-output pairs from multi-track recordings. Unfortunately, publicly available datasets are often rather small on the order of a few hundred tracks. This leads to overfitting and limits overall performance.

S. Ewert—Work was conducted at Queen Mary University of London.

D. Stoller—This work was funded by EPSRC grant EP/L01632X/1.

Informed source separation aims to circumvent this problem by providing additional information to the separation model, e.g. the musical score [6]. This way, the problem can be simplified, which often leads to improved results on small, annotated datasets. On the other hand, such approaches can only be employed if suitable side information is indeed available, which is often not the case for musical scores. In this paper, we thus focus on a more readily available and more easily created type of side information: vocal activity labels.

A joint separation-classification model [12] was proposed for the more general problem of *sound event detection* that employs a separation network whose output mask for each source is summarised with a mean or max operation to detect active sound events. While similar to our approach, it is designed for weak labels and might be more sensitive to dataset biases when training with different separation and detection datasets due to its simple detection component. Heittola *et al.* [8] use precise activity labels, but separation is used as a front-end for detection instead of performing joint estimation. Therefore, separation cannot be improved using mixtures with only activity labels.

To our knowledge, Chan *et al.* [4] provide the only work combining *singing voice separation (SVS)* in particular, with *singing voice detection (SVD)*. Vocal activity labels are used to construct a mask, which forces the corresponding parts of the mixture spectrogram to be modelled individually in a method based on robust principal component analysis (RPCA). For an increase in separation quality however, vocal activity labels are required during prediction. The labels also have to be quite precise as a false negative label would force the vocal estimate to be zero for vocal sections.

Schlüter [18] focusses solely on SVD, but also shows that the resulting network can be used for detecting the location of the singing voice in the time-frequency domain. This suggests it might be useful to integrate the information contained in the activity labels into separation models to improve their performance. A related method was introduced by Ikemiya *et al.* [10]. It produces a rough estimate of the vocals in a first step. After computing the fundamental frequency based on this estimate, the separation result is further refined. These two steps are repeated until convergence. We aim for a similar yet more integrated and joint estimation approach for the case of vocal activity labels.

Overall, vocal detection and separation are usually tackled as separate tasks despite their commonalities. Thus, a main goal in this paper is to explore how such information can be exploited in training audio-only models that can jointly detect and separate vocals. First, we use a simple approach for diversifying the training dataset for an SVS model, and observe that its implicit assumption that all data sources are from the same distribution is violated due to a bias specific to each dataset. Using a multi-task learning (MTL) approach, we then propose a model shown in Fig. 1 that performs SVS and SVD at the same time and can better account for such biases. The model can be trained on multi-track recordings in combination with mixtures with vocal activity labels, and yields predictions on completely unlabelled mixtures. By allowing the model to exploit correlations between the vocal activity labels and the source signals, performance

is improved for both tasks compared to baseline models trained with single-task learning (STL). While the overall improvement remained at a rather low level, we found the effect to be quite consistent – despite the small size of the datasets involved and their respective biases. We also found that the most commonly used evaluation metric [20] is flawed in the sense that capturing improvements on non-vocal sections are not captured, and propose a simple ad-hoc solution. As an additional contribution, we discuss the dataset biases we observed in some detail. Overall, based on these findings, we hypothesise that the joint prediction of source estimates along with side information such as musical scores in a multi-task setting could be a promising general direction for further research in music source separation.

2 Proposed Approaches

As a baseline system for SVS, we implemented a variant of the U-Net described in Sect. 3.2 and shown in Fig. 1. The approach is similar to [11, 19] and outputs a mask when given spectrogram magnitudes of a mixture excerpt. During training, audio excerpts are randomly selected from the multi-track dataset, and converted to a log-normalised spectrogram representation. The mean squared error (MSE) in spectral magnitudes between source estimates from the separator and the ground truth is used as a loss function.

2.1 Initial Approach to SVS: Using Additional Non-vocal Sections

Initially, we attempted to improve SVS performance by adding audio excerpts from instrumental sections of the SVD dataset to the SVS training set to increase its diversity: Standard supervised training on a multi-track dataset entails randomly selecting audio excerpts from the tracks to generate batches of samples.

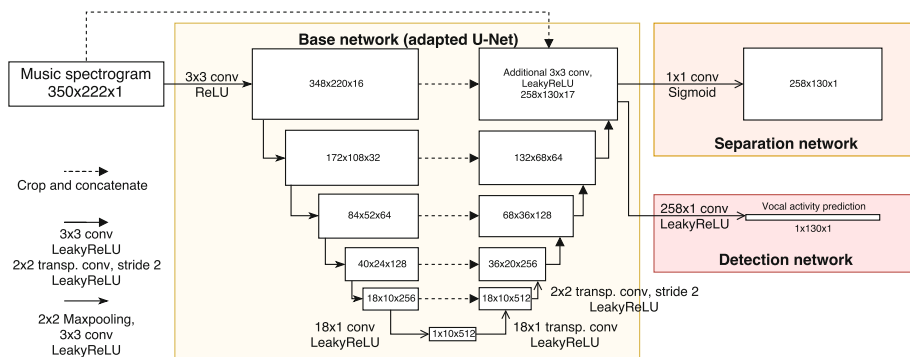


Fig. 1. Our multi-task model for jointly detecting and separating singing voice, given the spectrogram of a music piece as input. Tensor shapes are given in the order of frequency bins, time frames, and feature channels.

We changed this procedure so that when encountering an audio excerpt with silent vocals, it can be replaced with a randomly chosen non-vocal section from an additional music database with vocal activity labels. The replacement occurs with a probability of $\frac{N}{N+M}$, with N and M being the size of the SVS and SVD dataset, respectively, to ensure non-vocal sections are effectively randomly sampled from both datasets. To train from the additional non-vocal sections, we set their target accompaniment equal to the magnitudes of the respective mixture spectrogram, and all target magnitudes of the vocal spectrogram to zero.

The average MSE loss (see (3)) on the test set obtained when training the same model with and without this replacement technique was used to test whether separation performance improves. We performed the above training procedure with three different set-ups for the SVS and SVD dataset.

In the first experiment, we used the DSD100 [14] dataset for SVS training, testing and evaluation, and RWC [16] and Jamendo [17] as the SVD dataset. We also included a private collection of Dubstep, Hardstyle, Jazz, Classical and Trance music with 25 songs per genre. We found that the performance decreased compared to purely supervised learning. A first suspicion was that a bias in the test set might be responsible for inaccurate test performance measurements since only DSD100 is used (see Sect. 2.2 for details).

To investigate this issue more closely, we conducted a second experiment and additionally included the MedleyDB [2], CCMixer [13] and iKala [4] SVS datasets in the validation and test sets. Compared to the first experiment, the SVS training and test data is now less well matched, and the test performance gives a more accurate picture of generalisation capability. Here performance increased considerably using our technique, strongly indicating that a bias in the SVS training data can be alleviated by including extra non-vocal sections.

Finally, we distributed the DSD100, MedleyDB, CCMixer and iKala datasets in equal proportions into training, validation and test set for a more realistic set-up in which all available multi-track data is used, but in this experiment, separation performance again decreased using our approach.

These results suggest that the individual datasets are subject to different biases in the data distribution space, to which our approach is sensitive since it assumes that all samples come from the same distribution. These biases will be investigated in more detail in the next section. Another shortcoming of our approach is that we cannot learn from the additional vocal sections using this method since we do not have the source audio available.

2.2 Dataset Bias for Singing Voice Separation and Detection

Since we are combining data from different sources, it is important to consider the impact of dataset bias on the performance of models trained on such combined data. We hypothesised that datasets used for SVD and SVS are each uniquely biased, which can include properties such as the relative energy of the sources, overall energy levels and how often vocals occur on average. We computed metrics for the above for the MedleyDB, DSD100, CCMixer, iKala, Jamendo and

RWC datasets, as they are commonly used for SVD and SVS. Vocals were considered active if the average absolute amplitude in a 10 ms window exceeded $5 \cdot 10^{-4}$. Figure 2 shows the distribution of these properties for each dataset, where metrics have been averaged song-wise.

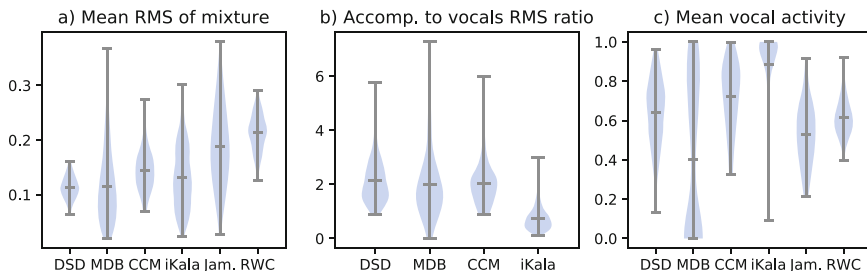


Fig. 2. Distribution of values for different collections of tracks, for different properties. Outliers for MedleyDB in (b) resulting from instrumental tracks have been excluded.

Clear dataset bias manifests itself in the uneven distribution of values across datasets. For example, iKala contains relatively loud vocals and very few instrumental sections, and CCMixer has louder tracks than DSD100 with more vocals on average. Additionally, even more dataset bias could be present in features which are more difficult to detect and quantify, such as timbre, language of the lyrics, music genre, recording conditions or the bleed level for multi-track recordings. Therefore, it is very difficult to directly prevent models from overfitting to these biases. We would like to highlight this as a critical problem for the field of SVS and SVD, since many models are trained on a single dataset source and thus may not generalise nearly as well as the test scores indicate.

2.3 Multi-task Learning Approach

To mitigate problems due to dataset biases, we employ a multi-task learning (MTL) approach [3] instead. We augment the separation model with a component that predicts vocal activity based on a hidden layer of the separation model. We train the combined model to output the source signals in the multi-track dataset and the vocal activity labels in the SVD dataset, respectively, with most parameters being shared for both tasks.

This approach has multiple benefits. Firstly, predicting both outputs based on a shared hidden representation only assumes that the source output has some relationship with human-annotated vocal activity labels, but we do not define it explicitly. For example, temporal inaccuracy in labels could mean that the beginnings of vocals are annotated as non-vocal. If we force the vocal output of the separator to be silent for all sections annotated as non-vocal, or use the approach from Sect. 2.1, we give incorrect information to the separator. Secondly, a different dataset bias for each task can be accounted for by the model to some

extent with its task-specific components. Thirdly, we exploit the information present in extra non-vocal and vocal sections. Finally, the trained model can be used for both SVS and SVD.

For the SVS task, we use the MSE between the separator prediction $f_\phi(\mathbf{m})$ for a mixture excerpt \mathbf{m} and the true sources \mathbf{s} as the loss:

$$L_{\text{MSE}} = \mathbb{E}_{(\mathbf{m}, \mathbf{s}) \sim p^1} \frac{1}{N} \|\mathbf{s} - f_\phi(\mathbf{m})\|^2 \quad (1)$$

where p^1 represents the multi-track dataset distribution, which is approximated by a batch of samples, and N denotes the dimensionality of the joint source vectors \mathbf{s} and $f_\phi(\mathbf{m})$. For output spectrograms with T time frames, F frequency bins and K sources, $N = T \cdot F \cdot K$.

For the SVD task, we use the binary cross-entropy at each time frame of the spectrogram excerpt, averaged over time and over data points:

$$L_{\text{CE}} = \mathbb{E}_{(\mathbf{m}, \mathbf{o}) \sim p^2} \frac{1}{T} \sum_{t=1}^T \log p_\phi^t(o_t | \mathbf{m}) \quad (2)$$

where p_ϕ^t denotes the probability of the vocal state the model assigns to time frame t of the audio excerpt with a total of T frames, and p^2 describes the SVD dataset distribution whose samples contain a binary vector \mathbf{o} with a vocal activity label o_t at each spectral frame t of the source output spectrogram.

For our MTL model, we combine the two above losses using a simple weighting scheme:

$$L_{\text{MTL}} = \alpha L_{\text{MSE}} + (1 - \alpha) L_{\text{CE}}. \quad (3)$$

We set $\alpha = 0.9$ so that experimentally the losses are approximately on the same scale during training. Although an optimisation of this hyper-parameter might improve results further, it is omitted here due to computational cost. We also experimented with a loss function derived from a Maximum Likelihood objective¹, but did not obtain better performance.

3 Evaluation

Next, we describe the experimental evaluation procedure for our MTL approach.

3.1 Datasets

For the SVS dataset, we use DSD100 with 50 songs each for training and testing, according to the predefined split. We use the Jamendo dataset for SVD, since it predominantly contains Western Pop and Rock music, similarly to DSD100, to avoid a large dataset bias. Jamendo’s validation and test partitions comprising

¹ See ancillary files at <https://arxiv.org/abs/1804.01650>.

30 songs are used for testing, leaving 60 songs for training. This set-up is intended as a proof of concept of the MTL approach – in this setting even slight improvements are promising, since vocal activity labels do not directly yield information on vocal structure, and should translate to larger improvements given larger SVS and particularly SVD datasets.

3.2 Model Architecture and Preprocessing

The audio input is converted to mono and down-sampled to 22050 Hz to reduce dimensionality, before the magnitude spectrogram is computed from a 512-point FFT with 50% overlap, and normalised by $x \rightarrow \log(1 + x)$. Excerpts comprised of 222 time frames each are used as input to our model shown in Fig. 1, which consists of a base network that branches off into a separation and a detection network.

The **base network** closely follows our previous implementation [19] of the U-Net [11]. The output of an initial 3×3 convolution with 16 filters and ReLU non-linearity is fed to a down-sampling block consisting of max-pooling with size and stride two followed by a 3×3 convolution with 32 filters. The down-sampling block is applied three more times, each time doubling the number of filters, finally yielding a $18 \times 10 \times 256$ feature map. We then use a 1D convolution with filter size 18×1 before applying the respectively transposed convolution, and concatenate it with the original $18 \times 10 \times 256$ feature map to capture frequency relationships. In the following up-sampling block, a 2×2 transposed convolution with 128 filters is applied, and the output concatenated with the output of the down-sampling block at the same network depth after centre-cropping it. Lastly, a 3×3 convolution with 128 filters is applied. After applying this up-sampling block another three times, each time with half as many filters for the convolutions, the resulting $258 \times 130 \times 16$ feature map is concatenated with the centre-cropped input. The resulting features are input to the SVS well as the SVD sub-network.

The output size is smaller than the input size since we use “valid” convolutions that do not employ implicit zero-padding. Therefore, the mixture naturally provides additional temporal context processed during convolution, and its magnitudes are zero-padded in frequency so that the separator output has the correct number of frequency bins. Unless otherwise stated, Leaky ReLU is used after all convolutions as non-linearities to allow for better gradient flow.

In the **SVS network**, the feature map from the base architecture is transformed into a filtering mask, which is multiplied point-wise with the original mixture spectrogram magnitudes to yield the source estimates. To generate the source audio, we use an inverse STFT using the mixture’s phase, and apply 10 iterations of the Griffin-Lim algorithm [7] to further refine the phase.

The **SVD network** takes the final feature map from the base architecture and applies a single $F \times 1$ filter, where F is the number of frequency bins, to reduce the time-frequency feature map to a single scalar for each time step. Application of a sigmoid non-linearity yields the probability of the presence of singing voice at each time step.

3.3 Experimental Set-Up and Metrics

To identify the impact of our proposed approach in comparison to solving separation and detection separately, we train and evaluate our network solely for either SVS or SVD, before comparing to training with the multi-task loss.

Model performance is evaluated on the test dataset every 1000 iterations and the model with the best performance is selected. Training is stopped after 10,000 iterations without performance improvement. For SVD, we use the *area under the receiver operating characteristic (AU-ROC)* to evaluate performance. For separation, we use the MSE training objective from (3) in the normalised magnitude space, as well as the track-wise SDR, SIR, and SAR metrics [20] on the audio signals. We select two MTL models with the best AU-ROC or MSE value, respectively, since best performance is reached at different training stages.

3.4 Results

Table 1 shows a performance comparison of the considered models. For both SVD and SVS, we achieve a slight improvement in both AU-ROC and MSE performance metrics using our model variants. This is promising since the SVD dataset is small and vocal activity labels are less informative training targets than the vocals themselves. Therefore, larger datasets could be used in future work to obtain larger performance increases.

Table 1. Performance comparison between SVS and SVD baseline and our approach. Results significantly better than the comparison model ($p < 0.05$) in bold. Significance of the AU-ROC difference determined with binary labels from all time frames as samples [5]. A paired Wilcoxon signed-rank test was used for all other metrics.

Model		Metric								
		AU-ROC	MSE	Non-voc. RMS	Vocals			Accompaniment		
					SDR	SIR	SAR	SDR	SIR	SAR
SVD	0.9239	-	-	-	-	-	-	-	-	-
SVS	-	0.01865	0.0194	2.83	5.27	6.88	6.71	14.75	13.25	
Ours	0.9250	0.01755	0.0155	2.86	5.56	6.23	6.69	13.24	14.11	

While the MSE on the normalised spectrogram magnitudes improves by about 6%, the mean SDR for vocals and accompaniment does not change significantly. To find the cause, we analyse the employed implementation for SDR computation on the DSD100 dataset² also used in the SiSec source separation evaluation campaign [14]. Tracks are partitioned into excerpts of 30 s duration, using 15 s of overlap, for which a local SDR value is computed. The final SDR is the average of the local SDR values. However, for excerpts where at least one source is completely silent, the SDR has an undefined value of $\log(0)$ and is

² <https://github.com/faroit/dsd100mat>.

excluded from the final SDR average, so that the model’s performance in these sections is ignored. This is the case for 79 of 736 excerpts due to non-vocal sections and is thus a practically relevant flaw of the evaluation metric.

More sophisticated methods such as [21] take audio perception more explicitly into account, but presumably suffer from the same issue with silent sources, as similar computations are used there as well. As an ad-hoc solution, we propose computing the source estimate’s energy or ideally loudness for silent sections of the source ground truth as a simple workaround and report it in addition to other metrics. Finding a consistent and perceptually accurate evaluation metric is thus an important unsolved problem, and listening tests arguably remain important to accurately assess separation quality.

A lower average MSE combined with a stagnating SDR suggests that our model improves especially on these non-vocal sections excluded from the SDR, potentially because negative vocal activity labels allow the separator to detect many different instruments as not being vocals. To test this more explicitly, we take the vocal estimates of the baseline and our model and compute the average RMS of the 79 excerpts excluded from SDR computation, as well as the average output over whole songs in the DSD100 dataset. We find that our model has less energy in its vocal output compared to the baseline, but also in the non-vocal sections (see Table 1). This demonstrates that our model performs better on non-vocal sections and about equally on vocal sections due to a similar SDR.

4 Conclusions

We demonstrated that jointly solving the task of singing voice detection and singing voice separation can improve performance in both tasks and alleviates the issue of dataset scarcity. Furthermore, we found biases specific to each dataset that could prevent source separation and detection models from generalising properly to unseen data. Finally, we discuss a major flaw in the most popular evaluation metric for source separation [20] related to the performance measurement in silent sections.

Therefore, further research into improved, perceptually relevant metrics is a definite need. As a workaround, we propose additionally measuring and reporting the loudness of the model’s source estimates for sections where the respective source is silent. Our multi-task approach could be generalised and applied to mixtures with pitch curve or phoneme annotations of the singing voice, or even to whole transcriptions of musical sources (see [1]). Performance increases can be expected to be larger especially for the latter case as correct predictions on one task greatly simplify solving the other one.

Acknowledgements. We thank Emmanouil Benetos for the useful comments and feedback, as well as Mi Tian for references on related literature.

References

1. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A.: Automatic music transcription: challenges and future directions. *J. Intell. Inf. Syst.* **41**(3), 407–434 (2013)
2. Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.: MedleyDB: a multitrack dataset for annotation-intensive MIR research. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) (2014)
3. Caruana, R.: *Multitask Learning*, pp. 95–133. Springer, Boston (1998). https://doi.org/10.1007/978-1-4615-5529-2_5
4. Chan, T.S., Yeh, T.C., Fan, Z.C., Chen, H.W., Su, L., Yang, Y.H., Jang, R.: Vocal activity informed singing voice separation with the iKala dataset. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 718–722 (2015)
5. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3), 837–845 (1988)
6. Ewert, S., Sandler, M.B.: Structured dropout for weak label and multi-instance learning and its application to score-informed source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2277–2281 (2017)
7. Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoust. Speech Sig. Process.* **32**(2), 236–243 (1984)
8. Heittola, T., Mesaros, A., Virtanen, T., Eronen, A.: Sound event detection in multi-source environments using source separation. In: *Machine Listening in Multisource Environments* (2011)
9. Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Singing-voice separation from monaural recordings using deep recurrent neural networks. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 477–482 (2014)
10. Ikemiya, Y., Yoshii, K., Itoyama, K.: Singing voice analysis and editing based on mutually dependent F0 estimation and source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 574–578 (2015)
11. Jansson, A., Humphrey, E.J., Montecchio, N., Bittner, R., Kumar, A., Weyde, T.: Singing voice separation with deep U-Net convolutional networks. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 323–332 (2017)
12. Kong, Q., Xu, Y., Wang, W., Plumbley, M.D.: A joint separation-classification model for sound event detection of weakly labelled data. CoRR abs/1711.03037 (2017). <http://arxiv.org/abs/1711.03037>
13. Liutkus, A., Fitzgerald, D., Rafii, Z.: Scalable audio separation with light kernel additive modelling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 76–80 (2015)
14. Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontecave, J.: The 2016 signal separation evaluation campaign. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) *LVA/ICA 2017*. LNCS, vol. 10169, pp. 323–332. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_31

15. Luo, Y., Chen, Z., Hershey, J.R., Roux, J.L., Mesgarani, N.: Deep clustering and conventional networks for music separation: stronger together. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 61–65 (2017)
16. Mauch, M., Dixon, S.: pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 659–663 (2014)
17. Ramona, M., Richard, G., David, B.: Vocal detection in music with support vector machines. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1885–1888 (2008)
18. Schlüter, J.: Learning to pinpoint singing voice from weakly labeled examples. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pp. 44–50 (2016)
19. Stoller, D., Ewert, S., Dixon, S.: Adversarial semi-supervised audio source separation applied to singing voice extraction. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2018)
20. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
21. Vincent, E.: Improved perceptual metrics for the evaluation of audio source separation. In: Theis, F., Cichocki, A., Yeredor, A., Zibulevsky, M. (eds.) *LVA/ICA 2012*. LNCS, vol. 7191, pp. 430–437. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28551-6_53



Multi-Resolution Fully Convolutional Neural Networks for Monaural Audio Source Separation

Emad M. Grais^(✉), Hagen Wierstorf, Dominic Ward, and Mark D. Plumbley

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
{grais,dominic.ward,m.plumbley}@surrey.ac.uk, hagen.wierstorf@posteo.de

Abstract. In deep neural networks with convolutional layers, all the neurons in each layer typically have the same size receptive fields (RFs) with the same resolution. Convolutional layers with neurons that have large RF capture global information from the input features, while layers with neurons that have small RF size capture local details with high resolution from the input features. In this work, we introduce novel deep multi-resolution fully convolutional neural networks (MR-FCN), where each layer has a range of neurons with different RF sizes to extract multi-resolution features that capture the global and local information from its input features. The proposed MR-FCN is applied to separate the singing voice from mixtures of music sources. Experimental results show that using MR-FCN improves the performance compared to feedforward deep neural networks (DNNs) and single resolution deep fully convolutional neural networks (FCNs) on the audio source separation problem.

Keywords: Multi-resolution features extraction
Fully convolutional neural networks · Deep learning
Audio source separation · Audio enhancement

1 Introduction

Monaural audio source separation (MASS) aims to separate audio sources from a single (mono) audio mixture [20]. A variety of deep neural networks with convolutional layers have been used recently to tackle this problem [2, 6, 11, 12, 18]. One of the main differences in those works relies on using either fully convolutional neural networks (FCN), where all the network layers are convolutional layers, or networks where some of the layers are convolutional and others are fully connected layers. The common aspect in those works is that each convolutional layer is composed of a set of neurons/filters that have the same receptive field (RF) size. The RF is the field of view of a neuron (filter in the FCN case) in a certain layer in the network [21]. In fully connected deep neural networks (DNNs), the output of each neuron in a certain layer depends on the entire input to that layer, while the output of a neuron in a convolutional layer only depends

on a region of the input: this region is the RF for that neuron. The RF size is a crucial issue in many audio and visual tasks, as the output must respond to areas with sizes correspond to the sizes of the different objects or patterns in the input data to extract useful information/features about each object [21]. The size of the RF equals the size of the filters in a convolutional layer. A large filter size captures the global structure of its input features [9, 17], while a small filter size captures the local details with high resolution but it does not capture the global structure of its input features. Intuitively, it might be useful to have sets of filters that can extract both the global and local details from the input features in each layer. This might be useful in the MASS problem, since the input signal is a mixture of different audio sources and useful features can be extracted for certain sources in certain time-frequency resolutions which may differ from one source to another [16].

The concept of extracting multi-resolution features has been proposed recently for many signal processing applications with different ways of extracting and combining the multi-resolution features from the input data [5, 9, 13, 23]. In this paper, we introduce a novel multi-resolution fully convolutional neural network (MR-FCN) model for MASS, where each layer in the MR-FCN is a convolutional layer that is composed of different sets of filters. All filters within a given set have the same size, which is different to the size of filters in other sets in the same layer. Thus, in each layer there are sets of filters with large and small sizes, which allows each layer to extract multi-resolution features that capture the global and local information from its input data. We believe that this is the first time that a deep neural network has been proposed with each layer composed of multi-resolution filters that extract multi-resolution features from the layer before, and the first time that the concept of extracting multi-resolution features has been used for MASS. The inputs and outputs of the MR-FCN are two-dimensional (2D) segments from the magnitude spectrogram of the mixed and target source signals respectively. The MR-FCN is trained to extract useful spectro-temporal features and patterns in different time-frequency resolutions to separate the target source from the input mixture.

This paper is organized as follows: Sect. 2 shows a brief introduction about the FCN and the proposed MR-FCN. The proposed approach of using MR-FCN for MASS is presented in Sect. 3. Section 4 introduces our experiment and discusses the results, and Sect. 5 draws conclusions and directions for future work.

2 Multi-Resolution Fully Convolutional Neural Networks

In this section we first give an introduction about the fully convolutional neural network (FCN) then we introduce the proposed MR-FCN.

2.1 Fully Convolutional Neural Networks (FCNs)

The FCN model that we propose here (Fig. 1) is somewhat similar to the convolutional denoising encoder-decoder (auto-encoder) network that was used in

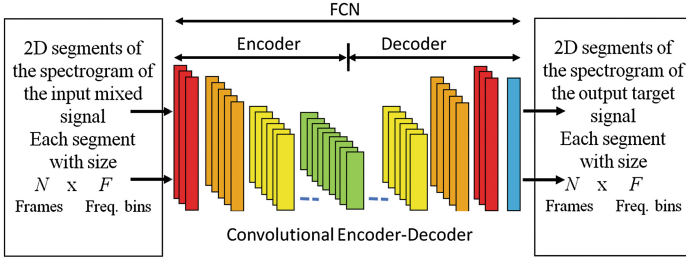


Fig. 1. Overview structure of a FCN that separates one target source from the mixed signal. Each layer consists of a single set of filters with the same size followed by an activation function. The sets of filters in the input and output layers have large filter sizes and small number of filters. The number of filters increases and the size decreases when getting further from the input and output layers [15]. There is symmetric in the filter sizes and numbers of filters between the encoder and decoder sides.

[6, 15], but without using either down-sampling (pooling) or up-sampling. The encoder part in the FCN is composed of repetitions of a convolutional layer and an activation layer. The decoder part consists of repetitions of a transpose convolutional layer [4] and an activation layer. Each layer in the FCN consists of a single set of filters with the same size to extract feature maps from its input layer, and the activation layer imposes nonlinearity to these feature maps.

The FCN can be trained from corrupted input signals and the encoder part is used to extract noise robust features that the decoder can use to reconstruct a cleaned-up version of the input data [15, 24]. In MASS, the input mixed signal can be seen as a sum of the target source that needs to be separated and background noise (the other sources in the mixture). The input and output data of the FCN are 2D signals (magnitude spectrograms) and the filtering is a 2D operator.

2.2 Multi-resolution FCN

Each layer in the FCN in Fig. 1 is composed of one set of filters that have the same RF size. The size of the RF is a very important parameter, as the output of each filter should respond to areas with sizes correspond to the sizes of the different objects/patterns in the input to extract useful information/features from the input data [21]. For example, if the size of the RF of a filter is much bigger than the size of the input pattern, the filter may capture blurred features from the input patterns, while if the RF of a unit is smaller than the size of the input patterns, the output of the filter loses the global structure of the input patterns [21].

In audio source separation problems, the spectrogram of the input mixed signal usually contains different combinations of different spectro-temporal patterns from different audio sources. There is a unique set of patterns associated with each source in the spectrogram of their mixture, and these patterns appear in different spectro-temporal sizes [1]. So, to use the FCN to extract useful

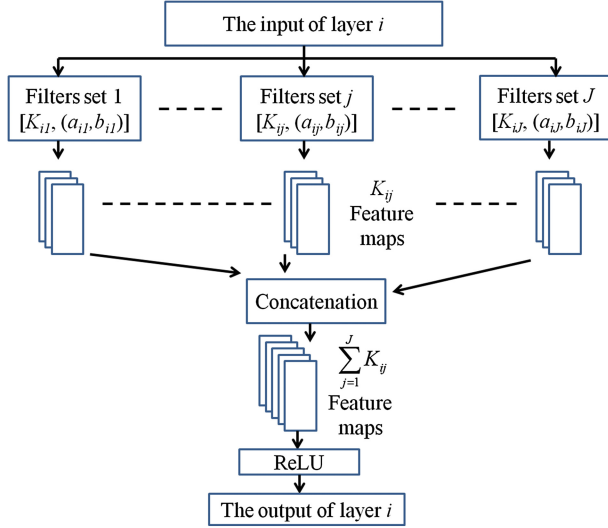


Fig. 2. Overview of the proposed structure of each layer of the MR-FCN. Where K_{ij} denotes the number of filters with size $a_{ij} \times b_{ij}$ in set j in layer i , a_{ij} is the dimension in the time direction of the filters, and b_{ij} is the dimension in the frequency direction of the filters in set j and layer i . The filters in different sets have different sizes and the filters within a set have the same size. Each set j in layer i generates K_{ij} feature maps. The number of feature maps that each layer i generates equal to the sum of the number of feature maps that all the sets in layer i generate ($\sum_{j=1}^J K_{ij}$). ReLU denotes a rectified linear unit as an activation function.

information about the individual sources in the spectrogram of their mixture, it might be useful to use filters with different RF sizes in each layer, where the different RF sizes are proportional to the diversity of the spectro-temporal sizes of the patterns in the spectrogram. Bearing these issues in mind, we propose a MR-FCN which is the FCN shown in Fig. 1 but with multi-resolution filters (filters with different sizes) in each layer. Thus, each layer in the MR-FCN has sets of 2D filters. The filters in one set have the same size which is different to the sizes of the filters in all other sets in the same layer. Each set of filters generates feature maps with certain time-frequency resolution. Fig. 2 shows the detailed structure for each layer in the MR-FCN. Each layer in the MR-FCN generates multi-resolution features from its input features and also combines the multi-resolution features from the previous layers to generate accurate patterns that compose the structure of the underlying data.

3 MR-FCN for MASS

Given a mixture of L sources as $y(t) = \sum_{l=1}^L s_l(t)$, the aim of MASS is to estimate the sources $s_l(t)$, $\forall l$, from the mixed signal $y(t)$ [20]. We work here in

the short-time Fourier transform (STFT) domain. Given the STFT of the mixed signal $y(t)$, the main goal is to estimate the STFT of each source in the mixture.

In this work, we propose to use as many MR-FCN as the number of sources to be separated from the mixed signal. Each MR-FCN sees the mixed signal as a combination of its target source and background noise. The main aim of each MR-FCN is to separate its corresponding target source from the other background sources that exist in the mixed signal. This is a challenging task since each MR-FCN deals with highly nonstationary background noise (other sources in the mixture). The inputs and outputs of the MR-FCNs are 2D-segments from the magnitude spectrograms of the mixed and target signals respectively. Therefore, the MR-FCNs span multiple spectral frames to capture multi-resolution spectro-temporal characteristics for each source. The number of spectral frames that each segment has is N and the number of frequency bins is F . In this work, F is the dimension of the whole spectral frame.

3.1 Training the MR-FCNs for Source Separation

We train each MR-FCN to map the magnitude spectrogram of the input mixture into the magnitude spectrogram of its corresponding target source. Let us assume that we have training data for the mixed signals and their corresponding clean/target sources. Let \mathbf{Y}_{tr} be the magnitude spectrogram of the mixed signal and \mathbf{S}_l be the magnitude spectrogram of the target source l . The subscript “tr” denotes the training data. The MR-FCN that separates source l from the mixture is trained to minimize the following cost function:

$$C_l = \sum_{n,f} (\mathbf{Z}_l(n, f) - \mathbf{S}_l(n, f))^2 \quad (1)$$

where \mathbf{Z}_l is the actual output of the last layer of the MR-FCN of source l , \mathbf{S}_l is the reference target output for source l , and n and f are the time and frequency indices respectively. The input to all the MR-FCNs is the magnitude spectrogram \mathbf{Y}_{tr} of the mixed signal. The input and output instants of the MR-FCNs are 2D-segments, where each segment is composed of N consecutive spectral frames taken from the magnitude spectrograms. This allows each MR-FCN to learn multi-resolution spectro-temporal patterns for its corresponding target source.

3.2 Testing the MR-FCNs for Source Separation

After training a MR-FCN for each source we wish to separate from the mixed signal, the magnitude spectrogram \mathbf{Y} of the mixed signal is passed through all the trained MR-FCNs. The output of the MR-FCN of source l is the estimate $\hat{\mathbf{S}}_l$ of the magnitude spectrogram of source l . The time domain estimate $\tilde{s}_l(t)$ is computed using the inverse STFT of the estimate $\hat{\mathbf{S}}_l$ and the phase of the STFT of the input mixture.

4 Experiments

We applied our proposed MASS using MR-FCN approach to separate the singing voice/vocal sources from a group of songs from the SiSEC-2015-MUS-task dataset [14]. The dataset has 100 stereo songs with different genres and instrumentations. To use the data for the proposed MASS approach, we converted the stereo songs into mono by computing the average of the two channels for all songs and sources in the data set. Each song is a mixture of vocals, bass, drums, and a group of other musical instruments. We used one MR-FCN to separate the vocal from each song.

The first 50 songs in the dataset were used as training and validation datasets, and the last 50 songs were used for testing. The data were sampled at 44.1 kHz. The magnitude spectrograms for the data were calculated using the STFT with Hanning window size 2048 points and hop size of 512 points. The FFT was computed with 2048 points and the first 1025 were used as features since they include the conjugate of the remaining points.

The quality of the separated sources was measured using the source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR) [19]. SIR indicates how well the sources are separated based on the remaining interference between the sources after separation. SAR indicates the artifacts caused by the separation algorithm in the estimated separated sources. SDR measures the overall distortion (interference and artifacts) of the separated sources. The SDR values are usually considered as the overall performance evaluation for any source separation approach [19]. Achieving high SDR, SIR, and SAR indicates good separation performance.

We compared the performance of using the proposed MR-FCN model with the performance of using feedforward deep neural networks (DNNs) and the (single-resolution) FCN for separating the vocal signals from each song in the test set. For the input and output data for the MR-FCN and FCN, we chose the number of spectral frames in each 2D-segment to be 15 frames. This means the dimension of each input and output instant for the MR-FCN and FCN is 15 (time frames) \times 1025 (frequency bins) as in [6]. Thus, each input and output instant (the 2D-segments from the spectrograms) spans around 209 ms of the waveforms of the data. Each input and output instant of the DNN is a single frame of the magnitude spectrograms of the input and output signals respectively.

4.1 Choosing the Parameters of the Models

As in many deep learning models, there are many parameters in the proposed MR-FCN to be chosen (number of layers, filter size, and the number of filters in each set) and usually these choices are data and application dependent. Choosing the parameters for the FCN is also not easy. In this work, we follow the same strategy as in [15] where the size of the filters decreases but the number of the filters increases as we progress through the layers of the encoder part. In contrast, we use fewer filters of increasing size as we develop through the decoder part. For MR-FCN, the number and size of the filters in each set in each layer are

need to be decided. We restricted ourselves in this work to use only three sets of filters for the whole network. The first set with size 15×39 (each filter in this set spans around 209 ms of the waveforms and a band of frequencies around 840 Hz in the spectrogram), the second set with size 9×19 (each filter in this set spans around 139 ms of the waveforms and a band of frequencies around 409 Hz in the spectrogram), and the third set with size 5×5 (each filter in this set spans around 93 ms of the waveforms and a band of frequencies around 108 Hz in the spectrogram). Which means each layer has sets of filters with three different time-frequency resolutions. Also following the same concept in [15] for choosing the number of filters, the layers towards the input and output layers have more filters with large size than the layers in the middle. The layers in the middle have more filters in the set with small filter size than the layers toward the input and output layers. For example, the first layer in MR-FCN has a set of 12 filters with size 15×39 , a set of 6 filters with size 9×19 , and a set of 6 filters with size 5×5 . Thus, the first layer generates 24 feature maps with three different resolutions. Each feature map is 15×1025 (the same size of the input and output segments).

Table 1. The filter specifications and the number of filters in each layer of the FCN and MR-FCN. For example “Conv2D[26,(15,39)]” denotes 2D convolutional layer with 26 filters and the size of each filter is 15×39 where 15 is the size of the filter in the time-frame direction and 39 in the frequency direction of the spectrogram.

FCN and MR-FCN model summary			
The input/output data with size 15 frames and 1025 frequency bins			
Layer number	FCN	MR-FCN	
1	Conv2D[26,(15,39)]	set 1	Conv2D[12,(15,39)]
		set 2	Conv2D[6,(9,19)]
		set 3	Conv2D[6,(5,5)]
2	Conv2D[42,(9,19)]	set 1	Conv2D[8,(15,39)]
		set 2	Conv2D[22,(9,19)]
		set 3	Conv2D[8,(5,5)]
3	Conv2D[66,(5,5)]	set 1	Conv2D[12,(15,39)]
		set 2	Conv2D[12,(9,19)]
		set 3	Conv2D[32,(5,5)]
4	Conv2D[42,(9,19)]	set 1	Conv2D[8,(15,39)]
		set 2	Conv2D[22,(9,19)]
		set 3	Conv2D[8,(5,5)]
5	Conv2D[26,(15,39)]	set 1	Conv2D[12,(15,39)]
		set 2	Conv2D[6,(9,19)]
		set 3	Conv2D[6,(5,5)]
6	Conv2D[1,(15,1025)]	Conv2D[1,(15,1025)]	
Total number of parameters	1,784,027	1,755,321	

To attempt to make a fair comparison between the proposed MR-FCN model and the FCN, we adjusted the number of filters and their sizes in each layer of both models to have total number of parameters in both models close to each other as shown in Table 1. Table 1 shows the number of layers, the number of filters in each layer, and the size of the filters for the FCN and MR-FCN models. The DNN has three hidden layers, and each hidden layer has 1025 nodes. The parameters of the DNN are tuned based on our previous work on the same dataset [7]. The rectified linear unit (ReLU) is used as the activation function for all the neural networks in this work. The DNN here has 4,206,600 parameters, the FCN has 1,784,027 parameters, and the MR-FCN has 1,755,321 parameters. This means the MR-FCN has the smallest number of parameters compared to the FCN and the DNN.

The parameters for all the networks were initialized randomly. They were trained using backpropagation with gradient descent optimization using Adam [10] with parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, a batch size 100, and an initial learning rate of 0.0001 which was reduced by a factor of 10 when the values of the cost function ceased to decrease on the validation set for 3 consecutive epochs. The maximum number of epochs was 100. We implemented our proposed algorithm using Keras with Tensorflow backend [3].

4.2 Results

Figure 3 shows boxplots of the SDR (a), SIR (b), and SAR (c) as measured on the vocals separated using three different deep learning models, namely DNN, FCN, and MR-FCN. The figure also shows the SDR and SIR values of the target vocal source in the mixed signal (denoted as Mix in Fig. 3). We did not show the SAR of the mixed signal because it is usually very high. From the figure we can see that the vocal signals in the input mixed signal (denoted as Mix in Fig. 3) have very low SDR and SIR values, which shows that we are dealing with a very challenging source separation problem.

As can be seen from Fig. 3, the three methods perform well on the SDR, SIR, and SAR values of the separated vocal signals. The proposed MR-FCN model outperforms the DNN and slightly outperforms the FCN in all measurements.

In the following, we consider the difference between a pair of models statistically significant if $p < 0.05$, Wilcoxon signed-rank test [22] and Bonferroni corrected [8]. Based on the shown results in Fig. 3, the difference between each pair of models for all the shown results of SDR is statistically significant with P values as follows. For SDR: $P(\text{DNN}, \text{FCN}) = 1.12 \times 10^{-7}$, $P(\text{DNN}, \text{MR-FCN}) = 1.22 \times 10^{-7}$, and $P(\text{FCN}, \text{MR-FCN}) = 0.004$. For SIR: $P(\text{DNN}, \text{FCN}) = 0.9$, $P(\text{DNN}, \text{MR-FCN}) = 0.04$, and $P(\text{FCN}, \text{MR-FCN}) = 4.9 \times 10^{-4}$. For SAR: $P(\text{DNN}, \text{FCN}) = 2.2 \times 10^{-7}$, $P(\text{DNN}, \text{MR-FCN}) = 3.8 \times 10^{-7}$, and $P(\text{FCN}, \text{MR-FCN}) = 2.02$. In particular, the MR-FCN is statistically significantly better than FCN in SDR and SIR ($p < 0.05$), and statistically significantly better than DNN in all the measurements.

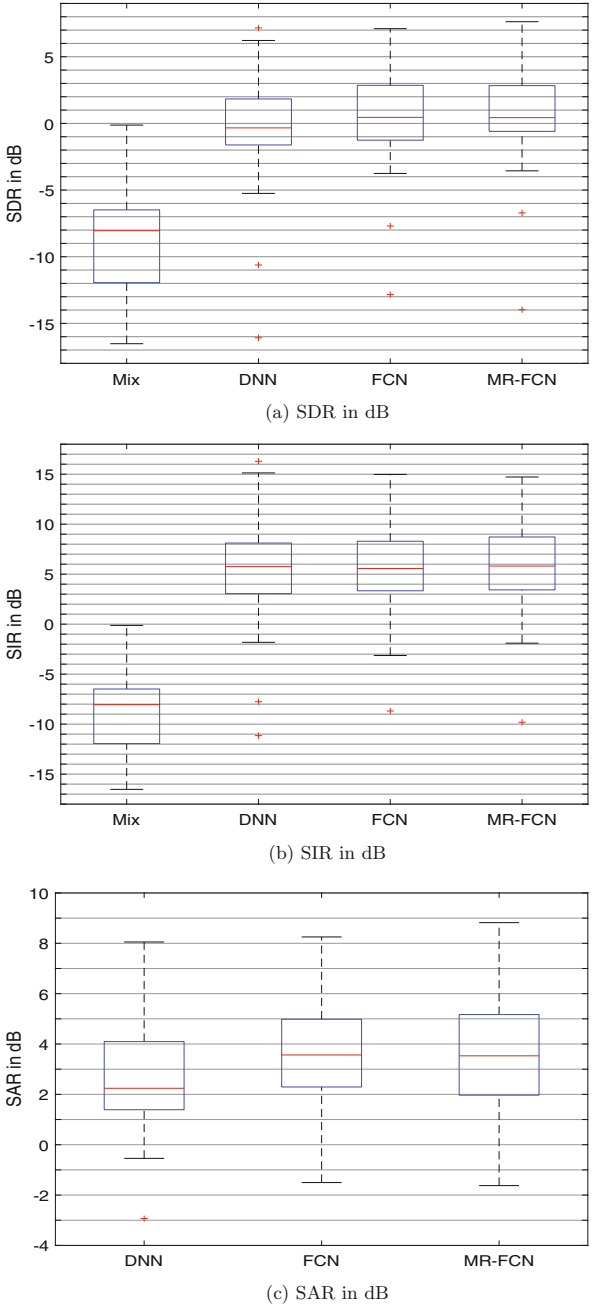


Fig. 3. (a) The SDR, (b) the SIR, and (c) the SAR (values in dB) for the separated vocal signals of using: deep fully connected feedforward neural networks (DNNs), deep fully convolutional neural networks (FCNs), and the proposed multi-resolution fully convolutional neural networks (MR-FCN). “Mix” denotes the input mixed signal.

4.3 Discussion

Although the difference between the results of MR-FCN and FCN is statistically significant ($p < 0.05$) in SDR and SIR, the improvement of using MR-FCN over FCN is marginal: the mean difference between MR-FCN and FCN is less than 1 dB in SDR and SIR. We believe that the filter sizes and the number of filters in each set should be refined to yield further improvements. These choices could be associated with the band of frequencies that each source covers in the input mixtures. Note that, FCN in this experiment has 28,706 more parameters than MR-FCN. In our future work, we will investigate different choices for the filter sizes and number of filters in each layer and each set.

5 Conclusions

In this work we proposed a new approach for monaural audio source separation (MASS). The new approach is based on using deep multi-resolution fully convolutional neural networks (MR-FCN). The MR-FCN learns multi-resolution patterns for each source and uses this information to separate the related components of each source from the mixed signal. The experimental results indicate that using MR-FCN for MASS is a promising approach and with a few number of parameters can achieve better results than the feedforward neural networks and the single resolution fully convolutional neural networks.

In our future work, we will investigate the possibility of applying the MR-FCN on raw audio data (time domain signals) to extract multi-resolution time-frequency features that can represent the input data better than the STFT features. Some audio sources require higher resolution in time than in frequency, and other audio sources require the opposite resolution of that. By applying MR-FCN on the raw audio data, we hope to extract useful features for each source according to its preferred time-frequency resolution which can improve the performance of many audio signal processing approaches.

Acknowledgement. This work is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).


References

1. Klapuri, A., Davy, M.: Signal Processing Methods for Music Transcription. Springer, Boston (2007). <https://doi.org/10.1007/0-387-32845-9>
2. Chandna, P., Miron, M., Janer, J., Gómez, E.: Monoaural audio source separation using deep convolutional neural networks. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 258–266. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_25
3. Chollet, F.: Keras (2015). <https://github.com/fchollet/keras>
4. Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning. [arXiv:1603.07285](https://arxiv.org/abs/1603.07285) (2016)
5. Espi, M., Fujimoto, M., Kinoshita, K., Nakatani, T.: Exploiting spectro-temporal locality in deep learning based acoustic event detection. EURASIP J. Audio Speech Music Process. **26**, 1–12 (2015)

6. Grais, E.M., Plumbley, M.D.: Single channel audio source separation using convolutional denoising autoencoders. In: Proceedings of GlobalSIP (2017)
7. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Combining mask estimates for single channel audio source separation using deep neural networks. In: Proceedings of InterSpeech (2016)
8. Hochberg, Y., Tamhane, A.C.: Multiple Comparison Procedures. Wiley, New York (1987). <https://doi.org/10.1002/9780470316672>
9. Kawahara, J., Hamarneh, G.: Multi-resolution-Tract CNN with hybrid pretrained and skin-lesion trained layers. In: Wang, L., Adeli, E., Wang, Q., Shi, Y., Suk, H.-I. (eds.) MLMI 2016. LNCS, vol. 10019, pp. 164–171. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47157-0_20
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) and presented at ICLR (2015)
11. Lim, W., Lee, T.: Harmonic and percussive source separation using a convolutional auto encoder. In: Proceedings of EUSIPCO (2017)
12. Miron, M., Janer, J., Gomez, E.: Monaural score-informed source separation for classical music using convolutional neural networks. In: Proceedings of ISMIR (2017)
13. Naderi, N., Nasersharif, B.: Multiresolution convolutional neural network for robust speech recognition. In: Proceedings of ICEE (2017)
14. Ono, N., Rafii, Z., Kitamura, D., Ito, N., Liutkus, A.: The 2015 signal separation evaluation campaign. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 387–395. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_45
15. Park, S.R., Lee, J.W.: A fully convolutional neural network for speech enhancement. In: Proceedings of Interspeech (2017)
16. Simpson, A.J.: Time-frequency trade-offs for audio source separation with binary masks. [arXiv:1504.07372](https://arxiv.org/abs/1504.07372) (2015)
17. Tang, Y., Mohamed, A.: Multi resolution deep belief networks. In: Proceedings of AISTATS (2012)
18. Venkataramani, S., Smaragdis, P.: End-to-end source separation with adaptive front-ends. In: Proceedings of WASPAA (2017)
19. Vincent, E., Gribonval, R., Fevotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
20. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process.* **15**, 1066–1074 (2007)
21. Wenjie, L., Yujia, L., Raquel, U., Richard, Z.: Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of NIPS, pp. 4898–4906 (2016)
22. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bulletin.* **1**(6), 80–83 (1945)
23. Xue, W., Zhao, H., Zhang, L.: Encoding multi-resolution two-stream CNNs for action recognition. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) ICONIP 2016. LNCS, vol. 9949, pp. 564–571. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46675-0_62
24. Zhao, M., Wang, D., Zhang, Z., Zhang, X.: Music removal by convolutional denoising autoencoder in speech recognition. In: Proceedings of APSIPA (2016)



Long-Term SNR Estimation Using Noise Residuals and a Two-Stage Deep-Learning Framework

Xuan Dong^(✉)  and Donald S. Williamson

Indiana University, Bloomington, IN 47408, USA
{xuandong,williads}@indiana.edu

Abstract. Knowing the signal-to-noise ratio of a noisy speech signal is important since it can help improve speech applications. This paper presents a two-stage approach for estimating the long-term signal-to-noise ratio (SNR) of speech signals that are corrupted by background noise. The first stage produces noise residuals from a speech separation module. The second stage then uses the residuals and a deep neural network (DNN) to predict long-term SNR. Traditional SNR estimation approaches use signal processing, unsupervised learning, or computational auditory scene analysis (CASA) techniques. We propose a deep-learning based approach, since DNNs have outperformed other techniques in several speech processing tasks. We evaluate our approach across a variety of noise types and input SNR levels, using the TIMIT speech corpus and NOISEX-92 noise database. The results show that our approach generalizes well in unseen noisy environments, and it outperforms several existing methods.

Keywords: Signal-to-noise ratio estimation · Speech separation
Deep neural networks

1 Introduction

The signal-to-noise ratio (SNR) is a strong indicator of the amount of noise interference in a given auditory environment. Knowledge of the SNR is useful for many speech-based applications, including hearing aids [1], automatic speech recognition (ASR) [2] and speech enhancement [3], where it can be used to select model parameters or optimization strategies [4]. For a given noisy speech signal, SNR is calculated from the speech and noise components, by comparing the energy of the speech signal to the energy of the noise. Unfortunately, in real environments, the SNR must be estimated since access to the speech and noise components is not possible.

There are typically two categories for SNR estimation algorithms. The first category performs SNR estimation at the time-frequency (T-F) unit level of a signal. This is known as instantaneous or short-time SNR [5], since SNR is

computed over smaller time segments. In [5], short-time SNR is computed from low-energy envelope estimates of noisy speech. In [6], a Gaussian mixture model (GMM) is used in the log-power domain to estimate the distributions of noise and noisy speech. The decision-directed (DD) approach estimates a priori SNR with a weighted sum of the a priori SNR estimate of the prior frame and the maximum likelihood SNR estimate of the current frame [7]. The accuracy of these approaches, however, degrades when estimates are computed over long durations.

The second category performs SNR estimation at the utterance level, referred to as global or long-term SNR. The widely used NIST SNR estimation algorithm uses the bimodal observation of the short-time energy histogram of noisy speech, to infer the distributions of noise and noisy speech [8]. It then uses these distributions to calculate the peak SNR, which erroneously overestimates the true SNR. The waveform amplitude distribution analysis (WADA) approach uses a gamma distribution to model the amplitudes of clean or noisy speech using a fixed shaping parameter, and a Gaussian distribution to model the background noise [9]. WADA estimates long-term SNR by computing the maximum likelihood estimate for the shaping parameter, but WADA only performs well when the above assumptions are met, which is not always the case. Long-term SNR is also calculated from a noise power spectral density (PSD) estimator [10] or a clean speech PSD estimator [11]. A computational auditory scene analysis (CASA) based approach is proposed in [12]. The algorithm uses an ideal binary mask (IBM) to segregate noisy speech into speech dominated and noise dominated T-F regions. The energy within each region is aggregated and used to compute the long-term SNR. This unsupervised approach, however, relies on the ability of the estimated IBM to correctly label T-F units as speech or noise dominated, which does not often occur at low SNR levels. This ultimately leads to performance degradations.

The goal of our work is to improve long-term SNR estimation of noisy speech in many complex environments, since current approaches do not always perform well. Unlike prior approaches, we propose a data-driven framework that uses deep learning to perform SNR estimation. Deep neural networks (DNNs) are used, largely due to their recent success in many speech processing tasks, including automatic speech recognition and speech separation [13–15], where they have outperformed alternative approaches and been shown to generalize in unseen environments. Environmental noise plays a dominant role in degrading SNR, so our idea is to use noise distortions as an indicator of long-term SNR. Specifically, we propose a two-stage long-term SNR estimation framework. In the first stage, a speech separation system separates noisy speech into enhanced speech and noise residuals. The residuals contain mostly noise energy and can be regarded as a reasonable noise indicator for the next stage. Then the second stage uses the residuals to estimate the long-term SNR of noisy speech in a supervised manner. Our results reveal that this strategy outperforms similar single- or two-stage DNN-based approaches.

This paper is organized as follows. A detailed description of our approach is given in Sect. 2. Experimental results and system comparisons are given in Sect. 3. Section 4 concludes the discussion of the proposed system.

2 System Description

The proposed two-stage long-term SNR estimation approach is shown in Fig. 1. It consists of a speech separation stage and a SNR estimation stage. The goal of speech separation is to separate the target speech from background interference. We view speech separation as our first stage, but the focus of this study is to use speech separation to assist in SNR estimation. Therefore, we investigate different front-end speech separation approaches, namely, IBM estimation, ideal ratio mask (IRM) estimation, complex ideal ratio masking (cIRM), and nonnegative matrix factorization (NMF) based speech separation. Each of these stages are described below.

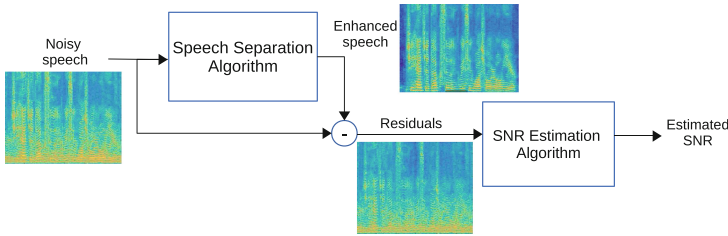


Fig. 1. The architecture of the proposed long-term SNR estimation system.

2.1 Speech Separation Stage

Recent approaches perform speech separation by estimating masking-based training targets [14, 15]. These approaches estimate T-F masks from the noisy speech signal, and use the estimated mask to separate speech from the noise: $\hat{S}(k, f) = M(k, f) * Y(k, f)$, where $\hat{S}(k, f)$ denotes the short-time Fourier transform (STFT) of the speech estimate, $M(k, f)$ denotes the estimated T-F mask, and $Y(k, f)$ is the STFT of the noisy speech. k and f index the time and frequency dimensions, respectively. The T-F domain speech estimate is then converted to a time-domain estimate, $\hat{s}(t)$, using overlap-add synthesis. We investigate three different DNN-based T-F mask estimation approaches, namely IBM [16], IRM [17] and cIRM [18] estimation, where definitions for these three mask are shown below:

$$\begin{aligned}
 \text{IBM}(k, f) &= \begin{cases} 1, & \text{if } |S(k, f)| > |N(k, f)|, \\ 0, & \text{otherwise} \end{cases} \\
 \text{IRM}(k, f) &= \left(\frac{|S(k, f)|^2}{|S(k, f)|^2 + |N(k, f)|^2} \right)^{0.5}, \\
 \text{cIRM}(k, f) &= \frac{S(k, f)}{Y(k, f)},
 \end{aligned} \tag{1}$$

$|S(k, f)|$ and $|N(k, f)|$ respectively denote the magnitude responses of the clean speech and noise. The cIRM involves complex division since $Y(k, f)$ and $S(k, f)$ are complex-valued numbers with real and imaginary components (e.g. $Y(k, f) = Y_r(k, f) + jY_i(k, f)$, $S(k, f) = S_r(k, f) + jS_i(k, f)$). The IBM is a binary matrix used to label T-F units of a signal as speech or noise dominant [16], and it has been shown to improve speech intelligibility, but not perceptual speech quality. An estimated IRM often outperforms an estimated IBM [15], since it gives soft values between 0 and 1. Intuitively, the IRM represents the percentage of energy that can be attributed to speech at each T-F unit. Unlike the IBM and IRM, the cIRM enhances the magnitude and phase response of speech, since it is complex-valued. Estimated cIRMs outperform IRM-based separation when evaluated with objective metrics and human evaluations [18]. Each T-F masks impact on estimating long-term SNR, however, is not known, so we elect to separately use each of them in our front-end speech separation module.

Separate DNNs are trained to estimate each of the above mentioned T-F masks and subsequently used to perform speech separation. The structures of the DNN match those described in [15, 18], where we omit details since our focus is on using speech separation to enhance long-term SNR estimation.

We alternatively use a NMF-based separation approach for our front-end speech separation stage. NMF is a model-based approach that uses trained speech and noise models (e.g. basis matrices) along with an activation matrix to separate speech from noise [19, 20]. The basis matrix represents the spectral features and the activation matrix linearly combines the spectral features to approximate a nonnegative signal. We first approximate a dictionary of clean speech signals, D , with the product of a trained basis matrix, W_{tr} , and a trained activation matrix, H_{tr} (e.g. $D \approx W_{tr}H_{tr}$). The basis and activation matrices are computed using a standard multiplicative update rule that minimizes the generalized Kullback-Leibler divergence between D and $W_{tr}H_{tr}$. To perform separation, the magnitude response of the speech estimate, $|\hat{S}(k, f)|$ is approximated as the product of W_{tr} and a new activation matrix, H_{new} , which is computed using the same multiplicative update rule and the fixed training basis matrix. Hence, $|\hat{S}| = W_{tr}H_{new}$. An estimate of the noise is computed and used along with the speech estimate to form a T-F mask. This mask is then applied to the noisy speech mixture to generate a speech estimate.

A noise residual is computed as $r(t) = y(t) - \hat{s}(t)$, where it is then provided as an input to the second stage of our approach.

2.2 Long-Term SNR Estimation Stage

We train a DNN to estimate the long-term SNR of the noisy speech signal from the noise residual. A depiction of this DNN is shown in Fig. 2. Complementary features [21] are extracted from the residuals and they are provided as inputs to the DNN. These features consist of amplitude modulation spectrogram (AMS), relative spectral transform perceptual linear prediction (RASTA-PLP), and mel-frequency cepstral coefficients (MFCC). We also add delta (Δ) features to capture the temporal dynamics of the residual. We use the same parameter

configuration for the complementary features as described in [18, 21], since they show success in modeling noisy speech. We tried to use log magnitude spectral features, Gammatone frequency (GF) features and Multi-resolution Cochleagram (MRCG) features separately as inputs, but they did not perform as well as the complementary set.

The training target is the true long-term SNR of the input noisy speech, which is calculated by the ratio of the energy of entire speech and the energy of corresponding noise, written as $\text{SNR}_{\text{global}} = 10 \log_{10}(E_{\text{speech}}/E_{\text{noise}})$. SNR estimation, however, occurs at the time frame level, so we label each time frame with this global SNR. The DNN estimates this long-term SNR in each of the 40 ms time frames of the signal. The final estimate is generated by averaging the estimated value in each time frame. The standard back-propagation algorithm with mean-square error cost function is used for training the DNN.

The DNN has three hidden layers where each has 512, 256, and 128 units, respectively. We experimented with different number of layers and units per layer, but empirical results indicate that this structure performs best. The rectified linear (ReLU) activation function is used for the hidden units, while a linear unit is used in the output layer. After DNN training, linear regression is used to learn a linear mapping between the DNN output and the true long-term SNR. This is often done to produce better predictions for long-term SNRs that are unseen during training [12, 22].

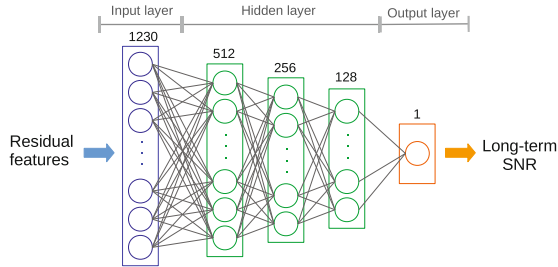


Fig. 2. Structure of the DNN that maps noise residuals to a SNR estimate.

3 Evaluation Results

3.1 Experimental Setup

All experiments are conducted with the TIMIT speech corpus [23] and NOISEX-92 noise database [24]. 600 TIMIT utterances are separately mixed with four noises: speech-shaped (SSN), cafeteria (Cafe), speech-babble (Babble) and factory floor (Factory) at 5 medium SNR levels (-6 , -3 , 0 , 3 , and 6 dB), resulting in a total of 12000 training mixtures. Random segments from the first half of each noise file is used in generating the training mixtures. These signals are used for training the

DNNs of two stages, and also for generating the speech and noise models of NMF at the first stage. A separate development set is used for model selection.

Two test sets are created for evaluating the generalization performance. The first test set mixes 200 different TIMIT utterances with the same matched noise signals and at the same SNR levels as defined above. Additional mismatched SNR signals are generated at unseen low SNRs (-15 , -12 , and -9 dB) and unseen high SNRs (9 , 12 , and 15 dB), using the same matched noise signals. This results in 8800 testing mixtures. Random segments from the second half of each matched noise signal is used in generating these testing mixtures. The second test set uses 200 different clean utterances that are mixed with six unmatched noise types: cockpit, destroyer engine (Engine), machine gun (Machine), pink, tank and white noise at 11 SNR levels ranging from -15 dB to 15 dB, in 3 dB increments, producing 13200 testing mixtures.

The STFTs in the speech separation stage are computed using a Hanning window length of 40 ms, a 640 point FFT and 50% overlap between adjacent frames. Each NMF basis matrix consists of 80 basis vectors.

The accuracy of long-term SNR estimation is measured with the mean absolute error (MAE) between the true SNR t_i and estimated SNR \hat{t}_i of the i -th mixture for all N testing mixtures [12].

$$\text{MAE}(t, \hat{t}) = \frac{1}{N} \sum_{i=1}^N |(t_i - \hat{t}_i)| \quad (2)$$

3.2 Results and Discussion

In the first stage, we separately employ and compare NMF, IBM, IRM and cIRM-based speech separation approaches, and investigate their influence on SNR estimation accuracy. In addition to using the residuals that result from the above separation approaches, we separately use the true noise signal as an input to the second DNN-stage of our approach. This assumes perfect separation and we regard it as an ideal case, since it provides upper bound performance capabilities.

Table 1 shows SNR estimation results in the matched noise case, but with seen and unseen SNRs. We find that in every case the system with cIRM separation gives the best estimation especially at low SNR conditions, and its performance is close to the ideal case. This occurs because cIRM estimation outperforms the other speech separation approaches, as indicated in [18]. This reveals that improving speech separation performance can clearly improve SNR estimation accuracy. Note that the average PESQ performance is 1.81 for noisy speech, 1.88 for NMF, 1.92 for IBM, 2.23 for IRM and 2.41 for cIRM separation. Although not trained in the system, the MAE performance at high SNRs achieves the lowest average error across all approaches. This occurs because separation performance in low SNR conditions is relatively not as good as in high SNR environments. Also notice that the performance in the unseen case is approximately the same as the seen training case on average, which indicates that the proposed approach generalizes well in unseen SNR environments.

Table 1. Avg. MAE for estimating seen and unseen SNR levels of matched noise types, when applying different separation approaches.

SNR level	Ideal	NMF	IBM	IRM	cIRM
Seen medium	1.78	4.38	4.98	3.83	1.85
Unseen high	1.42	2.77	2.33	1.96	1.64
Unseen low	1.34	5.93	9.15	4.85	2.01
All	1.56	4.36	5.39	3.60	1.86

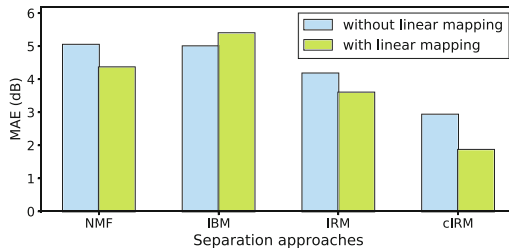
Table 2. Avg. MAE for SNR estimation under matched and unmatched noise conditions. The average is across all SNRs.

Noise type	NMF	IBM	IRM	cIRM
Matched	4.36	5.39	3.60	1.86
Unmatched	4.27	4.60	4.08	3.91
All	4.31	4.92	3.89	3.09

To further evaluate the generalization performance of our system, we test in matched and unmatched noise conditions. The average MAE of 11 SNR levels, ranging from -15 dB to 15 dB with a 3 dB step size, is reported for each noise type, see Table 2. Not surprisingly, cIRM estimation outperforms the other approaches across matched and unmatched noise conditions.

Our approach applies linear regression to the DNN output since this can expand the SNR prediction range, which is initially limited by the range of input SNRs that are used for training. Figure 3 shows MAE results when linear regression is and is not applied to the DNN output. Notice that the average MAE of NMF, IRM and cIRM reduce by 0.7 , 0.6 and 1.1 dB, respectively, when linear regression is applied, which shows that linear mapping improves performance.

We evaluate the importance of the speech separation stage by extracting features directly from the noisy speech and then by training the SNR-estimation DNN with the noisy speech features (e.g. no separation is performed). When this is done, SNR estimation is much worse, as it does not follow the trend of input

**Fig. 3.** Avg. MAE score when linear mapping is and is not applied.

SNRs as shown in Fig. 4 (left). Alternatively, we calculate SNR directly from the speech estimate and noise residual that are produced by the speech separation stage in order to determine how important the second stage is to long-term SNR estimation. Hence, the SNR estimation DNN is not used. These estimation results are severely worse than our proposed two-stage approach, see Fig. 4 (right). This occurs because the separation stage incorrectly places some speech energy in the estimated noise signal and noise energy in the estimated speech signal. The second SNR estimation DNN helps overcome this problem. Both experiments indicate that DNN-based speech separation followed by a SNR-estimation DNN is preferred.

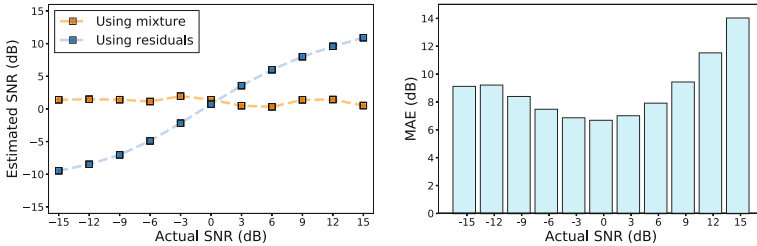


Fig. 4. Left: SNR estimation results with (e.g. residuals) and without (e.g. mixture) speech separation. Right: Avg. MAE when SNR is estimated directly from the speech and noise estimates from the speech separation stage.

Furthermore, we compare our system with four state-of-the-art long-term SNR estimation methods. The first algorithm is WADA [9], which has been proven to significantly outperform NIST [8]. The second method (e.g. Noise PSD) estimates long-term SNR by calculating the ratio of noisy speech power and the estimated noise PSD across all time frames and frequency bins [10]. Similarly, the third algorithm (e.g. Speech PSD) uses an MMSE estimator to estimate the PSD of clean speech. The energy ratio of speech PSD and noisy speech power is used to estimate SNR [11]. The last comparison approach is a CASA-based approach that uses an estimated IBM to identify the speech-dominant and noise-dominant T-F units in a unsupervised manner [12]. The estimated IBM is then used to approximate speech and noise energies for SNR calculation. Since the proposed system with cIRM separation shows advantages over other separation approaches, it is used in the comparison and is denoted as *P*-cIRM.

As shown in Table 3, *P*-cIRM achieves the lowest MAE under matched noise conditions, and it is better by about 0.7 dB compared to the CASA approach. Compared to noise PSD and speech PSD, it is better by 4 dB and 2.8 dB, respectively. *P*-cIRM works well in unmatched noise conditions, but it is slightly outperformed by the CASA-based approach. When evaluating by SNR, *P*-cIRM shows comparative advantages over WADA, noise PSD, speech PSD, and CASA. In low SNR levels, *P*-cIRM improves by 2.5 dB compared to CASA, which also has a SNR transformation process to reduce estimation errors in low SNR conditions. *P*-cIRM also outperforms CASA at high SNRs as well. Performance for

Table 3. Comparison of the proposed system with other SNR estimation methods. * indicates that the SNR was not seen during training.

Method	Mat. Noise	Unmat. Noise	High*	Medium	Low*
WADA	8.563	10.09	8.439	6.370	13.37
Noise PSD	5.866	7.274	5.911	2.581	11.28
Speech PSD	4.737	6.513	5.274	2.349	7.530
CASA	2.599	3.777	2.703	1.912	5.170
<i>P</i> -cIRM	1.864	3.913	2.359	2.131	2.596

the CASA-based approach depends on whether it can correctly label speech and noise regions, which does not always occur at low SNRs. WADA leads to poor estimation results, since its assumption on noisy speech and noise distributions are not satisfied. Similarly, noise PSD and speech PSD assume Gaussian distributions for the noise and speech. When the background noise is non-stationary or in very low SNR levels, both noise PSD and speech PSD make relative large estimation errors, and their results are not comparable to our best performing systems.

4 Conclusion

We propose a two-stage DNN-based approach for estimating long-term SNR. The first stage generates a noise residual, and the second stage uses the residual and a DNN to predict long-term SNR. The results show that our proposed approach accurately estimates long-term SNR residuals when compared to alternative options and existing unsupervised approaches, even when tested in seen and unseen testing environments.

The results further indicate that applying better separation algorithms will obtain lower mean absolute errors. Note that our system has two independent stages. Any state-of-the-art speech separation algorithm can be used in the first stage, and more sophisticated deep learning networks can also be used in the second stage to potentially produce more accurate estimation results.

References

1. May, T., Kowalewski, B., Fereczkowski, M., MacDonald, E.: Assessment of broadband SNR estimation for hearing aid applications. In: Proceedings of ICASSP, pp. 231–235 (2017)
2. Ris, C., Dupont, S.: Assessing local noise level estimation methods: application to noise robust ASR. *Speech Commun.* **34**, 141–158 (2001)
3. Cohen, I.: Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Trans. Speech Audio Process.* **13**, 870–881 (2005)
4. Tchorz, J., Kollmeier, B.: SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE Trans. Speech Audio Process.* **11**, 184–192 (2003)

5. Martin, R.: An efficient algorithm to estimate the instantaneous SNR of speech signals. In: Eurospeech, vol. 93, pp. 1093–1096 (1993)
6. Dat, T., Takeda, K., Itakura, F.: On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement. *Speech Commun.* **48**, 1515–1527 (2006)
7. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Audio Speech Lang. Process.* **32**, 1109–1121 (1984)
8. NIST speech signal to noise ratio measurements. <https://www.nist.gov/information-technology-laboratory/>
9. Kim, C., Stern, R.: Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In: Proceedings of Interspeech (2008)
10. Hendriks, R., Heusdens, R., Jensen, J.: MMSE based noise PSD tracking with low complexity. In: Proceedings of ICASSP, pp. 4266–4269 (2010)
11. Erkelens, J., Hendriks, R., Heusdens, R., Jensen, J.: Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Trans. Audio Speech Lang. Process.* **15**, 1741–1752 (2007)
12. Narayanan, A., Wang, D.L.: A CASA-based system for long-term SNR estimation. *IEEE Trans. Audio Speech Lang. Process.* **20**, 2518–2527 (2012)
13. Hinton, G., Deng, L., Yu, D., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Sig. Process. Mag.* **29**, 82–97 (2012)
14. Han, K., Wang, D.L.: A classification based approach to speech segregation. *J. Acoust. Soc. Amer.* **132**, 3475–3483 (2012)
15. Wang, Y., Narayanan, A., Wang, D.L.: On training targets for supervised speech separation. *IEEE Trans. Audio Speech Lang. Process.* **22**, 1849–1858 (2014)
16. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In: *Speech Separation by Humans and Machines*, pp. 181–197 (2005)
17. Srinivasan, S., Roman, N., Wang, D.L.: Binary and ratio time-frequency masks for robust speech recognition. *Speech Commun.* **48**, 1486–1501 (2006)
18. Williamson, D.S., Wang, Y., Wang, D.L.: Complex ratio masking for monaural speech separation. *IEEE Trans. Audio Speech Lang. Process.* **24**, 483–492 (2016)
19. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio Speech Lang. Process.* **15**, 1066–1074 (2007)
20. Gemmeke, J., Virtanen, T., Hurmalainen, A.: Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **19**, 2067–2080 (2011)
21. Wang, Y., Han, K., Wang, D.L.: Exploring monaural features for classification-based speech segregation. *IEEE Trans. Audio Speech Lang. Process.* **21**, 270–279 (2013)
22. Papadopoulos, P., Tsiartas, A., Narayanan, S.: Long-term SNR estimation of speech signals in known and unknown channel conditions. *IEEE Trans. Audio Speech Lang. Process.* **24**, 2495–2506 (2016)
23. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D.: DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA technical report 93 (1993)
24. Varga, A., Steeneken, H.J.M.: Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **12**(3), 247–251 (1993)



Improving Reverberant Speech Separation with Binaural Cues Using Temporal Context and Convolutional Neural Networks

Alfredo Zermini^(✉), Qiuqiang Kong, Yong Xu, Mark D. Plumbley, and Wenwu Wang

Centre for Vision, Speech and Signal Processing, University of Surrey,
Guildford, Surrey GU2 7XH, UK

{a.zermini,q.kong,yong.xu,m.plumbley,w.wang}@surrey.ac.uk

Abstract. Given binaural features as input, such as interaural level difference and interaural phase difference, Deep Neural Networks (DNNs) have been recently used to localize sound sources in a mixture of speech signals and/or noise, and to create time-frequency masks for the estimation of the sound sources in reverberant rooms. Here, we explore a more advanced system, where feed-forward DNNs are replaced by Convolutional Neural Networks (CNNs). In addition, the adjacent frames of each time frame (occurring before and after this frame) are used to exploit contextual information, thus improving the localization and separation for each source. The quality of the separation results is evaluated in terms of Signal to Distortion Ratio (SDR).

Keywords: Convolutional Neural Networks
Binaural cues · Reverberant rooms · Speech separation
Contextual information

1 Introduction

Sound source separation has been studied for a long time, with implementing methodologies such as independent component analysis [1], computational auditory scene analysis [2], and non-negative matrix factorization [3]. More recently, Deep Neural Networks (DNNs) [4] and Convolutional Neural Networks (CNNs) [5] have shown state-of-the-art performance in source separation [6–8]. This paper studies the problem of separating two speakers in rooms with different reverberation, which is a common scenario in real life. A target speech signal, corresponding to the main speaker, is disturbed by an interferer speaker, located in variable positions. This problem has already been studied in [8], where the target speech is separated by generating a time-frequency (T-F) mask, which is obtained by training a DNN by using binaural spatial cues such as Mixing Vectors (MV), Interaural Level Difference (ILD) and Interaural Phase Difference (IPD). The methods have

limitations for more reverberant rooms, in particular when the training room used is different from the room used in the testing set. In recent years, different types of approaches have been developed to overcome these issues. In [9], the introduction of spectral features such as the Log-Power Spectra (LPS) along the spatial cues, proved to be useful where one of the two speakers is replaced with noise. The last layer of the DNN is a softmax classifier, which estimates the Directions Of Arrival (DOAs) of the sources. This information is used to build a soft-mask for the target source. In [10, 11], the soft-mask is directly estimated through a regression approach by training a single DNN.

Other neural network structures, such as CNNs, are neural networks designed to process data in the form of multiple arrays (such as images with three colours channels) and contain convolutional and pooling layers [5]. CNNs have been used to estimate the DOA for speech separation in [12] and trained using synthesized noise signals, but recorded with a four-microphones array.

In this paper, we present a system that is able to perform source localization and source separation. Here, the relatively simple system of DNNs already introduced by Yu et al. in [8] is upgraded to a deeper system based on CNNs, in order to exploit the increased computational power available in modern GPUs, aiming for a better separation quality. In addition, contextual frame expanding [10] is introduced, which uses the information from neighbouring time frames before and after a given time frame. This gives a better estimation of each T-F point of the soft-mask because the DOA is estimated by checking if a speaker is still active in the time frames around the one that has been estimated.

The remainder of the paper is organized as follows. Section 2 introduces the proposed method, including the overall CNN architecture employed, the low-level feature extraction for the CNN input, and the output in the training stage and the system implementation. Section 3 describes how the soft-masks are generated starting from the output of each CNN. Experimental results are presented in Sect. 4, where evaluations are performed and analyses are given, followed by conclusions of our findings and insights for future work in Sect. 5.

2 Proposed Method

2.1 System Overview

A system of CNNs, shown in Fig. 1, is used to localize the direction of one or more speakers in a speech mixture. This system integrates the information from several CNNs, each one trained with the information from a narrow frequency band. The outputs are then merged together to get soft-masks, which are used to retrieve the speech source from the audio mixture, as shown in Fig. 1. The Short-Time Fourier Transform (STFT) on the left and right channels is calculated. The results are two spectrograms $X_L(m, f)$ and $X_R(m, f)$, where $m = 1, \dots, M$ and $f = 1, \dots, F$ are the time frame and frequency bin indices respectively. For each T-F point, low-level features (i.e. ILD and IPD) are calculated and used to train the CNNs. These features will be introduced in more detail in Sect. 2.2.

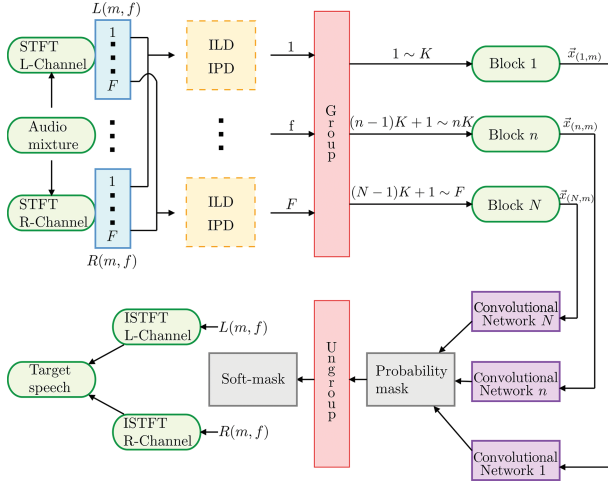


Fig. 1. Diagram of the system architecture using CNNs.

The low-level features are arranged into N blocks, each one containing the information from a small group of frequency bins and the output is a probability mask containing the information from just a narrow frequency band. Each of the $N = 128$ blocks, labelled n , includes $K = 8$ frequency bins in the range $((n-1)K+1, \dots, nK)$, small enough to reduce losses in resolution in the resulting probability output mask, where $K = F/N$ and N is the number of CNNs. Each block is used as the input of a different CNN for the training stage, each output is a softmax classifier, which gives the probability for a sound source to come from one of the possible J DOAs, so it contains J values between 0 and 1. As explained in Sect. 3, a series of soft-masks can be generated by stacking all the CNNs outputs, one for each test set j and by ungrouping each block into 8 frequency bins. The binaural soft-masks are multiplied element-wise by the mixture spectrograms and, after applying the inverse STFT (ISTFT), the target source can be recovered.

2.2 Low-Level Features

The binaural features used are IPD and ILD, have been already introduced for sound localization in [9,10]. They are used to derive high-level features which are easy to classify. IPD and ILD are the phase and the amplitude difference between the left and the right channels. By putting them in one vector, one can obtain, for each T-F unit:

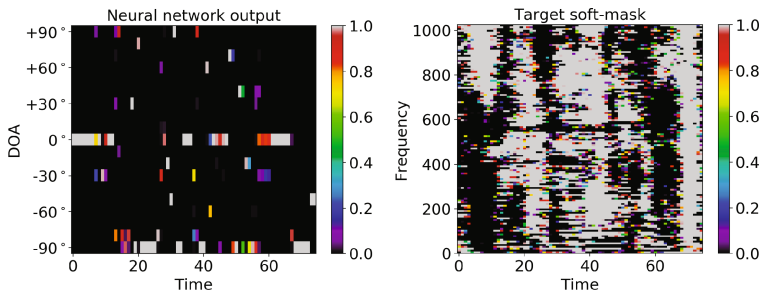
$$\mathbf{x}(m, f) = [ILD(m, f), IPD(m, f)]^T.$$

Each $\tilde{u}(m, f)$ is grouped into N blocks along the frequency bins, which represents the input vector of each CNN:

$$\mathbf{x}_{(n,m)} = [\mathbf{x}^T(m, (n - 1)K + 1), \dots, \mathbf{x}^T(m, nK)]^T.$$

3 Soft-Masks Construction

An output mask is created by exploiting the contextual time frame information from the neighbouring frames. A number of time frames τ is selected before and after a given central time frame $\tau_0 \in 1 \dots M$, where M is the number of time frames in the spectrogram. Each group of frames is thus composed of $C = 2 \times \tau + 1$ time frames. This operation is looped for all the $\tau_0 \in 1 \dots M$. All the M groups are concatenated and each frequency band is fed into a different CNN for training. In the output, the central time frames τ_0 are selected and concatenated to generate a probability mask with the correct size M . The probability mask for each CNN looks like the one shown in Fig. 2(a), representing the DOA probability as a function of the time frame. By averaging over all the time frames and the frequency bands, the highest value indicates the most probable DOA. The next step is selecting the entire row corresponding to the highest DOA probability. This row represents the target soft-mask for that specific frequency band. As last step, all the probability masks are stacked in order to build the T-F soft-mask for the target speech, shown in Fig. 2(b).



(a) Example of probability mask for one of the 128 CNNs.

(b) Target soft-mask.

Fig. 2. Probability mask and soft-mask.

4 Experiments

4.1 Experimental Setup

Binaural audio recordings are created by convolving a speech recording with Binaural Room Impulse Responses (BRIRs), captured in real echoic rooms [13].

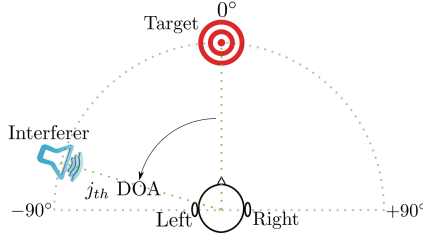


Fig. 3. The experimental setup.

The BRIRs dataset was recorded around a half-circular grid, ranging from -90° to 90° with steps of 10° , for a total of $J = 19$ DOAs. A dummy head located at the center of a given reverberant room has been used, with left and right microphones, as shown in Fig. 3. The training set has been produced by using speech samples from the TIMIT dataset, containing recordings of sentences from different male and female speakers, sampled at $f_s = 16$ kHz, high enough for our task. The training samples are randomly selected single reverberant speech recordings, 8 males and 8 female speakers, recorded at 19 different DOAs, each one being ≈ 2 s long. For the testing set, the same experimental setup and parameters as in [8] have been used. Two different speakers, named the target and the interferer, have been randomly selected from the TIMIT database for the two genders and mixed, for a total of 15 reverberant speech mixtures for each DOA, ≈ 2.3 s long each. The experimental setup is shown in Fig. 3. Both the target and interferer are located 1.5 m away from the dummy head, and the three objects have the same height. The amount of reverberation depends on the parameters of the room selected, listed in Table 1, where room ‘A’ is less reverberant and ‘D’ more reverberant. The STFT is performed where the Hann window is set to 2048 (128 ms) samples with 75% overlap between the neighbouring windows, so the resulting training and testing samples are 75 time frames long each. The parameters for each CNN in Fig. 4, are found empirically and gave the best performance in our experiments. The first part of the CNN is used for features learning. There is a convolutional input layer with 32 feature maps, kernel size (3, 3), batch normalization, followed by a max pooling layer with pooling size (2, 2) (or (1, 1) for $\tau = 0$, to keep the right dimensions) and a 10% dropout layer. The second part is for classification. We used a 1024 neurons dense layer, with batch normalization and 10% dropout. The output is another dense layer with 19 neurons. The rectified linear activation function has been used for both the convolutional and the hidden dense layer, while the softmax is used in the output. The number of epochs is set between 60 and 200, the batch size is set to 200 and the cost-function is the categorical cross-entropy.

4.2 Signal to Distortion Ratios (SDRs) Evaluation

Figures 5 and 6 show the Signal to Distortion Ratios (SDRs) evaluation for the target fixed at 0° or -90° , for variable positions of the interferer speaker.

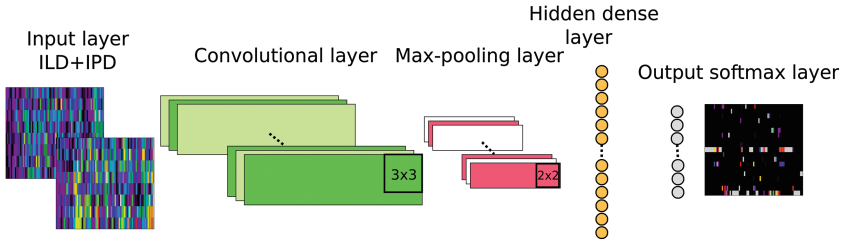


Fig. 4. Structure of a CNN.

Table 1. Rooms acoustic properties.

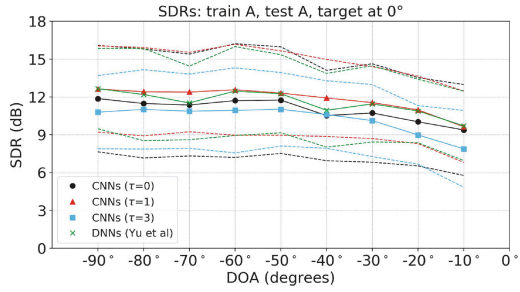
Room	Type	ITDG (ms)	DRR (dB)	T60 (s)
A	Medium office	8.73	6.09	0.32
D	Large seminar room	21.6	6.12	0.89

The dots indicate the average SDR over the test set at each DOA and are connected by continuous lines, dashed lines are the correspondent standard deviation. The cases where the interferer is in the range $[0^\circ, +90^\circ]$ will be omitted for a better visualization of the plots. When target and interferer are aligned (i.e. from the same direction), it is virtually impossible to separate the two speakers by using spatial features only, so they have been excluded from the plots as well.

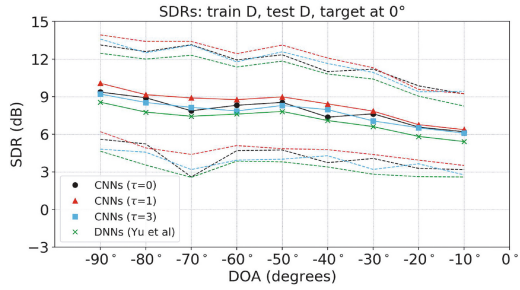
In Figs. 5 and 6, the system named CNNs $\tau = 0$ has been trained and tested without using any contextual information from the neighbouring time frames, while CNNs with $\tau = 1$ and $\tau = 3$ include τ contextual frames before and after each time frame. The last system, named DNNs, is a three dense layers DNNs system, similar to the one tested by Yu et al. in [8], here included as a baseline. The average improvement over all the DOAs compared to the baseline system, ΔSDR , is shown in Table 2.

Table 2. Average improvement on the SDRs for the CNNs at different τ compared to the DNNs baseline.

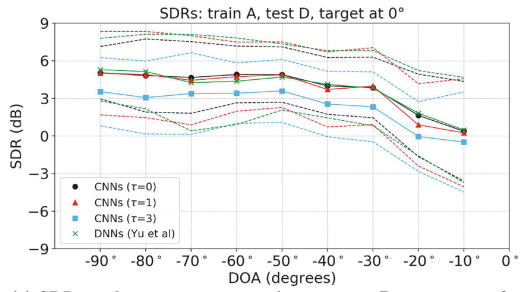
Target	Train	Test	$\Delta SDR(\tau = 0)$ (dB)	$\Delta SDR(\tau = 1)$ (dB)	$\Delta SDR(\tau = 3)$ (dB)
0°	A	A	-0.58	+0.25	-1.32
0°	D	D	+0.74	+1.24	+0.61
0°	A	D	+0.03	-0.13	-1.41
0°	D	A	+1.41	+0.79	-0.92
-90°	A	A	-2.03	+0.80	+0.23
-90°	D	D	-2.12	+1.63	+1.89
-90°	A	D	-0.09	-1.61	-1.82
-90°	D	A	+0.10	+0.60	+0.84



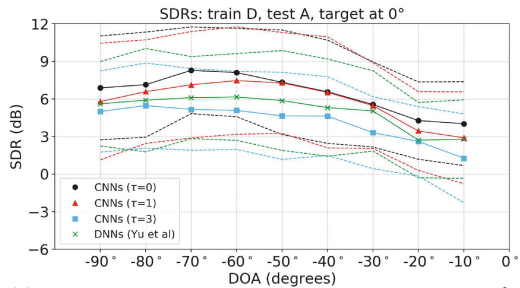
(a) SDRs evaluation: train room A, test room A, target at 0°.



(b) SDRs evaluation: train room D, test room D, target at 0°.

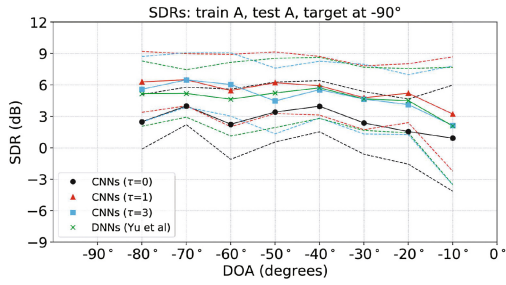


(c) SDRs evaluation: train room A, test room D, target at 0°.

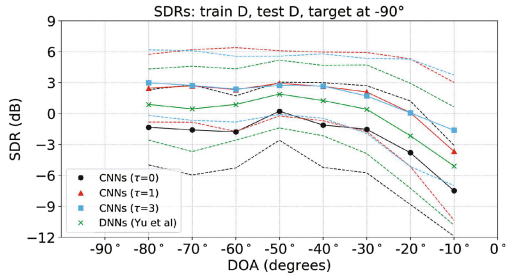


(d) SDRs evaluation: train room D, test room A, target at 0°.

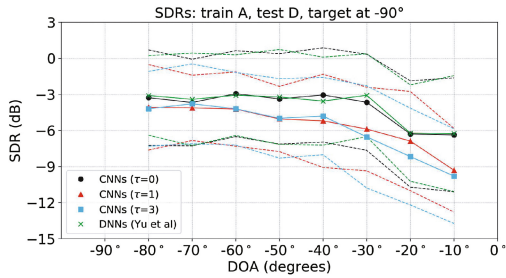
Fig. 5. SDR plotted against the DOA, target at 0°.



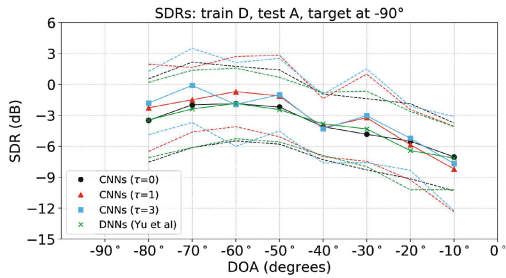
(a) SDRs evaluation: train room A, test room A, target at -90° .



(b) SDRs evaluation: train room D, test room D, target at -90° .



(c) SDRs evaluation: train room A, test room D, target at -90° .



(d) SDRs evaluation: train room D, test room A, target at -90° .

Fig. 6. SDR plotted against the DOA, target at -90° .

Figures 5(a) and (b) show the cases in which the room used for training and testing is the same. For room ‘A’, the CNNs with $\tau = 1$ system performs the best among the four systems tested, with $\Delta SDR \approx 0.25$ dB. The SDRs are in the range $\approx [10,13]$ dB in Fig. 5(a) for $\tau = 1$, giving a very good separation quality on the listening tests. The SDRs decrease while the interferer approaches 0° , because the binaural features contain less information when the differences in level and phase between left and right microphones are small. For room ‘D’, the CNNs with $\tau = 1$ give optimal results, as shown in Fig. 5(b), with $\Delta SDR \approx 1.23$ dB. The SDRs are in $\approx [6,10]$ dB, a good separation quality for a room with such a high reverberation level. The standard deviation, which is on average ≈ 3 dB, highly depends on the gender selection of the mixtures. In fact, where the speech recordings are from speakers of different genders, the frequency overlap is less compared to the case of same gender speakers, which means they are easier for the CNNs to localize.

Figures 5(c) and (d) show the cases where the training and testing room do not match. In this case, all the four systems perform slightly worse than the case in which training and testing rooms are the same, as they need to adapt to a type of reverberation that was not included in the training data. Figure 5(c) shows that DNNs and the CNNs with $\tau = 0$ and $\tau = 1$ have similar performances. Instead, in Fig. 5(d), the $\tau = 0$ CNNs system has the best separation quality, with $\Delta SDR \approx 1.41$ dB. Both in Fig. 5(c) and (d), the CNNs with $\tau = 3$ give by far the worst performance.

In all the Fig. 6 the target is fixed at -90° . In Figs. 6(a) and (b), training and testing rooms are the same. In Fig. 6(a) again, the case with $\tau = 1$ shows the best performance, with $\Delta SDR \approx 0.71$ dB and SDRs in $\approx [3,6]$ dB. In Fig. 6(b), unlike Fig. 6(a), the case $\tau = 3$ performs slightly better than $\tau = 1$, with $\Delta SDR \approx 1.68$ dB and SDRs in $\approx [0,3]$ dB. In both cases, $\tau = 0$ gives by far the worst separation results, suggesting that the contextual information improves the system in the localization task, especially in challenging scenarios when the target is located at wide angles. In the cases of room mismatch, plotted in Figs. 6(c) and (d), all the four systems have difficulty in retrieving the target, with SDRs on average below 0 dB.

5 Conclusions and Future Work

We presented a system of CNNs trained with binaural features and contextual information from the neighbouring time frames, where we used the outputs to build T-F masks. We applied these masks to speech mixtures to retrieve a target speaker. A system with a three dense layers DNNs had already been successfully tested for the same task in [8], showing some limitations, especially when the reverberation time of the testing room is long. As can be seen in Table 2, the systems of DNNs and CNNs with no contextual information, can be considered complementary, the separation quality depending on the training and testing rooms parameters. In general, when some contextual information is introduced, the CNNs outperform the DNNs baseline. In particular, when a small τ is chosen,

optimal results can be obtained, as summarized in Table 2. A possible explanation could be that introducing a large amount of contextual frames might include frames belonging to the interferer speaker, resulting in degradation in separation performance. Other works, such as [11], where a DNN is used for speech enhancement, suggest the use of a larger amount of contextual information, but they show how this is strictly related to the amount of training data, the neural network used and the task at hand. We have also tested the CNNs in more extreme conditions. In particular, when the target is fixed at -90° , its contribution arriving at the far-side ear is attenuated as compared to that of the near-side ear, which makes the separation task more challenging. Moreover, testing the networks in mismatched conditions, where the CNNs have to adapt to a new type of reverberation, in addition to the target located at wider angles, is a very challenging scenario, as shown in Figs. 6(c) and (d). Listening tests indicate that the target source is not separated, suggesting that none of the four systems tested has been effective.

As a future work, we believe that introducing the information from a regression model, along with the classification model presented in this paper, could further improve the separation performance, especially in rooms with longer reverberations and when the target is placed at wider angles. Moreover, we want to extend the system for the underdetermined case, with more interferer speakers.

Acknowledgements. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under grant agreement no 607290 SpaRTaN.

References

1. Comon, P., Jutten, C. (eds.): Handbook of Blind Source Separation: Independent Component Analysis and Applications. Elsevier, Amsterdam, Boston (2010)
2. Wang, D., Brown, G.J.: Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, Hoboken (2006)
3. Lee, D.D., Sebastian, S.H.: Algorithms for non-negative matrix factorization. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.) Advances in Neural Information Processing Systems 13, pp. 556–562. MIT Press, Cambridge (2001)
4. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
5. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436 (2015)
6. Jiang, Y., Wang, D., Liu, R., Feng, Z.: Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(12), 2112–2121 (2014)
7. Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(12), 2136–2147 (2015)
8. Yu, Y., Wang, W., Han, P.: Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. *EURASIP J. Audio Speech Music Process.* **2016**(1), 7 (2016)

9. Zermini, A., Liu, Q., Xu, Y., Plumbley, M.D., Betts, D., Wang, W.: Binaural and log-power spectra features with deep neural networks for speech-noise separation. In: IEEE 19th International Workshop on Multimedia Signal Processing, MMSP 2017, pp. 1–6. IEEE, October 2017
10. Xu, Y., Du, J., Dai, L., Lee, C.: A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(1), 7–19 (2015)
11. Xu, Y., Du, J., Dai, L.R., Lee, C.H.: An experimental study on speech enhancement based on deep neural networks. *IEEE Sign. Process. Lett.* **21**(1), 65–68 (2014)
12. Chakrabarty, S., Habets, E.A.P.: Multi-speaker localization using convolutional neural network trained with noise. In: 31st Conference on Neural Information Processing Systems (NIPS 2017) (2017)
13. Hummersone, C.: A psychoacoustic engineering approach to machine sound source separation in reverberant environments (2011). <https://github.com/IoSR-Surrey/RealRoomBRIRs/>



Generating Talking Face Landmarks from Speech

Sefik Emre Eskimez^(✉), Ross K. Maddox, Chenliang Xu, and Zhiyao Duan

University of Rochester, 500 Joseph C. Wilson Blvd., Rochester, NY 14627, USA
eeskimez@ur.rochester.edu,
{ross.maddox, chenliang.xu, zhiyao.duan}@rochester.edu

Abstract. The presence of a corresponding talking face has been shown to significantly improve speech intelligibility in noisy conditions and for hearing impaired population. In this paper, we present a system that can generate landmark points of a talking face from an acoustic speech in real time. The system uses a long short-term memory (LSTM) network and is trained on frontal videos of 27 different speakers with automatically extracted face landmarks. After training, it can produce talking face landmarks from the acoustic speech of unseen speakers and utterances. The training phase contains three key steps. We first transform landmarks of the first video frame to pin the two eye points into two predefined locations and apply the same transformation on all of the following video frames. We then remove the identity information by transforming the landmarks into a mean face shape across the entire training dataset. Finally, we train an LSTM network that takes the first- and second-order temporal differences of the log-mel spectrogram as input to predict face landmarks in each frame. We evaluate our system using the mean-squared error (MSE) loss of landmarks of lips between predicted and ground-truth landmarks as well as their first- and second-order temporal differences. We further evaluate our system by conducting subjective tests, where the subjects try to distinguish the real and fake videos of talking face landmarks. Both tests show promising results.

Keywords: Visual generation · Face landmarks
Audio-visual models · LSTM

1 Introduction

Speech is a natural way of communication, and understanding speech is essential in daily life. The auditory system, however, is not the only sensory system involved in understanding speech. The visual cues from a talker's face and articulators (lips, teeth, tongue) are also important for speech comprehension. Trained professionals are able to understand what is being said by purely looking

Z. Duan—This work is supported by the University of Rochester Pilot Award Program in AR/VR and the National Science Foundation grant No. 1741471.

at lip movements (lip reading) [9]. For ordinary people and the hearing impaired population, the presence of visual signals of speech has been shown to significantly improve speech comprehension, even if the visual signals are synthetic [14]. The benefits of adding the visual speech signals are more pronounced when the acoustic signal is degraded, due to background noise, communication channel distortion, and reverberation.

In many scenarios such as telephony, however, speech communication is still acoustical. The absence of the visual modality can be due to the lack of cameras, the limited bandwidth of communication channels, or privacy concerns. One way to improve speech comprehension in these scenarios is to synthesize a talking face from the acoustic speech in real time at the receiver’s side. A key challenge of this approach is to make sure that the generated visual signals, especially the lip movements, well coordinate with the acoustic signals, as otherwise more confusions will be introduced.

In this paper, we propose to use a long short-term memory (LSTM) network to generate landmarks of a talking face from acoustic speech. This network is trained on frontal videos of 27 different speakers of the Grid audio-visual corpus [6], with the face landmarks extracted using the Dlib toolkit [13]. The network takes the first- and second-order temporal differences of the log-mel spectra as the input, and outputs the x and y coordinates of 68 landmark points. To help the network capture the audio-visual coordination instead of the variation of face shapes across different people, we transform all training landmarks to those of a mean face across all talkers in the training set. After training, the network is able to generate face landmarks from an unseen utterance of an unseen talker. Objective evaluations of the generation quality are conducted on the LDC Audiovisual Database of Spoken American English dataset [18], which will be referred as the LDC dataset in the remaining of the paper. Subjective evaluation is also conducted to ask evaluators to distinguish speech videos with ground-truth and generated landmarks. Both the objective and subjective evaluations achieve promising results. The code and pre-trained talking face models are released to the community¹.

The remaining of the paper is structured as follows: Sect. 2 describes the related work. Section 3 describes the data and pre-processing steps. The architecture of the network is described in Sect. 4. Objective and Subjective evaluations are presented in Sects. 4.1 and 4.2. Finally, Sect. 5 concludes the paper.

2 Related Work

Generating a talking head automatically has been a great interest in the research community. Some researchers focused on text-driven generation [3, 10, 22, 23]. These methods map phonemes to talking face images. Compared to text, voice signals are surface-level signals that are more difficult to parse. Besides, voices of the same text show large variations across speakers, accents, emotions, and the recording environments. On the other hand, speech signals provide richer

¹ <http://www.ece.rochester.edu/projects/air/projects/talkingface.html>.

cues for generating natural talking faces. For text, any plausible face image sequence is sufficient to establish natural communication. For speech, it must be a plausible sequence that matches the speech audio. Therefore, text-driven generation and speech-driven generation are different problems and may require different approaches.

There exist a few approaches to speech-driven talking face generation. Early work in this field mostly used Hidden Markov Models (HMM) to model the correspondence between speech and facial movements [2, 4, 7, 8, 20, 24, 25]. One of the notable early work, Voice Puppetry [2], proposed an HMM-based talking face generation that is driven by only speech signal. In another work, Cosker et al. [7, 8] proposed a hierarchical model that animates sub-areas of the face independently from speech and merges them into a full talking face video. Xie et al. [24] proposed coupled HMMs (cHMMs) to model audio-visual asynchrony. Choi et al. [4] and Terissi et. al [20] used HMM inversion (HMMI) to estimate the visual parameters from speech. Zhang et al. [25] used a DNN to map speech features into HMM states, which further maps to generated faces.

In recent years, a few DNN-based approaches have also been proposed. Suwajanakorn et al. [19] designed an LSTM network to generate photo-realistic talking face videos of a target identity directly from speech. Their system requires several hours of face videos of the specific target identity, which greatly limits its application in many practical scenarios. Chung et al. [5] proposed a convolutional neural network (CNN) system to generate a photo-realistic talking face video from speech and a single face image of the target identity. Compared to [19], the reduction from several hours of face videos to a single face image for learning the target identity is a great advance.

While end-to-end speech-to-face-video generation is very useful in many scenarios, the main limitation of this approach is the lack of freedom for further manipulation of the generated face video. For example, within a generated video, one may want to vary the gestures, facial expressions, and lighting conditions, all of which can be relatively independent of the content of the speech. These end-to-end systems cannot accommodate such manipulations unless they can take these factors as additional inputs. However, that would significantly increase the amount and diversity of data required for training the systems.

A modular design that separates the generation of key parameters and the fine details of generated face images is more flexible for such manipulations. Ideally, the key parameters should just respond to the speech content, while the fine details should incorporate all other non-speech-content related factors. Pham et al. [16] adopted a modular design: the system first maps speech features to 3D deformable shape and rotation parameters using an LSTM network, and then generates a 3D animated face in real-time from the predicted parameters. In [17], they further improved this approach by replacing speech features with raw waveforms as the input and replacing the LSTM network with a convolutional architecture. However, compared to face landmarks used in our proposed approach, these shape and rotation parameters are less intuitive, and the mapping from these parameters to a certain gesture or facial expression is less clear.

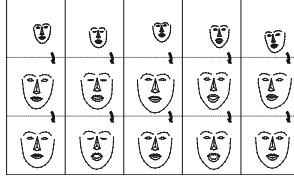


Fig. 1. Examples of extracted face landmarks from the training talking face videos. Certain landmarks are connected to make the shape of the face easier to recognize. The first row shows unprocessed landmarks of five unique talkers. The second row shows their landmarks after outer-eye-corner alignment. The third row shows their landmarks after alignment and the removal of identity information.

In addition, the landmarks generated by our system are for a normalized mean face instead of a certain target identity. This also helps remove factors that are not directly related to the voice.

3 Proposed Method

In this section, we describe our method to generate talking face landmarks. First, we extract face landmarks and align them across different speakers and transform their shapes into the mean shape to remove the identity information. We extract the first and second order temporal difference of the log-mel spectrogram and use them as the input to our system. Finally, we train an LSTM network to generate the face landmarks from the speech features.

3.1 Training Data and Feature Extraction

We employ the audio-visual GRID dataset [6] to train our system. There are in total 16 female and 18 male native English speakers, each of which has 1000 utterances that are 3 s long. The sentences are structured to contain a command, a color, a preposition, a letter, a digit, and an adverb, for example, “*set blue at C5 please*”.

The videos are provided in two resolutions, low (360×288) and high (720×576). In this work, we use the high-resolution videos. The videos use a frame rate of 25 frames per second (FPS), resulting in 75 frames for each video. The speech audio signal is extracted from the video with a sampling rate of 44.1 kHz.

We extract 68 face landmark points (x and y coordinates) using the DLIB library [13] from each frame for each video in the dataset. Examples are shown in the first row of Fig. 1. We calculate 64 bin log-mel spectra of the speech signal covering the entire frequency range using a 40 ms hanning window without any overlap to match the video frame rate. We then calculate the first- and second-order temporal differences of the log-mel spectra and use them as the input (128-d feature sequence) to our network. We experimented using log-mel

spectrogram with and without its first- and second-order derivatives as input to our network. The generated mouth for many speech utterances in these two setups, however, were almost always open even in silent segments, and the lip movements were less prominent than the current system. The first- and second-order temporal differences of the log-mel spectrogram may show less variations on the same syllable uttered by different speakers, and the mismatch problem is less pronounced.

3.2 Face Landmark Alignment

Since the talking face may appear in different regions with different sizes in different videos, we need to align them to reduce the complexity of training data. To do so, we follow the procedure described in [15] to simply pin the two outer corners of the eyes in the first frame of each video to two fixed locations, (180, 200) and (420, 200) in the image coordinate system, through an 6 DOF affine transformation. We then transform all of the landmarks in all video frames with the same transformation. Note that we do not align each video frame using their own affine transformation separately because we find that the eye-corner-based alignment is sensitive to eye blinks, which often results in zoom in/out effects of the transformed face shape. Also note that our approach assumes that the head does not move significantly within a video, as otherwise, the same affine transformation would not be able to align faces in different frames. The second row of Fig. 1 shows several examples of the aligned face landmarks.

3.3 Removing Identity Information from Landmarks

After alignment, faces of different speakers are of a similar size and general location; however, their shapes are still different as well as their mouth locations. This identity-related variation may pose challenges to the network for capturing the relation between speech and lip movement, especially when the amount and diversity of training data are small. Therefore, we propose to remove the identity information from the landmarks before training the network.

To do so, we apply the following steps. First, we calculate the mean face shape by averaging all aligned landmark locations across the entire training set. Second, for each face landmark sequence, we calculate the affine transform between the mean shape and the first frame of the sequence. Third, we calculate the difference between the current frame and the first frame and multiply with the scaling coefficients obtained from the second step with the result obtained in the third step. Finally, we add the mean shape to results obtained in fourth step to obtain the face landmark sequence that has no identity. The third row of Fig. 1 shows several examples of landmarks with the identity removed.

3.4 LSTM Network

Our proposed network, as shown in Fig. 2, uses four long short-term memory (LSTM) [12] layers with a sigmoid activation function. At each time step, the

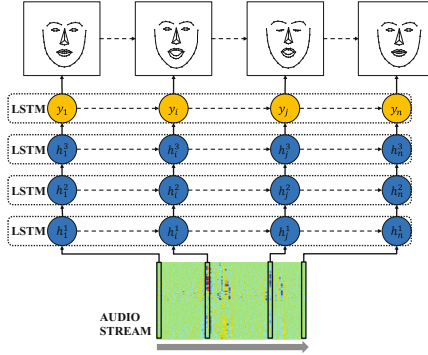


Fig. 2. The LSTM network architecture for generating landmarks of a talking face from the first and second order temporal differences of the log-mel spectrogram. h_t^l are the hidden layers, where t is the time step and l is the hidden layer index. y_t are the output face landmarks for the time step t .

input to the network is the first and second order temporal differences of the log-mel spectra of the current and the previous N frames. This provides short-term contextual information. The output is the predicted the x and y coordinates of face landmarks of the current frame (if no delay is added) or a previous frame (if a delay is added as described below). The reason for adding delay is because lips often move before the sound is produced. With a little delay, the network is able to “hear into the future” and can better prepare for those lip movements. The generated lip movements tend to be smoother. The amount of delay we introduce is between 1 (40 ms) and 5 frames (200 ms). This turns out to be enough for good generation results and is still tolerable in real-time speech communication.

During training, we use dropout between each layer and between recurrent connections, with a rate of 0.2. We use Adam optimizer to train our network. The training sequences are all 75 frames long. We set the batch size to 128 sequences and the learning rate to 0.001. Our network minimizes the following mean squared error (MSE) objective function J_{MSE} ,

$$J_{MSE} = \frac{1}{N} \sum_t \|\mathbf{s}_t - \hat{\mathbf{s}}_t\|^2, \quad (1)$$

where \mathbf{s} and $\hat{\mathbf{s}}$ are the x and y coordinates of ground-truth (GT) and predicted (PD) face landmarks sequences, respectively. N is the number of samples.

Finally, the predicted landmarks are further processed in order to fix the eye corner points to fixed points as described in Sect. 3.2, which produces more stable talking face landmarks.

Due to causality constraints, the bidirectional LSTM network is not considered in our experiments. We have also experimented with fully connected architecture instead of LSTM. However, the resulting face landmarks often show sudden jumps between frames, which looks unnatural. This is due to not having temporal connections in the architecture.

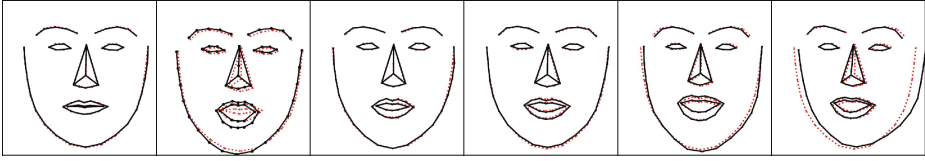


Fig. 3. Pair-wise comparison between ground-truth landmarks (black solid lines) and generated landmarks (red dotted lines) on unseen talkers and sentences. The second image shows a failure case for “oh” sound. (Color figure online)

Table 1. Objective evaluation results for different system configurations. The models are named according to the amount of delay and contextual information. For example, “D40-C5” describes a model trained with 40 ms delay and 5 frames of context. The lower value means better results, where the ideal result is zero.

	RMSE	RMSE First Diff	RMSE Second Diff
D0-C3	0.0954	0.0045	0.0073
D0-C5	0.0945	0.0042	0.0071
D40-C3	0.0932	0.0039	0.0068
D40-C5	0.0921	0.0032	0.0065
D80-C3	0.0946	0.0044	0.0072
D80-C5	0.0944	0.0043	0.0069

4 Experiments

We conduct our objective and subjective evaluations on a totally different audio-visual dataset, the LDC dataset [18]. It contains 10 female and 4 male speakers, where each speaker provides 94 samples, totaling to 1316 utterances. The duration of the videos is arbitrary, and the resolution of the samples are 720×480 . Since the frame rate of the videos is higher than the Grid dataset used to train our system, we resampled the videos to the same frame rate of 25 FPS. The vocabulary of the LDC dataset is much larger than that of the Grid dataset. There are various words and sentences from TIMIT sentences [11], Northwestern University Auditory Test No. 6 [21], and Central Institute for the Deaf (CID) Everyday Sentences [1]. The audio stream is provided at 48 kHz sampling rate, which we down-sampled to 44.1 kHz. Figure 3 shows examples of ground-truth and generated face landmarks in the first and second row, respectively. Examples of generated videos are publicly accessible².

4.1 Objective Evaluation

We report the root-mean-squared error (RMSE) results between the ground-truth (GT) and predicted (PD) face landmarks according to Eq. 1. The landmarks scale

² <http://www.ece.rochester.edu/projects/air/projects/talkingface.html>.

are between 0 and 1, therefore RMSE value of 0.01 approximately equivalent to 1% error. We also report the RMSE of the first and second order temporal differences of the GT and PD face landmarks to assess the movement. We report the results in Table 1. These results serve as a way of model selection. The best model according to these results is the model that has 40 ms delay and 5 frames of context information (D40-C5). We selected this model to conduct the subjective evaluations, which are described in the next section.

4.2 Subjective Evaluation

We conducted subjective tests to determine if our system can generate realistic face landmarks. 17 naive volunteer evaluators who are graduate students at the University of Rochester participated in the test. The test presented 25 real landmark videos and 25 generated landmark videos in a randomized order to each evaluator and asked the evaluator to label whether each presented video was real or fake. Each video was presented twice in the randomized video sequence. The real landmark videos were created from randomly selected LDC videos. Landmarks were extracted and aligned, and the identity information was removed, according to Sect. 3. Fake videos were generated from the audio signals of another 25 randomly selected LDC videos. The GT landmarks were noisy; hence we also added Gaussian noise to the PD landmarks to make them look more like the GT landmarks. In addition to a binary decision, the evaluators were asked to report their confidence level of each decision, between 0 and 100%.

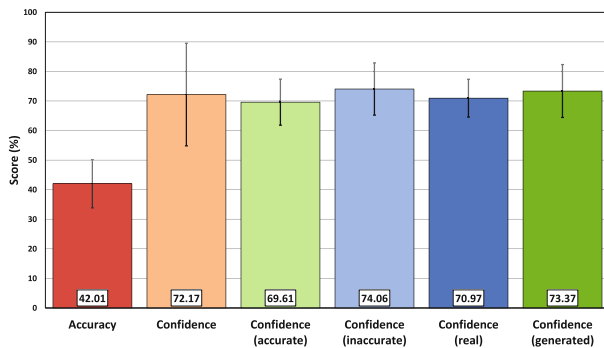


Fig. 4. Subjective evaluation results. The mean accuracy score and its standard deviation are averaged over all subjects. The mean confidence scores and their standard deviations are averaged over all subjects and videos.

The mean accuracy score of the evaluators are shown in Fig. 4, along with the overall mean confidence score and the mean confidence score for the correctly and incorrectly predicted samples. The results show that the evaluators struggled to distinguish real and generated samples, as the accuracy is 42.01% which is even below chance (50%). Another interesting observation of this test is that the

mean confidence score for accurately determined samples is lower than that for inaccurately determined samples. This suggests that the evaluators had a higher classification accuracy when they were more cautious. Another outcome is that the mean confidence score on answers for generated samples is more than the confidence score on answers for the ground truth samples.

5 Conclusion

In this work, we present a method to generate talking face landmarks from speech. We extract face landmarks from the Grid corpus, align them across different speakers, and transform their shapes into the mean shape to remove the identity information. The LSTM network predicts the face landmarks from the first and second order temporal differences of the log-mel spectrogram from any arbitrary voice. The network can produce face landmarks that look natural for the given speech input. The main limitation of this network is that it cannot produce “oh” and “oo” sounds right. We plan to balance the phonetic content of the dataset to enable the network to produce all phonemes correctly in our future work. We will evaluate the system against noise, and improve it to obtain a noise-resilient system in our future work. We report objective and subjective evaluation results that are promising. We release the code and example videos to the community.

References

1. Blamey, P.J., Pyman, B.C., Clark, G.M., Dowell, R.C., Gordon, M., Brown, A.M., Hollow, R.D.: Factors predicting postoperative sentence scores in postlinguistically deaf adult cochlear implant patients. *Ann. Otol. Rhinol. Laryngol.* **101**(4), 342–348 (1992)
2. Brand, M.: Voice puppetry. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 21–28. ACM Press/Addison-Wesley Publishing Co. (1999)
3. Cassidy, S., Stenger, B., Dongen, L.V., Yanagisawa, K., Anderson, R., Wan, V., Baron-Cohen, S., Cipolla, R.: Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. *Comput. Vis. Image Underst.* **148**, 193–200 (2016)
4. Choi, K., Luo, Y., Hwang, J.N.: Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **29**, 51–61 (2001)
5. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? (2017). arXiv preprint: [arXiv:1705.02966](https://arxiv.org/abs/1705.02966)
6. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **120**(5), 2421–2424 (2006)
7. Cosker, D., Marshall, D., Rosin, P.L., Hicks, Y.: Speech driven facial animation using a Hidden Markov coarticulation model. In: *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 128–131. IEEE (2004)

8. Cosker, D., Marshall, D., Rosin, P., Hicks, Y.: Video realistic talking heads using hierarchical non-linear speech-appearance models, *Mirage*, France, vol. 147 (2003)
9. Dodd, B.E., Campbell, R.E.: *Hearing by Eye: The Psychology of Lip-Reading*. Lawrence Erlbaum Associates, Inc., Hillsdale (1987)
10. Fan, B., Wang, L., Soong, F.K., Xie, L.: Photo-real talking head with deep bidirectional LSTM. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4884–4888. IEEE (2015)
11. Garofalo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: *The darpa timit acoustic-phonetic continuous speech corpus CD-ROM*. Linguistic Data Consortium (1993)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
13. King, D.E.: Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009)
14. Maddox, R.K., Atilgan, H., Bizley, J.K., Lee, A.K.: Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *eLife* **4** (2015)
15. Mallick, S.: *Face morph using opencv c++/python* (2016). <http://www.learnopencv.com/face-morph-using-opencv-cpp-python/>
16. Pham, H.X., Cheung, S., Pavlovic, V.: Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In: *The 1st DALCOM Workshop, CVPR* (2017)
17. Pham, H.X., Wang, Y., Pavlovic, V.: End-to-end learning for 3d facial animation from raw waveforms of speech (2017). arXiv preprint: [arXiv:1710.00920](https://arxiv.org/abs/1710.00920)
18. Richie, S., Warburton, C., Carter, M.: *Audiovisual database of spoken American English*. Linguistic Data Consortium (2009)
19. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing Obama: learning lip sync from audio. *ACM Trans. Graph. (TOG)* **36**(4), 95 (2017)
20. Terissi, L.D., Gómez, J.C.: Audio-to-visual conversion via HMM inversion for speech-driven facial animation. In: Zaverucha, G., da Costa, A.L. (eds.) *SBIA 2008. LNCS (LNAI)*, vol. 5249, pp. 33–42. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88190-2_9
21. Tillman, T.W., Carhart, R.: An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University auditory test no. 6. Technical report, Northwestern University Evanston Auditory Research Lab (1966)
22. Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., et al.: Photo-realistic expressive text to talking head synthesis. In: *INTERSPEECH*, pp. 2667–2669 (2013)
23. Wang, L., Han, W., Soong, F.K., Huo, Q.: Text driven 3d photo-realistic talking head. In: *Twelfth Annual Conference of the International Speech Communication Association* (2011)
24. Xie, L., Liu, Z.Q.: A coupled HMM approach to video-realistic speech animation. *Pattern Recogn.* **40**, 2325–2340 (2007)
25. Zhang, X., Wang, L., Li, G., Seide, F., Soong, F.K.: A new language independent, photo-realistic talking head driven by voice only. In: *Interspeech*, pp. 2743–2747 (2013)

Advances in Phase Retrieval and Applications



An Approximate Message Passing Approach for DOA Estimation in Phase Noisy Environments

Guillaume Beaumont¹(✉), Ronan Fablet², and Angélique Drémeau¹

¹ Lab-STICC UMR 6285, CNRS, ENSTA Bretagne, 29200 Brest, France
guillaume.beaumont@ensta-bretagne.org

² IMT-Atlantique, Lab-STICC UMR 6285, UBL, 29200 Brest, France

Abstract. In underwater acoustics, wave propagation can be greatly disrupted by random fluctuations in the ocean environment. In particular, phase measurements of the complex pressure field can be heavily noisy and can defeat conventional direction-of-arrival (DOA) estimation algorithms.

In this paper, we propose a new Bayesian approach to address such phase noise through an informative prior on the measurements. This is combined to a sparse assumption on the directions of arrival to achieve a highly-resolved estimation and integrated into a message-propagation algorithm referred to as the “paSAMP” algorithm (for Phase-Aware Swept Approximate Message Passing). Our algorithm can be seen as an extension of the recent phase-retrieval algorithm “prSAMP” to phase-aware priors.

Experiments on simulated data mimicking real environments demonstrate that paSAMP outperform conventional approaches (e.g. classic beamforming) in terms of DOA estimation. paSAMP also proves to be more robust to additive noise than other variational methods (e.g. based on mean-field approximation).

Keywords: DOA estimation · Sparse representation
Bayesian estimation · Variational Bayesian approximations
Message passing algorithms

1 Introduction

Common to many applications such as SONAR, RADAR, and telecommunications, direction-of-arrival (DOA) estimation aims at locating one or more sources emitting in some propagation media. Various methods have been proposed to address this problem. They can be distinguished by the assumptions made on the propagating medium and sources.

The beamforming approach [1] constitutes the most famous approach. As it implicitly assumes the noise to be Gaussian and additive, it leads to poor

G. Beaumont—This work has been supported by the DGA MRIS.

estimation performance for complex phase perturbations. The so-called “high-resolution” techniques consider additional assumptions over the number or the nature of the sources. This is the case of the well-known MUSIC method [2]. MUSIC assumes the number of sources to be known and the separability of the sub-spaces where the noise and the signal live. More recently, “compressive” beamforming approaches proposed *e.g.* in [3] benefit from an explicit sparse model on the sources.

While all the previously cited approaches rely on an additive Gaussian noise model, recent work has focused on the integration of phase-noise models better accounting for complex propagation processes. Such approaches aim to take into account the wave-front distortion occurring when waves travel through fluctuating media. This is of key interest for a wide range of application fields including as underwater acoustics [4, 5] or atmospheric sound propagation [6, 7]. These contributions mainly relate to recent advances in phase recovery (see *e.g.* [8–11]) and the use of informative priors on the missing phases. In this respect, we can mention the Bayesian approach “paVBEM” based on a mean-field approximation [12].

Here, we further explore a variational Bayesian approach. Knowing that higher-order approximations and associated message-passing algorithms outperform mean-field approximations for a wide range of inverse problems [13], we propose a novel approach based on the “swept approximate message passing” (SwAMP) algorithm introduced in [14]. Our algorithm is proven to be more robust to additive noise and multiplicative phase noise than previous approaches using phase-aware priors such as the paVBEM approach [12] and those using non-informative phase priors [9].

2 Problem Statement

In this section, we recall the Bayesian modeling introduced in [12], which we shall follow throughout of this paper, and introduce the estimation problem we propose to solve.

2.1 Observation Model

Our objective is to design an algorithm able to recover the directions of arrival of a few waves, despite a structured phase-noisy environment, exploiting one single temporal snapshot on a uniform linear sensor array. In underwater acoustics, this noise is mainly due to internal waves, changing the local sound-speed (see *e.g.* [4]). These internal waves and their impact on the acoustic measurements have been studied in different works (see [4, 5]), which leads to a statistical characterization of the phase noise.

In this context, we propose the following observation model: we consider a linear antenna composed of N regularly-spaced sensors and assume that the received signal at sensor n can be expressed as

$$y_n = e^{j\theta_n} \sum_{m=1}^M d_{nm}x_m + \omega_n, \tag{1}$$

where θ_n stands for the phase noise due to the propagation through the fluctuating medium and ω_n an additive noise. The scalar d_{nm} is the n -th element of the steering vector $\mathbf{d}_m = [e^{j\frac{2\pi}{\lambda} \Delta \sin(\phi_m)} \dots e^{j\frac{2\pi}{\lambda} \Delta N \sin(\phi_m)}]^T$ where the ϕ_m 's are some potential angles of arrival, Δ is the distance between two adjacent sensors and λ is the wavelength of the propagation waves.

Within model (1), at each sensor of the antenna, we assume that the received field is a combination of a few waves arriving from different angles ϕ_m . The DOA estimation problem then consists in identifying the positions of the non-zero coefficients in $\mathbf{x} \triangleq [x_1 \dots x_M]^T$. In underwater acoustics, the phase noise considered in (1) is well-suited to characterize phase perturbations of the wave front in a fluctuating ocean [5], especially in the case of the so-called ‘‘partially saturated’’ propagation regime defined in [4]. This regime focuses on far-field propagation at high frequency with no multipath. In this case, amplitude variations of the measured acoustic field can be neglected regarding the high sensibility to a structured phase-noise. Note that a similar fluctuation regime has been also identified in atmospheric sound propagation (see [7]).

2.2 Bayesian Formulation of the Problem

We address the estimation of \mathbf{x} from the measurements $\mathbf{y} \triangleq [y_1, \dots, y_N]^T$ in the presence of (unknown) additive noise $\boldsymbol{\omega} \triangleq [\omega_1, \dots, \omega_N]^T$ and multiplicative phase noise $\boldsymbol{\theta} \triangleq [\theta_1, \dots, \theta_N]^T$. To solve this problem, we consider a Bayesian framework and define some prior assumptions on the different variables in (1).

A first assumption is set on the number of sources (*i.e.* the non-zero coefficients in \mathbf{x}) that we suppose to be small in front of the number of sensors. To take into account this sparse hypothesis, we adopt a Bernoulli-Gaussian model $\forall m \in \{1, \dots, M\}$

$$p(x_m) = \rho \mathcal{CN}(x_m; m_x, \sigma_x^2) + (1 - \rho)\delta_0(x_m), \tag{2}$$

where ρ is the Bernoulli parameter and equals the probability for x_m to be non-zero¹, $\mathcal{CN}(x_m; m_x, \sigma_x^2)$ stands for the circular complex Gaussian distribution with mean m_x and variance σ_x^2 , and $\delta_0(x_m)$ for the Dirac distribution. The Bernoulli-Gaussian model is widely used when considering Bayesian inference methods for sparsity-constrained problems (see *e.g.* [15, 16]).

Previous studies of the statistical characterization of fluctuation phenomena [4, 5] provide the basis for the definition of a phase-noise prior. In underwater

¹ We assume the Bernoulli parameter to be the same for each $m \in \{1, \dots, M\}$.

acoustics, [4, 5] exhibited and characterized the existence of a spatial correlation of the measured field all along the antenna. To account for the resulting coherence length, we consider a Markovian model as

$$p(\theta_n | \theta_{n-1}) = \mathcal{N}(\theta_n; \beta \theta_{n-1}, \sigma_\theta^2), \quad \forall n \in \{2, \dots, N\}, \tag{3}$$

$$p(\theta_1) = \mathcal{N}(\theta_1; 0, \sigma_1^2), \tag{4}$$

with $\beta \in \mathbb{R}_+$. Variance σ_θ^2 is related to the coherence length and accounts for the strength of the fluctuations. As an example, a large σ_θ^2 models strong fluctuations of the medium and results in a small coherence length, such that the phase noise varies widely from a sensor to the neighboring ones.

We also introduce an additive noise ω to account for the combination of a large number of random parasitic phenomena. Based on the central limit theorem, we consider with a classic zero-mean Gaussian distribution with variance σ^2 .

Overall, our Bayesian formulation leads to the following Minimum Mean Square Error (MMSE) problem:

$$\hat{\mathbf{x}} = \underset{\tilde{\mathbf{x}}}{\operatorname{argmin}} \int_{\mathbf{x}} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 p(\mathbf{x} | \mathbf{y}) d\mathbf{x} \tag{5}$$

where $p(\mathbf{x} | \mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$.

To solve efficiently this problem, we propose to exploit a variational Bayesian inference strategy, that approximates the posterior joint distribution $p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y})$ by a distribution having a suitable factorization. In [12], a mean-field approximation $p(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \simeq q(\boldsymbol{\theta}) \prod_{m=1}^M q(x_m)$ was considered. Here, we address a different type of factorization, called the Bethe approximation, relating to the ‘‘approximate message passing’’ (AMP) algorithms [13]. This approximation exploits higher-order terms which result in better estimation performance [13].

We motivate and detail our approach in the next section.

3 The ‘‘paSAMP’’ Algorithm

In this section, we motivate and present the novel algorithm proposed to solve problem (5).

3.1 Motivation and Main Principles of the Approach

AMP algorithms have been considered for a few years as a serious solution to linear problems under sparsity constraints. First considered in the sole case of *i.i.d* (sub-)Gaussian matrices, they have been recently extended to random but more general matrices by the ‘‘vector approximate message passing’’ (VAMP) algorithm [17] and to highly correlated matrices by the ‘‘swept approximate message passing’’ (SwAMP) approach [14]. Both methods aim at alleviating the convergence issues of AMP (notably highlighted in [18]) due to its parallel update structure.

Algorithm 1. paSAMP Algorithm

Input: \mathbf{y} , \mathbf{D} , σ^2 , ρ , σ_x^2 , $\boldsymbol{\mu}_\theta$, $\boldsymbol{\Sigma}_\theta$, T_{max}
Define:
 $g_{out,n} \triangleq \frac{1}{\Sigma_{z_n}} (E_{Z|Y,P}\{z_n|y_n, \mu_{z_n}, \Sigma_{z_n}\} - \mu_{z_n})$
 $g'_{out,n} \triangleq \frac{1}{\Sigma_{z_n}} (\frac{var_{Z|Y,P}\{z_n|y_n, \mu_{z_n}, \Sigma_{z_n}\}}{\Sigma_{z_n}} - 1)$
 $g_{in,m} \triangleq E_{X|Y}\{x_m|\mu_{x_m}, \Sigma_{x_m}\}$
 $g'_{in,m} \triangleq var_{X|Y}\{x_m|\mu_{x_m}, \Sigma_{x_m}\}$

- 1: **while** $t < T_{max}$ **do**
- 2: **for** $n = 1 \dots N$ **do**
- 3: $\hat{z}_n(t) = \sum_{m=1}^M d_{nm} a_m(t)$
- 4: $\Sigma_{z_n}^1(t+1) = \sum_{m=1}^M |d_{nm}|^2 v_m(t)$
- 5: $\mu_{z_n}^1(t+1) = \hat{z}_n(t) - \Sigma_{z_n}^1(t) g_{out,n}$
- 6: **end for**
- 7: **for** $m = \text{permute}[1 \dots M]$ **do**
- 8: $\Sigma_{x_m}(t+1) = (-\sum_{n=1}^N |d_{nm}|^2 g'_{out,n})^{-1}$
- 9: $\mu_{x_m}(t+1) = a_m(t) + \Sigma_{x_m}(t+1) \sum_{n=1}^N d_{nm} g_{out,n}$
- 10: $v_m(t+1) = \Sigma_{x_m}(t+1) g'_{in,m}$
- 11: $a_m(t+1) = g_{in,m}$
- 12: **for** $n = 1 \dots N$ **do**
- 13: $\Sigma_{z_n}^{m+1}(t+1) = \Sigma_{z_n}^m(t+1) + |d_{nm}|^2 (v_m(t+1) - v_m(t))$
- 14: $\mu_{z_n}^{m+1}(t+1) = \mu_{z_n}^m(t+1) + d_{nm} (a_m(t+1) - a_m(t))$
 $-g_{out,n}(t) (\Sigma_{z_n}^{m+1}(t+1) - \Sigma_{z_n}^m(t+1))$
- 15: **end for**
- 16: update σ^2 according to [12]
- 17: update $[\theta_{m_n}, \Sigma_{\theta_n}]$ according to (14–15)
- 18: **end for**
- 19: **end while**
- 20: **Output:** $\{\hat{x}_m = a_m(T_{max})\}_m$

AMP, VAMP and SwAMP have been extended to generalized but component-wise measurement models [14, 19, 20]. They have been then successfully applied to the phase recovery task where $\theta_n \sim \mathcal{U}[0, 2\pi]$, $\forall n \in \{1, \dots, N\}$, giving raise to the so-called “prGAMP” [21], “prVAMP” [10] and “prSAMP” [9] algorithms. In particular, the latter was shown to outperform other state-of-the-art algorithms among which the mean-field approximation [8].

The prSAMP algorithm constitutes thus a promising approach for our DOA estimation² problem (5). However, here, the phases θ_n ’s are spatially-correlated (as represented in the Markov model). This prevents us from a direct application of prSAMP.

We thus propose an iterative algorithm based on the two following mathematical derivations:

- (i) the extension of prSAMP to a *i.i.d.* Gaussian prior on the phases,

² Note in addition that the DOA estimation problem involves a highly-correlated matrix. This further motivates a SwAMP-like approach.

(ii) the use of a mean-field approximation to estimate the (Gaussian) posterior distribution on the phases.

We detail both aspects in the next two sub-sections. In the following, we refer to the proposed procedure as “paSAMP” for “phase-aware SwAMP algorithm”. The pseudo-code of paSAMP is presented in Algorithm 1.

3.2 Extension of PrSAMP to *i.i.d* Gaussian phases

AMP algorithms are based on the propagation of two types of messages: the “outgoing” messages and the “ingoing” messages from and to variables’ nodes $\{x_m\}_{m=\{1\dots M\}}$. These messages are derived here for the prior distributions attached to the considered problem, namely (2) and (3) and (4)³. We first focus on the “outgoing messages”.

Considering $z_n \triangleq \sum_{m=1}^M d_{nm}x_m, \forall n \in \{1, \dots, N\}$, we assume that the z_n ’s follow Gaussian distributions with means μ_{z_n} and variances Σ_{z_n} as linear combinations of x_m ’s following Bernoulli-Gaussian distributions. By integrating over θ_n and resorting⁴ to an identification with a Von Mises distribution [22], we can write the moments of the posterior distribution as

$$E_{Z|Y}\{z_n|y_n, \mu_{z_n}, \Sigma_{z_n}\} = \frac{\Sigma_{z_n}}{\sigma^2 + \Sigma_{z_n}} \mathbf{R}_0\left(\frac{1}{\Sigma_{z_n}^z}\right) y_n e^{-j\mu_{\theta_n}^z} + \frac{\sigma^2}{\sigma^2 + \Sigma_{z_n}} \mu_{z_n}, \quad (6)$$

$$\begin{aligned} var_{Z|Y}\{z_n|y_n, \mu_{z_n}, \Sigma_{z_n}\} &= \frac{|\Sigma_{z_n} y_n e^{-j\mu_{\theta_n}^z} + \mu_{z_n} \sigma^2|^2}{|\sigma^2 + \Sigma_{z_n}|^2} \mathbf{R}_0\left(\frac{1}{\Sigma_{z_n}^z}\right) + \frac{\Sigma_{z_n} \sigma^2}{\sigma^2 + \Sigma_{z_n}} \\ &\quad - E_{Z|Y}\{z_n|y_n, \mu_{z_n}, \Sigma_{z_n}\}^2, \end{aligned} \quad (7)$$

with

$$\frac{1}{\Sigma_{z_n}^z} = \frac{1}{\alpha} + \frac{1}{\Sigma_{\theta_n}}, \quad \mu_{\theta_n}^z = \frac{-\frac{\arg(y_n^* \mu_{z_n})}{\alpha} + \frac{\mu_{\theta_n}}{\Sigma_{\theta_n}}}{\frac{1}{\alpha} + \frac{1}{\Sigma_{\theta_n}}}, \quad \alpha = \frac{\Sigma_{z_n} + \sigma^2}{|y_n| |\mu_{z_n}|},$$

μ_{θ_n} (resp. Σ_{θ_n}) is the marginalized posterior mean (resp. variance) of the phase noise θ_n as discussed in the next section, and $\mathbf{R}_0(\cdot) = \frac{I_1(\cdot)}{I_0(\cdot)}$ where $I_n(\cdot)$ is the modified Bessel function of the first kind at order n . We refer the reader to our technical report [23] which details the derivations of the computations.

Regarding the “ingoing” messages, which carry the prior information on the $\{x_m\}_{m=\{1\dots M\}}$, the Bernoulli-Gaussian case has already been considered within the AMP context, in particular in [15]. Similarly to the “outgoing” messages, the moments of the “ingoing” messages resort to intermediary parameters μ_{x_m} and Σ_{x_m} resp. homogeneous to the mean and variance of a Gaussian distribution:

$$E_{X|Y}(x_m|\mu_{x_m}, \Sigma_{x_m}) = \frac{\rho\sqrt{2\pi\nu^2}}{Z_{nor}} e^{-\frac{|m_x - \mu_{x_m}|^2}{2(\sigma^2 + \Sigma_{x_m})}} \gamma, \quad (8)$$

$$var_{X|Y}(x_m|\mu_{x_m}, \Sigma_{x_m}) = \frac{\rho\sqrt{2\pi\nu^2}}{Z_{nor}} e^{-\frac{|m_x - \mu_{x_m}|^2}{2(\sigma^2 + \Sigma_{x_m})}} |\gamma^2 + \nu^2| - E_{X|Y}(x_m|\mu_{x_m}, \Sigma_{x_m})^2 \quad (9)$$

³ We refer the reader to papers [9, 14] for a more general presentation of the approach.

⁴ We justify and develop this point in the technical report [23].

with

$$Z_{nor} = \rho \sqrt{2\pi\nu^2} e^{-\frac{|m_x - \mu_{x_m}|^2}{2(\sigma^2 + \Sigma_{x_m})}} + (1 - \rho) e^{-\frac{|\mu_{x_m}|^2}{2\Sigma_{x_m}}}, \quad (10)$$

$$\gamma = \frac{\sigma^2 \mu_{x_m} + \Sigma_{x_m} m_x}{\Sigma_{x_m} + \sigma^2}, \quad \nu^2 = \frac{\sigma^2 \Sigma_{x_m}}{\Sigma_{x_m} + \sigma^2}. \quad (11)$$

We implement those calculations to paSAMP as new definitions of the two functions g_{in} and g'_{in} defined in the pseudo-code Algorithm 1. We remind the reader that, as an extended implementation of the SwAMP algorithm, the paSAMP algorithm will conserve the structure described in [9, 14]. For sake of clarity, we use the notations introduced in [21] except for the scalar d_{nm} .

3.3 Mean-Field Approximation for the Phase Noise

The above expressions call on the knowledge of the moments of the posterior distribution on θ . To simplify the latter computation, we propose in this step to resort to a mean-field approximation. Following a similar reasoning as in [12], we get

$$q(\theta) = \mathcal{N}(\theta; \mu_\theta, \Sigma_\theta), \quad (12)$$

where $\Sigma_\theta^{-1} = \Lambda_\theta^{-1} + \text{diag}\left(\frac{2}{\sigma^2}|\boldsymbol{\eta}|\right), \quad (13)$

$$\mu_\theta = \Sigma_\theta \left(\text{diag}\left(\frac{2}{\sigma^2}|\boldsymbol{\eta}|\right) \arg(\boldsymbol{\eta}) \right), \quad (14)$$

with $\eta_n = y_n \sum_{m=1}^M d_{nm}^* E_{X|Y}^* \{x_m | \mu_{x_m}, \Sigma_{x_m}\}$, the n th element in $\boldsymbol{\eta}$ ($|\boldsymbol{\eta}|$ stands here for the element-wise absolute value of $\boldsymbol{\eta}$ and $*$ for the complex conjugate), and Λ_θ^{-1} is the precision matrix attached to the prior distribution (4) on θ , *i.e.*

$$\Lambda_\theta^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} + \frac{\beta^2}{\sigma_\theta^2} & -\frac{\beta}{\sigma_\theta^2} & 0 & 0 \\ -\frac{\beta}{\sigma_\theta^2} & \frac{1+\beta^2}{\sigma_\theta^2} & \ddots & 0 \\ 0 & \ddots & \ddots & -\frac{\beta}{\sigma_\theta^2} \\ 0 & 0 & -\frac{\beta}{\sigma_\theta^2} & \frac{1}{\sigma_\theta^2} \end{pmatrix}. \quad (15)$$

Note that since the distribution $q(\theta)$ is Gaussian, marginals $q(\theta_n)$ used in the previous “prSAMP-step” of the algorithm come as

$$q(\theta_n) = \mathcal{N}(\theta_n; \mu_{\theta_n}, \Sigma_{\theta_n}) \quad (16)$$

where μ_{θ_n} (resp. Σ_{θ_n}) is the n th element in μ_θ (resp. in the diagonal of Σ_{θ_n}).

Finally, following [12], we insert an estimation of the variance σ^2 of the additive noise as a maximization step of an Expectation-Maximization (EM) algorithm. Due to space limitation, we omit here the derivation of the computation, but we refer again the reader to our technical report [23].

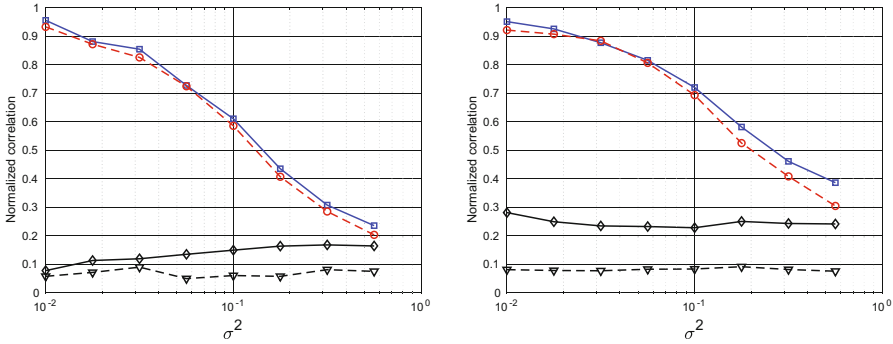


Fig. 1. Evolution of the (averaged) normalized correlation as a function of the variance σ^2 for $K = 2$ (left) and $K = 5$ (right), Comparison of the performance of conventional (delay-and-sum) beamforming (triangle mark), “prSAMP” (diamond mark), “paVBEM”(circle mark) and “paSAMP” (square mark). Experiments show that “paSAMP” provides better results and successfully integrates the phase noisy observation model.

4 Numerical Experiments

In this section, we perform a quantitative and qualitative evaluation of the proposed approach with respect to state-of-the-art algorithms.

We consider the problem of the identification of the directions of arrival of K plane waves from an antenna composed of $N = 256$ sensors. We assume that the angles of the K incident waves can be written as $\phi_k = \frac{\pi}{2} + i_k \frac{\pi}{50}$ with $i_k \in [1, 50]$. A set of $M = 50$ steering vectors is defined from a set of angles $\{\phi_i = -\pi + i \frac{\pi}{50}\}_{i \in \{1, \dots, 50\}}$ and the choice of the parameter $\lambda/\Delta = 4$. For each of the K incident waves, the coefficient x_{i_k} is initialized with $m_x = 0.5 + j0.5$, $\rho = K/M$ and $\sigma_x^2 = 0.1$. Finally, we set the following parameters for the phase Markov model (3): $\sigma_0^2 = 10$, $\sigma_\theta^2 = 0.1$ and $\beta = 0.8$. This corresponds to a situation where we have a high uncertainty on the initial value but a physical link between two space-consecutive angle measurements is taken into account.

We compare the performance of the following 4 different algorithms: (i) the standard beamforming introduced in [1] (dashed black curve, triangle mark); (ii) the prSAMP algorithm proposed in [9] as a solution to the phase retrieval problem (continuous black curve, diamond mark); (iii) the paVBEM procedure proposed in [12] exploiting the same prior models (dashed red curve, circle mark); (iv) the paSAMP algorithm described in Sect. 3 (continuous blue curve, square mark). To evaluate the performance of these procedures, we consider the normalized correlation between the ground truth \mathbf{x} and its reconstruction $\hat{\mathbf{x}}$, that is $\frac{|\mathbf{x}^H \hat{\mathbf{x}}|}{\|\mathbf{x}\| \|\hat{\mathbf{x}}\|}$, as a function of the additive noise variance σ^2 . This quantity is averaged over 100 realizations for each point of simulation.

The results achieved by the 4 procedures are presented in Fig. 1, resp. for $K = 2$ (left) and $K = 5$ (right) sources. In both cases, we see that the conventional beamforming and the prSAMP algorithm fail to reconstruct \mathbf{x} properly.

These results illustrate the benefits of carefully accounting for the phase noise in fluctuating environments. We can also notice the superiority of paSAMP over its mean-field counterpart paVBEM, especially in presence of a strong additive noise. This comes in the continuity of previous work [9], where prSAMP proved to outperform prVBEM. Finally, it is interesting to compare the performance of both paSAMP and paVBEM algorithms with regard to the number of sources. Both achieve better performance when confronting to $K = 5$ sources than to $K = 2$ sources. As mentioned in [12], this behavior is typical for the phase retrieval problems, where the loss information on the phases can be compensated by a larger number of sources. In addition, we observe that the performance gap between paSAMP and paVBEM tends to increase with the number of sources. This is in accordance with previous work [13] demonstrating the relevance of the Bethe approximation over the mean-field approximation when the signal to recover exhibits a low sparsity (*i.e.* a high number of non-zero coefficients).

5 Conclusion

We have presented here a novel AMP algorithm able to perform DOA estimation in a corrupted phase-noisy environment. This approach exploits both a sparsity prior on the sources and a structured prior on the phase noise. Compared to state-of-the-art algorithms, the approach presents a good behaviour illustrating a successful inclusion of the different assumptions. In particular, it outperforms a recent algorithm dealing with the same DOA estimation problem in fluctuating environments. Future work will include further assessment on real data.

Acknowledgment. The authors thank Boshra Rajaei for sharing her MATLAB implementation of the prSAMP algorithm.

References

1. Johnson, D.H., Dudgeon, D.E.: Array Signal Processing: Concepts and Techniques. P T R Prentice Hall, Englewood Cliffs (1993)
2. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**, 276–280 (1986)
3. Xenaki, A., Gerstoft, P., Mosegaard, K.: Compressive beamforming. J. Acoust. Soc. Am. **136**, 260–271 (2014)
4. Dashen, R., Flatté, S.M., Munk, W.H., Watson, K.M., Zachariasen, F.: Sound Transmission Through a Fluctuating Ocean. Cambridge University Press, London (2010)
5. Colosi, J.A.: Sound Propagation Through the Stochastic Ocean. Cambridge University Press, Cambridge (2016)
6. Cheinet, S., Ehrhardt, L., Juve, D., Blanc-Benon, P.: Unified modeling of turbulence effects on sound propagation. J. Acoust. Soc. Am. **132**, 2198–2209 (2012)
7. Ehrhardt, L., Cheinet, S., Juve, D., Blanc-Benon, P.: Evaluating a linearized Euler equations model for strong turbulence effects on sound propagation. J. Acoust. Soc. Am. **133**, 1922–1933 (2013)

8. Dremeau, A., Krzakala F.: Phase recovery from a Bayesian point of view: the variational approach. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3661–3665 (2015)
9. Rajaei, B., Gigan, S., Krzakala, F., Daudet, L.: Robust phase retrieval with the swept approximate message passing (prSAMP) algorithm. *IPOL* **7**, 43–55 (2016)
10. Metzler, C.A., Sharma, M.K., Nagesh, S., Baraniuk, R.G., Cossairt, O., Veeraghavan, A.: Coherent inverse scattering via transmission matrices: efficient phase retrieval algorithms and a public dataset. In: Proceedings of the IEEE International Conference on Computational Photography, pp. 1–16 (2017)
11. Waldspurger, I., d’Aspremont, A., Mallat, S.: Phase recovery, maxcut and complex semidefinite programming. *Math. Program.* **149**, 47–81 (2015)
12. Dremeau, A., Herzet, C.: DOA estimation in structured phase noisy environments: technical report (2016)
13. Krzakala, F., Manoel, A., Tramel, E.W., Zdeborova, L.: Variational free energies for compressed sensing. In: 2014 Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 1499–1503 (2014)
14. Manoel, A., Krzakala, F., Tramel, E., Zdeborova, L.: Swept approximate message passing for sparse estimation. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 1123–1132 (2015)
15. Vila, J., Schniter, P.: Expectation-maximization Bernoulli-Gaussian approximate message passing. In: Proceedings of the Signals, Systems and Computers (ASILOMAR), pp. 799–803 (2011)
16. Soussen, C., Idier, J., Brie, D., Duan, J.: From Bernoulli-Gaussian deconvolution to sparse signal restoration. *IEEE Trans. Signal Process.* **59**, 4572–4584 (2011)
17. Rangan, S., Schniter, P., Fletcher, A.K. : Vector approximate message passing. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 1588–1592 (2017)
18. Caltagirone, F., Zdeborova, L., Krzakala, F.: On convergence of approximate message passing. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 1812–1816 (2014)
19. Rangan, S.: Generalized approximate message passing for estimation with random linear mixing. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 2168–2172 (2011)
20. Schniter, P., Rangan, S., Fletcher, A.K.: Vector approximate message passing for the generalized linear model. In: Asilomar Conference on Signals, Systems and Computers, pp. 1525–1529 (2016)
21. Schniter, P., Rangan, S.: Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.* **63**, 1043–1055 (2015)
22. Mardia, K.V., Jupp, P.E.: *Directional Statistics*. Wiley, Chichester (2009)
23. Beaumont, G., Fablet, R., Dremeau, A.: DOA estimation in fluctuating environments: an approximate message-passing approach. Technical report (2018). <https://hal.archives-ouvertes.fr/hal-01624855v4/document>



An Expectation-Maximization Approach to Tuning Generalized Vector Approximate Message Passing

Christopher A. Metzler¹(✉), Philip Schniter², and Richard G. Baraniuk¹

¹ Department of Electrical and Computer Engineering, Rice University,
6100 Main Street, Houston, TX 77005, USA
cam6@rice.edu

² Department of Electrical and Computer Engineering, The Ohio State University,
2015 Neil Avenue, Columbus, OH 43210, USA

Abstract. Generalized Vector Approximate Message Passing (GVAMP) is an efficient iterative algorithm for approximately minimum-mean-squared-error estimation of a random vector $\mathbf{x} \sim p_{\mathbf{x}}(\mathbf{x})$ from generalized linear measurements, i.e., measurements of the form $\mathbf{y} = Q(\mathbf{z})$ where $\mathbf{z} = \mathbf{A}\mathbf{x}$ with known \mathbf{A} , and $Q(\cdot)$ is a noisy, potentially nonlinear, componentwise function. Problems of this form show up in numerous applications, including robust regression, binary classification, quantized compressive sensing, and phase retrieval. In some cases, the prior $p_{\mathbf{x}}$ and/or channel $Q(\cdot)$ depend on unknown deterministic parameters $\boldsymbol{\theta}$, which prevents a direct application of GVAMP. In this paper we propose a way to combine expectation maximization (EM) with GVAMP to jointly estimate \mathbf{x} and $\boldsymbol{\theta}$. We then demonstrate how EM-GVAMP can solve the phase retrieval problem with unknown measurement-noise variance.

Keywords: Expectation maximization · Generalized linear model
Compressive sensing · Phase retrieval

1 Introduction

We consider the problem of estimating a random vector $\mathbf{x} \in \mathbb{R}^N$ from observations $\mathbf{y} \in \mathbb{R}^M$ generated as shown in Fig. 1, which is known as the *generalized linear model* (GLM) [1]. Under this model, \mathbf{x} has a prior density $p_{\mathbf{x}}$ and \mathbf{y} obeys a likelihood function of the form $p(\mathbf{y}|\mathbf{x}) = p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{A}\mathbf{x})$, where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a known linear transform and $\mathbf{z} \triangleq \mathbf{A}\mathbf{x}$ are hidden transform outputs. The conditional density $p_{\mathbf{y}|\mathbf{z}}$ can be interpreted as a probabilistic measurement channel that accepts a vector \mathbf{z} and outputs a random vector \mathbf{y} . Although we have assumed real-valued quantities for the sake of simplicity, it is straightforward to generalize the methods in this paper to complex-valued quantities.

The GLM has many applications in statistics, computer science, and engineering. For example, in *statistical regression*, \mathbf{A} and \mathbf{y} contain experimental

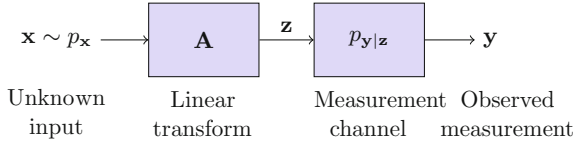


Fig. 1. Generalized Linear Model (GLM): An unknown random vector \mathbf{x} is observed through a linear transform \mathbf{A} followed by a probabilistic measurement channel $p_{\mathbf{y}|\mathbf{z}}$, yielding the measured vector \mathbf{y} .

features and outcomes, respectively, and \mathbf{x} are coefficients that best predict \mathbf{y} from \mathbf{A} . The relationship between \mathbf{y} and the optimal scores $\mathbf{z} = \mathbf{A}\mathbf{x}$ is then characterized by $p_{\mathbf{y}|\mathbf{z}}$. In *imaging*-related inverse problems, \mathbf{x} is an image to recover, \mathbf{A} is often Fourier-based, and $p_{\mathbf{y}|\mathbf{z}}$ models the sensor(s). In *communications* problems, \mathbf{x} may be a vector of discrete symbols to recover, in which case \mathbf{A} is a function of the modulation/demodulation scheme and the propagation physics. Or, \mathbf{x} may contain propagation-channel parameters to recover, in which case \mathbf{A} is a function of the modulation/demodulation scheme and the pilot symbols. In both cases, $p_{\mathbf{y}|\mathbf{z}}$ models receiver hardware and interference.

Below we give some examples of the measurement channels $p_{\mathbf{y}|\mathbf{z}}$ that are encountered in these applications.

- *Regression* often models $\mathbf{y} = \mathbf{z} + \mathbf{w}$ with additive noise \mathbf{w} , and so $p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) = p_{\mathbf{w}}(\mathbf{y} - \mathbf{z})$, where $p_{\mathbf{w}}$ is the density of \mathbf{w} . The “standard linear model” treats \mathbf{w} as additive white Gaussian noise (AWGN) but is not robust to outliers. Robust methods typically use heavy-tailed models for \mathbf{w} .
- *Binary linear classification* can be modeled using $y_m = \text{sgn}(z_m + w_m)$, where $\text{sgn}(v) = 1$ for $v \geq 0$ and $\text{sgn}(v) = -1$ for $v < 0$, and w_m are errors. Gaussian w_m yields the “probit” model and logistic w_m yields the “logistic” model.
- *Quantized compressive sensing* models $y_m = Q(z_m + w_m)$, where $Q(\cdot)$ is a scalar quantizer and w_m is additive, often AWGN.
- *Phase retrieval* models $y_m = |z_m|$ in the noiseless case, where $z_n \in \mathbb{C}$. When noise is present, one approach is to model $y_m = |z_m + w_m|$ with $w_m \in \mathbb{C}$ and another is to model $y_m = |z_m|^2 + w_m$ with real-valued w_m .

In this work, we focus on the case that the prior $p_{\mathbf{x}}$ and the likelihood $p_{\mathbf{y}|\mathbf{z}}$ depend on parameters $\boldsymbol{\theta}$ that are a priori unknown. For example, the prior $p_{\mathbf{x}}$ might be Bernoulli-Gaussian with unknown sparsity rate and variance, and the likelihood might involve an additive noise of an unknown variance. We are interested in jointly estimating \mathbf{x} and $\boldsymbol{\theta}$ from \mathbf{y} , where $\boldsymbol{\theta}$ are treated as deterministic. In particular, we aim to compute the ML estimate of $\boldsymbol{\theta}$ and the MMSE estimate of \mathbf{x} under $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\text{ML}}$:

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} p(\mathbf{y}; \boldsymbol{\theta}) \tag{1a}$$

$$\hat{\mathbf{x}}_{\text{MMSE}} = \mathbb{E}\{\mathbf{x}|\mathbf{y}; \hat{\boldsymbol{\theta}}_{\text{ML}}\}, \tag{1b}$$

sometimes referred to as the “empirical Bayesian” approach.

For most priors and likelihoods of interest, exact computation of the conditional mean in (1b) is intractable. Thus we might settle for an approximation of the MMSE estimate $\hat{\mathbf{x}}_{\text{MMSE}}$. In the case that \mathbf{A} is well modeled as a realization of a large rotationally invariant random matrix, the generalized vector approximate message passing (GVAMP) algorithm [2] is a computationally efficient approach to approximate-MMSE inference under the GLM in Fig. 1. In the large system limit (i.e., $M, N \rightarrow \infty$ with $M/N \rightarrow \delta \in (0, 1)$), it is rigorously characterized by state-evolution whose fixed points, when unique, are Bayes optimal [3].

For the special case of an AWGN likelihood, i.e., $p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}; \mathbf{z}, \nu_w \mathbf{I})$ for some $\nu_w > 0$, GVAMP reduces to the simpler VAMP algorithm [4]. By merging VAMP with expectation maximization (EM) [5], one obtains the ‘‘EM-VAMP’’ approach [6] to the empirical-Bayesian estimation problem (1). In fact, with large right-rotationally invariant \mathbf{A} , EM-VAMP is rigorously characterized by state-evolution [7]. Furthermore, under some identifiability conditions, it is possible to show that EM-VAMP yields an asymptotically efficient estimate of $\boldsymbol{\theta}$.

In this paper, we propose a way to merge EM and GVAMP to tackle GLMs of the form discussed above. This yields, for example, a way to handle phase retrieval with unknown measurement-noise variance. The proposed ‘‘EM-GVAMP’’ approach is described in the next section.

2 EM-GVAMP

In the sequel we assume a GLM of the form

$$p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z) = \prod_{i=1}^M p(y_i|z_i; \boldsymbol{\theta}_z), \quad \mathbf{z} = \mathbf{A}\mathbf{x}, \quad p(\mathbf{x}; \boldsymbol{\theta}_x) = \prod_{j=1}^N p(x_j; \boldsymbol{\theta}_x), \quad (2)$$

where $\boldsymbol{\theta} \triangleq [\boldsymbol{\theta}_x, \boldsymbol{\theta}_z]$ are unknown deterministic parameters, and where $\mathbf{z} \in \mathbb{R}^M$ and $\mathbf{x} \in \mathbb{R}^N$.

2.1 The EM Algorithm

Recalling the empirical-Bayesian methodology (1), the maximum-likelihood estimate of $\boldsymbol{\theta}$ given \mathbf{y} can be written as

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} \{ -\ln p(\mathbf{y}; \boldsymbol{\theta}) \}, \quad (3)$$

where

$$\begin{aligned} p(\mathbf{y}; \boldsymbol{\theta}) &= \int p(\mathbf{y}, \mathbf{z}, \mathbf{x}; \boldsymbol{\theta}) \, d\mathbf{z} \, d\mathbf{x} = \int p(\mathbf{y}|\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}) p(\mathbf{z}, \mathbf{x}; \boldsymbol{\theta}) \, d\mathbf{z} \, d\mathbf{x} \\ &= \int p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}_x) \, d\mathbf{z} \, d\mathbf{x}. \end{aligned} \quad (4)$$

Although $p(\mathbf{y}; \boldsymbol{\theta})$ is difficult to work with directly, the expectation-maximization (EM) algorithm [8] offers an alternative. There, the idea is to write

$$-\ln p(\mathbf{y}; \boldsymbol{\theta}) = J(b; \boldsymbol{\theta}) - D(b \parallel p(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta})) \tag{5}$$

for an arbitrary belief $b(\mathbf{x})$, where $D(\cdot \parallel \cdot)$ is KL divergence,

$$J(b; \boldsymbol{\theta}) \triangleq D(b \parallel p(\mathbf{x}; \boldsymbol{\theta}_x)) + D(b \parallel p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_z)) + H(b) \tag{6}$$

is known as the Gibbs free energy, and $H(b)$ is the entropy of b . Because $D(b \parallel p(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta})) \geq 0$ for any b , we have that $J(b; \boldsymbol{\theta})$ is an upper bound on $-\ln p(\mathbf{y}; \boldsymbol{\theta})$, the quantity that ML seeks to minimize. Thus, if it is tractable to construct and minimize $J(b; \boldsymbol{\theta})$, it makes sense to iterate the following two steps (over $k = 1, 2, \dots$):

E step: $b^k(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}; \hat{\boldsymbol{\theta}}^k)$ (7)

M step: $\hat{\boldsymbol{\theta}}^{k+1} = \arg \min_{\boldsymbol{\theta}} J(b^k; \boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta}} D(b^k \parallel p(\mathbf{x}; \boldsymbol{\theta}_x)) + D(b^k \parallel p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_z)),$ (8)

which together constitute the EM algorithm. The ‘‘E’’ step creates an upper bound on $-\ln p(\mathbf{y}; \boldsymbol{\theta})$ that is tight at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^k$, and the ‘‘M’’ step finds the estimate of $\boldsymbol{\theta}$ that minimizes this bound.

Unfortunately, however, the posterior density required by the E-step (7),

$$p(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}) = \frac{p(\mathbf{x}; \boldsymbol{\theta}_x)p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_z)}{p(\mathbf{y}; \boldsymbol{\theta})} = \frac{p(\mathbf{x}; \boldsymbol{\theta}_x) \int p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)\delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \, d\mathbf{z}}{\int p(\mathbf{x}; \boldsymbol{\theta}_x)p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)\delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \, d\mathbf{z} \, d\mathbf{x}}, \tag{9}$$

is difficult to compute due to the high-dimensional integration. Thus we consider an approximation afforded by the GVAMP algorithm [2]. For this, we first reparameterize the GLM (2) as a standard linear model (SLM).

2.2 An SLM Equivalent

The GLM (2) can be written as an SLM using the following formulation:

$$\bar{\mathbf{y}} = \bar{\mathbf{A}}\bar{\mathbf{x}} + \bar{\mathbf{w}} \quad \text{with} \quad \bar{\mathbf{y}} \triangleq \mathbf{0}, \quad \bar{\mathbf{A}} \triangleq [\mathbf{A} \quad -\mathbf{I}_M], \quad \bar{\mathbf{x}} \triangleq \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix}, \quad \bar{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \epsilon \mathbf{I}_M) \text{ s.t. } \epsilon \rightarrow 0. \tag{10}$$

Here, \mathbf{x} is a priori independent of \mathbf{z} ; the dependence between \mathbf{x} and \mathbf{z} manifests only a posteriori, i.e., after the measurement $\bar{\mathbf{y}}$ is observed. For \mathbf{x} , we assign the prior $p(\mathbf{x}; \boldsymbol{\theta}_x)$, and for \mathbf{z} we assign the *improper* (i.e., unnormalized) prior $p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)$. The lack of normalization will not be an issue in GVAMP, because the ‘‘prior’’ $p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)$ is used only to compute posteriors of the form

$$p(\mathbf{z}|\mathbf{y}; \hat{\mathbf{p}}, \tau, \boldsymbol{\theta}_z) \propto p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)\mathcal{N}(\mathbf{z}; \hat{\mathbf{p}}, \mathbf{I}/\tau), \tag{11}$$

which are well defined because the right side is always integrable over \mathbf{z} .

Let us first consider direct ML estimation of θ in the above SLM. The θ -likelihood function is

$$\begin{aligned} p(\bar{\mathbf{y}}; \theta) &= \int p(\bar{\mathbf{y}}, \bar{\mathbf{x}}; \theta) d\bar{\mathbf{x}} = \int p(\bar{\mathbf{y}}|\bar{\mathbf{x}})p(\bar{\mathbf{x}}; \theta) d\bar{\mathbf{x}} = \int \mathcal{N}(\bar{\mathbf{y}}; \bar{\mathbf{A}}\bar{\mathbf{x}}, \epsilon\mathbf{I})p(\bar{\mathbf{x}}; \theta) d\bar{\mathbf{x}} \\ &= \int \underbrace{\mathcal{N}(\mathbf{z}; \mathbf{A}\mathbf{x}, \epsilon\mathbf{I})}_{\rightarrow \delta(\mathbf{z} - \mathbf{A}\mathbf{x})} p(\mathbf{x}; \theta_x)p(\mathbf{y}|\mathbf{z}; \theta_z) d\mathbf{x} d\mathbf{z}, \end{aligned} \tag{12}$$

which is consistent with (4) as $\epsilon \rightarrow 0$. Likewise, for any belief $b(\bar{\mathbf{x}})$, we can upper bound the negative log-likelihood by a Gibbs free energy $\bar{J}(b; \theta)$ of the form

$$\bar{J}(b; \theta) \triangleq D(b \| p(\bar{\mathbf{x}}; \theta)) + D(b \| p(\bar{\mathbf{y}}|\bar{\mathbf{x}})) + H(b), \tag{13}$$

since $-\ln p(\bar{\mathbf{y}}; \theta) = \bar{J}(b; \theta) - D(b \| p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta))$ with $D(b \| p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta)) \geq 0$. The corresponding EM algorithm is

$$\text{E step: } b^k(\bar{\mathbf{x}}) = p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \hat{\theta}^k) \tag{14}$$

$$\text{M step: } \hat{\theta}^{k+1} = \arg \min_{\theta} \bar{J}(b^k; \theta) = \arg \min_{\theta} D(b^k \| p(\bar{\mathbf{x}}; \theta)). \tag{15}$$

As before, the posterior density required by the E-step (14)

$$p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta) = \frac{p(\bar{\mathbf{x}}; \theta)p(\bar{\mathbf{y}}|\bar{\mathbf{x}})}{p(\bar{\mathbf{y}}; \theta)} = \frac{p(\mathbf{x}; \theta_x)p(\mathbf{y}|\mathbf{z}; \theta_z)\delta(\mathbf{z} - \mathbf{A}\mathbf{x})}{\int p(\mathbf{x}; \theta_x)p(\mathbf{y}|\mathbf{z}; \theta_z)\delta(\mathbf{z} - \mathbf{A}\mathbf{x}) d\mathbf{z} d\mathbf{x}}, \tag{16}$$

is difficult to compute due to the high-dimensional integral. Thus we consider an approximation afforded by the GVAMP algorithm [2], as described in the next section.

2.3 GVAMP

Recall that the exact posterior can (in principle) be found by solving the variational optimization problem

$$p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta) = \arg \min_b D(b \| p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta)) \tag{17}$$

$$= \arg \min_b \bar{J}(b; \theta) \tag{18}$$

$$= \arg \min_b D(b \| p(\bar{\mathbf{x}}; \theta)) + D(b \| p(\bar{\mathbf{y}}|\bar{\mathbf{x}})) + H(b), \tag{19}$$

where (18) follows from $\bar{J}(b; \theta) = D(b \| p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta)) - \ln p(\bar{\mathbf{y}}; \theta)$ and (19) follows from (13). But since the posterior computation problem is NP hard in general, (19) is no more tractable than any other approach. The GVAMP algorithm computes a posterior approximation using the expectation-consistent (EC) method [9, 10]. In this application of EC, we first split $b(\bar{\mathbf{x}})$ into three copies, i.e.,

$$p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \theta) = \arg \min_{b_1=b_2=q} D(b_1 \| p(\bar{\mathbf{x}}; \theta)) + D(b_2 \| p(\bar{\mathbf{y}}|\bar{\mathbf{x}})) + H(q), \tag{20}$$

and then relax the density-matching constraint $b_1 = b_2 = q$ to a moment-matching constraint:

$$p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \boldsymbol{\theta}) \approx \arg \min_{b_1, b_2, q} D(b_1 \| p(\bar{\mathbf{x}}; \boldsymbol{\theta})) + D(b_2 \| p(\bar{\mathbf{y}}|\bar{\mathbf{x}})) + H(q) \quad (21)$$

$$\begin{aligned} \text{s.t. } \mathbb{E}[\bar{\mathbf{x}}|b_1] &= \mathbb{E}[\bar{\mathbf{x}}|b_2] = \mathbb{E}[\bar{\mathbf{x}}|q] \text{ and } \text{tr}_2\{\text{Cov}[\bar{\mathbf{x}}|b_1]\} \\ &= \text{tr}_2\{\text{Cov}[\bar{\mathbf{x}}|b_2]\} = \text{tr}_2\{\text{Cov}[\bar{\mathbf{x}}|q]\}, \end{aligned} \quad (22)$$

where $\mathbb{E}[\bar{\mathbf{x}}|b_i]$ and $\text{Cov}[\bar{\mathbf{x}}|b_i]$ denote the expectation and covariance of $\bar{\mathbf{x}}$ under $\bar{\mathbf{x}} \sim b_i(\bar{\mathbf{x}})$, and where

$$\text{tr}_2 \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \right) \triangleq \begin{bmatrix} \text{tr}(\mathbf{A}) \\ \text{tr}(\mathbf{C}) \end{bmatrix} \text{ for } \mathbf{A} \in \mathbb{R}^{N \times N} \text{ and } \mathbf{C} \in \mathbb{R}^{M \times M}. \quad (23)$$

Essentially, $\text{tr}_2\{\text{Cov}[\bar{\mathbf{x}}]\}$ separately computes the trace of the covariance of \mathbf{x} and the trace of the covariance of \mathbf{z} . The right side of (21) yields three different approximations of the posterior:

$$b_1(\bar{\mathbf{x}}; \boldsymbol{\theta}) \propto p(\bar{\mathbf{x}}; \boldsymbol{\theta}) \mathcal{N} \left(\bar{\mathbf{x}}; \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{p}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{I}_N/\gamma_1 & \\ & \mathbf{I}_M/\tau_1 \end{bmatrix} \right) \quad (24)$$

$$b_2(\bar{\mathbf{x}}; \boldsymbol{\theta}) \propto p(\bar{\mathbf{y}}|\bar{\mathbf{x}}) \mathcal{N} \left(\bar{\mathbf{x}}; \begin{bmatrix} \mathbf{r}_2 \\ \mathbf{p}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{I}_N/\gamma_2 & \\ & \mathbf{I}_M/\tau_2 \end{bmatrix} \right) \quad (25)$$

$$q(\bar{\mathbf{x}}; \boldsymbol{\theta}) \propto \mathcal{N} \left(\bar{\mathbf{x}}; \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{z}} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_N/\eta & \\ & \mathbf{I}_M/\zeta \end{bmatrix} \right), \quad (26)$$

where the form of (24)–(26) can be deduced by analyzing the stationary points of the Lagrangian of (21), as shown in [9].

The GVAMP algorithm is an iterative approach to finding the values of $\mathbf{r}_1, \gamma_1, \mathbf{p}_1, \tau_1, \mathbf{r}_2, \gamma_2, \mathbf{p}_2, \tau_2, \hat{\mathbf{x}}, \eta, \hat{\mathbf{z}}, \zeta$ under which the three beliefs in (24)–(26) obey the moment constraints in (21). When \mathbf{A} is large and rotationally invariant, GVAMP is rigorously characterized by a state evolution [3]. Empirically, we find that the algorithm converges quickly in this scenario (e.g., on the order of 10 iterations).

Note that the values of $\mathbf{r}_1, \gamma_1, \mathbf{p}_1, \tau_1, \mathbf{r}_2, \gamma_2, \mathbf{p}_2, \tau_2, \hat{\mathbf{x}}, \eta, \hat{\mathbf{z}}, \zeta$ that satisfy the moment constraints are interdependent, and thus they all depend on the assumed value of $\boldsymbol{\theta}$ through (24).

2.4 EM-GVAMP

Recall that our current motivation for using GVAMP is to compute an approximation to the posterior $b^k(\bar{\mathbf{x}}) = p(\bar{\mathbf{x}}|\bar{\mathbf{y}}; \hat{\boldsymbol{\theta}}^k)$ in the EM algorithm (14) and (15). Of the three posterior approximations produced by GVAMP, the Gaussian approximation from (26) is the simplest to use for this purpose. Plugging the Gaussian

approximation into (14) and (15) yields

$$\text{E step: } b^k(\bar{\mathbf{x}}) = \mathcal{N}\left(\bar{\mathbf{x}}; \begin{bmatrix} \hat{\mathbf{x}}^k \\ \hat{\mathbf{z}}^k \end{bmatrix}, \begin{bmatrix} \mathbf{I}_N/\eta^k & \\ & \mathbf{I}_M/\zeta^k \end{bmatrix}\right) \text{ found via GVAMP with } \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^k \tag{27}$$

$$\text{M step: } \hat{\boldsymbol{\theta}}^{k+1} = \arg \min_{\boldsymbol{\theta}} D(b^k \| p(\bar{\mathbf{x}}; \boldsymbol{\theta})). \tag{28}$$

The difference between the EM algorithm (14), (15) and the EM algorithm (27) and (28) is that, in the former case, the bound is tight at each EM iteration k , whereas in the latter case the bound is only approximately tight.

Due to the form of $b^k(\bar{\mathbf{x}})$ in (27), the M-step is relatively easy to compute:

$$\hat{\boldsymbol{\theta}}^{k+1} = \arg \min_{\boldsymbol{\theta}} D(b^k \| p(\bar{\mathbf{x}}; \boldsymbol{\theta})) \tag{29}$$

$$= \arg \min_{\boldsymbol{\theta}} D(\mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}^k, \mathbf{I}_N/\eta^k) \mathcal{N}(\mathbf{z}; \hat{\mathbf{z}}^k, \mathbf{I}_M/\zeta^k) \| p(\mathbf{x}; \boldsymbol{\theta}_x) p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z)) \tag{30}$$

$$= \arg \max_{\boldsymbol{\theta}} \int \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}^k, \mathbf{I}_N/\eta^k) \ln p(\mathbf{x}; \boldsymbol{\theta}_x) d\mathbf{x} + \int \mathcal{N}(\mathbf{z}; \hat{\mathbf{z}}^k, \mathbf{I}_M/\zeta^k) \ln p(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_z) d\mathbf{z} \tag{31}$$

$$= \arg \max_{\boldsymbol{\theta}} \sum_{j=1}^N \int \mathcal{N}(x_j; \hat{x}_j^k, 1/\eta^k) \ln p(x_j; \boldsymbol{\theta}_x) dx_j + \sum_{i=1}^M \int \mathcal{N}(z_i; \hat{z}_i^k, 1/\zeta^k) \ln p(y_i|z_i; \boldsymbol{\theta}_z) dz_i. \tag{32}$$

The resulting $(\hat{\boldsymbol{\theta}}_x^{k+1}, \hat{\boldsymbol{\theta}}_z^{k+1})$ are necessarily values of $(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)$ that zero the gradient of the right side of (32) with respect to $\boldsymbol{\theta}_x$ and to $\boldsymbol{\theta}_z$.

3 Application to Noise-Variance Estimation in Phase Retrieval

In this section we will demonstrate how the EM procedure can be used to estimate noise variances in the context of phase retrieval. Noise variance estimation in this setting has also been performed in [11, 12]. The below derivation is related to, but distinct from, these previous works.

Phase retrieval is a problem that can be formulated in the GLM setting [11], allowing application of the GVAMP algorithm [13]. We denote the special case of GVAMP applied to phase retrieval as prVAMP.

One way to model the i th measured intensity y_i is via

$$y_i = |z_i + w_i| \text{ for i.i.d. } w_i \sim \mathcal{N}(0, \nu_w), \tag{33}$$

where $z_i, w_i \in \mathbb{C}$ and $\mathcal{N}(w_i; \mu, \nu) = \frac{1}{\pi\nu} \exp(-|w_i - \mu|^2/\nu)$ represents a circular complex-Gaussian density with mean $\mu \in \mathbb{C}$ and variance $\nu > 0$. In this case, the

measurement noise variance ν_w may be unknown in practice, and so we might try to estimate it using the methods described in this report. In that case, the unknown \mathbf{z} -likelihood parameters “ $\boldsymbol{\theta}_z$ ” reduce to ν_w . In the sequel, we will use the notation ν_w instead of $\boldsymbol{\theta}_z$.

It was shown [11] that, under (33), the z_i -likelihood function $p(y_i|z_i; \nu_w)$ takes the form

$$p(y_i|z_i; \nu_w) = 1_{y_i \geq 0} y_i \int_0^{2\pi} \mathcal{N}(y_i e^{j\theta_i}; z_i, \nu_w) d\theta_i \quad (34)$$

$$= \frac{2y_i}{\nu_w} \exp\left(-\frac{y_i^2 + |z_i|^2}{\nu_w}\right) I_0\left(\frac{2y_i|z_i|}{\nu_w}\right) 1_{y_i \geq 0}, \quad (35)$$

where $I_0(\cdot)$ is the 0th-order modified Bessel function of the first kind. If we view $p(y_i|z_i; \nu_w)$ as a density on y_i , then y_i is Rician (conditional on z_i). Note that θ_i above denotes the (hidden) phase on $z_i + w_i$; it should not be confused with the statistical parameters $\boldsymbol{\theta}$ described earlier in this paper.

From (32), we see that the EM estimate $\widehat{\nu}_w^{k+1}$ of ν_w must obey

$$0 = \frac{\partial}{\partial \nu_w} \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \widehat{z}_i^k, 1/\zeta^k) \ln p(y_i|z_i; \widehat{\nu}_w^{k+1}) dz_i \quad (36)$$

$$= \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \widehat{z}_i^k, 1/\zeta^k) \frac{\partial}{\partial \nu_w} \ln \int_0^{2\pi} \mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta_i dz_i \quad (37)$$

$$= \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \widehat{z}_i^k, 1/\zeta^k) \frac{\int_0^{2\pi} \frac{\partial}{\partial \nu_w} \mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta_i}{\int_0^{2\pi} \mathcal{N}(y_i e^{j\theta'_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta'_i} dz_i. \quad (38)$$

Plugging in the derivative expression (see [14])

$$\frac{\partial}{\partial \nu_w} \mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1}) = \frac{\mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1})}{2(\widehat{\nu}_w^{k+1})^2} (|y_i e^{j\theta_i} - z_i|^2 - \widehat{\nu}_w^{k+1}) \quad (39)$$

into (38) and multiplying both sides by $2(\widehat{\nu}_w^{k+1})^2$, we find

$$\widehat{\nu}_w^{k+1} = \frac{1}{M} \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \widehat{z}_i^k, 1/\zeta^k) \frac{\int_0^{2\pi} |y_i e^{j\theta_i} - z_i|^2 \mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta_i}{\int_0^{2\pi} \mathcal{N}(y_i e^{j\theta'_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta'_i} dz_i \quad (40)$$

$$= \frac{1}{M} \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \widehat{z}_i^k, 1/\zeta^k) \int_0^{2\pi} |y_i e^{j\theta_i} - z_i|^2 p(\theta_i; z_i, \widehat{\nu}_w^{k+1}) d\theta_i dz_i \quad (41)$$

with the newly defined pdf

$$p(\theta_i; z_i, \widehat{\nu}_w^{k+1}) \triangleq \frac{\mathcal{N}(y_i e^{j\theta_i}; z_i, \widehat{\nu}_w^{k+1})}{\int_0^{2\pi} \mathcal{N}(y_i e^{j\theta'_i}; z_i, \widehat{\nu}_w^{k+1}) d\theta'_i} \propto \exp\left(-\frac{|z_i - y_i e^{j\theta_i}|^2}{\widehat{\nu}_w^{k+1}}\right) \quad (42)$$

$$\propto \exp(\kappa_i \cos(\theta_i - \phi_i)) \text{ for } \kappa_i \triangleq \frac{2|z_i|y_i}{\widehat{\nu}_w^{k+1}}, \quad (43)$$

where ϕ_i denotes the phase of z_i . The expression (43) identifies this pdf as a von Mises distribution [15], which can be stated in normalized form as

$$p(\theta_i; z_i, \hat{\nu}_w^{k+1}) = \frac{\exp(\kappa_i \cos(\theta_i - \phi_i))}{2\pi I_0(\kappa_i)}. \tag{44}$$

Expanding the quadratic in (41) and plugging in (44), we get

$$\begin{aligned} \hat{\nu}_w^{k+1} &= \frac{1}{M} \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \hat{z}_i^k, 1/\zeta^k) \left(y_i^2 + |z_i|^2 \right. \\ &\quad \left. - 2y_i|z_i| \int_0^{2\pi} \cos(\theta_i - \phi_i) \frac{\exp(\kappa_i \cos(\theta_i - \phi_i))}{2\pi I_0(\kappa_i)} d\theta_i \right) dz_i \end{aligned} \tag{45}$$

$$= \frac{1}{M} \sum_{i=1}^M \int_{\mathbb{C}} \mathcal{N}(z_i; \hat{z}_i^k, 1/\zeta^k) \left(y_i^2 + |z_i|^2 - 2y_i|z_i| R_0 \left(\frac{2|z_i|y_i}{\hat{\nu}_w^{k+1}} \right) \right) dz_i, \tag{46}$$

where $R_0(\cdot)$ is the modified Bessel function ratio $R_0(\kappa_i) \triangleq I_1(\kappa_i)/I_0(\kappa_i)$ and (46) follows from [16, 9.6.19].

Simplifying approximations of (46) could be taken as needed. For example, in the high-SNR case, the expansion $R_0(\kappa) = 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} + o(\kappa^{-3})$ from [17, Lemma 5] could be used to justify

$$R_0(\kappa) \approx 1 - \frac{1}{2\kappa}, \tag{47}$$

which, when applied to (46), yields

$$\hat{\nu}_w^{k+1} \approx \frac{2}{M} \sum_{i=1}^M \int_{\mathbb{C}} (y_i - |z_i|)^2 \mathcal{N}(z_i; \hat{z}_i^k, 1/\zeta^k) dz_i. \tag{48}$$

Approximation (48) can be reduced to an expression that involves the mean of a Rician distribution. In particular, using $z_i = \rho_i e^{j\phi_i}$, the integral in (48) can be converted to polar coordinates as follows:

$$\begin{aligned} \int_0^\infty (y_i - \rho_i)^2 \underbrace{\int_0^{2\pi} \mathcal{N}(\rho_i e^{j\phi_i}; \hat{z}_i^k, 1/\zeta^k) d\phi_i}_{\frac{2\rho_i}{1/\zeta^k} \exp\left(-\frac{\rho_i^2 + |\hat{z}_i^k|^2}{1/\zeta^k}\right) I_0\left(\frac{2\rho_i|\hat{z}_i^k|}{1/\zeta^k}\right)} d\rho_i &= y_i^2 - 2y_i\mathbb{E}[\rho_i] + \mathbb{E}[\rho_i^2], \tag{49} \\ &1_{\rho_i \geq 0} \end{aligned}$$

where, for the expectations, ρ_i has the Rician density under the brace. For this density, it is known that

$$\mathbb{E}[\rho_i] = \sqrt{\frac{\pi}{4\zeta^k}} L_{1/2}(-\zeta^k|\hat{z}_i^k|^2) \tag{50}$$

$$\mathbb{E}[\rho_i^2] = 1/\zeta^k + |\hat{z}_i^k|^2, \tag{51}$$

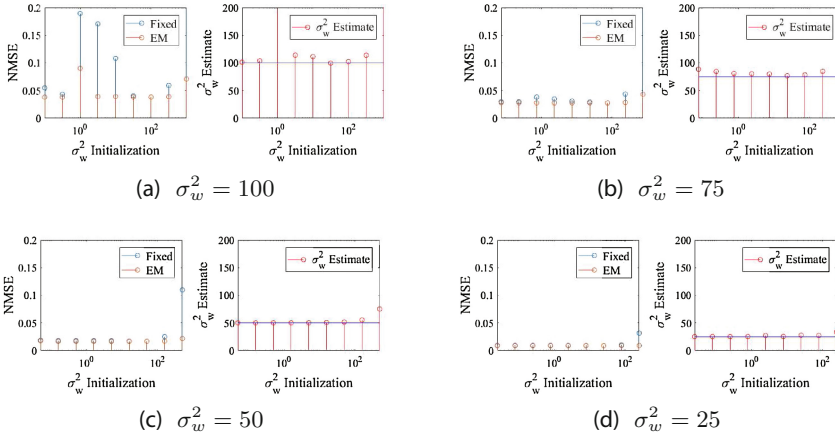


Fig. 2. Reconstruction errors (left subplots) and estimates of σ_w^2 (right subplots) with different initial estimates of σ_w^2 . The EM procedure is capable of estimating the true noise variance over a range of operating conditions. Using this estimate of the noise variance incrementally improves recovery accuracy.

where the Laguerre polynomial $L_{1/2}(x)$ can be computed as

$$L_{1/2}(x) = \exp\left(\frac{x}{2}\right) \left[(1-x)I_0\left(-\frac{x}{2}\right) - xI_1\left(-\frac{x}{2}\right) \right]. \tag{52}$$

Note that, for reasons of numerical precision, $\exp(x/2)I_d(-x/2)$ is computed using “`besseli(d, -x/2, 1)`” in Matlab, not “`exp(x/2) .* besseli(d, -x/2)`.”

4 Simulations

In this section, we demonstrate the effectiveness of the EM procedure in simulation. In particular, we show how EM can approximately recover the noise variance even when initialized by estimates far from the ground truth. This in turn enables improved signal reconstruction when the noise variance is apriori unknown.

We set up our simulations as follows. We aim to recover an i.i.d. circular Gaussian random vector $x \in \mathbb{C}^n$, with variance $\sqrt{2}$, from phaseless noisy measurements of the form $y = |\mathbf{A}x + w|$. Our measurement matrix \mathbf{A} is 8192×1024 and the elements of A are i.i.d. circular Gaussian with variance $\sqrt{2}$. The elements of the noise vector w also follow an i.i.d. circular Gaussian distribution, but with variance σ_w^2 . We test the cases of $\sigma_w^2 = 100$, $\sigma_w^2 = 75$, $\sigma_w^2 = 50$, and $\sigma_w^2 = 25$. prVAMP was provided with initial estimates of σ_w^2 ranging from 1% to 10× the true variance. Using these initializations, we reconstructed the signal with and without the EM procedure.

Figure 2 presents our reconstructions. The results demonstrate that EM can be used to estimate σ_w^2 . Moreover, it shows that this estimate lets prVAMP accurately reconstruct the signal even when σ_w is not known apriori.

Code demonstrating the EM procedure will be made available at http://gampmatlab.wikia.com/wiki/Generalized_Approximate_Message_Passing.

5 Conclusion

This paper combines EM and GVAMP to estimate the unknown channel parameters associated with GLMs. This in turn enables GVAMP to estimate signals from their generalized linear measurements. In this paper we applied the proposed technique to phase retrieval and showed that it is effective at estimating unknown noise variances, thus enabling noise robust phase retrieval over a range of operating conditions.

References

1. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall, New York (1989)
2. Schniter, P., Rangan, S., Fletcher, A.K.: Vector approximate message passing for the generalized linear model. In: Asilomar Conference on Signals, Systems, and Computers, pp. 1525–1529 (2016)
3. Fletcher, A.K., Rangan, S., Schniter, P.: Inference in deep networks in high dimensions. arXiv preprint [arXiv:1706.06549](https://arxiv.org/abs/1706.06549) (2017)
4. Rangan, S., Schniter, P., Fletcher, A.K.: Vector approximate message passing. In: Proceedings of the IEEE ISIT, pp. 1588–1592 (2017)
5. Dempster, A., Laird, N.M., Rubin, D.B.: Maximum-likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–17 (1977)
6. Fletcher, A.K., Schniter, P.: Learning and free energies for vector approximate message passing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4247–4251 (2017)
7. Fletcher, A.K., Sahraee-Ardakan, M., Rangan, S., Schniter, P.: Rigorous dynamics and consistent estimation in arbitrarily conditioned linear systems. In: Proceedings of NIPS, pp. 2542–2551 (2017)
8. 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*, pp. 355–368. MIT Press, Cambridge (1998)
9. Opper, M., Winther, O.: Expectation consistent approximate inference. *J. Mach. Learn. Res.* **1**, 2177–2204 (2005)
10. Fletcher, A.K., Sahraee-Ardakan, M., Rangan, S., Schniter, P.: Expectation consistent approximate inference: generalizations and convergence. In: Proceedings of IEEE ISIT, pp. 190–194 (2016)
11. Schniter, P., Rangan, S.: Compressive phase retrieval via generalized approximate message passing. *IEEE Trans. Signal Process.* **63**(4), 1043–1055 (2015)
12. Drémeau, A., Krzakala, F.: Phase recovery from a Bayesian point of view: the variational approach. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3661–3665. IEEE (2015)
13. Metzler, C.A., Sharma, M.K., Nagesh, S., Baraniuk, R.G., Cossairt, O., Veeraraghavan, A.: Coherent inverse scattering via transmission matrices: efficient phase retrieval algorithms and a public dataset. In: Proceedings of the International Conference on Computational Photography (ICCP) (2017)

14. Vila, J.P., Schniter, P.: Expectation-maximization Gaussian-mixture approximate message passing. *IEEE Trans. Signal Process.* **61**, 4658–4672 (2013)
15. Mardia, K.V., Jupp, P.E.: *Directional Statistics*. Wiley, New York (2000)
16. Abramowitz, M., Stegun, I.A. (eds.): *Handbook of Mathematical Functions*. Dover, New York (1964)
17. Robert, C.: Modified Bessel functions and their applications in probability and statistics. *Stat. Prob. Lett.* **9**, 155–161 (1990)



A Study on the Benefits of Phase-Aware Speech Enhancement in Challenging Noise Scenarios

Martin Krawczyk-Becker^(✉) and Timo Gerkmann

Universität Hamburg, 20148 Hamburg, Germany
{martin.krawczyk-becker,timo.gerkmann}@uni-hamburg.de

Abstract. In recent years, there has been a renaissance of research on the role of the spectral phase in single-channel speech enhancement. One of the recent proposals is to not only estimate the clean speech phase but also use this phase estimate as an additional source of information to facilitate the estimation of the clean speech magnitude. To assess the potential benefit of such approaches, in this paper we systematically explore in which situations additional information about the clean speech phase is most valuable. For this, we compare the performance of phase-aware and phase-blind clean speech estimators in different noise scenarios, i.e. at different signal to noise ratios (SNRs) and for noise sources with different degrees of stationarity. Interestingly, the results indicate that the greatest benefits can be achieved in situations where conventional magnitude-only speech enhancement is most challenging, namely in highly non-stationary noises at low SNRs.

Keywords: Phase · Speech enhancement · Noise reduction

1 Introduction

The enhancement of speech that is corrupted by noise is a long-standing research topic that has seen many new ideas and improvements over the last decades. In this paper, we focus on single-channel speech enhancement, i.e. approaches that are applied to a single microphone signal or to the output of a multichannel preprocessing stage. Specifically, we consider minimum mean square error (MMSE) optimal Bayesian estimators of the clean speech in the short-time discrete Fourier transform (STFT) domain. Well-known examples of this class of estimators are the Wiener filter and Ephraim and Malah's short-time spectral amplitude estimator (STSA) [2]. Over the years, numerous improvements have been proposed, including the use of super-Gaussian speech priors [4, 21] and/or different optimization criteria [1, 3, 28]. See e.g. [13] for a concise overview. What the vast majority of mainstream approaches have in common is that they are magnitude-centric, meaning that the spectral phase is neither used as a source of information nor is the noise corrupted spectral phase enhanced, which is frequently justified by the statement that the enhancement of the spectral phase

is unimportant [27]. However, contrary to the widespread believe at the time, more recent studies, including [9, 24], showed that the spectral phase is indeed important for speech enhancement. These findings sparked a renewed interest in the estimation of the clean speech spectral phase for speech enhancement, e.g. [10, 16, 22].

With the availability of phase estimates, also the interest in how these phase estimates can best be utilized has risen. A straight forward way is to simply exchange the noisy phase with the phase estimate and combine it with a clean speech magnitude estimate that has been obtained with an existing of-the-shelf estimator. A more elaborate way to utilize the newly available phase estimate is to use it as an additional source of information that facilitates the estimation of the clean speech magnitudes [8, 18] or even the complex-valued coefficients [6]. We denote such approaches as being *phase-aware*, while conventional magnitude-centric approaches like the Wiener filter or the STSA are considered *phase-blind*.

Phase-aware approaches have been shown to be capable of generally outperforming their phase-blind counterparts in terms of instrumental measures, e.g. in [8, 18, 23], and also by means of formal listening experiments [17]. To assess the potential of phase-aware speech enhancement in more detail, in this paper we systematically investigate in which acoustic situations it provides the largest benefits. For this, we directly compare the performance of two phase-aware estimators based on [6, 18] to that of their phase-blind counterparts, namely the STSA and the Wiener filter, at different SNRs and for noise sources with different degrees of stationarity. First, we consider pink noise and modulate it with an increasing modulation frequency, which allows us to adjust the amount of non-stationarity in a very controlled way. As a second, practically very relevant example, we use babble noise, where the non-stationarity is adjusted by deliberately changing the number of talkers. The results indicate that the greatest benefits can be achieved in situations where conventional phase-blind speech enhancement is most challenging, i.e. in highly non-stationary noises at low SNRs.

2 Signal Model and Notation

In each time-frequency point of the STFT domain we have a additive superposition of mutually independent speech and noise,

$$Y = S + V = Ae^{j\phi^S} + De^{j\phi^V} = Re^{j\phi^Y}, \quad (1)$$

where Y , S , and V denote the complex coefficients of the observed noisy speech, the desired clean speech, and the additive noise, respectively. The spectral phases are denoted by ϕ^Y , ϕ^S , and ϕ^V , while the spectral magnitudes are denoted by R , A , and D . Here we make the common assumption that the noise coefficients V follow a circular symmetric zero-mean complex Gaussian distribution with a power spectral density (PSD) of σ_v^2 , where the circular symmetry implicates a uniformly distributed noise phase ϕ^V . The PSD of speech is denoted as σ_s^2 . We use the hat-symbol to distinguish estimates from their true counterparts, i.e. \hat{S} is an estimate of S .

3 Conventional Phase-Blind Clean Speech Estimation

Commonly, MMSE estimators of the clean speech S , or more generally any function $f(S)$, are derived by finding the expected value of $f(S)$ given the noisy observation and the PSDs of speech and noise:

$$\widehat{f(S)} = \mathbb{E}(f(S) | Y, \sigma_s^2, \sigma_v^2) = \int_0^\infty \int_0^{2\pi} f(S) p(A, \Phi^S | Y, \sigma_s^2, \sigma_v^2) d\Phi^S dA \quad (2)$$

$$= \frac{\int_0^\infty \int_0^{2\pi} f(S) p(y|A, \Phi^S, \sigma_v^2) p(A | \sigma_s^2) p(\Phi^S) d\Phi^S dA}{\int_0^\infty \int_0^{2\pi} p(y|A, \Phi^S, \sigma_v^2) p(A | \sigma_s^2) p(\Phi^S) d\Phi^S dA}, \quad (3)$$

where the second line is obtained by applying Bayes' rule. For complex Gaussian noise, the likelihood is given as $p(y|A, \Phi^S, \sigma_v^2) = \mathcal{N}(S, \sigma_v^2)$, see e.g. [2]. For a uniform *phase* prior, i.e. $p(\Phi^S) = 1/(2\pi)$ for $\Phi^S \in [-\pi, \pi)$, Eq. (3) has been solved analytically for different *magnitude* priors $p(A | \sigma_s^2)$ and functions $f(S)$. Assuming a Rayleigh distribution for $p(A | \sigma_s^2)$, for instance, the Wiener filter is obtained as the MMSE optimal estimator of the complex clean speech coefficients ($f(S) = S$) and Ephraim and Malah's STSA as the MMSE optimal estimators of the clean speech magnitudes ($f(S) = A$). Also more elaborate super-Gaussian clean speech estimators have been derived via (3) by using, e.g., a χ distribution [1] or a generalized gamma distribution [4] for $p(A | \sigma_s^2)$ with different functions $f(S)$. However, in all these approaches the phase prior $p(\Phi^S)$ is modeled as a uniform distribution, which implies that the complex clean speech coefficients are circularly-complex distributed. Without any prior information about the clean speech spectral phase, the uniform distribution is indeed a reasonable assumption that is supported by long term histogram data [4].

4 Phase-Aware Clean Speech Estimation

In contrast to the conventional phase-blind approaches discussed above, phase-aware estimators such as the ones in [6, 8, 18] assume that besides the speech and noise PSDs also a prior estimate of the clean speech spectral phase is available. Such a phase estimate can be obtained from the noisy signal for instance based on a harmonic model such as in [16, 22] or using an iterative approach similar to Griffin and Lim [12] and its successors [19, 25]. To derive MMSE optimal phase-aware estimators, we propose to compute the expected value of $f(S)$ conditioned not only on Y , σ_s^2 , and σ_v^2 as for conventional estimators, but also on the prior phase estimate $\widehat{\Phi^S}$ [6, 8, 18]:

$$\widehat{f(S)} = \mathbb{E}(f(S) | Y, \sigma_s^2, \sigma_v^2, \widehat{\Phi^S}) \quad (4)$$

$$= \frac{\int_0^\infty \int_0^{2\pi} f(S) p(y|A, \Phi^S, \sigma_v^2) p(A | \sigma_s^2) p(\Phi^S | \widehat{\Phi^S}) d\Phi^S dA}{\int_0^\infty \int_0^{2\pi} p(y|A, \Phi^S, \sigma_v^2) p(A | \sigma_s^2) p(\Phi^S | \widehat{\Phi^S}) d\Phi^S dA}, \quad (5)$$

where the second line is again obtained using Bayes' rule and making only mild assumptions. Comparing the phase-aware estimator in (5) and the phase-blind estimator in (3), it can be seen that the only difference is the replacement of $p(\Phi^S)$ by $p(\Phi^S|\widehat{\Phi}^S)$. If the prior phase estimate $\widehat{\Phi}^S$ is informative, the true clean speech phase Φ^S does not follow a uniform distribution anymore. Instead, $p(\Phi^S|\widehat{\Phi}^S)$ reflects uncertain information about the true clean speech phase. Similar to [6, 18] we employ a von Mises distribution with mean direction $\widehat{\Phi}^S$ to model this uncertainty in the prior phase estimate:

$$p(\Phi^S|\widehat{\Phi}^S) = \exp(\kappa \cos(\Phi^S - \widehat{\Phi}^S)) / (2\pi I_0(\kappa)), \quad (6)$$

where κ is the concentration parameter and $I_0(\cdot)$ is the modified Bessel function of the first kind and zeroth-order. Examples for $p(\Phi^S|\widehat{\Phi}^S)$ for $\widehat{\Phi}^S = 0$ and different concentration parameters κ are presented in Fig. 1. The larger κ , the more $p(\Phi^S|\widehat{\Phi}^S)$ is concentrated around the prior phase estimate $\widehat{\Phi}^S$. Accordingly, $\widehat{\Phi}^S$ is modeled as an increasingly accurate estimate of the true clean speech phase Φ^S . For the extreme case of $\kappa \rightarrow \infty$, the distribution reduces to a single peak at $\widehat{\Phi}^S$, i.e. the prior phase estimate is assumed to be exactly the true clean speech phase Φ^S . On the contrary, the lower κ , the wider $p(\Phi^S|\widehat{\Phi}^S)$, which corresponds to modeling less accurate prior estimates. For $\kappa = 0$, $p(\Phi^S|\widehat{\Phi}^S)$ reduces to a uniform distribution and the prior phase estimate $\widehat{\Phi}^S$ does not provide any useful information about the true phase Φ^S , i.e. $p(\Phi^S|\widehat{\Phi}^S) = p(\Phi^S)$. In this special case, the phase-aware estimator in (5) degenerates to a conventional phase-blind estimator similar to (3).

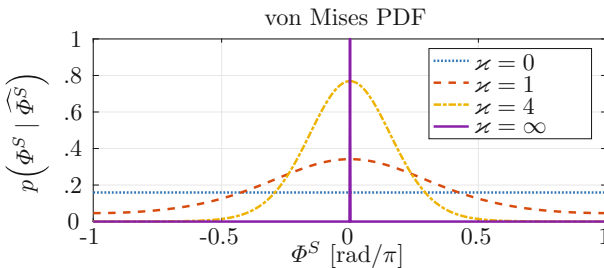


Fig. 1. Von Mises distribution for $p(\Phi^S|\widehat{\Phi}^S)$ with a mean direction of $\widehat{\Phi}^S = 0$ and an increasing concentration parameter κ .

Very general super-Gaussian phase-aware estimators of the clean speech spectral magnitudes $f(S) = A^\beta$ and the complex clean speech coefficients $f(S) = A^\beta e^{j\phi^S}$ have been derived in [6, 18] by solving (5) using a flexible χ distribution for the magnitude prior $p(A | \sigma_s^2)$. The magnitude estimator uses the prior phase estimate $\widehat{\Phi}^S$ only to facilitate the estimation of the clean speech magnitude. Similar to the phase-blind estimators, the spectral phase is not modified and the estimated magnitude is combined with the noisy phase Φ^Y to obtain the final estimate. The complex estimator, however, not only enhances the spectral magnitude but also jointly enhances the spectral phase.

For simplicity, in this paper we consider only two special cases of the general phase-aware estimators in [6, 18]. Specifically, we set $\beta = 1$, i.e. we estimate $f(S) = Ae^{j\phi^S} = S$ and $f(S) = A$, and choose the parameter of the χ distribution such that it reduces to a Rayleigh distribution. Both, the simplified estimator of the clean speech coefficients $f(S) = S$ as well as the simplified estimator of the clean speech magnitudes $f(S) = A$ have well-known phase-blind counterparts: If the prior phase estimate is assumed to provide no useful information, i.e. $\varkappa = 0$, it has been shown in [18] that the simplified estimator of S reduces to the Wiener filter, while the simplified magnitude estimators reduces to the STSA [2]. This direct relation between phase-aware and well-known phase-blind estimators allow to investigate the effects of phase-aware speech enhancement in isolation. We denote the simplified phase-aware magnitude estimator ($f(S) = A$) as PAM and the simplified phase-aware complex estimator ($f(S) = S$) as PAC.

5 Evaluation

In this section, we evaluate in which acoustic scenarios phase-aware speech enhancement is most beneficial. For this, the two simplified phase-aware clean speech estimators are compared to their respective phase-blind counterparts. Specifically, we compare Ephraim and Malah’s STSA [2] to the PAM and the conventional Wiener filter to the PAC. The evaluation is performed on 128 gender balanced utterances taken from the TIMIT database [5] at a sampling rate of 16 kHz. In the first part, the clean speech utterances are deteriorated by stationary pink noise and pink noise modulated with an increasing modulation frequency, i.e. 0.5 Hz, 1 Hz, and 2 Hz. This allows us to investigate how the performance of phase-aware speech enhancement depends on the non-stationarity of the noise in a very controlled manner. Furthermore, to assess the influence of the SNR on phase-aware speech enhancement, this experiment is conducted for two SNRs, namely 0 dB and 10 dB. We present three measures: global SNR, raw wideband ‘Perceptual Evaluation of Speech Quality’ (WB-PESQ) scores [15], and raw ‘Short-Time Objective Intelligibility Measure’ (STOI) values [26].

As a less controlled but practically very relevant example, in the second experiment we deteriorate the clean speech utterances with babble noise, where the amount of non-stationarity is controlled by the number of speakers. The noise is created by randomly superimposing TIMIT sentences that have not been used as clean speech material, with the number of speakers ranging from 40, which

represents the most stationary example, to a single competing talker as the most non-stationary noise.

In both experiments, the STFT representation is obtained with a segment length of 32 ms and a segment shift of 8 ms with a square-root Hann window for spectral analysis and synthesis without zero padding. The speech PSD is estimated using the decision-directed approach [2] with a smoothing parameter of 0.96. The noise PSD is estimated via [7]. The maximum attenuation in each time frequency point is set to -15 dB, which is a common way to reduce artifacts in the enhanced signal by introducing a residual noise floor. To assess the full potential of phase-aware speech enhancement without the shortcomings of current phase estimators, here the true clean speech phase Φ^S is provided as the prior phase estimate $\widehat{\Phi}^S$. The concentration parameter in (6) is accordingly set to $\kappa \rightarrow \infty$. Please note that for this specific choice of \varkappa , the only difference between PAM and PAC is that PAC combines the magnitude estimate with the noisy phase, while PAM uses the prior phase $\widehat{\Phi}^S$. In practice, the clean speech phase is however not available. Therefore, we finally also present results for the case that the prior phase is blindly estimated via [16] to further confirm the outcome of the oracle experiments.

5.1 Modulated Pink Noise

In Fig. 2, we present global SNR, WB-PESQ, and STOI for pink noise with an increasing amount of non-stationarity. For a better accessibility, we do not present absolute values but rather the improvement of the phase-aware estimator over its conventional phase-blind counterpart. First, it can be seen that the benefit of phase-aware speech enhancement is generally larger at low SNRs (top) than at higher SNRs (bottom). Second, independent of the SNR, the benefit of phase-aware speech enhancement increases with increasing non-stationarity. Generally, in non-stationary noises at low SNRs, speech enhancement is most challenging, specifically because the estimation of the speech PSD σ_s^2 and the noise PSD σ_v^2 becomes increasingly difficult. For instance, most noise PSD estimators, including minimum statistics [20] and the estimator based on speech presence probability [7] that is employed here, rely on the assumption that noise is more stationary than speech. Such approaches consequently become less accurate for highly non-stationary noise. This is also reflected in an increasing log distortion error [14]

$$\text{LOG-Err}_{\text{seg}} = \text{mean} \left| 10 \log_{10} \frac{\sigma_v^2}{\sigma_s^2} \right|, \quad (7)$$

where the mean is taken over all time-frequency points. In the noise-only case, $\text{LOG-Err}_{\text{seg}}$ gradually increases from 1.5 for stationary pink noise to 3.1 for pink noise modulated with 2 Hz. See e.g. [7, 14] for a more detailed discussion on this topic.

Since the conventional phase-blind estimators (3) rely solely on the PSD estimates, PSD estimation errors propagate through to the final estimate, leading to

noise leakage and/or speech distortions. These artifacts can substantially reduce the speech enhancement performance. The fact that phase-aware speech enhancement provides the most benefit specifically in these situations highlights its relevance and potential. Furthermore, comparing the complex estimator PAC to the magnitude estimator PAM it can be seen that the phase enhancement of PAC leads to an additional improvement in all three measures and all acoustic situations.

The consistently small gains in STOI at 10 dB SNR on the bottom right of Fig. 2 can be explained by the fact that the speech intelligibility, which is predicted by STOI, is already close to 100% even for the noisy signal at high SNRs. Thus there is only little room for improvement.

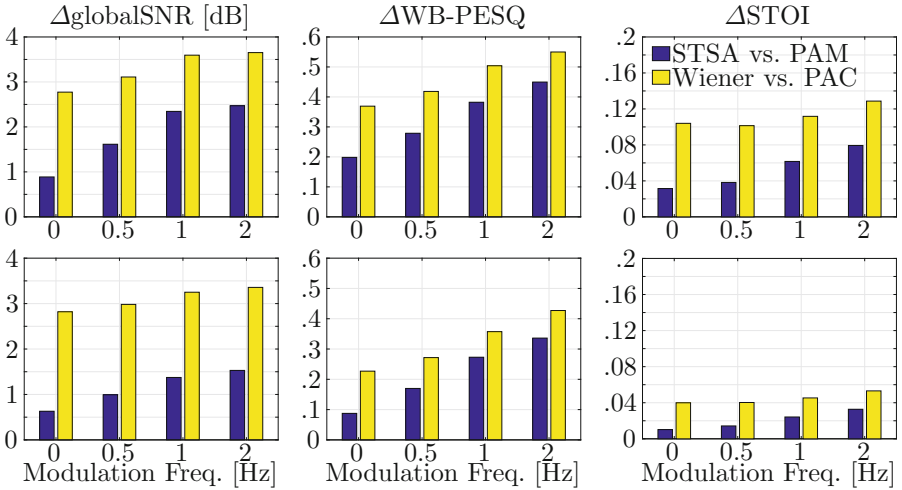


Fig. 2. Improvement of the phase-aware estimators over their respective phase-blind counterparts in SNR, WB-PESQ, and STOI for pink noise modulated with an increasing modulation frequency, representing an increasing degree of non-stationarity. SNR: 0 dB (top) and 10 dB (bottom).

5.2 Babble Noise

In Fig. 3, we present the results for babble noise with a varying number of talkers. The fewer talkers the noise is comprised of, the less stationary it is. Similar to the first experiment in Fig. 2, the benefit of phase-aware speech enhancement is most prominent in highly non-stationary noise at low SNRs. The largest improvements are achieved for 5-talker babble noise, while for a single interfering talker the improvement in WB-PESQ and STOI is somewhat lower, especially for the complex estimator PAC.

Finally, in Fig. 4, we present results that are achieved when the prior phase $\widehat{\phi}^S$ is estimated blindly on the noisy signal Y via [16]. No oracle information is used. Although the improvements are substantially smaller than for the oracle

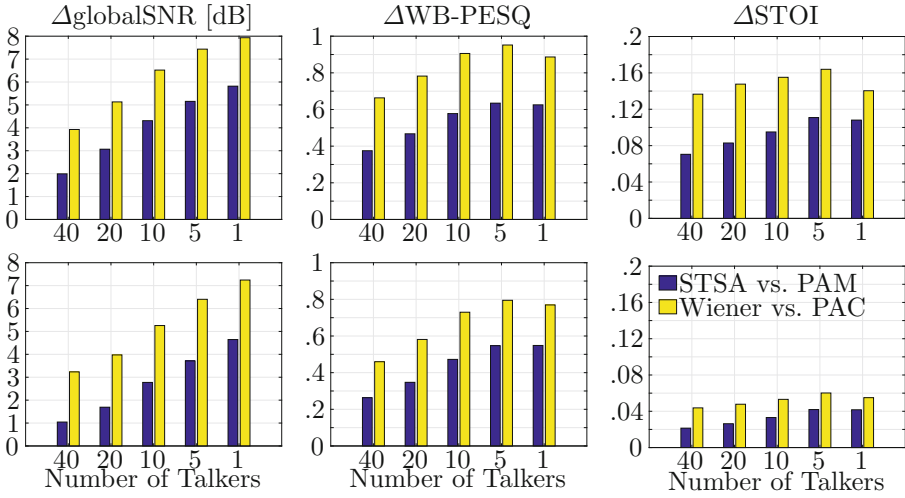


Fig. 3. Improvement of the phase-aware estimators over their respective phase-blind counterparts in SNR, WB-PESQ, and STOI for babble noise with a decreasing number of talkers, representing an increasing degree of non-stationarity. SNR: 0 dB (top) and 10 dB (bottom).

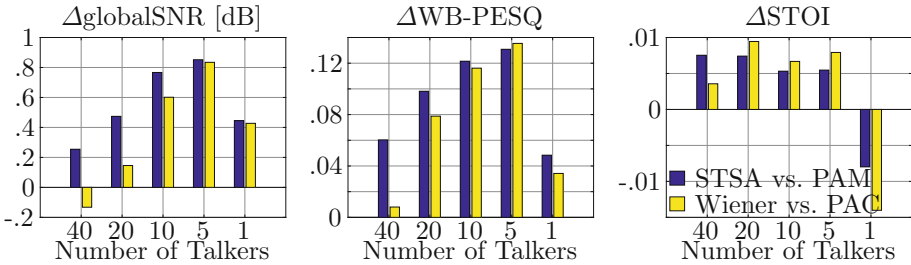


Fig. 4. Improvement of the phase-aware estimators over their respective phase-blind counterparts in SNR, WB-PESQ, and STOI for babble noise with a decreasing number of talkers at 0 dB SNR. The prior phase is blindly estimated via [16].

experiment in Fig. 3, similar trends can be observed. While the improvement in STOI is generally small, SNR and WB-PESQ improvements are again the largest for the highly non-stationary 5-talker babble. The reduced performance for the single competing talker is likely a consequence of the phase estimation process: In [16], the spectral phase of voiced speech is estimated based on a harmonic signal model, which in turn relies on a fundamental frequency estimate obtained via [11]. For a single competing talker at 0 dB SNR, estimating the fundamental frequency only of the desired speaker becomes extremely challenging. Thus the prior phase estimate can strongly deviate from the true clean speech phase in

this situation and might even resemble the phase of the competing talker at times, which limits the overall speech enhancement performance.

While the impressive performance gains for the oracle experiments in Figs. 2 and 3 clearly highlight the potential of phase-aware speech enhancement, the current gap between the oracle performance and the one in Fig. 4 makes research into more robust and accurate phase estimation techniques a relevant and promising topic for single-channel speech enhancement.

6 Conclusions

In this paper, we investigated in which situations additional information of the clean speech phase is most valuable. The results show that the greatest benefits can be achieved in situations where conventional magnitude-only speech enhancement is most challenging, namely in highly non-stationary noises at low SNRs. The current gap between the optimal performance of phase-aware speech enhancement and the performance obtained using blindly estimated prior phases highlight the importance of ongoing research into robust and accurate phase estimation techniques.

References

1. Breithaupt, C., Gerkmann, T., Martin, R.: A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, pp. 4897–4900 (2008)
2. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984)
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. Erkelens, J.S., Hendriks, R.C., Heusdens, R., Jensen, J.: Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors. *IEEE Trans. Audio Speech Lang. Process.* **15**(6), 1741–1752 (2007)
5. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.: DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993)
6. Gerkmann, T.: Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase. *IEEE Trans. Signal Process.* **62**(16), 4199–4208 (2014)
7. Gerkmann, T., Hendriks, R.C.: Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1383–1393 (2012)
8. Gerkmann, T., Krawczyk, M.: MMSE-optimal spectral amplitude estimation given the STFT-phase. *IEEE Signal Process. Lett.* **20**(2), 129–132 (2013)
9. Gerkmann, T., Krawczyk, M., Rehr, R.: Phase estimation in speech enhancement – unimportant, important, or impossible? In: IEEE Convention of Electrical and Electronics Engineers in Israel, Eilat, Israel (2012)

10. Gerkmann, T., Krawczyk-Becker, M., Le Roux, J.: Phase processing for single channel speech enhancement: history and recent advances. *IEEE Signal Process. Mag.* **32**(2), 55–66 (2015)
11. Gonzalez, S., Brookes, M.: PEFACT - a pitch estimation algorithm robust to high levels of noise. *IEEE Trans. Audio Speech Lang. Process.* **22**(2), 518–530 (2014)
12. Griffin, D.W., Lim, J.S.: Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **32**(2), 236–243 (1984)
13. Hendriks, R.C., Gerkmann, T., Jensen, J.: DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art. Morgan & Claypool, Colorado (2013)
14. Hendriks, R.C., Jensen, J., Heusdens, R.: Noise tracking using DFT domain subspace decompositions. *IEEE Trans. Audio Speech Lang. Process.* **16**(3), 541–553 (2008)
15. ITU-T: Perceptual evaluation of speech quality (PESQ). ITU-T Recommendation P.862 (2001)
16. Krawczyk, M., Gerkmann, T.: STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(12), 1931–1940 (2014)
17. Krawczyk-Becker, M., Gerkmann, T.: An evaluation of the perceptual quality of phase-aware single-channel speech enhancement. *J. Acoust. Soc. Am.* **140**(4), EL364–EL369 (2016)
18. Krawczyk-Becker, M., Gerkmann, T.: On MMSE-based estimation of spectral speech coefficients under phase-uncertainty. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(12), 2251–2262 (2016)
19. Le Roux, J., Vincent, E.: Consistent Wiener filtering for audio source separation. *IEEE Signal Process. Lett.* **20**(3), 217–220 (2013)
20. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* **9**(5), 504–512 (2001)
21. Martin, R.: Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.* **13**(5), 845–856 (2005)
22. Mowlae, P., Kulmer, J.: Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(9), 1521–1532 (2015)
23. Mowlae, P., Saeidi, R.: Iterative closed-loop phase-aware single-channel speech enhancement. *IEEE Signal Process. Lett.* **20**(12), 1235–1239 (2013)
24. Paliwal, K., Wójcicki, K., Shannon, B.: The importance of phase in speech enhancement. *ELSEVIER Speech Commun.* **53**(4), 465–494 (2011)
25. Sturmel, N., Daudet, L.: Signal reconstruction from STFT magnitude: a state of the art. In: International Conference on Digital Audio Effects (DAFx), Paris, France, pp. 375–386 (2011)
26. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2125–2136 (2011)
27. Wang, D.L., Lim, J.S.: The unimportance of phase in speech enhancement. *IEEE Trans. Acoust. Speech Signal Process.* **30**(4), 679–681 (1982)
28. You, C.H., Koh, S.N., Rahardja, S.: β -order MMSE spectral amplitude estimation for speech enhancement. *IEEE Trans. Speech Audio Process.* **13**(4), 475–486 (2005)



Phase Reconstruction for Time-Frequency Inpainting

A. Marina Krémé^{1,2}(✉), Valentin Emiya², and Caroline Chaux¹

¹ Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France

² Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
{A.Marina.Kreme,Valentin.Emiya,Caroline.Chaux}@univ-amu.fr

Abstract. We address the problem of phase inpainting, i.e. the reconstruction of partially-missing phases in linear measurements. We thus aim at reconstructing missing phases of some complex coefficients assuming that the phases of the other coefficients as well as the modulus of all coefficients are known. The mathematical formulation of the inverse problem is first described and then, three methods are proposed: a first one based on the well known Griffin and Lim algorithm and two other ones based on positive semidefinite programming (SDP) optimization methods namely PhaseLift and PhaseCut, that are extended to the case of partial phase knowledge. The three derived algorithms are tested with measurements from a short-time Fourier transform (STFT) in two situations: the case where the missing data are distributed uniformly and independently at random and the case where they constitute holes with a given width. Results show that the knowledge of a subset of phases contributes to improve the signal reconstruction and to shorten the convergence of the optimization process.

Keywords: Audio · Time-frequency · Missing data · Inpainting
Phase reconstruction · SDP optimization
Short-time Fourier transform · PhaseLift · PhaseCut

1 Introduction

Time-frequency inpainting is an inverse problem where the goal is to estimate a subset of masked coefficients in a time-frequency complex-valued matrix from the observation of the remaining coefficients. A natural strategy consists in performing a spectrogram inpainting stage, where the amplitude of the missing coefficients are estimated, followed by a phase inpainting stage, where the missing phases are estimated. While spectrogram inpainting has been addressed in several works [11, 13, 18], phase inpainting has not been addressed by advanced methods and thus remains a challenge. Indeed, phase reconstruction is known to be a difficult task generally posed as a non-convex problem. Many works have

This work was supported by ANR JCJC program MAD (ANR-14-CE27-0002).

been proposed to reconstruct the phase of all the time-frequency coefficients from their amplitude and may be extended to the phase inpainting problem. A first set of phase reconstruction methods relies on alternate projections [7–10] among which the Griffin and Lim (GL) algorithm [10] is widely used in audio processing. Its success may be due to the simplicity of its implementation and the low computational cost of its iterations. However, its performance is limited by a slow convergence towards a local minimum. Higher reconstruction performance has been reached by semidefinite programming (SDP) approaches, at the cost of much higher time and space complexities. In particular, PhaseLift [4] and PhaseCut [19] methods have been proposed for any linear operator and further studies [3, 12] have established their good performance in the case of the short-time Fourier transform (STFT). While yet other phase reconstruction algorithms have been recently proposed [1, 5, 6, 14–17], we focus on extending original GL and SDP approaches to phase inpainting.

The organization of the paper is as follows. In Sect. 2, the phase inpainting problem is formalized and we propose three dedicated algorithms: Griffin and Lim for phase inpainting (GLI), PhaseLift for phase inpainting (PLI) and PhaseCut for phase inpainting (PCI). These three algorithms are the extensions of existing algorithms, in which we add the knowledge of the partially observed phases. While the algorithms are introduced in the general case of any linear operator, Sect. 3 is dedicated to their specific implementation with the STFT operator. In Sect. 4, some experiments in small dimensions with various ratios of missing data and several mask shapes illustrate their performance and their limitations. Finally, conclusions and perspectives are drawn in Sect. 5.

2 Proposed Phase Inpainting Algorithms

2.1 Phase Inpainting Problem

For a signal $\mathbf{x} \in \mathbb{C}^N$, we consider K complex linear measurements $\mathbf{A}\mathbf{x} = [\langle \mathbf{a}_k, \mathbf{x} \rangle]_{k=1}^K \in \mathbb{C}^K$ where $\mathbf{a}_1, \dots, \mathbf{a}_K \in \mathbb{C}^N$ and $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_K]^H$. While the specific case of the STFT operator is used in Sects. 3 and 4, the general case of any linear operator is addressed throughout Sect. 2. We assume that we observe both the magnitude and the phase of a subset of measurements while only the magnitude of the remaining measurements is available. The location of these subsets is given by a binary mask $\mathbf{m} \in \{0, 1\}^K$: $\mathbf{m}[k] = 1$ if both the magnitude and the phase of measurement k are known and $\mathbf{m}[k] = 0$ if only its magnitude is known.

Throughout the document we will adopt the following notations: for $\mathbf{b} \in \mathbb{C}^K$ $\angle \mathbf{b}$ denotes its phase, $\bar{\mathbf{b}}$ its conjugate. For $\mathbf{m} \in \{0, 1\}^K$, \neg denotes the negation. Denoting by $\text{supp}(\mathbf{m})$ the support of \mathbf{m} , let $\mathbf{b} \in \mathbb{C}^K$ be the vector containing the fully known coefficients $\mathbf{b}[k]$ for $k \in \text{supp}(\mathbf{m})$, and the known amplitudes $\mathbf{b}[k]$ for $k \in \text{supp}(\neg \mathbf{m})$. Then the phase inpainting problem is given by

$$\text{Find } \mathbf{x} \in \mathbb{C}^N \text{ s.t. } \begin{cases} \langle \mathbf{a}_k, \mathbf{x} \rangle &= \mathbf{b}[k], \forall k \in \text{supp}(\mathbf{m}) \\ |\langle \mathbf{a}_k, \mathbf{x} \rangle| &= \mathbf{b}[k], \forall k \in \text{supp}(\neg \mathbf{m}) \end{cases} \quad (1)$$

2.2 Griffin and Lim Algorithm for Phase Inpainting (GLI)

We propose an extension of the Griffin and Lim algorithm [10] to solve approximately problem (1) by taking into account the known phases. The algorithm is described in Algorithm 1, \circ denoting the Hadamard product. It mainly relies on alternating a projection onto the span of the linear operator using projector $\Pi_{\mathbf{a}}$ and a projection onto the known magnitude and phase constraints. The initialization of this algorithm may be done with random phases for coefficients with unknown phase.

Algorithm 1. Griffin and Lim algorithm for phase inpainting (GLI)

Require:

binary mask $\mathbf{m} \in \{0, 1\}^K$

observation $\mathbf{b} \in \mathbb{C}^K$ such that $\begin{cases} \mathbf{b}[k] \in \mathbb{C}, & \forall k \in \text{supp}(\mathbf{m}) \text{ (fully known coefficients)} \\ \mathbf{b}[k] \in [0, \infty[, & \forall k \in \text{supp}(-\mathbf{m}) \text{ (known magnitudes)} \end{cases}$

projector onto the span of the linear operator $\Pi_{\mathbf{a}}$

initial phases $\varphi_0 \in [0, 2\pi]^K$

number of iterations n_{iter}

Output: complete estimated measurements $\mathbf{y}^{(n_{\text{iter}})}$

$\varphi \leftarrow \mathbf{m} \circ \angle \mathbf{b} + (1 - \mathbf{m}) \circ \varphi_0 \quad \forall k \in \text{supp}(\mathbf{m})$

$\mathbf{y}^{(0)} \leftarrow \mathbf{b} \circ \exp(i\varphi) \quad \forall k \in \text{supp}(-\mathbf{m})$

for $i \in \{1, 2, \dots, n_{\text{iter}}\}$ **do**

$\mathbf{z}^{(i)} \leftarrow \Pi_{\mathbf{a}}(\mathbf{y}^{(i-1)})$

$\varphi^{(i)} \leftarrow \mathbf{m} \circ \angle \mathbf{b} + (1 - \mathbf{m}) \circ \angle \mathbf{z}^{(i)}$ {Project onto phase constraints}

$\mathbf{y}^{(i)} \leftarrow \mathbf{b} \circ \exp(i\varphi^{(i)})$ {Project onto magnitude constraints}

end for

2.3 PhaseLift for Phase Inpainting (PLI)

The second proposed approach is based on lifting and SDP. The quadratic constraints in problem (1) become linear by means of a projection in a large dimensional space where the variable is a semidefinite positive matrix $\mathbf{X} \succeq 0$. The PhaseLift method [4] is adapted in order to address phase inpainting, which results in Proposition 1.

Proposition 1. *With notations of problem (1), let $\mathbf{A}_{lk} = \mathbf{a}_l \mathbf{a}_k^H$ for $l, k \in \{1, \dots, K\}$. Using the lifting $\mathbf{X} = \mathbf{x}\mathbf{x}^H$, problem (1) is equivalent to:*

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \text{Rank}(\mathbf{X}) \text{ s.t. } \begin{cases} \text{Trace}(\mathbf{A}_{lk}\mathbf{X}) = \mathbf{b}[k]\bar{\mathbf{b}}[l], & \forall l, k \in \text{supp}(\mathbf{m}) \\ \text{Trace}(\mathbf{A}_{kk}\mathbf{X}) = \mathbf{b}^2[k], & \forall k \in \text{supp}(-\mathbf{m}) \\ \mathbf{X} \succeq 0 \end{cases} \quad (2)$$

and can be relaxed as :

$$\min_{\mathbf{X} \in \mathbb{C}^{N \times N}} \text{Trace}(\mathbf{X}) \text{ s.t. } \begin{cases} \text{Trace}(\mathbf{A}_{lk}\mathbf{X}) = \mathbf{b}[k]\bar{\mathbf{b}}[l], & \forall l, k \in \text{supp}(\mathbf{m}) \\ \text{Trace}(\mathbf{A}_{kk}\mathbf{X}) = \mathbf{b}^2[k], & \forall k \in \text{supp}(-\mathbf{m}) \\ \mathbf{X} \succeq 0 \end{cases} \quad (3)$$

Proof. The proof can be conducted in three steps:

1. Assume that \mathbf{x} satisfies (1). For $k, l \in \text{supp}(\mathbf{m})$, the phase constraint is obtained by considering that

$$\mathbf{b}[k]\bar{\mathbf{b}}[l] = \text{Trace}(\mathbf{a}_k^H \mathbf{x} \mathbf{x}^H \mathbf{a}_l) = \text{Trace}(\mathbf{a}_l \mathbf{a}_k^H \mathbf{x} \mathbf{x}^H) = \text{Trace}(\mathbf{A}_{lk} \mathbf{X})$$

For $k \in \text{supp}(-\mathbf{m})$, the magnitude constraint is obtained similarly.

2. Problem (1) can then be reformulated as

$$\text{Find } \mathbf{X} \in \mathbb{C}^{N \times N} \quad \text{s.t.} \quad \begin{cases} \text{Trace}(\mathbf{A}_{lk} \mathbf{X}) = \mathbf{b}[k]\bar{\mathbf{b}}[l], & \forall l, k \in \text{supp}(\mathbf{m}) \\ \text{Trace}(\mathbf{A}_{kk} \mathbf{X}) = \mathbf{b}^2[k], & \forall k \in \text{supp}(-\mathbf{m}) \\ \text{Rank}(\mathbf{X}) = 1 \\ \mathbf{X} \succeq 0 \end{cases}$$

which is equivalent to problem (2).

3. Since the rank is not convex, one may finally relax the rank by the nuclear norm to obtain Problem (3). \square

Formulation (3) is called PhaseLift for phase inpainting (PLI). The objective function and equality constraints are linear and the domain $\mathbf{X} \succeq 0$ is a convex cone. One may notice that only phase differences appear, in the first constraint, to exploit the known phases. In the particular case $\text{supp}(\mathbf{m}) = \emptyset$, the original PhaseLift problem [4] is obtained.

Finally, from the solution \mathbf{X} of problem (3), \mathbf{x} can be estimated as $\sqrt{\lambda_{\max}} \mathbf{z}_{\max}$ where \mathbf{z}_{\max} is the eigenvector associated with the largest eigenvalue λ_{\max} of \mathbf{X} .

In order to solve the PLI problem (3), we use Matlab toolbox TFOCS [2]. Two solvers may be used: `solver_sSDP` that performs trace minimization under linear constraints as in (3), or `solver_TraceLS` that solves unconstrained problems of the form $\min_{\mathbf{X} \succeq 0} \lambda \text{Trace}(\mathbf{X}) + \frac{1}{2} \|\mathcal{A}(\mathbf{X}) - \beta\|^2$ with

$$\mathcal{A} : \mathbf{X} \mapsto \begin{bmatrix} \text{vec} \left([\text{Trace}(\mathbf{A}_{lk} \mathbf{X})]_{l,k \in \text{supp}(\mathbf{m})} \right) \\ [\text{Trace}(\mathbf{A}_{kk} \mathbf{X})]_{k \in \text{supp}(-\mathbf{m})} \end{bmatrix}, \beta = \begin{bmatrix} \text{vec} \left([\mathbf{b}[k]\bar{\mathbf{b}}[l]]_{l,k \in \text{supp}(\mathbf{m})} \right) \\ [\mathbf{b}[k]]_{k \in \text{supp}(-\mathbf{m})} \end{bmatrix}. \tag{4}$$

2.4 PhaseCut for Phase Inpainting (PCI)

The third and last proposed algorithm is also an SDP optimization algorithm, namely PhaseCut for phase inpainting (PCI), which is an extension of the original PhaseCut designed for phase retrieval [19].

As in [19], problem (1) is reformulated by explicitly splitting the amplitude and phase variables, so that one may optimize only on the phase vector $\mathbf{u} \in \mathbb{C}^K$ such that $\forall k, |\mathbf{u}[k]| = 1$. We use the lifting $\mathbf{U} = \mathbf{u} \mathbf{u}^H$ to obtain Proposition 2.

Proposition 2. Using notations of problem (1), let $\mathbf{\Gamma} = \text{Diag}(\mathbf{c}^H)(\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)\text{Diag}(\mathbf{c})$ with $\mathbf{c} \in \mathbb{C}^K$ is defined by $\mathbf{c}[k] = |\mathbf{b}[k]|, \forall k$. Then problem (1) is equivalent to

$$\min_{\mathbf{U} \in \mathbb{C}^{K \times K}} \text{Trace}(\mathbf{U}\mathbf{\Gamma}) \text{ s.t. } \begin{cases} \text{Diag}(\mathbf{U}) = \mathbf{1} \\ \mathbf{U}[k_1, k_2] = \frac{\mathbf{b}[k_1]}{|\mathbf{b}[k_1]|} \frac{\bar{\mathbf{b}}[k_2]}{|\mathbf{b}[k_2]|}, \forall k_1, k_2 \in \text{supp}(\mathbf{m}) \\ \text{Rank}(\mathbf{U}) = 1 \\ \mathbf{U} \succeq 0 \end{cases} \quad (5)$$

and may be relaxed into a convex problem by dropping the rank constraint as

$$\min_{\mathbf{U} \in \mathbb{C}^{K \times K}} \text{Trace}(\mathbf{U}\mathbf{\Gamma}) \text{ s.t. } \begin{cases} \text{Diag}(\mathbf{U}) = \mathbf{1} \\ \mathbf{U}[k_1, k_2] = \frac{\mathbf{b}[k_1]}{|\mathbf{b}[k_1]|} \frac{\bar{\mathbf{b}}[k_2]}{|\mathbf{b}[k_2]|}, \forall k_1, k_2 \in \text{supp}(\mathbf{m}) \\ \mathbf{U} \succeq 0 \end{cases} \quad (6)$$

Proof. Using the amplitude vector \mathbf{c} and the phase vector \mathbf{u} , problem (1) becomes

$$\text{Find } \mathbf{x} \in \mathbb{C}^N, \mathbf{u} \in \mathbb{C}^K \text{ s.t. } \begin{cases} \mathbf{A}\mathbf{x} &= \text{Diag}(\mathbf{c})\mathbf{u} \\ \mathbf{u}[k] &= e^{i\angle \mathbf{b}[k]} \forall k \in \text{supp}(\mathbf{m}) \\ |\mathbf{u}[k]| &= 1 \forall k \end{cases} \quad (7)$$

which is equivalent to

$$\min_{\mathbf{x} \in \mathbb{C}^N, \mathbf{u} \in [0, 2\pi]^{|\text{supp}(\mathbf{m})|}} \|\mathbf{A}\mathbf{x} - \text{Diag}(\mathbf{c})\mathbf{u}\|_2^2 \text{ s.t. } \begin{cases} \mathbf{u}[k] &= e^{i\angle \mathbf{b}[k]} \forall k \in \text{supp}(\mathbf{m}) \\ |\mathbf{u}[k]| &= 1 \forall k \end{cases} \quad (8)$$

Given that $\mathbf{A}\mathbf{x} = \text{Diag}(\mathbf{c})\mathbf{u}$ implies $\mathbf{x} = \mathbf{A}^\dagger \text{Diag}(\mathbf{c})\mathbf{u}$, then $\|\mathbf{A}\mathbf{x} - \text{Diag}(\mathbf{c})\mathbf{u}\|_2^2 = \mathbf{u}^H \mathbf{\Gamma} \mathbf{u}$, thus (8) is equivalent to (5) which can be relaxed into (6). \square

Formulation (6) is called PhaseCut for phase inpainting (PCI). As for PLI, phase differences appear in the constraints that involve known phases. In the particular case where all phases are unknown ($\text{supp}(\mathbf{m}) = \emptyset$), constraints $\mathbf{U}[k_1, k_2] = \frac{\mathbf{b}[k_1]}{|\mathbf{b}[k_1]|} \frac{\bar{\mathbf{b}}[k_2]}{|\mathbf{b}[k_2]|}$ disappear and the original PhaseCut problem [19] is obtained \mathbf{x} .

Finally, from the solution \mathbf{U} of problem (6), signal \mathbf{x} is estimated as $\mathbf{x} = \mathbf{A}^\dagger \text{Diag}(\mathbf{c})e^{i\angle \mathbf{u}_{\max}}$ where \mathbf{u}_{\max} is an eigenvector associated to the largest eigenvalue of \mathbf{U} .

In order to solve PCI problem (6), we adapt the block coordinate descent algorithm proposed in [19] from [20], as given in Algorithm 2. By picking coordinates i in $\text{supp}(\mathbf{m})$ instead of $\{1, \dots, K\}$, all unknown coefficients in \mathbf{U} , and only them, are updated.

Algorithm 2. PhaseCut for phase inpainting (PCI) : BCD algorithm

Require:
 binary mask $\mathbf{m} \in \{0, 1\}^K$
 observation $\mathbf{b} \in \mathbb{C}^K$ such that $\begin{cases} \mathbf{b}[k] \in \mathbb{C}, & \forall k \in \text{supp}(\mathbf{m}) \text{ (fully known coefficients)} \\ \mathbf{b}[k] \in [0, \infty[, & \forall k \in \text{supp}(\neg\mathbf{m}) \text{ (known magnitudes)} \end{cases}$
 number of iterations n_{iter}
 barrier parameter $\nu > 0$
Output: $\mathbf{U} \in \mathbb{C}^{K \times K}$
 {Initialization}
 $\mathbf{c} \leftarrow \mathbf{m} \circ \mathbf{b} + (1 - \mathbf{m}) \circ \mathbf{b}$
 $\Gamma \leftarrow \text{Diag}(\mathbf{c}^H)(I - \mathbf{A}\mathbf{A}^\dagger)\text{Diag}(\mathbf{c})$
 for $1 \leq k, l, \leq K$, $\mathbf{U}[k, l] \leftarrow \begin{cases} 1 & \text{if } k = l \\ \frac{\mathbf{b}[k]}{\|\mathbf{b}[k]\|} \frac{\bar{\mathbf{b}}[l]}{\|\bar{\mathbf{b}}[l]\|} & \text{if } k, l \in \text{supp}(\mathbf{m}) \\ 0 & \text{otherwise} \end{cases}$
 {Main loop}
for n_{iter} iterations **do**
 pick $i \in \{1, \dots, K\} \setminus \text{supp}(\mathbf{m})$
 $\mathbf{x} \leftarrow \mathbf{U}_{i^c, i^c} \Gamma_{i^c, i}$ and $\gamma \leftarrow \mathbf{x}^H \Gamma_{i^c, i}$
 $\mathbf{U}_{i^c, i}, \mathbf{U}_{i^c, i}^H \leftarrow \begin{cases} -\sqrt{\frac{1-\nu}{\gamma}} \mathbf{x} & \text{if } \gamma > 0 \\ 0 & \text{otherwise} \end{cases}$
end for

3 Implementation Issues Specific to the STFT

Phase Inpainting Problem with STFT Measurements. The STFT of a signal $\mathbf{x} \in \mathbb{C}^N$ is defined for frame index $t \in \{0, \dots, T - 1\}$ and frequency index $\nu \in \{0, \dots, F - 1\}$ as $\text{STFT}[t, \nu] = \langle \mathbf{x}, \mathbf{a}_{t, \nu} \rangle = \mathbf{a}_{t, \nu}^H \mathbf{x}$ where $\mathbf{a}_{t, \nu} = [\mathbf{w}[n - th]e^{2i\pi \frac{\nu}{F} n}]_{n=0}^{N-1} \in \mathbb{C}^K$, \mathbf{w} being the analysis window and h the so-called *hop size* between two successive frames. Hence the $K = FT$ measurements are indexed by $k = (t, \nu)$: measurements may be seen equivalently either as a doubly-indexed vector or as a matrix. A simple reshaping operation can be used to switch between representations, and with a small abuse of notations, both of them are used without explicit distinction in this paper. The STFT phase inpainting problem in time-frequency is thus given by

$$\text{Find } \mathbf{x} \in \mathbb{C}^N \text{ s.t. } \begin{cases} \langle \mathbf{x}, \mathbf{a}_{t, \nu} \rangle = \mathbf{b}[t, \nu], & \forall (t, \nu) \in \text{supp}(\mathbf{m}) \\ |\langle \mathbf{x}, \mathbf{a}_{t, \nu} \rangle| = \mathbf{b}[t, \nu], & \forall (t, \nu) \in \text{supp}(\neg\mathbf{m}) \end{cases} \quad (9)$$

GLI Implementation. The GLI algorithm is obtained by setting $\mathbf{\Pi}_a : \mathbf{y} \mapsto \text{STFT}(\text{STFT}^{-1}(\mathbf{y}))$ where STFT^{-1} is the (pseudo-)inverse operator for the STFT computed from the canonical dual window of \mathbf{w} .

PLI Implementation. We used `solver_TraceLS` of TFOCS library, which happened to be faster than `solver_sSDP`. The implementation of the direct operator \mathcal{A} defined in (4) and of its adjoint can be more efficient using fast Fourier transforms (FFT) as follows. We have $\text{Trace}(\mathbf{A}_{lk}\mathbf{X}) = \left(\text{STFT}_{\text{row}}\left(\overline{\text{STFT}_{\text{col}}(\mathbf{X})}\right)\right)^H [k, l]$ for $k, l \in \{1, \dots, K\}$, where $\text{STFT}_{\text{row}}(\mathbf{X})$ denotes the STFT on the columns of \mathbf{X} and $\text{STFT}_{\text{col}}(\mathbf{X})$ the STFT on the

rows of \mathbf{X} . Hence one may compute $\mathcal{A}(\mathbf{X})$ from only $2N$ STFT's. By denoting by $k_0 = \#\text{supp}(\mathbf{m})$ the number of known phases, the adjoint operator $\mathcal{A}^* : \mathbb{C}^{k_0^2+K-k_0} \rightarrow \mathbb{C}^{N \times N}$ is such that $\mathcal{A}^*(\mathbf{y}) = \left(\text{STFT}_{row}^* \left(\overline{\text{STFT}_{col}^*(\mathbf{Y})} \right) \right)^H$ where STFT^* is the adjoint of the STFT operator and $\mathbf{Y} \in \mathbb{C}^{K \times K}$ is defined by $\mathbf{Y}(\mathbf{m}, \mathbf{m}) = \text{reshape}(\mathbf{y}(1 : k_0^2), k_0, k_0)$ and $\mathbf{Y}(\sim \mathbf{m}, \sim \mathbf{m}) = \text{Diag}(\mathbf{y}(k_0^2+1 : K))$. It thus requires $2K$ calls to STFT^* .

PCI Implementation. Each iteration of Algorithm 2 for PCI requires K calls to one direct STFT and one inverse STFT, using FFT's.

4 Experiments

All experiences are available on mad.lis-lab.fr. Experiments in small dimensions are conducted on a signal with length $N = 128$ composed of the sum of two linear chirps with normalized frequency ranges $(0, 0.8)$ and $(0.8, 0.6)$, a dirac located at sample 64 and white Gaussian noise at a signal-to-noise ratio of 10 dB. The STFT is generated with a Hann window with length 16, a hop size of 8 samples (i.e., $T = 16$ frames) and $F = 32$ frequency bins, resulting in $K = 512$ measurement in a 32×16 time-frequency matrix. In a first experiment, masks for missing phases are generated randomly and uniformly among the measures, with various ratios of missing phases. In a second experiment, the ratio of missing phases is fixed at 30% and missing phases are grouped in holes of a given width, with randomly distributed centers, the widths varying between 1 and 9 coefficients. Figure 1 illustrates the STFT of the signal and of one generated mask.

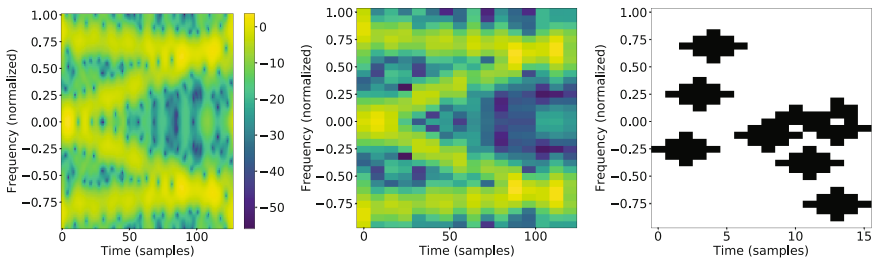


Fig. 1. Spectrogram of the signal (left, smoothed with $T = F = 128$; middle, with $T = 16$ and $F = 32$ as set in the experiment) and example of a mask with random holes of width 5 in black (right).

Algorithms are used with the following settings. For GLI, $n_{\text{iter}} = 6000$. For PLI, $\lambda = 10^{-30}$ and TFOCS is used with a maximum of 5000 iterations, no restart, $\text{tol} = 10^{-10}$. For PCI, $\nu = 10^{-14}$ and $n_{\text{iter}} = 10^5$. A baseline approach is also used, denoted as Random Phase Inpainting (RPI) and consisting in filling the missing phases by drawing random values independently and uniformly in $[0, 2\pi]$.

Performance is assessed in terms of relative reconstruction error up to a global phase shift, defined by $E_{dB}(\mathbf{x}, \hat{\mathbf{x}}) = 20 \log_{10} \min_{\theta} \frac{\|\mathbf{x} - e^{i\theta} \hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$ where \mathbf{x} denotes the original signal and $\hat{\mathbf{x}}$ the reconstructed one.

Results are shown in Fig. 2, where the reconstruction errors from all methods can be compared, as a function of the ratio of missing phases, for each experiment. The known phases clearly contribute to improve the signal reconstruction. For isolated missing phases (left figure), one can see that below 40% missing phases, GLI and PCI achieves perfect reconstruction while PLI performs very good but not perfect. Beyond 40% missing phases, SDP methods PLI and PCI perform better than GLI, with a much better performance for PLI. For holes with a larger width at 30% missing phases (right figure), one can see that GLI may achieve poor reconstruction due to local minima. SDP methods offer very good performance, with a reconstruction error generally below -50 dB.

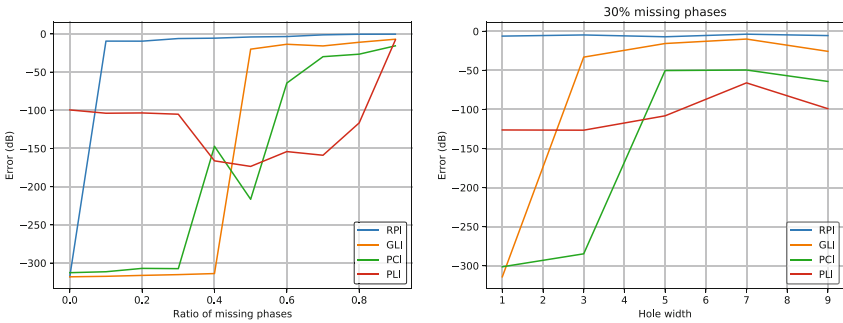


Fig. 2. Reconstruction error as a function of the ratio of missing phases randomly distributed (left) or, for 30% missing phases, as a function of the width of randomly distributed holes (right).

The convergence and running time of each method have been checked as follows. For GLI, it was checked visually and manually that the algorithm converges before the maximum number of iterations, with a running time lower than one second for each call. For PLI, similarly, it has been checked that the algorithm stops before the maximum number of iterations is reached, with a running time up to 6 h for one call when the number of missing phases is large. For PCI, convergence may be observed in Fig. 3 by representing the reconstruction error as a function of the iterations. As for PLI, the running time until convergence is all the more reduced as many phases are known, lasting about 4 h for 10^5 iterations.

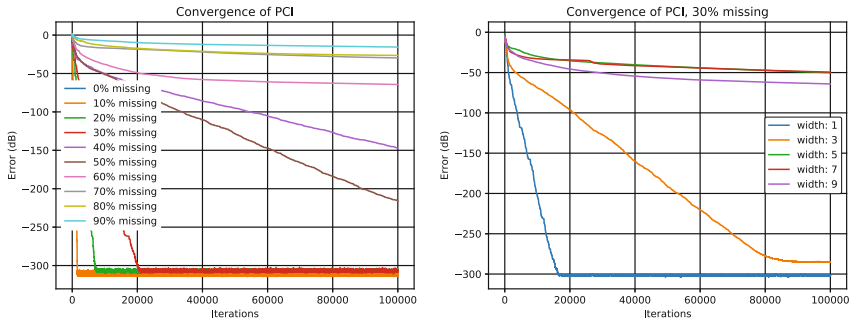


Fig. 3. Illustration of the convergence of PCI by representing the reconstruction error as a function of the iterations in the same two settings as in Fig. 2.

5 Conclusion and Perspectives

We have considered the phase inpainting problem in which a subset of measurements have missing phases that must be recovered. We have proposed three dedicated algorithms, which are extensions of existing algorithms, namely Griffin and Lim, PhaseLift and PhaseCut, adapted to the phase inpainting problem by incorporating the partial phase information as constraints in the optimization process. Those algorithms have been implemented using fast transforms in the case of the STFT. Experiments in small dimensions confirm that SDP methods perform better than Griffin and Lim algorithm, in particular when the problem is difficult (more unknown phases, larger holes). Even if those methods are very time consuming, it also appears that the knowledge of a subset of phases result in a faster convergence.

Experiments may be extended to the noisy case where only approximate values are available known phases and amplitudes. Time and space complexity of SDP approaches being very large, they cannot be applied to typical audio signals for which dimensions are higher than those used in the proposed experiments. In order to benefit from SDP results, one may investigate the adaptation of SDP algorithms to process only a local time-frequency region instead of the whole STFT matrix. Other algorithms may be designed for phase inpainting. In particular, some recent contributions to phase retrieval [1, 5, 6, 14–17] may be adapted and may give good performance without the computational limits of SDP methods.

References

1. Bahmani, S., Romberg, J.: Phase retrieval meets statistical learning theory: a flexible convex relaxation. [arXiv:1610.04210v2](https://arxiv.org/abs/1610.04210v2) (2017)
2. Becker, S., Candès, E.J., Grant, M.: Templates for convex cone problems with applications to sparse signal recovery. Technical report, Department of Statistics, Stanford University (2010)

3. Bendory, T., Eldar, Y.C., Boumal, N.: Non-convex phase retrieval from STFT measurements. *IEEE Trans. Inform. Theory* **64**(1), 467–484 (2018)
4. Candès, E.J., Eldar, Y., Strohmer, T., Vershynina, V.: Phase retrieval via matrix completion. *SIAM Rev.* **57**(2), 225–251 (2015)
5. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory* **61**(4), 1985–2007 (2015)
6. A. Dremeau and F. Krzakala. Phase recovery from a bayesian point of view: The variational approach. In *IEEE Trans. Acous., Speech Signal Process.* IEEE, Apr 2015
7. Fienup, J.R.: Reconstruction of an object from the modulus. *Opt. Lett.* **3**, 27–29 (1978)
8. Fienup, J.R.: Phase retrieval algorithms : a comparison. *Appl. Opt.* **21**(15), 2758–2769 (1982)
9. Gerchberg, R.W., Saxton, W.: A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972)
10. Griffin, D., Lim, J.: Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acous. Speech Signal Process.* **32**(2), 236–243 (1984)
11. Hamon, R., Emiya, V., Févotte, C.: Convex nonnegative matrix factorization with missing data. In: *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (2016)
12. Jaganathan, K., Eldar, Y.C., Hassibi, B.: STFT phase retrieval: uniqueness guarantees and recovery algorithms. *IEEE J. Sel. Topics Signal Process.* **10**(4), 770–781 (2016)
13. 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 Commun.* **53**, 658–676 (2010)
14. Metzler, C.A., Sharma, M.K., Nagesh, S., Baraniuk, R.G., Cossairt, O., Veeraraghavan, A.: Coherent inverse scattering via transmission matrices: efficient phase retrieval algorithms and a public dataset. In: *2017 IEEE International Conference on Computational Photography (ICCP)*. IEEE, May 2017
15. Netrapalli, P., Jain, P., Sanghavi, S.: Phase retrieval using alternating minimization. *Adv. Neural Inf. Process. Syst.* **26**, 2796–2804 (2013)
16. Prusa, Z., Balazs, P., Sondergaard, P.: A non-iterative method for reconstruction of phase from stft magnitude. *IEEE Trans. Audio Speech Lang. Process.* **25**, 1154–1164 (2017)
17. Rajaei, B., Gigan, S., Krzakala, F., Daudet, L.: Robust phase retrieval with the swept approximate message passing (prSAMP) algorithm. *Image Process. On Line* **7**, 43–55 (2017)
18. Smaragdis, P., Raj, B., Shashanka, M.: Missing data imputation for spectral audio signals. In: *Proceedings of MLSP, Grenoble, France* (2009)
19. Waldspurger, I., d’Aspremont, A., Mallat, S.: Phase recovery maxcut and complex semidefinite programming. *Math. Prog.* **149**(12), 47–81 (2015)
20. Wen, Z., Goldfarb, D., Scheinberg, K.: Block coordinate descent methods for semidefinite programming. In: Anjos, M., Lasserre, J. (eds.) *Handbook on Semidefinite, Conic and Polynomial Optimization*. International Series in Operations Research & Management Science, vol. 166, pp. 533–564. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-0769-0_19

Sparsity-Related Methods



Revisiting Synthesis Model in Sparse Audio Declipper

Pavel Záváška^{1(✉)}, Pavel Rajmic^{1(✉)}, Zdeněk Průša², and Vítězslav Veselý³

¹ Signal Processing Laboratory, Brno University of Technology,
Brno, Czech Republic
{xzavis01,rajmic}@vutbr.cz

² Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
zdenek.prusa@oeaw.ac.at

³ Faculty of Mechanical Engineering, Brno University of Technology,
Brno, Czech Republic

Abstract. The state of the art in audio declipping has currently been achieved by SPADE (SParse Audio DEclipper) algorithm by Kitić et al. Until now, the synthesis/sparse variant, S-SPADE, has been considered significantly slower than its analysis/cosparse counterpart, A-SPADE. It turns out that the opposite is true: by exploiting a recent projection lemma, individual iterations of both algorithms can be made equally computationally expensive, while S-SPADE tends to require considerably fewer iterations to converge. In this paper, the two algorithms are compared across a range of parameters such as the window length, window overlap and redundancy of the transform. The experiments show that although S-SPADE typically converges faster, the average performance in terms of restoration quality is not superior to A-SPADE.

Keywords: Clipping · Declipping · Audio · Sparse
Cosparse · SPADE · Projection · Restoration

1 Introduction

Clipping is a non-linear form of signal distortion which appears in the context of signal acquisition, processing or transmission. In general, clipping occurs when the signal amplitude gets outside of the allowed dynamic range. Along with missing samples and additive noise, clipping is one of the most common types of audio signal degradation. Not only does clipping have a negative effect on perceived audio quality [35], it also degrades the accuracy of automatic speech recognition [19, 25, 34]. This motivates a restoration task usually termed *declipping*, i.e. the recovery of signal samples that originally lay outside the recognized range.

In this work, we concentrate on the case of the so-called *hard clip* degradation, where the waveform of the signal is simply truncated such that the signal value

cannot leave the interval $[-\theta_c, \theta_c]$. If the vector $\mathbf{x} \in \mathbb{R}^N$ denotes the original discrete-time signal, then the respective hard-clipped signal is

$$\mathbf{y}[n] = \begin{cases} \mathbf{x}[n] & \text{for } |\mathbf{x}[n]| < \theta_c, \\ \theta_c \cdot \text{sgn}(\mathbf{x}[n]) & \text{for } |\mathbf{x}[n]| \geq \theta_c, \end{cases} \quad (1)$$

i.e. hard clipping acts elementwise, wasting information in the peaks of \mathbf{x} that exceed the *clipping threshold* θ_c .

In the past, several attempts were made to perform declipping. Since declipping is inherently ill-posed, any method attacking the problem must introduce an assumption about the signal. As a short review of the field, we mention a method based on autoregressive signal modelling [21], a method based on the knowledge of the original signal bandwidth [1], statistical approaches [15, 17], and simple, even if not quite effective algorithms in [11, 26, 32]. The quality of restoration was significantly elevated when models involving the *sparsity* of the audio signal were introduced. In such models, it is assumed that there is a transform which either approximates the signal well using a low number of nonzero coefficients (the synthesis/sparse model), or the transform applied to the signal produces a low number of nonzero coefficients (the analysis/cosparse model) [8, 13, 27]. Suitable transforms are usually time-frequency operators such as the Discrete Fourier Transform (DFT), the Discrete Cosine Transform (DCT), or the Discrete Gabor Transform (DGT), also known as the Short-time Fourier Transform (STFT) [9, 14, 18].

The very first method for sparse declipping was published in [2]; it was based on the greedy approximation of a signal within the reliable (i.e. not clipped) parts. Many alternative approaches appeared after this successful paper, such as in [36] that brought convex optimization into play, or in [33] where the authors forced a structure into the sparse coefficients (known as “structured” or “social” sparsity). Article [12] shows that the introduction of a psychoacoustic masking model (although very simple) improves the perceived quality of the restored signal. Besides [6], which relies on non-negative matrix factorization, all the mentioned papers process the signal from the synthesis viewpoint. More recently, a series of papers have considered the declipping problem from the analysis side as well [22–24], while [24] is considered a state-of-the-art declipper.

In this paper, we show that using a novel projection lemma we were able to derive a synthesis-based algorithm which is even faster than the analysis-based algorithm in [24]. Our experiments show that our algorithm, nevertheless, does not outperform the analysis version in terms of quality of restoration.

In Sect. 2, the declipping problem is formalized. Then in Sect. 3, the two versions of the SPADE algorithm [24] are reviewed, and the new projection lemma is exploited to develop a fast synthesis-based algorithm. Sect. 4 reports on experiments that have been run.

2 Problem Formulation

Assume that a signal $\mathbf{x} \in \mathbb{R}^N$ has been clipped according to (1). We observe the clipped signal $\mathbf{y} \in \mathbb{R}^N$. We suppose that it is possible to divide the signal

samples into three sets R , H and L , which correspond to “reliable” samples and samples that have been clipped to the “high” and “low” clipping thresholds, respectively. To select only samples of a specific set, linear *restriction operators* M_R , M_H and M_L will be used. Note that if these sets are not known in advance, they can be trivially induced from the particular values of \mathbf{y} .

Denote the declipped signal by $\hat{\mathbf{x}}$. While performing any declipping algorithm, it is natural to enforce that the samples $M_R\hat{\mathbf{x}}$ match the reliable samples $M_R\mathbf{y}$. The authors of the C-IHT algorithm [22] call this approach *consistent*. Our approach obeys full *consistency*, meaning that in addition, the samples $M_H\hat{\mathbf{x}}$ should lie at or above θ_c and the samples from $M_L\hat{\mathbf{x}}$ should not lie above $-\theta_c$. These requirements are formalized by defining a set of signals Γ , consistent with the three conditions:

$$\Gamma = \Gamma(\mathbf{y}) = \{\hat{\mathbf{x}} \mid M_R\hat{\mathbf{x}} = M_R\mathbf{y}, M_H\hat{\mathbf{x}} \geq \theta_c, M_L\hat{\mathbf{x}} \leq -\theta_c\}. \quad (2)$$

In line with the recent literature, the fact that many musical signals are sparse with respect to a (time-)frequency transform will be exploited. To put it in words, one would like to find signal $\hat{\mathbf{x}}$ that is the most sparse among all signals belonging to the consistency set Γ . The state of the art declipping results are achieved by the SPADE algorithm, which will be described in the next section. It comes in two variants, based on either the synthesis (sparse) or the analysis (cosparsity) understanding of “sparsity” [27].

3 The SPADE Algorithm

SPADE (SParse Audio DEclipper) [24] is a heuristic declipping algorithm, approximating the solution of the following non-convex, NP-hard synthesis- or analysis-regularized inverse problems:

$$\min_{\mathbf{x}, \mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s. t.} \quad \mathbf{x} \in \Gamma(\mathbf{y}) \quad \text{and} \quad \|\mathbf{x} - \mathbf{D}\mathbf{z}\|_2 \leq \epsilon, \quad (3)$$

$$\min_{\mathbf{x}, \mathbf{z}} \|\mathbf{z}\|_0 \quad \text{s. t.} \quad \mathbf{x} \in \Gamma(\mathbf{y}) \quad \text{and} \quad \|\mathbf{A}\mathbf{x} - \mathbf{z}\|_2 \leq \epsilon. \quad (4)$$

Here $\|\mathbf{z}\|_0$ is the ℓ_0 pseudonorm measuring the sparsity, i.e. counting the nonzero elements of \mathbf{z} . The ℓ_2 constraint delimits the distance between the estimate and its sparse approximation. The linear operator $\mathbf{D} : \mathbb{C}^P \mapsto \mathbb{R}^N$ is the synthesis operator, with $N \leq P$; if regarded as a matrix in (3), it takes coefficients \mathbf{z} and forms the signal as the linear combination of its columns. Matrix \mathbf{D} is often called the *dictionary* [8]. In (4), the analysis operator $\mathbf{A} : \mathbb{R}^N \mapsto \mathbb{C}^P$ is considered, which analyses the signal and produces its transform coefficients. In order to be able to compare the two approaches, we naturally restrict ourselves to the case when the operators are mutually adjoint, $\mathbf{A} = \mathbf{D}^*$.

Note that problems (3) and (4) both seek the signal and its coefficients simultaneously and that both of them fall into a common, recently introduced general signal restoration framework, see [16]. Both \mathbf{A} and \mathbf{D} are assumed full rank,

N , and both formulations produce equal results when \mathbf{D} is a unitary operator $\mathbf{A} = \mathbf{D}^{-1}$ (the same will hold for the approximate solutions by SPADE).

It should be noted that in SPADE, the above optimization problems are solved frame-by-frame, i.e. the signal is segmented into possibly overlapping time chunks, and windowed. Problems (3) and (4) are then solved individually on each such segment, and the output is formed using a common overlap-add procedure. This allows real-time processing, and at the same time the time-frequency structure of the processing is preserved. Specifically, the windowed (I)DFT is used in place of the operators \mathbf{A} and \mathbf{D} , possibly with frequency oversampling [9].

SPADE addresses the above two problems by a modified ADMM algorithm [5, 7] resulting in the synthesis SPADE (S-SPADE) as shown in Algorithm 1, and the analysis SPADE (A-SPADE) as given in Algorithm 2.

Algorithm 1. S-SPADE	Algorithm 2. A-SPADE
Require: $\mathbf{D}, \mathbf{y}, M_R, M_H, M_L, s, r, \epsilon$	Require: $\mathbf{A}, \mathbf{y}, M_R, M_H, M_L, s, r, \epsilon$
1 $\hat{\mathbf{z}}^{(0)} = \mathbf{D}^* \mathbf{y}, \mathbf{u}^{(0)} = \mathbf{0}, i = 1, k = s$	1 $\hat{\mathbf{x}}^{(0)} = \mathbf{y}, \mathbf{u}^{(0)} = \mathbf{0}, i = 1, k = s$
2 $\bar{\mathbf{z}}^{(i)} = \mathcal{H}_k \left(\hat{\mathbf{z}}^{(i-1)} + \mathbf{u}^{(i-1)} \right)$	2 $\bar{\mathbf{z}}^{(i)} = \mathcal{H}_k \left(\mathbf{A} \hat{\mathbf{x}}^{(i-1)} + \mathbf{u}^{(i-1)} \right)$
3 $\hat{\mathbf{z}}^{(i)} = \arg \min_{\mathbf{z}} \ \mathbf{z} - \bar{\mathbf{z}}^{(i)} + \mathbf{u}^{(i-1)}\ _2^2$ s.t. $\mathbf{D}\mathbf{z} \in \Gamma$	3 $\hat{\mathbf{x}}^{(i)} = \arg \min_{\mathbf{x}} \ \mathbf{A}\mathbf{x} - \bar{\mathbf{z}}^{(i)} + \mathbf{u}^{(i-1)}\ _2^2$ s.t. $\mathbf{x} \in \Gamma$
4 if $\ \hat{\mathbf{z}}^{(i)} - \bar{\mathbf{z}}^{(i)}\ _2 \leq \epsilon$ then	4 if $\ \mathbf{A}\hat{\mathbf{x}}^{(i)} - \bar{\mathbf{z}}^{(i)}\ _2 \leq \epsilon$ then
5 terminate	5 terminate
6 else	6 else
7 $\mathbf{u}^{(i)} = \mathbf{u}^{(i-1)} + \hat{\mathbf{z}}^{(i)} - \bar{\mathbf{z}}^{(i)}$	7 $\mathbf{u}^{(i)} = \mathbf{u}^{(i-1)} + \mathbf{A}\hat{\mathbf{x}}^{(i)} - \bar{\mathbf{z}}^{(i)}$
8 $i \leftarrow i + 1$	8 $i \leftarrow i + 1$
9 if $i \bmod r = 0$ then	9 if $i \bmod r = 0$ then
10 $k \leftarrow k + s$	10 $k \leftarrow k + s$
11 end	11 end
12 go to 2	12 go to 2
13 end	13 end
14 return $\hat{\mathbf{x}} = \mathbf{D}\hat{\mathbf{z}}^{(i)}$	14 return $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(i)}$

Both SPADE algorithms rely on two principal steps. The first of them is the hard thresholding \mathcal{H}_k . This operator enforces sparsity by setting all but k largest components of the input vector to zero. In practice, the sparsity k of signals is unknown, therefore SPADE performs *sparsity relaxation*: in every r -th iteration the variable k is incremented by s until the constraint embodied by the ℓ_2 norm is smaller than ϵ . The second main step is the projection onto Γ in order to keep the consistency given by (2) and will be discussed in the following.

3.1 Projection in A-SPADE

The projection in SPADE (row 3 in both Algorithms 1 and 2) constitutes the computationally most demanding step. For general \mathbf{A} and \mathbf{D} , such projections are achievable only via iterative algorithms.

The projection in A-SPADE is written as an optimization problem where one has to find a consistent signal \mathbf{x} such that its analysis coefficients $\mathbf{A}\mathbf{x}$ are

the closest possible with respect to the given $(\bar{\mathbf{z}}^{(i)} - \mathbf{u}^{(i-1)})$. The authors of [24] exploit the advantage that when \mathbf{A}^* is a tight Parseval frame, i.e. $\mathbf{A}^*\mathbf{A} = \mathbf{D}\mathbf{D}^* = \mathbf{D}\mathbf{A}$ are all identity operators [9], then the projection can be done elementwise in the time domain, such that

$$\hat{\mathbf{x}}^{(i)} = \text{proj}_\Gamma \left(\mathbf{A}^* (\bar{\mathbf{z}}^{(i)} - \mathbf{u}^{(i-1)}) \right) \quad (5)$$

where proj_Γ is the operator of orthogonal projection onto a convex set Γ , in our case defined as

$$[\text{proj}_\Gamma(\mathbf{w})]_n = \begin{cases} [\mathbf{y}]_n & \text{for } n \in R, \\ \max\{[\mathbf{w}]_n, \theta_c\} & \text{for } n \in H, \\ \min\{[\mathbf{w}]_n, -\theta_c\} & \text{for } n \in L, \end{cases} \quad (6)$$

where $[\cdot]_n$ denotes the n -th element of a vector.

We now rewrite the projection into a more convenient form. Let $\tilde{\mathbb{R}}$ denote the extended real line, i.e. $\tilde{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. Define the lower and upper bounding vectors $\mathbf{b}_L, \mathbf{b}_H \in \tilde{\mathbb{R}}$ such that

$$[\mathbf{b}_L]_n = \begin{cases} [\mathbf{y}]_n & \text{for } n \in R, \\ \theta_c & \text{for } n \in H, \\ -\infty & \text{for } n \in L, \end{cases} \quad [\mathbf{b}_H]_n = \begin{cases} [\mathbf{y}]_n & \text{for } n \in R, \\ \infty & \text{for } n \in H, \\ -\theta_c & \text{for } n \in L. \end{cases} \quad (7)$$

Recognizing that the multidimensional interval $[\mathbf{b}_L, \mathbf{b}_H]$ matches the set of feasible solutions (2), specifically $\Gamma = \{\mathbf{x} \mid \mathbf{b}_L \leq \mathbf{x} \leq \mathbf{b}_H\}$, the final A-SPADE projection formula (5) can be written as

$$\hat{\mathbf{x}}^{(i)} = \text{proj}_{[\mathbf{b}_L, \mathbf{b}_H]} (\mathbf{A}^* \mathbf{v}) \quad \text{with} \quad \mathbf{v} = \bar{\mathbf{z}}^{(i)} - \mathbf{u}^{(i-1)}. \quad (8)$$

The projection onto the interval can be implemented as

$$\text{proj}_{[\mathbf{b}_L, \mathbf{b}_H]}(\mathbf{w}) = \min\{\max\{\mathbf{b}_L, \mathbf{w}\}, \mathbf{b}_H\}, \quad (9)$$

with the min and max functions returning pairwise extremes element by element.

Note that restricting to the Parseval tight frames in applications is not an issue [3, 4, 20, 24, 28, 30, 31].

3.2 Projection in S-SPADE

For S-SPADE the situation is different. The projection has to be done in the domain of coefficients. The authors of [24] claim that the projection needs to be computed iteratively and that a somewhat efficient implementation can be achieved with \mathbf{D} forming a tight Parseval frame. Still, [24] reports many times higher computational time for S-SPADE compared to A-SPADE.

We will show that it is possible to use an explicit formula to compute the projection in S-SPADE, making the two algorithms identical from the point of view of complexity per iteration. Our goal is to find the optimizer

$$\hat{\mathbf{z}}^{(i)} = \arg \min_{\mathbf{z}} \|(\bar{\mathbf{z}}^{(i)} - \mathbf{u}^{(i-1)}) - \mathbf{z}\|_2^2 \text{ s.t. } \mathbf{D}\mathbf{z} \in \Gamma. \quad (10)$$

The following lemma can be found in several variations, see, for example, [29] or [10]. We introduce a real-setting version for simplicity.

Lemma: *Let the operator $\mathbf{D} : \mathbb{R}^P \mapsto \mathbb{R}^N$, $N \leq P$, full-rank, $\mathbf{D}\mathbf{D}^\top$ identity. Let the multidimensional interval bounds $\mathbf{b}_L, \mathbf{b}_H \in \mathbb{R}^N$, $\mathbf{b}_L \leq \mathbf{b}_H$. Then the projection of a vector $\mathbf{v} \in \mathbb{R}^N$, respectively denoted and defined by*

$$\text{proj}_{\{\mathbf{x} \mid \mathbf{D}\mathbf{x} \in [\mathbf{b}_L, \mathbf{b}_H]\}}(\mathbf{v}) := \arg \min_{\mathbf{u}} \|\mathbf{v} - \mathbf{u}\|_2 \text{ s.t. } \mathbf{D}\mathbf{u} \in [\mathbf{b}_L, \mathbf{b}_H],$$

can be evaluated as

$$\text{proj}_{\{\mathbf{x} \mid \mathbf{D}\mathbf{x} \in [\mathbf{b}_L, \mathbf{b}_H]\}}(\mathbf{v}) = \mathbf{v} - \mathbf{D}^\top \left(\mathbf{D}\mathbf{v} - \text{proj}_{[\mathbf{b}_L, \mathbf{b}_H]}(\mathbf{D}\mathbf{v}) \right). \tag{11}$$

In our application, we will need a complex \mathbf{D} with a special (time-)frequency structure. Indeed, our \mathbf{D} will be the synthesis operator of (possibly redundant) discrete Fourier and Gabor tight frames [9]. In such cases, it is only necessary to substitute \mathbf{D}^\top by \mathbf{D}^* in (11). The proof of such an extended lemma, however, gets much more involved by switching to the complex case, and therefore we omit it for simplicity of presentation, as we plan to publish it in a separate paper (currently in preparation).

Using \mathbf{b}_L and \mathbf{b}_H as defined above, the projection (10) can be written as

$$\hat{\mathbf{z}}^{(i)} = \mathbf{v} - \mathbf{D}^* \left(\mathbf{D}\mathbf{v} - \text{proj}_{[\mathbf{b}_L, \mathbf{b}_H]}(\mathbf{D}\mathbf{v}) \right) \quad \text{with} \quad \mathbf{v} = \bar{\mathbf{z}}^{(i)} - \mathbf{u}^{(i-1)}. \tag{12}$$

3.3 Comparing Computational Complexity

In both SPADE algorithms, the computational cost is dominated by the analysis and synthesis operators. Returning to Algorithms 1 and 2, we see that A-SPADE requires one analysis in step 2 and one synthesis in the projection (5). In the case of S-SPADE, the projection is the only demanding calculation, and according to the new formula (12), it requires one synthesis and one analysis. This shows that *per iteration*, the two algorithms are equally demanding. This breaks down the disadvantage of S-SPADE as presented in [24].

4 Experiments and Results

Experiments are designed to compare A-SPADE and S-SPADE algorithms in terms of quality of restoration and computational time. The quality of restoration is evaluated using ΔSDR , which expresses the signal-to-distortion ratio improvement, according to the following formula:

$$\Delta\text{SDR} = \text{SDR}(\mathbf{x}, \hat{\mathbf{x}}) - \text{SDR}(\mathbf{x}, \mathbf{y}) \tag{13}$$

where \mathbf{x} represents the original signal (known in our study), \mathbf{y} is the clipped signal and $\hat{\mathbf{x}}$ is the reconstructed signal, while the SDR itself is defined as

$$\text{SDR}(\mathbf{u}, \mathbf{v}) = 10 \log_{10} \frac{\|\mathbf{u}\|_2^2}{\|\mathbf{u} - \mathbf{v}\|_2^2} \text{ [dB]}. \tag{14}$$

The results below are usually presented in *average* Δ SDR values, taking the arithmetic mean of the particular values from all the tested audio signals in dB.

The advantage of using Δ SDR over the plain SDR is that the Δ SDR value remains the same irrespective of whether the SDR is computed on the whole signal or on the clipped samples only. (This can be easily shown directly from (13), using the fact that our algorithms are consistent, i.e. the reliable samples of the recovered signal and of the clipped signal match.)

Experiments were performed on five audio samples with an approximate duration of 5 s at a sampling frequency of 16 kHz. These excerpts were thoroughly selected to be diverse enough in tonal content and in sparsity with respect to the time-frequency transform. As a preprocessing step, the signals under consideration were peak-normalized and then artificially clipped using multiple clipping thresholds, $\theta_c \in \{0.1, 0.2, \dots, 0.9\}$. The algorithms were implemented in MATLAB R2017a and ran on a PC with Intel i7-3770, 16 GB RAM in single thread mode.

Note that some authors [16, 22–24] evaluate the quality of restoration depending on the *input SDR*, while in this paper we plot the results against the *clipping threshold* θ_c . To get a notion of their relationship, we attach Table 1 which shows both the θ_c and the average input SDR values.

Table 1. Average SDR values for a particular clipping threshold θ_c computed on the test signals as a whole and on the clipped samples only.

θ_c [-]	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
SDR [dB] whole signal	3.71	7.49	11.40	15.54	20.15	25.32	31.44	38.74	48.11
SDR [dB] clipped samples	3.46	6.30	8.68	10.78	13.16	15.22	18.04	20.37	23.63

Although the original SPADE algorithms [24] were purely designed to process individual windowed time-frames, one after another, we also include an experiment using SPADE on the whole signal, considering the DGT coefficients all at once (Sect. 4.1). Then in Sect. 4.2, the classical SPADE setup is investigated, and in later sections the influence of the window length, transform redundancy and window overlap is considered. Note that in this paper, the term *redundancy* specifies the rate of oversampling in the frequency domain—for example, using an oversampled Fourier analysis with 2048 frequency channels applied to a signal of length 1024 means redundancy 2.

4.1 SPADE Applied to Whole Signal

Figure 1 presents the SDR improvement (Δ SDR) for signals processed with SPADE as a whole. The relaxation parameters of both algorithms are set to

$r = 1, s = 100$ and $\epsilon = 0.1$. In this experiment, the most common DGT declipping setting such as 1024-sample-long Hann window and 75% overlap is used, although according to Sect. 4.3 such setting favors the analysis approach, which performs better with shorter windows. Redundancy levels 1, 2 and 4 are achieved by setting the number of frequency channels M to 1024, 2048 and 4096. The black line in Fig. 1 denotes the result of S-SPADE with redundancy 1. However, this is identical to A-SPADE results with the same redundancy—in such a case, $\mathbf{D}^{-1} = \mathbf{A}$ and the algorithms perform equally (see Sect. 3).

An iteration of S-SPADE is typically slightly slower (approximately by 2%) than an iteration of A-SPADE. However, in general, S-SPADE needs fewer iterations to converge. The algorithm is considered converged if the condition on row 4 in both Algorithms 1 and 2 gets fulfilled, i.e. the termination function falls under a prescribed ϵ . Figure 2 presents the computation times; it is clear that S-SPADE converges significantly faster than A-SPADE does, especially at higher redundancies. The average course of the termination function is presented in Fig. 7.

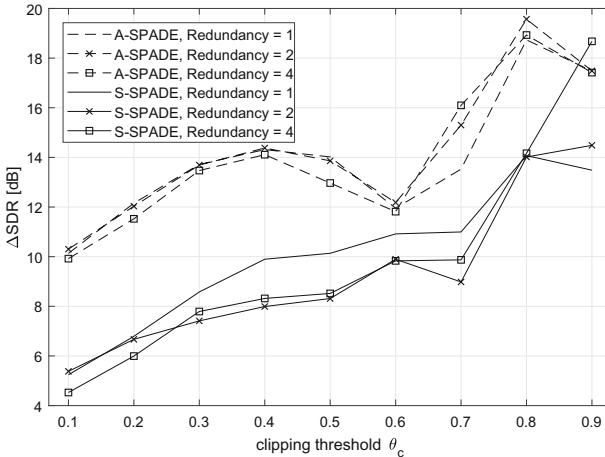


Fig. 1. Declipping performance in terms of ΔSDR performed on the whole signal.

4.2 SPADE with Signal Segmentation

The disadvantage of the approach in Sect. 4.1 is that the largest time-frequency coefficients are selected from the whole signal, and the placement of the coefficients over time is not taken into account. This can easily result in selecting a group of significant coefficients from a short time period and ignoring coefficients that are significant rather locally. Thus, (as will be confirmed by experiments) it is more beneficial to process the signal with SPADE *block by block*.

For this experiment, the relaxation parameters are set according to the original paper [24], i.e. $r = 1, s = 1$ and $\epsilon = 0.1$. The transform parameters are set as

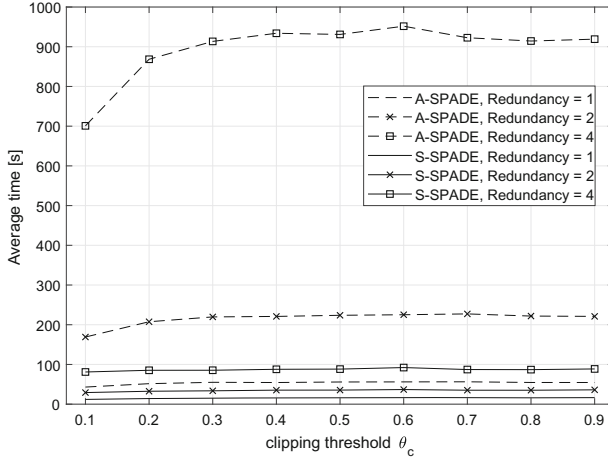


Fig. 2. Average computational times for declipping performed on the whole signal.

in the previous experiment, i.e. the sliding Hann window 1024 samples long with 75% overlap and DFT with redundancy 1, 2 and 4 is used in each time-block.

Figure 3 presents Δ SDR results of both the A-SPADE and S-SPADE algorithms with processing by blocks. Even in this experiment, A-SPADE performs slightly better but it is worth repeating that the choice of the window length suits better the A-SPADE. Interestingly, A-SPADE performs somewhat better with more redundant DFT, while S-SPADE, on the contrary, performs best with no redundancy at all.

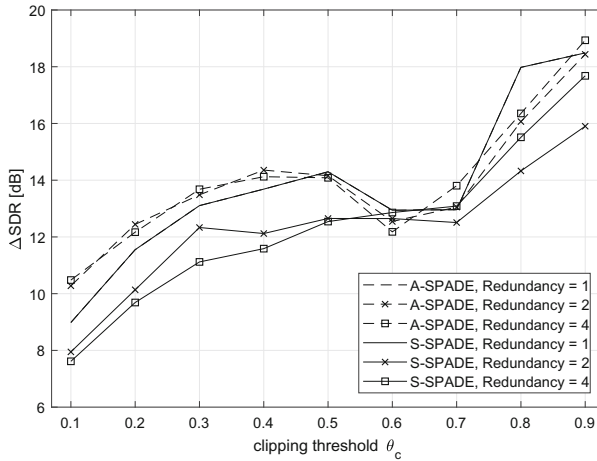


Fig. 3. Declipping performance in terms of Δ SDR performed with signal segmentation.

Apart from the overall performance, we also evaluated the two algorithms locally—we wanted to know whether A-SPADE or S-SPADE better recovers the signal *within a short time range*. Figures 4 and 5 demonstrate SDR results on two audio signals using the Hann window 1024 samples long with 75% overlap and DFT with redundancy 2. For each 2048-sample-long block we computed two corresponding SDR values, which are represented by a marker in the scatter plot. For clarity, we only used clipping thresholds from 0.1 to 0.5. The SDR values were computed using formula (14) on the whole signals; computing SDRs on clipped samples would only reflect in a pure shift of axes in the scatter plot.

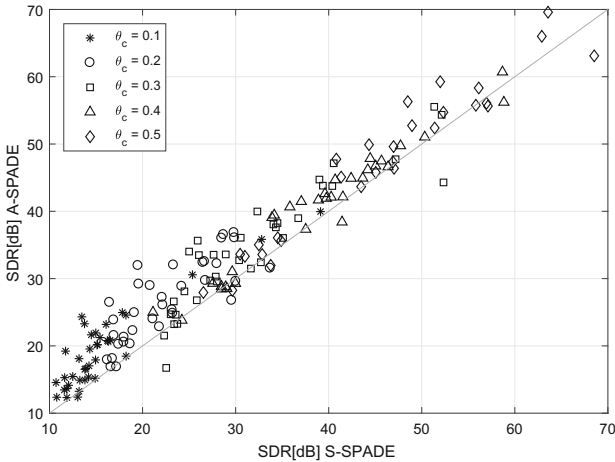


Fig. 4. Scatter plot of SDR values for both S-SPADE and A-SPADE computed locally on sliding blocks 2048 samples long. It is clear that in most time chunks, A-SPADE results are better than those of S-SPADE. Results shown here are for the acoustic guitar signal, but nevertheless such a scatter plot is obtained for most of our test signals.

When redundancy 1 is used, the two algorithms perform identically, and they also terminate after the same number of iterations. In light of this, computation times presented in Fig. 6 show that in such a case, A-SPADE is marginally faster. For more redundant transforms, S-SPADE needs fewer iterations to fulfill the termination criterion and its solution is obtained more quickly.

Figure 7 presents the average course of the termination function (row 4 in both Algorithm 1 and 2). For S-SPADE, this function decreases faster, causing the whole algorithm to converge in fewer iterations.

4.3 Window Length

In many declipping algorithms where processing by blocks or via STFT is done, such as [2, 22, 23, 33], the usual block (or window) length is set to 1024 samples. This experiment is designed to compare both SPADE algorithms depending

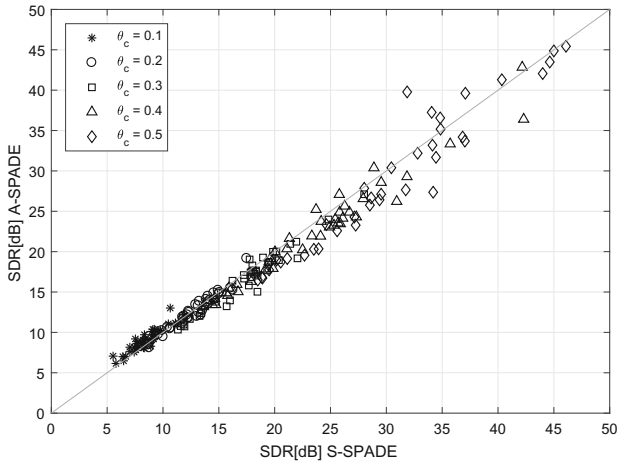


Fig. 5. Scatter plot of SDR values for both S-SPADE and A-SPADE computed locally on sliding blocks 2048 samples long. More than half the time chunks resulted in markers below the identity line, indicating that S-SPADE returned better results. The audiosignal used here is a heavy metal song; note that it was hard to find a signal with such a scatter plot. The result may seem optimistic for S-SPADE—however considering that already the input signal has been clipped on purpose (as is commonly done in this music genre, making the signal far from being sparse), it in turn means that the A-SPADE in effect outperforms the S-SPADE again.

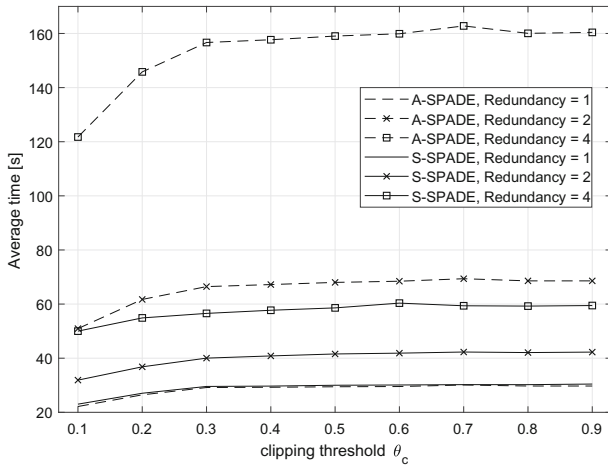


Fig. 6. Declipping performance in terms of average computational time performed block-by-block.

on selected window length. Signals are processed block by block similarly to the previous experiment, except that the redundancy of the DFT is 2 and the length of the sliding window is set to 512, 1024, 2048 and 4096 samples. In all four cases, the window overlap is fixed to 75%.

Figures 8 and 9 present Δ SDR results depending on the window length for A-SPADE and S-SPADE respectively. For the analysis approach, the length of 2048 samples seems to give the best results for most clipping thresholds. When using shorter (512 samples) or longer (4096 samples) windows, the SDR performance drops down approximately by 2 dB. On the other hand, according

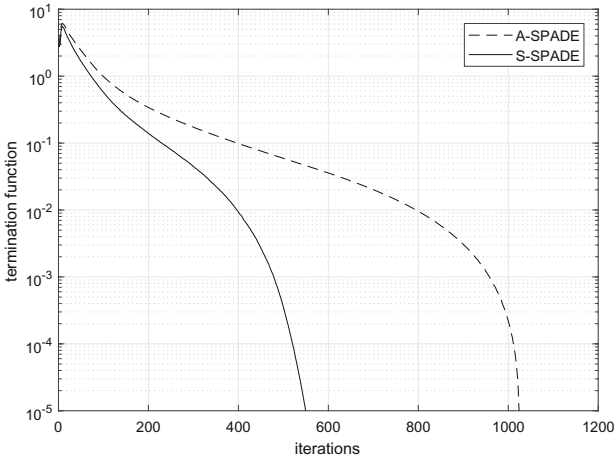


Fig. 7. Average course of the termination function during iterations for redundancy 2.

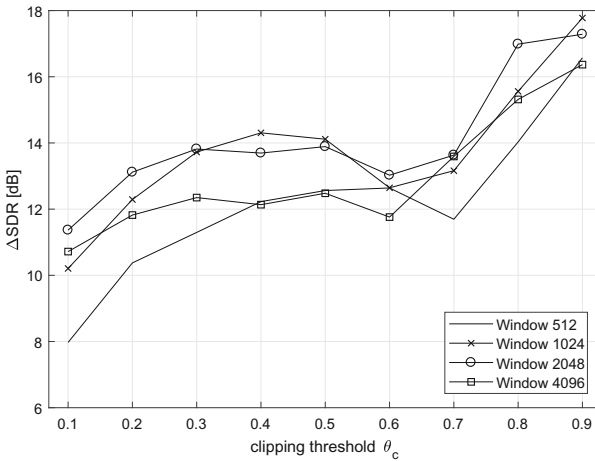


Fig. 8. Declipping performance of A-SPADE in terms of Δ SDR for different window lengths.

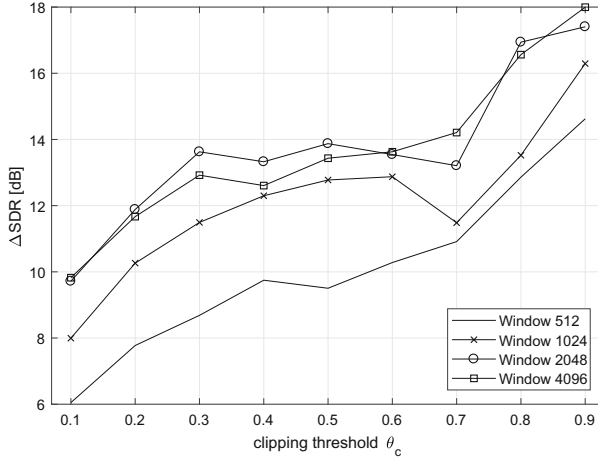


Fig. 9. Declipping performance of S-SPADE in terms of Δ SDR for different window lengths.

to Fig. 9 the synthesis approach performs better with longer windows. The length of 2048 samples seems to be optimal for S-SPADE as well, but the 4096-sample-long window performs by 2 dB better than the 1024 long one.

As far as the computation time is concerned, a longer window means longer computation time. Average computational times for the window lengths 512, 1024, 2048 and 4096 are listed in Table 2.

Table 2. Average computation times in seconds depending on window length using overcomplete DFT with redundancy 2.

Window length	512	1024	2048	4096
A-SPADE	53	68	104	207
S-SPADE	34	41	60	115

4.4 Window Overlap

Window overlap is also an important parameter of the transform; it affects not only the quality of restoration but also the computational time. Therefore, in this experiment, the restoration quality depending on window overlap is explored. As in the previous experiment, DFT with redundancy 2 and Hann window 1024 samples long is used.

Figure 10 shows an expectable fact that the bigger the overlap is set, the better in terms of SDR the results are produced. In line with the results given above, A-SPADE performs slightly better than S-SPADE due to the chosen window length. More interestingly, the performance of the synthesis version drops

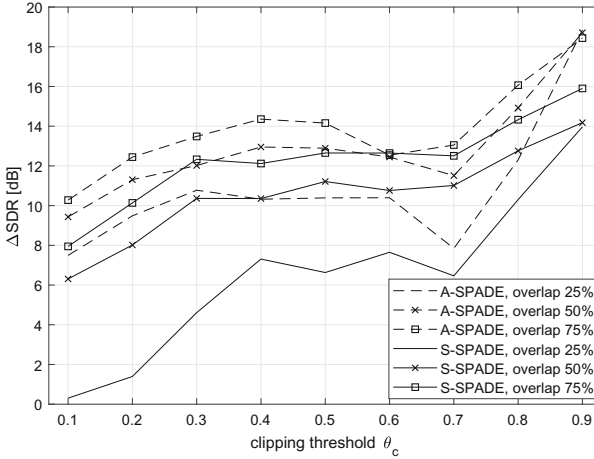


Fig. 10. Declippling performance in terms of ΔSDR for different window overlaps.

significantly when the overlap is set to 25%. Thus, for a good reconstruction, it is necessary to set the window overlap at least to 50% of the window length.

Table 3 shows the average computation times for overlaps of 25, 50 and 75% of the window length. We note that an overlap larger than 75% further increases the computation time but does not bring much improvement in terms of SDR (these facts are not shown).

Table 3. Average computation times in seconds depending on window overlap.

Overlap	25%	50%	75%
A-SPADE	22	34	68
S-SPADE	14	21	41

5 Implementation

Developing the idea of reproducible research, we make our Matlab codes publicly available. The bundle is downloadable from URL http://www.utko.feec.vutbr.cz/~rajmic/software/aspade_vs.sspade.zip. The main file of the SPADE package is a batch file `declipping_main.m`, reading the audio, normalizing and clipping the signal by calling `hard_clip.m`. It is possible to set transform parameters, such as the window length, overlap, window type, and the transform redundancy.

To process signals block-by-block, `spade_segmentation.m` is used. This function performs signal padding, dividing into blocks, multiplying by the analysis window and, after processing, multiplying by the synthesis window and folding

blocks back together (in the common “overlap-add” manner). The SPADE algorithm itself is implemented in two m-files: `aspade.m` for the analysis version and `sspade.m` for the synthesis version.

Recall that the spectrum of a real signal is provided with the complex-conjugate structure. Hard thresholding, performed by `hard_thresholding.m`, takes therefore the oversampled spectrum and thresholds the respective *pairs* of complex entries, in order to keep the signal real. Projections onto the set of feasible solutions are implemented in two m-files. Projection in the time domain for A-SPADE according to (6) is implemented in `proj_time.m`. S-SPADE uses `proj_parse_frame.m` according to (11).

6 Conclusion

We exploited a novel projection lemma to speed up the synthesis version of declipping algorithm SPADE. Use the explicit projection formula, the computational cost, dominated by synthesis and analysis operators, is identical for both versions (per iteration). However, S-SPADE needs fewer iterations to converge, thus turning it to be significantly faster than A-SPADE. As a result, S-SPADE is preferable in real-time processing. On average, A-SPADE performs better in terms of ΔSDR than S-SPADE does.

Experiments involving the parameters of the DGT/DFT show that the optimal window size differs for the algorithms. Whereas A-SPADE performs best with shorter windows, S-SPADE, on the contrary, prefers slightly longer windows. The influence of the window overlap is not negligible either—we have shown that the bigger the overlap is, the better the restoration results are obtained, in both algorithms.

Unfortunately, our results for S-SPADE differ from what the original paper [24] reports. The authors of [24] claim that S-SPADE performs slightly better than A-SPADE does in terms of ΔSDR , and also that S-SPADE performs best with redundancy 4. Our results indicate quite the opposite; in particular, our S-SPADE performs worse in terms of ΔSDR and performs best when redundancy is set to 1.

Our future work will be to investigate in greater depth the differences between the synthesis and the analysis model (and their influence on audio restoration methods). We also believe that introducing a psychoacoustic model could lead to higher declipping quality.

Acknowledgement. The authors thank S. Kitić for providing us with his implementation of the SPADE algorithms and for discussion. The work was supported by the joint project of the FWF and the Czech Science Foundation (GAČR): numbers I 3067-N30 and 17-33798L, respectively. Research described in this paper was financed by the National Sustainability Program under grant LO1401. Infrastructure of the SIX Center was used.

References

1. Abel, J., Smith, J.: Restoring a clipped signal. In: 1991 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91, vol. 3, pp. 1745–1748, April 1991
2. Adler, A., Emiya, V., Jafari, M., Elad, M., Gribonval, R., Plumbley, M.: A constrained matching pursuit approach to audio declipping. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 329–332 (2011)
3. Bayram, I., Kamasak, M.: A simple prior for audio signals. *IEEE Trans. Acoust. Speech Signal Process.* **21**(6), 1190–1200 (2013)
4. Bayram, I., Akyıldız, D.: Primal-dual algorithms for audio decomposition using mixed norms. *Signal Image Video Process.* **8**(1), 95–110 (2014)
5. Bertsekas, D.: *Nonlinear Programming*. Athena Scientific, Belmont (1999)
6. Bilen, C., Ozerov, A., Perez, P.: Audio declipping via nonnegative matrix factorization. In: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–5, October 2015
7. Boyd, S.P., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122 (2011)
8. 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)
9. Christensen, O.: *Frames and Bases, An Introductory Course*. Birkhäuser, Boston (2008)
10. Combettes, P., Pesquet, J.: Proximal splitting methods in signal processing. In: Bauschke, H., Burachik, R., Combettes, P., Elser, V., Luke, D., Wolkowicz, H. (eds.) *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer Optimization and Its Applications, pp. 185–212. Springer, New York (2011)
11. Dahimene, A., Noureddine, M., Azrar, A.: A simple algorithm for the restoration of clipped speech signal. *Informatica* **32**, 183–188 (2008)
12. Defraene, B., Mansour, N., Hertogh, S.D., van Waterschoot, T., Diehl, M., Moonen, M.: Declipping of audio signals using perceptual compressed sensing. *IEEE Trans. Audio Speech Lang. Process.* **21**(12), 2627–2637 (2013)
13. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci.* **100**(5), 2197–2202 (2003)
14. Duhamel, P., Vetterli, M.: Fast fourier transforms: a tutorial review and a state of the art. *Signal Process.* **19**, 259–299 (1990)
15. Fong, W., Godsill, S.: Monte carlo smoothing for non-linearly distorted signals. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), vol. 6, pp. 3997–4000 (2001)
16. Gaultier, C., Bertin, N., Kitić, S., Gribonval, R.: A modeling and algorithmic framework for (non)social (co)sparse audio restoration, November 2017
17. Godsill, S.J., Wolfe, P.J., Fong, W.N.: Statistical model-based approaches to audio restoration and analysis. *J. New Music Res.* **30**(4), 323–338 (2001)
18. Gröchenig, K.: *Foundations of Time-Frequency Analysis*. Birkhäuser, Boston (2001)
19. Harvilla, M.J., Stern, R.M.: Least squares signal declipping for robust speech recognition (2014)

20. Holighaus, N., Wiesmeyer, C.: Construction of warped time-frequency representations on nonuniform frequency scales, part i: Frames. [arXiv:1409.7203](https://arxiv.org/abs/1409.7203) (2016)
21. Janssen, A.J.E.M., Veldhuis, R.N.J., Vries, L.B.: Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Trans. Acoust. Speech Signal Process.* **34**(2), 317–330 (1986)
22. Kitić, S., Jacques, L., Madhu, N., Hopwood, M., Spriet, A., De Vleeschouwer, C.: Consistent iterative hard thresholding for signal declipping. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5939–5943, May 2013
23. Kitić, S., Bertin, N., Gribonval, R.: Audio declipping by cosparse hard thresholding. In: 2nd Traveling Workshop on Interactions Between Sparse Models and Technology (2014)
24. Kitić, S., Bertin, N., Gribonval, R.: Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) *LVA/ICA 2015*. LNCS, vol. 9237, pp. 243–250. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_28
25. Málek, J.: Blind compensation of memoryless nonlinear distortions in sparse signals. In: 21st European Signal Processing Conference (EUSIPCO 2013), pp. 1–5, September 2013
26. Miura, S., Nakajima, H., Miyabe, S., Makino, S., Yamada, T., Nakadaï, K.: Restoration of clipped audio signal using recursive vector projection. In: *TENCON 2011–2011 IEEE Region 10 Conference*, pp. 394–397, November 2011
27. Nam, S., Davies, M., Elad, M., Gribonval, R.: The cosparse analysis model and algorithms. *Appl. Comput. Harmonic Anal.* **34**(1), 30–56 (2013)
28. Necciari, T., Balazs, P., Holighaus, N., Sondergaard, P.: The erblet transform: An auditory-based time-frequency representation with perfect reconstruction. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 498–502, May 2013
29. Sorel, M., Bartoš, M.: Efficient jpeg decompression by the alternating direction method of multipliers. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 271–276, December 2016
30. Průša, Z., Søndergaard, P., Balazs, P., Holighaus, N.: LTFAT: A Matlab/Octave toolbox for sound processing. In: *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR 2013)*, Marseille, France, Laboratoire de Mécanique et d’Acoustique, Publications of L.M.A., pp. 299–314, October 2013
31. Rajmic, P., Bartlová, H., Průša, Z., Holighaus, N.: Acceleration of audio inpainting by support restriction. In: 7th International Congress on Ultra Modern Telecommunications and Control Systems (2015)
32. Selesnick, I.: Least squares with examples in signal processing, April 2013
33. Siedenburg, K., Kowalski, M., Dorfler, M.: Audio declipping with social sparsity. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1577–1581. IEEE (2014)
34. Tachioka, Y., Narita, T., Ishii, J.: Speech recognition performance estimation for clipped speech based on objective measures. *Acoust. Sci. Technol.* **35**(6), 324–326 (2014)
35. Tan, C.T., Moore, B.C.J., Zacharov, N.: The effect of nonlinear distortion on the perceived quality of music and speech signals. *J. Audio Eng. Soc.* **51**(11), 1012–1031 (2003)
36. Weinstein, A.J., Wakin, M.B.: Recovering a clipped signal in sparseland. *CoRR abs/1110.5063* (2011)



Consistent Dictionary Learning for Signal Declipping

Lucas Rencker¹(✉), Francis Bach², Wenwu Wang¹, and Mark D. Plumbley¹

¹ Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, UK
{l.rencker,w.wang,m.plumbley}@surrey.ac.uk
² SIERRA-Project Team, INRIA, Paris, France
francis.bach@inria.fr

Abstract. Clipping, or saturation, is a common nonlinear distortion in signal processing. Recently, declipping techniques have been proposed based on sparse decomposition of the clipped signals on a fixed dictionary, with additional constraints on the amplitude of the clipped samples. Here we propose a dictionary learning approach, where the dictionary is directly learned from the clipped measurements. We propose a soft-consistency metric that minimizes the distance to a convex feasibility set, and takes into account our knowledge about the clipping process. We then propose a gradient descent-based dictionary learning algorithm that minimizes the proposed metric, and is thus consistent with the clipping measurement. Experiments show that the proposed algorithm outperforms other dictionary learning algorithms applied to clipped signals. We also show that learning the dictionary directly from the clipped signals outperforms consistent sparse coding with a fixed dictionary.

1 Introduction

Clipping is a common nonlinear distortion in analog and digital systems, that often happens due to dynamic range limitations. When the signal energy is too high, the waveform is truncated above a certain level, and samples above that level are lost. Declipping is the task of recovering the clipped samples from the surrounding, unclipped samples. Early strategies to recover a clipped signal include autoregressive modelling [1], bandwidth limited models [2], or Bayesian estimation [3]. More recently, sparsity-based declipping techniques have attracted a lot of interest. The idea is that the original signal can be sparsely represented using a known dictionary of atoms. Declipping can be treated as a simple signal *inpainting* problem, i.e. by discarding the clipped samples and solving a sparse decomposition problem on the unclipped samples [4]. However it was noted in [4, 5] that the reconstruction can be greatly improved by using

L. Rencker—The research leading to these results has received funding from the European Union’s H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement no 642685 MacSeNet.

extra information in the reconstruction process: indeed, we know that the clipped samples should have an amplitude that is *greater* than the clipping threshold. Several approaches have been proposed in the literature in order to enforce clipping consistency, i.e. taking into account the clipping threshold. Sparse decomposition with amplitude constraints were proposed in [4–9], and solved using a two-step algorithm [4], Alternating Direction Method of Multipliers (ADMM) [6, 10], or using general purpose convex optimization toolboxes [7–9, 11]. However these approaches can be computationally intensive, and possibly non robust to measurement noise. Smooth regularizers were proposed in [12, 13], which lead to simple unconstrained cost functions, and can be optimized using variants of well known algorithms such as Iterative Hard Thresholding (IHT) or Iterative Shrinkage/Thresholding (ISTA). Additional information has also been used, such as perceptual weights [8], social sparsity priors [13], or multichannel data [14].

Sparsity-based declipping techniques proposed in the literature use fixed dictionaries such as discrete cosine transform (DCT) or Gabor. However, dictionary learning has proved to perform better in a variety of signal reconstruction tasks, such as denoising [15] or inpainting [16]. Well known dictionary learning algorithms have been proposed for denoising or inpainting [15–17], however dictionary learning from clipped measurements has not been addressed in the literature. In this paper we propose a dictionary learning algorithm that is able to learn directly from nonlinearly clipped measurements. We formulate the declipping problem as a problem of minimizing the distance between a sparse signal and a convex feasibility set. This provides a convex and smooth cost function which generalizes the Euclidean distance commonly used in sparse coding and dictionary learning. We then propose a gradient-descent based sparse coding and dictionary learning algorithm, that takes into account our knowledge about the clipping process. Experiments show that the proposed consistent dictionary learning algorithm performs better on the task of declipping than state-of-the-art dictionary learning algorithms for signal inpainting. We also show that the proposed consistent dictionary learning improves the reconstruction, compared to consistent sparse coding with a fixed dictionary.

The paper is organized as follows: in Sect. 2 we briefly give an overview of sparsity-based declipping techniques, and of dictionary learning. In Sect. 3 we propose a new formulation of the declipping problem, and a consistent dictionary learning algorithm for signal declipping. Experiments are presented in Sect. 4, before the conclusion is drawn.

2 Background

2.1 Signal Declipping

Let $\mathbf{x} \in \mathbb{R}^N$ be a clean input signal, and $\mathbf{y} \in \mathbb{R}^N$ its clipped measurement. In this paper we consider the case of *hard* clipping, where each sample y_i is measured as:

$$y_i = \begin{cases} \theta^+ & \text{if } x_i \geq \theta^+ \\ \theta^- & \text{if } x_i \leq \theta^- \\ x_i & \text{otherwise,} \end{cases} \quad (1)$$

where $\theta^+ > 0$ and $\theta^- < 0$ are positive and negative clipping thresholds respectively, and x_i is the input sample. This can be written in vector form as:

$$\mathbf{y} = \mathbf{M}^r \mathbf{x} + \theta^+ \mathbf{M}^{c^+} \mathbf{1} + \theta^- \mathbf{M}^{c^-} \mathbf{1}, \quad (2)$$

where $\mathbf{1}$ is the all-ones vector in \mathbb{R}^N , and \mathbf{M}^r , \mathbf{M}^{c^+} and \mathbf{M}^{c^-} are diagonal *sensing* matrices in $\{0, 1\}^{N \times N}$ that define the *reliable*, positive and negative clipped samples respectively. These matrices can be estimated by detecting samples that have reached the clipping threshold, e.g. $[\mathbf{M}^{c^+}]_{i,i} = 1$ if $y_i = \theta^+$, or 0 otherwise. Since the clipped samples are missing, a simple way to treat declipping is to formulate it as an *inpainting* problem, i.e. a problem of interpolating missing samples [4]. Assuming that the original signal can be sparsely represented in a known dictionary $\mathbf{D} \in \mathbb{R}^{N \times M}$, the inpainting problem can be formulated as:

$$\min_{\alpha \in \mathbb{R}^M} \|\mathbf{M}^r(\mathbf{y} - \mathbf{D}\alpha)\|_2^2 \quad \text{s.t.} \quad \|\alpha\|_0 \leq K, \quad (3)$$

where $\|\cdot\|_0$ is the ℓ_0 pseudo-norm, and K is a parameter that controls the sparsity level. Eq. (3) is a classical sparse coding problem, which can be solved using well known algorithms like IHT [18]. However, we can use extra information about the clipping process. Indeed, we know that the clipped samples should have an amplitude that is above (resp. below) the clipping threshold θ^+ (resp. θ^-). This can be enforced using amplitude constraints in the reconstruction process [4,5]:

$$\min_{\alpha \in \mathbb{R}^M} \|\mathbf{M}^r(\mathbf{y} - \mathbf{D}\alpha)\|_2^2 \quad \text{s.t.} \quad \begin{cases} \|\alpha\|_0 \leq K \\ \mathbf{M}^{c^+} \mathbf{D}\alpha \succeq \theta^+ \mathbf{M}^{c^+} \mathbf{1} \\ \mathbf{M}^{c^-} \mathbf{D}\alpha \preceq \theta^- \mathbf{M}^{c^-} \mathbf{1} \end{cases} \quad (4)$$

Equation (4) is a difficult non-convex and constrained optimization problem, which cannot be readily solved using off-the-shelf sparse decomposition solvers such as IHT. A two-step algorithm was proposed in [4], where the support of non-zero atoms is first estimated using (3), and the signal is then estimated using a constrained least squares on the estimated support. However, the support selection does not take into account the clipping constraints and is thus suboptimal. A similar constraint-based formulation was proposed in [6]:

$$\min_{\alpha \in \mathbb{R}^M} \|\alpha\|_0 + \mathbb{1}_{\mathcal{C}(\mathbf{y})}(\mathbf{D}\alpha), \quad (5)$$

where $\mathbb{1}_{\mathcal{C}(\mathbf{y})}$ is the indicator function of the set $\mathcal{C}(\mathbf{y})$, and:

$$\mathcal{C}(\mathbf{y}) \triangleq \{\mathbf{x} \mid \mathbf{M}^r \mathbf{y} = \mathbf{M}^r \mathbf{x}, \mathbf{M}^{c^+} \mathbf{x} \succeq \mathbf{M}^{c^+} \mathbf{y}, \mathbf{M}^{c^-} \mathbf{x} \preceq \mathbf{M}^{c^-} \mathbf{y}\} \quad (6)$$

is the set of *feasible* signals, i.e. the set of signals that are consistent with the observation \mathbf{y} . The authors in [6] proposed an ADMM based algorithm

to solve (5). The ADMM-based declipper [6] leads to good performance, but proves to be computationally expensive since it involves non-orthogonal projections which need to be computed iteratively¹. Similar ℓ_1 -based constrained formulation were also proposed in [7–9], and solved using general purpose convex optimization toolboxes [11], which can also be time consuming. Moreover, constrained formulation like (5) might not be robust to measurement noise, as will be discussed in the experimental section. Several authors proposed to enforce consistency with the clipped samples in a more tractable way. A smooth regularizer that penalizes clipped samples was proposed in [12]:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^M} \quad & \| \mathbf{M}^r(\mathbf{y} - \mathbf{D} \alpha) \|_2^2 + \| \mathbf{M}^{c+}(\theta^+ \mathbf{1} - \mathbf{D} \alpha)_+ \|_2^2 \\ & + \| \mathbf{M}^{c-}(\theta^- \mathbf{1} - \mathbf{D} \alpha)_- \|_2^2 \quad \text{s.t.} \quad \| \alpha \|_0 \leq K, \end{aligned} \tag{7}$$

where $(u)_+ = \max(0, u)$ and $(u)_- = -(-u)_+$. Since the cost in (7) is smooth, gradient-based sparse coding algorithms can easily be extended to the clipping consistent model (7). A *consistent* IHT was proposed in [12] in order to enforce clipping consistency. A similar formulation with an ℓ_1 norm was proposed in [13] along with ISTA-like algorithms. Although the algorithm in [12] did not perform as well as the ADMM based declipper [6], the soft consistency metric in (7) provides a simple, unconstrained way to enforce consistency with the clipped samples. Moreover, simple iterative thresholding algorithms can be derived, which are computationally faster than solving constrained optimization problems like (5).

2.2 Dictionary Learning

Previously mentioned declipping techniques use fixed dictionaries, such as DCT or Gabor. However in many applications, learning a dictionary that is adaptive to the data has proved to lead to much better signal estimates [15, 16]. A dictionary learning problem (from clean signals) is often formulated as [19]:

$$\min_{\mathbf{D} \in \mathcal{D}, \alpha_t} \sum_t \| \mathbf{x}_t - \mathbf{D} \alpha_t \|_2^2 \quad \text{s.t.} \quad \forall t, \| \alpha_t \|_0 \leq K \tag{8}$$

where $\{ \mathbf{x}_t \}_{1 \dots T}$ is a collection of T signals in \mathbb{R}^N . The dictionary is often constrained to be in $\mathcal{D} = \{ \mathbf{D} \in \mathbb{R}^{N \times M} | \forall i, \| \mathbf{d}_i \|_2 \leq 1 \}$ in order to avoid scaling ambiguity [19]. Many dictionary learning algorithms have been proposed to learn from clean or noisy data, such as MOD [17] or K-SVD [15]. In the case of inpainting, a weighted K-SVD (wK-SVD) has been proposed in order to deal with missing samples [16]. Dictionary learning from nonlinearly clipped data has not been addressed in the literature. Since dictionary learning usually alternates between several iterations of sparse coding and dictionary learning update over large datasets, a

¹ An *analysis* sparsity version of (5) was also proposed in [6], which proved to be computationally more tractable. In this paper we focus on the synthesis sparsity model, and leave the analysis sparsity counterpart for future work.

computationally tractable and stable formulation is needed. In the next section, we propose a soft data-consistency metric, that provides a simple optimization problem for dictionary learning. We then propose a consistent dictionary learning algorithm that is able to learn from the clipped measurements.

3 Consistent Dictionary Learning for Signal Declipping

3.1 Proposed Problem Formulation

We first reformulate declipping as a problem of minimizing the distance between the approximated signal, and the feasible set $\mathcal{C}(\mathbf{y}_t)$ defined in (6):

$$\min_{\mathbf{D} \in \mathcal{D}, \alpha_t} \sum_t d(\mathbf{D} \alpha_t, \mathcal{C}(\mathbf{y}_t))^2 \quad \text{s.t.} \quad \forall t, \|\alpha_t\|_0 \leq K, \quad (9)$$

where $d(\mathbf{x}, \mathcal{C}(\mathbf{y}))$ is the Euclidean distance between \mathbf{x} and the set $\mathcal{C}(\mathbf{y})$, defined as:

$$d(\mathbf{x}, \mathcal{C}(\mathbf{y})) = \min_{\mathbf{z} \in \mathcal{C}(\mathbf{y})} \|\mathbf{x} - \mathbf{z}\|_2. \quad (10)$$

The formulation (9) thus enforces the estimated signals to be “close” to their feasibility sets $\mathcal{C}(\mathbf{y}_t)$ in a Euclidean-distance sense, unlike the formulation in (5) which constrains the signals to be exactly in $\mathcal{C}(\mathbf{y}_t)$. We thus have proposed here a problem of minimizing the distance to a set, which differs from classical sparse coding and dictionary learning approaches which minimize the distance to a point in \mathbb{R}^N . Using (9) and (10), can further be reformulated as a “min-min” problem:

$$\min_{\mathbf{D} \in \mathcal{D}, \alpha_t} \sum_t \min_{\mathbf{z} \in \mathcal{C}(\mathbf{y}_t)} \|\mathbf{D} \alpha_t - \mathbf{z}\|_2^2 \quad \text{s.t.} \quad \forall t, \|\alpha_t\|_0 \leq K. \quad (11)$$

Note that as a minimum of a family of convex functions $\|\cdot\|_2$ over a non-empty and convex set $\mathcal{C}(\mathbf{y})$, $d(\mathbf{x}, \mathcal{C}(\mathbf{y}))$ is a convex cost function [20, Sect. 3.2.5]. Moreover, using Danskin’s Min-Max theorem ([21, Theorem 4.1], originally proposed in [22]), it can be shown that $d(\mathbf{x}, \mathcal{C}(\mathbf{y}))^2$ is differentiable with gradient [23]:

$$\nabla_{\mathbf{x}} d(\mathbf{x}, \mathcal{C}(\mathbf{y}))^2 = 2(\mathbf{x} - \Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{x})), \quad (12)$$

where $\Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{x})$ is the Euclidean projection of \mathbf{x} onto $\mathcal{C}(\mathbf{y})$. The proposed formulation in (9) is thus a problem of minimizing a smooth and convex cost function, with a sparsity constraint, which is similar to the classical dictionary learning problem (8). The proposed cost function thus generalizes the linear least-squares commonly used in sparse coding and dictionary learning.

3.2 Algorithm

We propose a simple gradient descent-based algorithm, which we present in Algorithm 1. The proposed algorithm alternates between a sparse coding step and

a dictionary update step. Similarly to IHT, the sparse coding step alternates between gradient descent (13) and a hard thresholding (14). The dictionary is updated using projected gradient descent (15). n_i and μ_i ($i = 1, 2$) are parameters that control the number of gradient descent steps and step sizes respectively.

Algorithm 1. Dictionary learning for declipping

Require: $\{\mathbf{y}_t\}_{1..T}$, \mathbf{D}^0 , n_1 , n_2 , μ_1 , μ_2

initialize: $\mathbf{D} \leftarrow \mathbf{D}^0$, $\boldsymbol{\alpha}_t \leftarrow \mathbf{0}$

while stopping criterion not reached **do**

for $t = 1..T$ **do**

 ▷ Sparse coding step

for $i = 1, \dots, n_1$ **do**

$$\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}_t + \mu_1 \mathbf{D}^T (\Pi_{\mathcal{C}(\mathbf{y}_t)}(\mathbf{D} \boldsymbol{\alpha}_t) - \mathbf{D} \boldsymbol{\alpha}_t) \quad (13)$$

$$\boldsymbol{\alpha}_t \leftarrow \mathcal{H}_K(\boldsymbol{\alpha}_t) \quad (14)$$

for $j = 1, \dots, n_2$ **do**

 ▷ Dictionary update step

$$\mathbf{D} \leftarrow \Pi_{\mathcal{D}}(\mathbf{D} + \mu_2 \sum_t (\Pi_{\mathcal{C}(\mathbf{y}_t)}(\mathbf{D} \boldsymbol{\alpha}_t) - \mathbf{D} \boldsymbol{\alpha}_t) \boldsymbol{\alpha}_t^T) \quad (15)$$

return $\hat{\mathbf{D}}, \{\hat{\boldsymbol{\alpha}}_t\}_{1..T}$

3.3 Computation of the Residuals, and Interpretation

Algorithm 1 involves the computation of residuals $\Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{D} \boldsymbol{\alpha}) - \mathbf{D} \boldsymbol{\alpha}$ at every step. These residuals can be easily computed in closed form. It can be easily verified that the projection operator $\Pi_{\mathcal{C}(\mathbf{y})}$ is computed as:

$$\Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{x}) = \mathbf{M}^r \mathbf{y} + \mathbf{M}^{c+} \max(\mathbf{y}, \mathbf{x}) + \mathbf{M}^{c-} \min(\mathbf{y}, \mathbf{x}). \quad (16)$$

Note that this is a simple 1-dimensional orthogonal projection, that can be computed at a negligible cost. The residuals can be computed as:

$$\Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{D} \boldsymbol{\alpha}) - \mathbf{D} \boldsymbol{\alpha} = \mathbf{M}^r (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha}) + \mathbf{M}^{c+} (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha})_+ + \mathbf{M}^{c-} (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha})_-. \quad (17)$$

This also shows that the proposed soft-consistency metric (9) can be written in closed form as:

$$\begin{aligned} d(\mathbf{D} \boldsymbol{\alpha}, \mathcal{C}(\mathbf{y}))^2 = & \|\mathbf{M}^r (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha})\|_2^2 + \|\mathbf{M}^{c+} (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha})_+\|_2^2 \\ & + \|\mathbf{M}^{c-} (\mathbf{y} - \mathbf{D} \boldsymbol{\alpha})_-\|_2^2, \end{aligned} \quad (18)$$

which (noticing that $\mathbf{M}^{c+} \mathbf{y} = \theta^+ \mathbf{M}^{c+} \mathbf{1}$) is the same as the cost (7) used in [12, 13]. The proposed approach is thus a different way to motivate the soft-consistency metric (7), and the sparse coding step in Algorithm 1 is equivalent to the ‘‘consistent IHT’’ [12]. Note also that when no sample is clipped, we have $\mathcal{C}(\mathbf{y}) = \{\mathbf{y}\}$, $d(\mathbf{D} \boldsymbol{\alpha}, \mathcal{C}(\mathbf{y}))^2 = \|\mathbf{D} \boldsymbol{\alpha} - \mathbf{y}\|_2^2$, and $\Pi_{\mathcal{C}(\mathbf{y})}(\mathbf{D} \boldsymbol{\alpha}) = \mathbf{y}$. Thus (9) becomes a classical dictionary learning problem, and Algorithm 1 a classical dictionary learning algorithm. The proposed method is thus a generalization of dictionary learning to nonlinearly clipped measurements.

4 Evaluation

We evaluate the performance of the proposed algorithm on audio declipping tasks². The test set consists of 10 speech and 10 music signals of 10 s each, sampled at 16 kHz. The signals were processed with Hamming windows of size $N = 256$ samples, with 75% overlap, for a total of approximately $T = 2500$ frames per signal. The dictionary learning algorithm was initialized with a DCT dictionary of $M = 512$ atoms, and the sparse coefficients initialized to zero. Each gradient descent step was then initialized with a *warm restart* strategy, i.e. using the estimate from the previous iteration [19]. We performed 50 iterations of gradient descent, with 20 iterations for each inner sparse coding and dictionary update step. The gradient descent steps were chosen as $\mu_1 = 1/\|\mathbf{D}\|_2^2$ and $\mu_2 = 1/\|\mathbf{A}\|_2^2$ (with $\mathbf{A} = [\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_T]$), and updated at each iteration using the current estimates \mathbf{D} and \mathbf{A} . When no noise is present, the estimated signals $\hat{\mathbf{x}}$ can be re-projected on the set $\{\mathbf{x} | \mathbf{M}^r \mathbf{x} = \mathbf{M}^r \mathbf{y}\}$ as a final step, in order to avoid approximation errors. The quality of the estimated signal can then be evaluated using the signal to distortion ratio (SDR) computed on the *clipped* samples:

$$\text{SDR}_c(\hat{\mathbf{x}}, \mathbf{x}) = 20 \log \frac{\|(\mathbf{M}^{c+} + \mathbf{M}^{c-})\mathbf{x}\|_2}{\|(\mathbf{M}^{c+} + \mathbf{M}^{c-})(\mathbf{x} - \hat{\mathbf{x}})\|_2}.$$

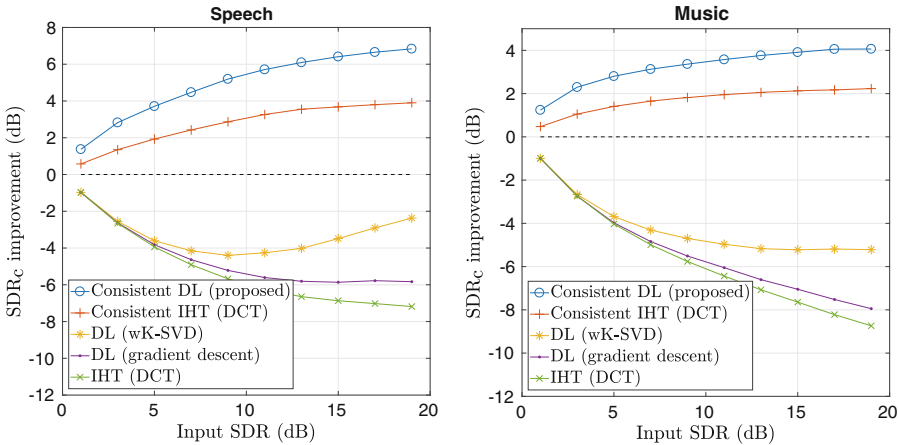


Fig. 1. Comparison with state-of-the-art dictionary learning algorithms

Figure 1 shows the performance of the proposed consistent dictionary learning (DL) algorithm compared to other dictionary learning algorithms for inpainting. We show the average performance for different clipping levels, ranging from severe clip (SDR = 1 dB) to light clip (SDR = 19 dB). As a baseline, we show the performance of IHT, computed on the unclipped samples and with a fixed DCT dictionary. We show the performance of two dictionary learning algorithms computed on the unclipped samples: a gradient descent-based algorithm similar to

² The MATLAB code and some examples are available at <http://www.cvssp.org/Personal/LucasRencker/software.html>.

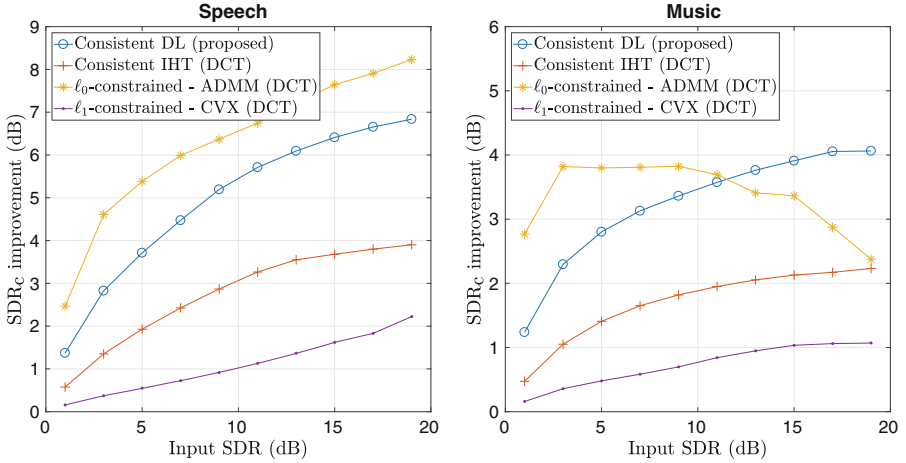


Fig. 2. Comparison with state-of-the-art declipping algorithms

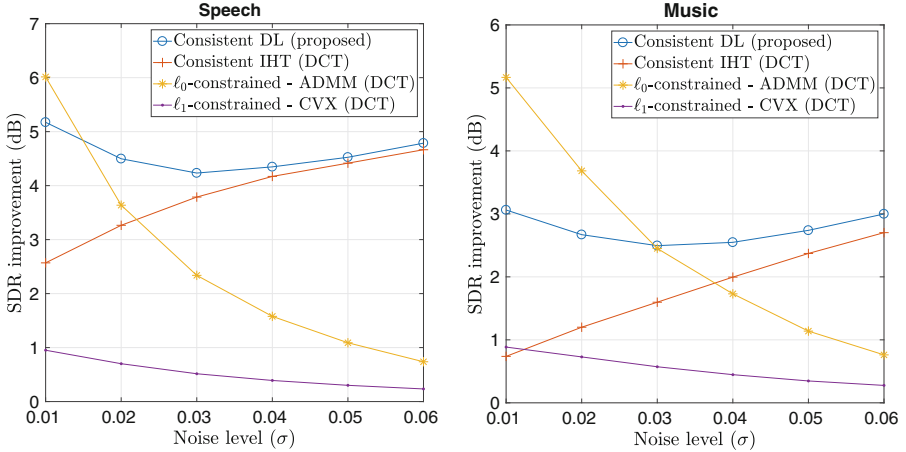


Fig. 3. Robustness to measurement noise

Algorithm 1, and wK-SVD which achieved state-of-the-art performance in signal inpainting [16]. Although these two algorithms slightly improve the reconstruction compared to IHT, the overall performance is quite poor. Consistent IHT [12] clearly outperforms methods that discard clipped samples. The proposed consistent DL algorithm further improves the reconstruction, with an improvement of up to 3dB in the case of speech signals. This shows that learning the dictionary directly from the clipped signals outperforms fixed dictionaries, and that the learned dictionary generalizes well to the clipped samples. Note also that while standard dictionary learning algorithms fail to improve the reconstruction when the data is heavily clipped ($\text{SDR} \leq 5$ dB), the proposed dictionary learning algorithm is still able to improve compared to consistent IHT with fixed DCT.

Figure 2 shows the performance comparison with other declipping algorithms proposed in the literature. We compare with consistent IHT, the ℓ_1 -constrained formulation proposed in [9] solved using CVX [11], and the ℓ_0 -constrained formulation (5) solved using the ADMM-based algorithm proposed in [6], considered as the current state-of-the-art. All ℓ_0 -based algorithms were computed with a fixed $K = 32$, except the ADMM algorithm which we have found does not converge when K is fixed. We have thus implemented ADMM with the adaptive sparsity strategy proposed in [6], which might favor it. Although the proposed consistent DL algorithm does not match ADMM's performance on average, our algorithm bridges the gap between consistent IHT and ADMM, and outperforms ADMM in the case of music signals when $\text{SDR} \geq 12\text{dB}$. However as shown in the next experiment, the proposed algorithm is more robust to measurement noise. Figure 3 shows the reconstruction performance for signals contaminated with additive Gaussian noise with variance σ^2 , and clipped at $\theta = 0.3$. Figure 3 shows that algorithms based on soft-consistency metric such as consistent IHT or the proposed algorithm are more robust to noise than constrained-based formulation. In particular, the proposed algorithm outperforms every other algorithms for noise levels above 0.01 in speech, and 0.03 in music. From a computational point of view, consistent IHT takes about 5 s to process a 10 s signal, the proposed consistent DL and ADMM about 2–3 min, and CVX about an hour.

5 Conclusion

We proposed a smooth and convex cost function for signal declipping, and a dictionary learning algorithm that is able to learn from clipped measurements. The proposed algorithm outperforms classical dictionary learning algorithms, and improves the declipping performance compared to consistent sparse coding with a fixed dictionary. The proposed algorithm is simple and efficient, and the model proposed in (9) could potentially be applied to other nonlinear measurements, such as quantization or 1-bit measurements, which will be addressed in a future publication. Analysis sparsity has shown promising results in [6], so future work will also investigate analysis dictionary learning for declipping.

References

1. Janssen, A., Veldhuis, R.N., Vries, L.B.: Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Trans. Acoust. Speech Signal Process.* **34**(2), 317–330 (1986)
2. Abel, J.S., Smith, J.O.: Restoring a clipped signal. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, pp. 1745–1748 (1991)
3. Godsill, S.J., Wolfe, P.J., Fong, W.N.: Statistical model-based approaches to audio restoration and analysis. *J. New Music Res.* **30**(4), 323–338 (2001)
4. Adler, A., Emiya, V., Jafari, M., Elad, M., Gribonval, R., Plumbley, M.D.: Audio inpainting. *IEEE Trans. Audio Speech Lang. Process.* **20**(3), 922–932 (2012)

5. Adler, A., Emiya, V., Jafari, M.G., Elad, M., Gribonval, R., Plumbley, M.D.: A constrained matching pursuit approach to audio declipping. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 329–332 (2011)
6. Kitić, S., Bertin, N., Gribonval, R.: Sparsity and cosparsity for audio declipping: a flexible non-convex approach. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) LVA/ICA 2015. LNCS, vol. 9237, pp. 243–250. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_28
7. Mansour, H., Saab, R., Nasiopoulos, P., Ward, R.: Color image desaturation using sparse reconstruction. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 778–781 (2010)
8. Defraene, B., Mansour, N., Hertogh, S.D., van Waterschoot, T., Diehl, M., Moonen, M.: Declipping of audio signals using perceptual compressed sensing. *IEEE Trans. Audio Speech Lang. Process.* **21**(12), 2627–2637 (2013)
9. Foucart, S., Needham, T.: Sparse recovery from saturated measurements. *Inf. Inference J. IMA* **6**(2), 196–212 (2016)
10. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122 (2011)
11. Grant, M., Boyd, S., Ye, Y.: CVX: Matlab software for disciplined convex programming (2008)
12. Kitić, S., Jacques, L., Madhu, N., Hopwood, M.P., Spriet, A., Vleeschouwer, C.D.: Consistent iterative hard thresholding for signal declipping. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5939–5943, May 2013
13. Siedenburg, K., Kowalski, M., Dörfler, M.: Audio declipping with social sparsity. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, pp. 1577–1578, May 2014
14. Ozerov, A., Bilen, Ç., Pérez, P.: Multichannel audio declipping. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 659–663 (2016)
15. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**(12), 3736–3745 (2006)
16. Mairal, J., Elad, M., Sapiro, G.: Sparse representation for color image restoration. *IEEE Trans. Image Process.* **17**(1), 53–69 (2008)
17. Engan, K., Aase, S.O., Husoy, J.H.: Method of optimal directions for frame design. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. 2443–2446 (1999)
18. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27**(3), 265–274 (2009)
19. Mairal, J., Bach, F., Ponce, J.: Sparse modeling for image and vision processing. *Found. Trends Comput. Graph. Vis.* **8**(2–3), 85–283 (2014)
20. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, New York (2004)
21. Bonnans, J.F., Shapiro, A.: Optimization problems with perturbations: a guided tour. *SIAM Rev.* **40**(2), 228–264 (1998)
22. Danskin, J.M.: *The Theory of Max-Min and its Application to Weapons Allocation Problems*. *Ökonometrie und Unternehmensforschung*. Springer, Heidelberg (1967). <https://doi.org/10.1007/978-3-642-46092-0>
23. Holmes, R.B.: Smoothness of certain metric projections on Hilbert space. *Trans. Am. Math. Soc.* **184**, 87–100 (1973)



Learning Fast Dictionaries for Sparse Representations Using Low-Rank Tensor Decompositions

Cássio F. Dantas^{1(✉)}, Jérémy E. Cohen², and Rémi Gribonval¹

¹ Univ Rennes, Inria, CNRS, IRISA, Rennes, France
`cassio.fraga-dantas@inria.fr`

² University of Mons, Mons, Belgium
`jeremy.cohen@umons.ac.be`

Abstract. A new dictionary learning model is introduced where the dictionary matrix is constrained as a sum of R Kronecker products of K terms. It offers a more compact representation and requires fewer training data than the general dictionary learning model, while generalizing Tucker dictionary learning. The proposed Higher Order Sum of Kroneckers model can be computed by merging dictionary learning approaches with the tensor Canonic Polyadic Decomposition. Experiments on image denoising illustrate the advantages of the proposed approach.

Keywords: Kronecker product · Tensor data · Dictionary learning

1 Introduction

Multi-dimensional data arise in a large variety of applications such as telecommunications, biomedical sciences, image and video processing to name a few [10].

Explicitly accounting for this tensorial structure of the data can be more advantageous than relying on its vectorized version and losing the original neighboring relations, besides providing more efficient and economic representation and subsequent processing. The Kronecker product structure arises naturally when dealing with multi-dimensional (tensorial) data, since it manages to recover the underlying tensorial nature of vectorized data samples. Indeed, when applied to a vectorized tensor, each composing factor of a Kronecker-structured linear operator acts independently on each mode of the data. Conveniently, such operators are more compact to store and can be applied much more efficiently than their unstructured counterpart.

Nevertheless, relatively little attention has been paid to exploiting this type of structure on representation learning methods such as dictionary learning

J. E. Cohen—Author acknowledges the support by the F.R.S.-FNRS (incentive grant for scientific research n° F.4501.16).

algorithms, which aim at obtaining a set of explanatory variables (the dictionary) capable of sparsely approximating an input dataset. Conventional methods impose no particular structure to the dictionary matrix learned from the vectorized input samples, therefore completely disregarding a potential multi-dimensional characteristic of the data.

In this paper, we provide a novel method to induce a generalized version of the Kronecker structure on the dictionary (namely, a sum of Kronecker products). To this end, we draw a parallel between the problem of approximating an arbitrary linear operator by a Kronecker-structured one and the low-rank tensor approximation problem.

The learned dictionary is constrained to the following structure:

$$\mathbf{D} = \sum_{r=1}^R \mathbf{D}_1^r \otimes \cdots \otimes \mathbf{D}_K^r = \sum_{r=1}^R \bigotimes_{k=1}^K \mathbf{D}_k^r;$$

which we refer to as a *rank- R K -Kronecker-structured* matrix (or simply (R, K) -KS and even K -KS when $R = 1$). In [17] R is referred to as the *separation rank*. As we will see in Sect. 1.1, this structure may lead to significant memory and computational savings when compared to an unstructured matrix. At the same time, because we allow sums of several Kronecker terms, we make the structure more flexible thus increasing its approximation capabilities with respect to the conventional Tucker dictionary model [4].

1.1 Motivations

The quest for Kronecker-structured linear operators has various motivations besides the suitability to multi-dimensional signals: (1) reduced computational complexity; (2) diminished memory requirements and (3) smaller sample complexity on learning applications.

The complexity savings on matrix-vector multiplications are explained as follows. For an $(n \times m)$ K -KS matrix $\mathbf{D} = \mathbf{D}_K \otimes \cdots \otimes \mathbf{D}_1$ with factors $\mathbf{D}_k \in \mathbb{R}^{n_k \times m_k}$, $n = \prod_{k=1}^K n_k$ and $m = \prod_{k=1}^K m_k$, the matrix-vector product $\mathbf{D}\mathbf{x}$ can be rewritten as

$$\mathbf{y} = (\mathbf{D}_K \otimes \cdots \otimes \mathbf{D}_1)\mathbf{x} \quad \rightarrow \quad \underline{\mathbf{Y}} = \underline{\mathbf{X}} \times_1 \mathbf{D}_1 \times_2 \cdots \times_K \mathbf{D}_K \quad (1)$$

where $\underline{\mathbf{Y}} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ and $\underline{\mathbf{X}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ are the tensorized versions of $\mathbf{y} \in \mathbb{R}^{n_1 \cdots n_K}$ and $\mathbf{x} \in \mathbb{R}^{m_1 \cdots m_K}$ respectively [10] and \times_k denotes the mode- k tensor matrix product. In other words, it comes down to multiplying each factor by the corresponding mode on the tensorized version $\underline{\mathbf{X}}$ of the vector \mathbf{x} .

Since the composing factors \mathbf{D}_k are much smaller than \mathbf{D} , the total complexity for computing (1) can be significantly smaller than the usual $\mathcal{O}(nm)$. In particular, when the factors are all square (i.e. $n_k = m_k \forall k$) the total number of operations is given by $(\prod_{k=1}^K n_k)(\sum_{k=1}^K n_k)$ [16] compared to $(\prod_{k=1}^K n_k)^2$ operations for an unstructured matrix of the same size. For a sum of R Kronecker products, the mentioned complexity is simply multiplied by R .

In addition, the total storage cost for the structured operator is proportional to $(\sum_{k=1}^K n_k m_k)$ instead of $(\prod_{k=1}^K n_k m_k)$. Similarly, recent studies [15] on the minimax risk for the dictionary identifiability problem showed that the necessary number of samples for reliable reconstruction, up to a given mean squared error, of a KS dictionary within its local neighborhood scales with $(m \sum_{k=1}^K n_k m_k)$ compared to $(m \prod_{k=1}^K n_k m_k)$ for unstructured dictionaries of the same size [9].

1.2 Related Work

The Kronecker structure was introduced in the Dictionary Learning domain by [8,13] both addressing only 2-dimensional data (i.e. 2-KS dictionaries). The model was extended to the 3rd-order (3-KS dictionaries) [12,19] and even for an arbitrary tensor order [4,7] based on the Tucker decomposition, a model coined as Tucker Dictionary Learning. However, none of these works include a sum of Kronecker terms. Even though the formulation in [7] would allow it, they restrict their analysis to the rank-one ($R = 1$) case.

The sum of Kronecker products model was initially explored in the covariance estimation community [3,17] and was recently used in [5] as an extension of the existing Kronecker-structured dictionaries. But once again, these works addressed only the 2nd-order case ($K = 2$). We now extend [5] for an arbitrary tensor order, thus allowing for both $R \geq 1$ and $K \geq 2$. An advantage of our approach w.r.t [5,7] is that here we can choose the desired number of summing terms beforehand without needing to empirically adjust a regularization parameter.

Finally, similarly to what is introduced in this manuscript, Bastelier *et al.* have recently discussed factorizations strategies for (R, K) -KS matrices, but with the goal of preserving the data structure and thus resorting to orthogonality constraints [2].

1.3 Notation

Throughout this document, we denote \otimes the Kronecker Product and \circ the outer product. Matrices (resp. tensors) are represented by bold uppercase (resp. underscored) letters and the (i, j) th entry of a matrix \mathbf{D} is denoted $d(i, j)$. The vectorization operation, denoted $\text{vec}(\cdot)$, consists in stacking the columns of a matrix and $\text{unvec}(\cdot)$ stands for the converse operation.

2 A Dictionary Learning Algorithm for Tensorial Data

2.1 (R, K) -KS Dictionary Learning Model

Given a training data set in a tensor format $\underline{\mathbf{Y}}$, a naive approach to learn a dictionary from $\underline{\mathbf{Y}}$ is to rearrange its entries into a matrix and apply a workhorse two-way dictionary learning algorithm. However, such a matricization of $\underline{\mathbf{Y}}$ erases mode-wise information, such as the 2D structure of images or color information.

In [5], we have derived a dictionary learning algorithm for dictionaries as sums of 2-Kronecker products, which account for the two-way structure of the original training data. We now wish to extend this approach to tensors $\underline{\mathbf{Y}}$ of any order. For this, we constrain the learned dictionary to be (R, K) -Kronecker-structured.

The dictionary learning model may be computed by solving the following minimization problem:

$$\min_{\mathbf{D} \in \mathcal{C}_K^R, \mathbf{X}} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + g(\mathbf{X}) \quad (2)$$

where \mathcal{C}_K^R denotes the set of all (R, K) -KS matrices of size $(n \times m)$ and the training data, arranged as the columns of the matrix $\mathbf{Y} \in \mathbb{R}^{n \times N}$, is to be approximated as the product between the dictionary $\mathbf{D} \in \mathbb{R}^{n \times m}$ and the sparse representation matrix $\mathbf{X} \in \mathbb{R}^{m \times N}$. The sparsity of the columns of \mathbf{X} is enforced by the penalty function g (classical examples are the ℓ_0 and ℓ_1 norms). We also impose the columns of the dictionary to have unit ℓ_2 -norm.

Provided that the rank can be set to arbitrarily large values, any linear operator \mathbf{D} may actually be written as a (R, K) -KS matrix [17]. On the other hand, a K -KS matrix has a very specific structure that may not be appropriate for learning a good dictionary in a generic scenario. Therefore, for small rank values, the proposed approach may be understood as a trade-off between precision and complexity, storage and robustness to small training sets.

2.2 Dictionary Learning Algorithm

Following the literature, we employ a sub-optimal alternating minimization strategy to tackle problem (2). Since the problem is not jointly convex, there are no optimality guarantees for this classical approach. At each step, respectively *dictionary update* and *sparse coding*, we optimize with respect to one variable, resp. \mathbf{D} and \mathbf{X} , while fixing the other. The procedure is repeated N_{it} times.

The *sparse coding* step can be performed by any existing sparse regression algorithm. In our experiments we use the OMP [11] algorithm.

The difficulty in computing (2) lies in the (R, K) -KS structure imposed on \mathbf{D} . In fact, in Sects. 3 and 4, we show that this constraint is equivalent to imposing a Canonical Polyadic Decomposition model on a rearranged tensor $\mathcal{R}_K(\mathbf{D})$ obtained from dictionary \mathbf{D} . Therefore, for the *dictionary update* step, we propose a projected gradient algorithm, where the projection onto the set of \mathcal{C}_K^R (set of matrices written as a sum of R K -Kronecker products) is approximated by the CPD algorithm applied to the rearranged tensor $\mathcal{R}_K(\mathbf{D})$ as shown in Algorithm 1. The projection step is detailed in Sect. 4.

The step-size γ_t is determined with a backtracking line search. For acceleration purposes, the CPD can be initialized with the results of the previous iteration. Finally, note that the column normalization (line 5) can break the imposed structure. This is not a real concern, since the normalization coefficients can be stored separately, implying only m additional products in a matrix-vector operation. Put differently, it is equivalent to storing a dictionary in the form $\mathbf{D}\mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is a diagonal matrix containing the inverse norm of each column of \mathbf{D} .

Algorithm 1. $\mathbf{D} = \text{DictionaryUpdate}(\mathbf{Y}, \mathbf{X}, R, \mathbf{n}, \mathbf{m})$

- 1: Initialize $\mathbf{D}_0, t = 0$
 - 2: **while** $\|\mathbf{D}_{t+1} - \mathbf{D}_t\|_F > \text{tol}$ **do**
 - 3: $\mathbf{D}_{t+1} = \mathbf{D}_t - \gamma_t(\mathbf{D}_t\mathbf{X} - \mathbf{Y})\mathbf{X}^T$ ▷ Gradient step
 - 4: $\mathbf{D}_{t+1} = \text{HO-SuKroApprox}(\mathbf{D}_{t+1}, R, \mathbf{n}, \mathbf{m})$ ▷ Projection into \mathcal{C}_K^R
 - 5: Normalize columns of \mathbf{D}
-

3 Rearrangement: Transforming a K -Kronecker-Structured Matrix into a Low-Rank Tensor of Order K

This section introduces a rearrangement from a matrix to a multidimensional array that links the inference of the Kronecker structure for matrices to the low-rank approximation problem for multidimensional arrays *i.e.* higher-order tensors.

3.1 The Second Order Case

Consider a matrix $\mathbf{D} \in \mathbb{R}^{n_1 n_2 \times m_1 m_2}$ such that $\mathbf{D} = \mathbf{D}_1 \otimes \mathbf{D}_2$, where $\mathbf{D}_1 \in \mathbb{R}^{n_1 \times m_1}$ and $\mathbf{D}_2 \in \mathbb{R}^{n_2 \times m_2}$. Then one can define an operator $\mathcal{R}_2 : \mathbb{R}^{n_1 n_2 \times m_1 m_2} \rightarrow \mathbb{R}^{n_2 m_2 \times n_1 m_1}$ which rearranges the elements of \mathbf{D} in such way that the output $\mathbf{D}^\pi = \mathcal{R}_2(\mathbf{D})$ is a rank-1 matrix [18] given by

$$\mathbf{D}^\pi = \mathcal{R}_2(\mathbf{D}) = \text{vec}(\mathbf{D}_2) \text{vec}(\mathbf{D}_1)^T = \text{vec}(\mathbf{D}_2) \circ \text{vec}(\mathbf{D}_1) \tag{3}$$

It consists in vectorizing the j th-block in \mathbf{D} to form the j th-column of $\mathcal{R}(\mathbf{D})$, running through the blocks column-wise. This process is illustrated in Fig. 1.

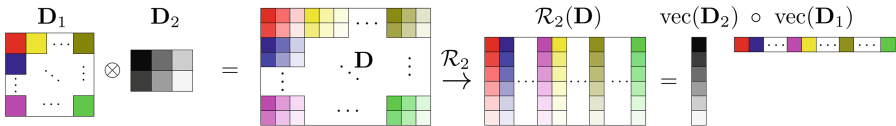


Fig. 1. Rearrangement operation

Since the isomorphism \mathcal{R}_2 is linear, when applied to a $(R, 2)$ -KS matrix $\mathbf{D} = \sum_{r=1}^R \mathbf{D}_1^r \otimes \mathbf{D}_2^r$, it outputs a rank- R matrix $\mathbf{D}^\pi = \sum_{r=1}^R \text{vec}(\mathbf{D}_2^r) \circ \text{vec}(\mathbf{D}_1^r)$.

3.2 Generalizing to Higher Order

Now, suppose a K -KS matrix $\mathbf{D} \in \mathbb{R}^{n_1 n_2 \dots n_K \times m_1 m_2 \dots m_K}$ given by

$$\mathbf{D} = \mathbf{D}_1 \otimes \dots \otimes \mathbf{D}_K = \bigotimes_{k=1}^K \mathbf{D}_k \tag{4}$$

with $\mathbf{D}_k \in \mathbb{R}^{n_k \times m_k}$ for $k \in \{1, \dots, K\}$.

We can generalize the rearrangement operator \mathcal{R}_2 (for 2-KS matrices) to an operator $\mathcal{R}_K : \mathbb{R}^{n_1 n_2 \dots n_K \times m_1 m_2 \dots m_K} \rightarrow \mathbb{R}^{n_K m_K \times \dots \times n_1 m_1}$ (for K -KS matrices) in a recursive manner. For this, let's suppose \mathcal{R}_{K-1} known, such that

$$\mathbf{D} = \bigotimes_{k=2}^K \mathbf{D}_k \quad \Rightarrow \quad \underline{\mathbf{D}}^\pi = \mathcal{R}_{K-1}(\mathbf{D}) = \text{vec}(\mathbf{D}_K) \circ \dots \circ \text{vec}(\mathbf{D}_2), \quad (5)$$

which outputs a $(K-1)$ th-order rank-1 tensor for an input $(K-1)$ -KS matrix, and try to define \mathcal{R}_K from it.

Note that we can rewrite $\mathbf{D} = \mathbf{D}_1 \otimes (\mathbf{D}_2 \otimes \dots \otimes \mathbf{D}_K)$ which means that \mathbf{D} is composed of n_1 by m_1 blocks given by $d_1(i, j)$ ($\mathbf{D}_2 \otimes \dots \otimes \mathbf{D}_K$) in which we can directly apply \mathcal{R}_{K-1} . If we again run through all these blocks columnwise applying \mathcal{R}_{K-1} and stacking the resulting $(K-1)$ th-order tensors along dimension K , we will obtain a (K) -order rank-1 tensor given by:

$$\underline{\mathbf{D}}^\pi = \mathcal{R}_K(\mathbf{D}) = (\text{vec}(\mathbf{D}_K) \circ \dots \circ \text{vec}(\mathbf{D}_2)) \circ \text{vec}(\mathbf{D}_1)$$

Therefore, we can define a recursive algorithm to calculate \mathcal{R}_K which has \mathcal{R}_2 defined in Sect. 3.1 as a base case. Actually, we can go even further and decompose \mathcal{R}_2 recursively in the exact same way, in which case the base case \mathcal{R}_1 becomes a simple vectorization operation. The described procedure is illustrated in Fig. 2 for the particular case of $K=3$.

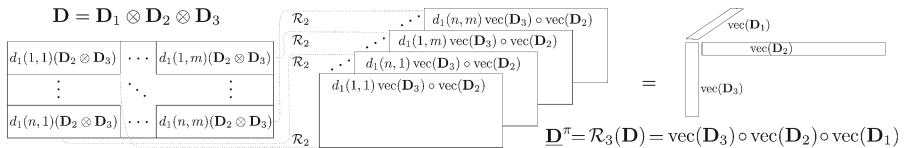


Fig. 2. Rearrangement operation \mathcal{R}_K for $K=3$.

Just like in Sect. 3.1, note that for an input (R, K) -KS matrix, the rearrangement outputs a sum of R rank-1 tensors with factors $\text{vec}(\mathbf{D}_i)$ sorted in the reverse lexicographic order: $\underline{\mathbf{D}}^\pi = \sum_{r=1}^R \text{vec}(\mathbf{D}_K^r) \circ \dots \circ \text{vec}(\mathbf{D}_1^r)$.

3.3 Inverse Rearrangement

The inverse rearrangement \mathcal{R}_K^{-1} consists simply in switching the input and output indexes of the previous operator so to yield $\mathbf{D} = \mathcal{R}_K^{-1}(\underline{\mathbf{D}}^\pi)$. In practical terms, the resulting recursive algorithm consists in progressively reconstructing the blocks of \mathbf{D} from their vectorized versions in $\underline{\mathbf{D}}^\pi$. The base case is now a matricization (unvec) operation.

4 From SVD to CPD

Let $\mathbf{D} \in \mathbb{R}^{n \times m}$ and consider the following constrained approximation problem:

$$\min_{\hat{\mathbf{D}} \in \mathcal{C}_K^R} \|\hat{\mathbf{D}} - \mathbf{D}\|_F^2 \tag{6}$$

With the help of the rearrangement operators defined in Sect. 3, the task of approximating any given matrix \mathbf{D} comes down to low-rank approximation of the K -th-order rearranged tensor $\underline{\mathbf{D}}^\pi = \mathcal{R}_K(\mathbf{D})$.

When the targeted structure is a (sum of) 2-Kronecker product(s), i.e. $K=2$, we are interested in the low-rank approximation of $\mathbf{D}^\pi = \mathcal{R}_2(\mathbf{D})$ which is still a matrix (2nd-order tensor). This is easily achieved – and also optimally, from Eckart-Young theorem – via the SVD.

The task gets harder if we want to generalize to a (sum of) K -Kronecker product(s), since a low-rank approximation of an order- K tensor $\underline{\mathbf{D}}^\pi = \mathcal{R}_K(\mathbf{D})$ is required. In this work, we propose to use the a Canonical Polyadic Decomposition (CPD) [10] to approximate the tensor $\underline{\mathbf{D}}^\pi$ with a sum of R rank-one tensors.

$$\{\hat{\mathbf{d}}_k^r\}_{k=\{1,\dots,K\}}^{r=\{1,\dots,R\}} = \text{CPD}(\underline{\mathbf{D}}^\pi, R) \quad \text{such that} \quad \hat{\mathbf{D}}^\pi = \sum_{r=1}^R \hat{\mathbf{d}}_K^r \circ \dots \circ \hat{\mathbf{d}}_1^r \tag{7}$$

We can see that each composing $\hat{\mathbf{D}}_k^r$ can be obtained from the corresponding vector $\hat{\mathbf{d}}_k^r$ through a simple matricization operation: $\hat{\mathbf{D}}_k^r = \text{unvec}(\hat{\mathbf{d}}_k^r)$. The resulting approximation is thus a (R, K) -KS matrix with factors $\hat{\mathbf{D}}_k^r$.

The proposed procedure to obtain $\hat{\mathbf{D}}$ and its factors $\hat{\mathbf{D}}_k^r$ is summarized in Algorithm 2, which we call HO-SuKro (**H**igher **O**rders **S**um of **K**roneckers) Approximation algorithm. It is parametrized by the targeted rank R and two vectors $\mathbf{n} = [n_1, \dots, n_K]$ and $\mathbf{m} = [m_1, \dots, m_K]$ with $(n_k \times m_k)$ being the dimensions of the k -th factor $\hat{\mathbf{D}}_k^r (\forall r \in \{1, \dots, R\})$, such that $n = \prod_{k=1}^K n_k$ and $m = \prod_{k=1}^K m_k$.

Algorithm 2. $\left[\hat{\mathbf{D}}, \{\hat{\mathbf{D}}_k^r\}_{k=\{1,\dots,K\}}^{r=\{1,\dots,R\}} \right] = \text{HO-SuKroApprox}(\mathbf{D}, R, \mathbf{n}, \mathbf{m})$

1: $\mathcal{R}_K(\mathbf{D}) = \text{Rearrange}(\mathbf{D}, \mathbf{n}, \mathbf{m})$	▷ Rearranging input matrix
2: $\{\hat{\mathbf{d}}_k^r\}_{k=\{1,\dots,K\}}^{r=\{1,\dots,R\}} = \text{CPD}(\mathcal{R}_K(\mathbf{D}), R)$	▷ CPD on rearranged tensor
3: $\{\hat{\mathbf{D}}_k^r\}_{k=\{1,\dots,K\}}^{r=\{1,\dots,R\}} = \text{unvec}(\hat{\mathbf{d}}_k^r)$	▷ Recovering factors $\hat{\mathbf{D}}_k^r$
4: return $\left[\hat{\mathbf{D}} = \sum_{r=1}^R \hat{\mathbf{D}}_1^r \otimes \dots \otimes \hat{\mathbf{D}}_K^r, \{\hat{\mathbf{D}}_k^r\}_{k=\{1,\dots,K\}}^{r=\{1,\dots,R\}} \right]$	

5 Experiments

To evaluate the proposed dictionary learning algorithm, we have performed some color image denoising experiments following the set-up introduced in [6]. The test images, $512 \times 512 \times 3$ color images, are corrupted by a white Gaussian noise with different standard deviations σ . In these experiments we thus have $K = 3$.

The dictionary is trained from a set of vectorized $(6 \times 6 \times 3)$ -pixel patches extracted uniformly from the noisy image itself and then used to reconstruct all overlapping patches with a one-pixel step. The denoised image is obtained by averaging the pixel values of the overlapping patches. The simulation parameters are as follows: sample dimension (n) is 108 (with $n_1 \times n_2 \times n_3 = 6 \times 6 \times 3$), number of atoms (m) is 864 with $m_1 \times m_2 \times m_3 = 12 \times 12 \times 6$, number of training samples (N) is 4×10^4 , convergence tolerance (tol) is $\|\mathbf{D}\|_F \times 10^{-4}$ and number of iterations (N_{it}) is 20.

The sparse coding step is performed by the OMP algorithm. The PSNR of the reconstructed images are evaluated as follows: $PSNR = 255^2 N_{px} / \left(\sum_{i=1}^{N_{px}} (y_i - \hat{y}_i)^2 \right)$ where 255 is the maximum pixel value, N_{px} is the total number of pixels on the input image, y_i and \hat{y}_i are respectively the i -th pixel value on the input and reconstructed image. The results are averaged over five noise realizations.

Figure 3 shows the denoised image PSNR (in dB) as a function of the number of Kronecker summing terms in the dictionary (i.e. the rank R of the rearranged tensor). We compare our results to an unstructured dictionary with the same size learned by the K-SVD [1] algorithm and to the 3D-ODCT analytic dictionary, which is actually a 3-Kronecker dictionary as well (although not trained from the data). 3D-ODCT is also used for initializing the proposed HO-SuKro¹.

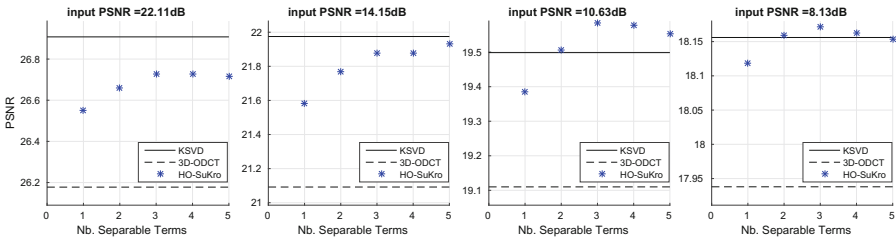


Fig. 3. PSNR vs. number of separable terms (R) for the mandrill image.

Compared to the fixed 3D-ODCT dictionary, our algorithm achieves considerably better denoising results, even for one single separable term ($R = 1$) which

¹ The 1-D $n \times m$ overcomplete DCT dictionary, as defined in [14], is a cropped version of the orthogonal $m \times m$ DCT matrix. The K -dimensional ODCT is the Kronecker product of K 1-D ODCT.

is the exact same structure as the 3D-ODCT. It remains slightly inferior to a K-SVD dictionary in most cases, but with the advantage of the reduced application complexity due to the Kronecker structure (see Table 1). The chosen structure proved well suited to this kind of application, since its introduction compromised very little the performance. Actually, it even enhanced the denoising capabilities in higher noise scenarios, which we attribute to an overfitting prevention due to the structure constraint. The addition of separable terms tends to improve the performance until a certain point after which it saturates – or even deteriorates at higher noise, indicating the onset of overfitting.

Table 1 shows the theoretical complexity savings provided by the Kronecker structure for matrix vector operations as well as its storage cost compared to an unstructured dictionary. The gains are around one and two orders of magnitude in this case. Obviously, for the complexity gains to be observed in practice, the matrix-vector multiplications should be performed according to Eq. (1).

Table 1. Complexity costs (for matrix-vector multiplications) and storage costs^a.

	HO-SuKro	Unstructured	Ratio (Unstructured/HO-SuKro)
Complexity (# operations)	$12960 \times R$	186624	$14.4/R$
Storage (# parameters)	$162 \times R$	93312	$576/R$

^aComplexity cost for HO-SuKro is trivially obtained considering Eq. (1). Storage cost is the total number of elements in all factors: $R(\sum_k n_k m_k)$.

In Fig. 4 we evaluate the robustness of the learning algorithm to a reduction on the training set size. It shows the Δ PSNR, defined as the difference with respect to the PSNR obtained by ODCT.

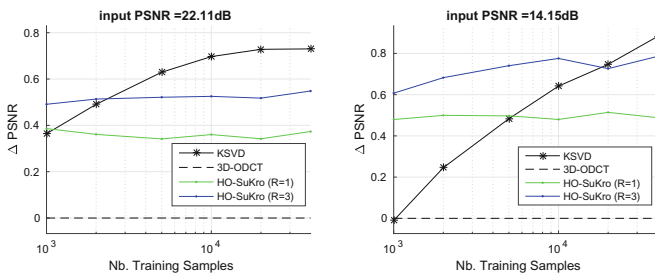


Fig. 4. PSNR vs. Number of training samples for the mandrill image.

Note that the Kronecker-structured dictionaries become more competitive as the size of the training set decreases, to the point of consistently outperforming K-SVD for small enough training sets. This result goes in line with the theoretical results in [15] suggesting a smaller sample complexity for KS dictionaries.

6 Conclusion

To improve on storage, robustness to sample size and computational complexity, a new dictionary learning model was introduced where the dictionary is constrained as a sum of R Kronecker products of K terms. Such a constraint arises naturally when the original data are contained in a K th-order tensor, such as a collection of color images. We have drawn a parallel between this sum of Kroneckers constraint and the tensor Canonical Polyadic Decomposition, the latter being used as a projection algorithm for imposing the structure constraint. Encouraging results are shown on color image denoising, and future works will focus on properly exploiting the Kronecker structure to accelerate the training phase of the dictionary, which is currently quite time consuming (about one order of magnitude slower than the K-SVD, as a comparison). Nevertheless, it remains interesting for cases where the dictionary is to be repeatedly applied afterwards.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing over-complete dictionaries for sparse representation. *IEEE Trans. Sig. Process.* **54**(11), 4311–4322 (2006)
2. Batselier, K., Wong, N.: A constructive arbitrary-degree Kronecker product decomposition of tensors. *Numer. Linear Algebra Appl.* **24**(5), e2097 (2017). <https://doi.org/10.1002/nla.2097>
3. Bijma, F., De Munck, J., Heethaar, R.M.: The spatiotemporal MEG covariance matrix modeled as a sum of Kronecker products. *NeuroImage* **27**, 402–15 (2005)
4. Caiafa, C.F., Cichocki, A.: Multidimensional compressed sensing and their applications. *Wiley Interdisc. Rev. Data Min. Knowl. Discov.* **3**(6), 355–380 (2013)
5. Dantas, C., da Costa, M.N., Lopes, R.: Learning dictionaries as a sum of Kronecker products. *IEEE Sig. Process. Lett.* **24**, 559–563 (2017)
6. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* **15**(12), 3736–3745 (2006)
7. Ghassemi, M., Shakeri, Z., Sarwate, A.D., Bajwa, W.U.: STARK: structured dictionary learning through rank-one tensor recovery. In: 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (2017)
8. Hawe, S., Seibert, M., Kleinstueber, M.: Separable dictionary learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 438–445 (2013)
9. Jung, A., Eldar, Y.C., Görtz, N.: On the minimax risk of dictionary learning. *IEEE Trans. Inf. Theory* **62**(3), 1501–1515 (2016)
10. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. *SIAM Rev.* **51**(3), 455–500 (2009)
11. Pati, Y.C., Rezaifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: 1993 Conference Record of the 27th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 40–44 (1993)

12. Peng, Y., Meng, D., Xu, Z., Gao, C., Yang, Y., Zhang, B.: Decomposable non-local tensor dictionary learning for multispectral image denoising. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2949–2956 (2014)
13. Roemer, F., Del Galdo, G., Haardt, M.: Tensor-based algorithms for learning multidimensional separable dictionaries. In: 2014 IEEE Conference on International Acoustics, Speech and Signal Processing (ICASSP), pp. 3963–3967. IEEE (2014)
14. Rubinstein, R., Zibulevsky, M., Elad, M.: Double sparsity: learning sparse dictionaries for sparse signal approximation. *IEEE Trans. Sig. Process.* **58**(3), 1553–1564 (2010)
15. Shakeri, Z., Bajwa, W.U., Sarwate, A.D.: Sample complexity bounds for dictionary learning of tensor data. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4501–4505 (2017)
16. Tadonki, C., Philippe, B.: Parallel multiplication of a vector by a Kronecker product of matrices. In: *Parallel Numerical Linear Algebra*, pp. 71–89. Nova Science Publishers Inc., Commack (2001)
17. Tsiligkaridis, T., Hero, A.O.: Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Trans. Sig. Process.* **61**(21), 5347–5360 (2013)
18. Van Loan, C.F., Pitsianis, N.: Approximation with Kronecker products. In: Moonen, M.S., Golub, G.H., De Moor, B.L.R. (eds.) *Linear Algebra for Large Scale and Real-Time Applications*. NATO ASI Series, vol. 232, pp. 293–314. Springer, Dordrecht (1993). https://doi.org/10.1007/978-94-015-8196-7_17
19. Zubair, S., Wang, W.: Tensor dictionary learning with sparse tucker decomposition. In: 2013 18th International Conference on Digital Signal Processing (DSP), pp. 1–6 (2013)



Truncated Variational Sampling for ‘Black Box’ Optimization of Generative Models

Jörg Lücke^{1(✉)}, Zhenwen Dai², and Georgios Exarchakis³

¹ Universität Oldenburg, Oldenburg, Germany

joerg.luecke@uni-oldenburg.de

² Amazon Research, Cambridge, UK

³ École Normale Supérieure, Paris, France

Abstract. We investigate the optimization of two probabilistic generative models with binary latent variables using a novel variational EM approach. The approach distinguishes itself from previous variational approaches by using latent states as variational parameters. Here we use efficient and general purpose sampling procedures to vary the latent states, and investigate the “black box” applicability of the resulting optimization approach. For general purpose applicability, samples are drawn from approximate marginal distributions as well as from the prior distribution of the considered generative model. As such, sampling is defined in a generic form with no analytical derivations required. As a proof of concept, we then apply the novel procedure (A) to Binary Sparse Coding (a model with continuous observables), and (B) to basic Sigmoid Belief Networks (which are models with binary observables). Numerical experiments verify that the investigated approach efficiently as well as effectively increases a variational free energy objective without requiring any additional analytical steps.

Keywords: Maximum likelihood · Variational EM
Generative models

1 Introduction

The use of expectation maximization (EM) for advanced probabilistic data models requires approximations because exact E-steps (computing full posteriors) are typically intractable. Many models of recent interest have binary latents [1–4], and for such models these intractabilities are primarily computational: exact E-steps can be computed but they scale exponentially with the number of latents. To overcome intractabilities for models with binary latents there are typically two types of approaches applied: sampling approaches or variational EM with the latter having been dominated by factored variational approaches in the past [e.g. 5]. Variational approaches and sampling have also often been combined [4, 6–8] to leverage

Z. Dai—Work has started prior to joining Amazon.

the advantages of both methods. However, given a generative model, both approximations require often cumbersome derivations either to obtain efficient posterior samplers or to obtain update equations for variational parameter optimization. The question how procedures can be defined that automatize the development of learning algorithms for generative models has therefore shifted into the focus of recent research [4, 9–12]. In this paper, we make use of truncated approximations to EM which have repeatedly been applied before [4, 13, 14]. Here we show how novel theoretical results on truncated variational distributions [15] can be used to couple variational EM and sampling exceptionally tightly. This coupling then enables “black box” applicability.

2 Truncated Posteriors and Sampling

Let us consider generative models with H binary latent variables, $\mathbf{s} = (s_1, \dots, s_H)$ with $s_h \in \{0, 1\}$. Let $p(\mathbf{s}, \mathbf{y} | \Theta)$ be a model’s joint probability with Θ denoting its parameters. Truncated approximations have been motivated by the observation that the exponentially large sums over states required for expectation values w.r.t. posteriors are typically dominated by summands corresponding to very few states. If for a given data point $\mathbf{y}^{(n)}$ these few states are contained in a set $\mathcal{K}^{(n)}$, we can define a posterior approximation as follows [e.g., as in 4, 13]:

$$q^{(n)}(\mathbf{s}; \mathcal{K}, \Theta) = \frac{p(\mathbf{s} | \mathbf{y}^{(n)}, \Theta) \delta(\mathbf{s} \in \mathcal{K}^{(n)})}{\sum_{\mathbf{s}' \in \mathcal{K}^{(n)}} p(\mathbf{s}' | \mathbf{y}^{(n)}, \Theta)}, \quad (1)$$

where $\delta(\mathbf{s} \in \mathcal{K}^{(n)}) = 1$ if $\mathcal{K}^{(n)}$ contains \mathbf{s} and zero otherwise. It is straightforward to derive expectation values w.r.t. these approximate posteriors simply by inserting (1) into the definition of expectation values and by multiplying numerator and denominator by $p(\mathbf{y}^{(n)} | \Theta)$, which yields:

$$\langle g(\mathbf{s}) \rangle_{q^{(n)}} = \frac{\sum_{\mathbf{s} \in \mathcal{K}^{(n)}} p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta) g(\mathbf{s})}{\sum_{\mathbf{s}' \in \mathcal{K}^{(n)}} p(\mathbf{s}', \mathbf{y}^{(n)} | \Theta)} \quad (2)$$

where $g(\mathbf{s})$ is a function of the hidden variables. As the dominating summands are different for each data point $\mathbf{y}^{(n)}$, the sets $\mathcal{K}^{(n)}$ are different. If a set $\mathcal{K}^{(n)}$ now contains those states \mathbf{s} which dominate the sums over the joints w.r.t. the exact posterior, then Eq. 2 is a very accurate approximation.

Truncated posterior approximations have successfully been applied to a number of elementary and more advanced generative models, and they do not suffer from potential biases introduced by posterior independence assumptions. Previously, the sets $\mathcal{K}^{(n)}$ were defined based on sparsity assumptions and/or latent preselection [13, 16]. The approach followed here, in contrast, uses sets $\mathcal{K}^{(n)}$ which

contain samples from model and data dependent distributions. By treating the truncated distribution (1) as variational distributions within a free energy framework [15], we can then derive efficient procedures to update the samples in $\mathcal{K}^{(n)}$ such that the variational free energy is always monotonically increased. For this we use the following theoretical results: (A) We use that the M-step equations remain unchanged if instead of full posteriors the truncated posteriors (1) are used; (B) We make use of the result that after each M-step the free energy corresponding to truncated variational distributions is given by the following simplified and computationally tractable form:

$$\mathcal{F}(\mathcal{K}, \Theta) = \sum_n \log \left(\sum_{\mathbf{s} \in \mathcal{K}^{(n)}} p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta) \right), \quad (3)$$

where $\mathcal{K} = (\mathcal{K}^{(1)}, \dots, \mathcal{K}^{(N)})$. The variational E-step then consists of finding a set \mathcal{K}_{new} which increases $\mathcal{F}(\mathcal{K}, \Theta)$ w.r.t. \mathcal{K} . For any larger scale multiple-cause model we can not exhaustively iterate through all latent states. We therefore here seek to find new sets $\tilde{\mathcal{K}}$ using sampling such that the free energy is increased: $\mathcal{F}(\tilde{\mathcal{K}}, \Theta) > \mathcal{F}(\mathcal{K}, \Theta)$. To keep the computational demand limited, we will take the sets \mathcal{K} and $\tilde{\mathcal{K}}$ to be of constant size after each E-step by demanding $|\mathcal{K}^{(n)}| = |\tilde{\mathcal{K}}^{(n)}| = S$ for all n (where S can be relatively small for most data). Instead of explicitly computing and comparing the free-energies (3) for \mathcal{K} and $\tilde{\mathcal{K}}$, we can use a comparison of joint probabilities $p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta)$ as a criterion for free energy increase. The following can be shown [15]:

For a replacement of $\mathbf{s} \in \mathcal{K}^{(n)}$ by a new state $\mathbf{s}_{\text{new}} \notin \mathcal{K}^{(n)}$ the free energy $\mathcal{F}(\mathcal{K}, \Theta)$ is increased if and only if

$$p(\mathbf{s}_{\text{new}}, \mathbf{y}^{(n)} | \Theta) > p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta), \quad (4)$$

i.e., the free energy is guaranteed to increase if we replace, e.g., the state with the lowest joint in $\mathcal{K}^{(n)}$ by a newly sampled state $\mathbf{s}_{\text{new}} \notin \mathcal{K}^{(n)}$ with a higher joint. Instead of comparing single joints, a computationally more efficient procedure is to

Algorithm 1. Sampling-based TV-E-step

for $n = 1, \dots, N$ **do**

draw M samples $\mathbf{s} \sim p_{\text{var}}^{(n)}(\mathbf{s})$;
 define $\mathcal{K}_{\text{new}}^{(n)}$ to contain all M samples;
 set $\mathcal{K}^{(n)} = \mathcal{K}^{(n)} \cup \mathcal{K}_{\text{new}}^{(n)}$;
 remove those $(|\mathcal{K}^{(n)}| - S)$ samples $\mathbf{s} \in \mathcal{K}^{(n)}$
 with the lowest $p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta)$;

use batches of many newly sampled states, and then to use criterion (4) to increase $\mathcal{F}(\mathcal{K}, \Theta)$ as much as possible. Such a procedure is given by Algorithm 1: For each datum n , we first draw M new samples from a yet to be specified distribution $p_{\text{var}}^{(n)}(\mathbf{s})$. These samples are then united with the states already in $\mathcal{K}^{(n)}$.

Of this union, we then take the S states with highest joints to define the new state set $\mathcal{K}^{(n)}$. The last step selects because of (4) the best possible subset of the union. Furthermore, the selection of the S states with largest joints can be solved by selection algorithms of linear time complexity, i.e., in $\mathcal{O}(M + S)$ time in our case. Instead of selecting the S largest joints, we can also remove the $(|\mathcal{K}^{(n)}| - S)$ lowest ones (last line in Algorithm 1). For *any* distribution $p_{\text{var}}^{(n)}(\mathbf{s})$, Algorithm 1 is guaranteed to monotonically increase the free energy $\mathcal{F}(\mathcal{K}, \Theta)$ w.r.t. \mathcal{K} .

3 Posterior, Prior, and Marginal Sampling

While the partial E-step of Algorithm 1 monotonically increases the free energy for any distribution $p_{\text{var}}^{(n)}(\mathbf{s})$ used for sampling, the specific choice for $p_{\text{var}}^{(n)}(\mathbf{s})$ is of central importance for the efficiency of the procedure. If the distribution is not chosen well, any significant increase of $\mathcal{F}(\mathcal{K}, \Theta)$ may require unreasonable amounts of time, e.g., because new samples which increase $\mathcal{F}(\mathcal{K}, \Theta)$ are sampled too infrequently. By considering Algorithm 1, the requirement for $p_{\text{var}}^{(n)}(\Theta)$ is to provide samples with high joint probability $p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta)$ for a given $\mathbf{y}^{(n)}$. The first distribution that comes to mind for $p_{\text{var}}^{(n)}(\Theta)$ is the posterior distribution $p(\mathbf{s} | \mathbf{y}^{(n)}, \Theta)$. Samples from the posterior are likely to have high posterior mass and therefore high joint mass relative to the other states because all states share the same normalizer $p(\mathbf{y}^{(n)} | \Theta)$. On the downside, however, sampling from the posterior may not be an easy task for models with binary latents and a relatively high dimensionality as we intend to aim at here. Furthermore, the derivation of posterior samplers requires additional analytical efforts for any new generative model we apply the procedure to, and requires potentially additional design choices such as definitions of proposal distributions. All these points are contrary to our goal of a ‘black box’ procedure which is applicable as generally and generically as possible. Instead, we therefore seek distributions $p_{\text{var}}^{(n)}(\Theta)$ for Algorithm 1 that can efficiently optimize the free energy but that can be defined without requiring model-specific analytical derivations. Candidates for $p_{\text{var}}^{(n)}(\Theta)$ are consequently the prior distribution of the given generative model, $p(\mathbf{s} | \Theta)$, or the marginal distribution. A prior sampler is usually directly given by the generative model but may have the disadvantage that finally new samples only very rarely increase the free energy because the prior sampler is independent of a given data point (only the average over data points has high posterior mass). Marginal samplers, on the other hand, are data driven but the computation of activation probabilities $p(s_h = 1 | \mathbf{y}^{(n)}, \Theta)$ is unfortunately not computationally efficient. To obtain data-driven but efficient samplers, we will for our purposes, therefore, use approximate marginal samplers.

1st Approximation. First observe that we can obtain an efficiently computable approximation to a marginal sampler by using the truncated distributions $q^{(n)}(\mathbf{s})$ in (1) themselves. For binary latents s_h we can approximate:

$$p(s_h = 1 | \mathbf{y}^{(n)}, \Theta) = \langle s_h \rangle_{p(\mathbf{s} | \mathbf{y}^{(n)}, \Theta)} \approx \langle s_h \rangle_{q^{(n)}(\mathbf{s})}, \text{ and} \quad (5)$$

accordingly $p(s_h = 0 | \mathbf{y}^{(n)}, \Theta)$. Because of the arguments given above the expectations (2) w.r.t. $q^{(n)}(\mathbf{s})$ are efficiently computable using (2) with $g(\mathbf{s}) = \mathbf{s}$. Using (5) we can consequently define for each latent h an approximation of the marginal $p(s_h = 1 | \mathbf{y}^{(n)}, \Theta)$. Given a directed generative model, no derivations are required to efficiently generate samples from this approximation because the joint probabilities to estimate $p(s_h = 1 | \mathbf{y}^{(n)}, \Theta)$ using (2) can directly be computed. The truncated marginal sampler defined by Eq. 5 becomes increasingly similar to an exact marginal sampler the better the truncated distributions approximate the exact posteriors.

2nd Approximation. To further improve efficiency and convergence times, we optionally apply a second approximation by using the approximate marginal distributions (Eq. 5) as target objective for a parametric function $f_h(\mathbf{y}^{(n)}; \Lambda)$ which approximates the truncated marginal. A parametric function from data to marginal probabilities of the latents has the advantage of modeling data similarities by mapping similar data to similar marginal distributions. The mapping incorporates information across the data points, which can facilitate training and, e.g., avoids more expensive $\mathcal{K}^{(n)}$ updates of some data points due to bad initialization. The mapping $f_h(\mathbf{y}^{(n)}; \Lambda)$ is estimated with the training data and the current approximate marginal $q_{\text{mar}}(s_h = 1 | \mathbf{y}, \Theta)$ defined by (5) with (2). For simplicity, we use a Multi-Layer Perceptron (MLP) for the function mapping and trained with cross-entropy. We use a generic MLP with one hidden layer. As such, the MLP itself is independent of the generative model considered but optimized for the generic truncated approximation (5) which contains the model’s joint. The idea of using a parametric function to approximate expectations w.r.t. intractable posteriors is an often applied technique [e.g., 17, 18, and refs therein].

For our experiments we combine prior and (approx.) marginal sampling to suggest new variational states. The easy to use prior samplers are not data driven and rather represent *exploration*. Marginal sampling, on the other hand, is rather an *exploitation* strategy that produces good results when sufficiently much from the data is already known. Mixing the two has turned out best for our purposes. Posterior samplers do require additional derivations but,

Algorithm 2. TVS.

```

initialize model parameters  $\Theta$ ;
for all  $n$  init  $\mathcal{K}^{(n)}$  such that  $|\mathcal{K}^{(n)}| = S$ ;
set  $M_p$ ; (# samples from prior distribution)
set  $M_q$ ; (# samples from marginal distr.)
repeat
  update  $M_p$  and  $M_q$  (sampler adjustment)
  for  $(n = 1, \dots, N)$  do
    draw  $M_p$  samples from  $p(\mathbf{s} | \Theta) \rightarrow \mathcal{K}_p^{(n)}$ ;
    draw  $M_q$  samples from  $q_{\text{mar}}^{(n)}(\mathbf{s}; \Theta) \rightarrow \mathcal{K}_q^{(n)}$ ;
     $\mathcal{K}^{(n)} = \mathcal{K}^{(n)} \cup \mathcal{K}_p^{(n)} \cup \mathcal{K}_q^{(n)}$ ;
    remove those  $(|\mathcal{K}^{(n)}| - S)$  elements
     $\mathbf{s} \in \mathcal{K}^{(n)}$  with lowest  $p(\mathbf{s}, \mathbf{y}^{(n)} | \Theta)$ ;
   $\mathcal{K} = (\mathcal{K}^{(1)}, \dots, \mathcal{K}^{(N)})$ ;
  use M-steps with (2) to change  $\Theta$ ;
until  $\Theta$  has sufficiently converged;

```

to our experience so far, are also not necessarily better than combined prior/marginal sampling in optimizing the free energy.

Before we consider concrete generative models, let us summarize the general novel procedure in the form of the pseudo code given by Algorithm 2. First, we have to initialize the model parameters Θ and the sets $\mathcal{K}^{(n)}$. While initializing Θ can be done as for other EM approaches, one option for an initialization of $\mathcal{K}^{(n)}$ would be the use of samples from the prior given Θ (more details are given below). The inner loop (the variational E-step) of Algorithm 2 is then based on a mix of prior and marginal samplers, and each of these samplers is directly defined in terms of a considered generative model, no model-specific derivations are used. The same does not apply for the M-step but we will consider two examples how this point can be addressed: (A) either by using well-known standard M-step or (B) by applying automatic differentiation. Algorithm 2 will be referred to as *truncated variational sampling* (TVS).

4 Applications of TVS

Exemplarily, we consider two generative models which are complementary in many aspects.

Binary Sparse Coding. In the first example we will consider dictionary learning – a typical application domain of variational EM approaches and sampling approaches in general. Probabilistic sparse coding models are not computationally tractable and common approximations such as maximum a-posteriori approximations can result in suboptimal solutions. Factored variational EM as well as sampling approaches have therefore been routinely applied to sparse coding. Of particular interest for our purposes are sparse coding models with discrete or semi-discrete latents [e.g. 1, 4, 19, 20], where binary sparse coding [BSC; 19, 20] represents an elementary example.

BSC assumes independent and identically distributed (iid) binary latent variables following a Bernoulli prior distribution, and it uses a Gaussian noise model:

$$p(\mathbf{s} | \Theta) = \prod_{h=1}^H \pi^{s_h} (1 - \pi)^{1-s_h}, \quad p(\mathbf{y} | \mathbf{s}, \Theta) = \mathcal{N}(\mathbf{y}; W\mathbf{s}, \sigma^2 \mathbf{1}), \quad (6)$$

where $\pi \in [0, 1]$ and where $\Theta = (\pi, W, \sigma^2)$ is the set of model parameters.

As TVS is an approximate EM approach, let us first consider exact EM which seeks parameters Θ that optimize the data likelihood for the BSC data model (6). Parameter update equations are canonically derived and given by [e.g. 20]:

$$\begin{aligned} \pi &= \frac{1}{N} \sum_{n=1}^N \sum_{h=1}^H \langle s_h \rangle_{q_n}, \quad W = \left(\sum_{n=1}^N \mathbf{y}^{(n)} \langle \mathbf{s} \rangle_{q_n}^T \right) \left(\sum_{n=1}^N \langle \mathbf{s} \mathbf{s}^T \rangle_{q_n} \right)^{-1} \quad (7) \\ \sigma^2 &= \frac{1}{ND} \sum_{n=1}^N \langle \|\mathbf{y}^{(n)} - W\mathbf{s}\|^2 \rangle_{q_n} \quad (8) \end{aligned}$$

where the q_n are equal to the exact posteriors for exact EM, $q_n = p(\mathbf{s} | \mathbf{y}^{(n)}, \Theta)$.

A standard variational EM approach for BSC would now replace these posteriors by variational distributions q_n . Applications of (mean-field) variational

distributions as, e.g., applied by [19], entails (A) a choice which family of distributions to use; and (B) additional derivations in order to derive update equations for the introduced variational parameters. Also the application of sampling based approaches would require derivations. The same is not the case for the application of TVS (Algorithm 2). In order to obtain a TVS learning algorithm for BSC, we do (for the update equations) just have to replace the expectation values in Eqs. 7 to 8 by (2). For the E-step, we then use the generative model description (Eq. 6) in order to update the sets $\mathcal{K}^{(n)}$ using prior and approximate marginal distributions as described by Algorithm 2.

Artificial Data. Firstly, we verify and study the novel approach using artificial data generated by the BSC data model using ground-truth generating parameters Θ_{gt} . We use $H = 10$ latent variables, s_h , sampled independently by a Bernoulli distribution parameterized by $\pi_{gt} = 0.2$. We set the ground truth parameters for the dictionary matrix, $W \in \mathbb{R}^{D \times H}$, to appear like vertical and horizontal bars [compare 20] when rasterized to 5×5 images, see Fig. 1, with a value of 10 for a pixel that belongs to the bar and 0 for a pixel that belongs to the background. We linearly combine the latent variables with the dictionary elements to generate a $D = 25$ -dimensional datapoint, \mathbf{y} , to which we add mean-free Gaussian noise with standard deviation $\sigma_{gt} = 2.0$. In this way we generate $N = 10\,000$ datapoints that form our artificial dataset. We now use TVS for BSC to fit another instance of the BSC model to the generated data. The model is initialized with a noise parameter σ equal to the average standard deviation of

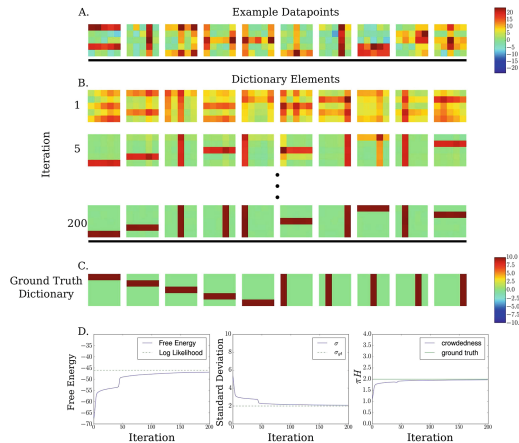


Fig. 1. Linear Bars Test. **A.** A subset of the generated datapoints. **B.** The evolution of the dictionary over TVS iterations. Note that permutations of the dictionary elements would yield the same likelihood. **C.** The ground truth dictionary. **D.** The evolution of free energy over TVS iterations plotted next to the exact log-likelihood (left). Evolution of the model standard deviation plotted next to the ground truth (middle). Evolution of the expected number of active units πH plotted against the ground truth (right).

each observation in the data $\mathbf{y}^{(n)}$, the prior parameter is initialized as $\pi = 1/H$ where the latent variable $H = 10$ is maintained from the generating model. We initialize the columns of the dictionary matrix with the mean datapoint plus mean Gaussian samples with a standard deviation $\sigma/4$.

We train the model using the TVS algorithm for 200 TV-EM iterations maintaining the number of variational states at $|\mathcal{K}^{(n)}| = S = 64$ for all datapoints throughout the duration of the training. We use $M_q = 32$ samples drawn from the marginal distribution (only 1st approximation) and $M_p = 32$ samples drawn from the prior to vary \mathcal{K} according to Algorithm 2. The evolution of the parameters during training is presented in Fig. 1. We were able to extract very precise estimates of the ground truth parameters of the dataset. Convergence is faster for the dictionary elements W while we finally also achieve very good estimates for the noise scale σ and prior π . We also appear to achieve a very close approximation of the exact log-likelihood using the truncated free energy (Fig. 1), which shows that our free energy bound is very tight for this data.

Image Patches. For training, we now use $N = 100\,000$ patches of size $D = 16 \times 16$ from a subset of the Van Hateren image dataset [21] that excludes images containing artificial structures. We used the same preprocessing as in [22]. We trained BSC with TVS for 2000 EM iterations and used a sampler adjustment (see Algorithm 2): the first 100 iterations used $M_p = 200$ samples from the prior and $M_q = 0$ samples from the marginal distribution (only 1st approximation); from iteration 100 to iteration 200 we then linearly decreased the number of prior samples to $M_p = 0$ and increased the number of marginal samples to $M_q = 200$ (at all times $M_p + M_q = 200$). Figure 2 shows the basis functions W to converge to represent, e.g., Gabor functions [compare 20].

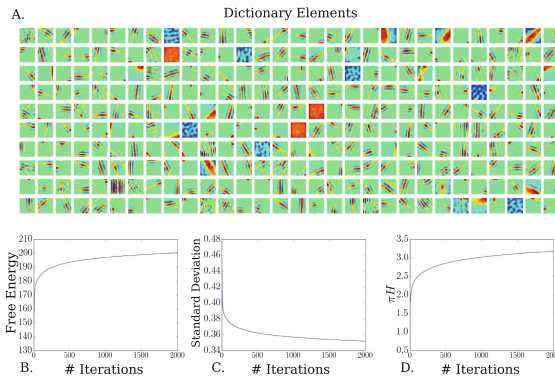


Fig. 2. Image Patches. **A.** The dictionary at convergence. **B.** The evolution of the free energy over TVS iterations. **C.** The evolution of the model standard deviation over TVS iterations. **D.** Evolution of the expected number of active units πH over TVS iterations.

Sigmoid Belief Networks. The second example we will consider here is a typical representative of a Bayesian Network: Sigmoid Belief Networks [SBNs; 23]. While sparse coding approaches are applied to continuous (Gaussian distributed) observed variables, SBNs have binary observed and hidden variables. A further difference is that SBNs require gradients for parameter updates (partial M-steps), while parameter updates of sparse coding models including BSC have well-known updates that fully maximize a corresponding free energy (full M-steps). SBNs thus serve as an example complementary to BSC, and are well suited for our purposes of studying generality and effectiveness of TVS.

For simplicity, we will here consider an SBN with the same graphical model architecture as BSC: one observed and one hidden layer. The SBN generative model is then given by:

$$p(s_h) = \prod_h \pi_h^{s_h} (1 - \pi_h)^{(1-s_h)}, p(\mathbf{y}|\mathbf{s}) = \prod_d g_d^{y_d} (1 - g_d)^{(1-y_d)} \quad (9)$$

where π_h parameterizes the prior distribution and where $g_d = \sigma(\sum_h W_{dh}s_h + b_d)$ is a post-linear non-linearity with sigmoid function σ .

In general, inference for SBNs is challenging because of potentially complex dependencies among its variables. Because of this, direct applications of standard variational approaches [e.g. 23] are challenging, and also popular recent variational methods applying reparameterization [24, 25] are not directly applicable. Also (variational) sampling approaches require additional mechanisms, e.g., the score function based approach needs to reduce the variance of estimations [3].

Artificial Data. As for BSC, the optimization of SBNs by applying Algorithm 2 does not require additional derivations. We again first use a bars test as for BSC but the linearly superimposed bars now go through the sigmoid function and produce binary representations, i.e., generation according to (9). We optimized an SBN with $H = 10$ hidden units on $N = 2000$ data points of such a bars test.

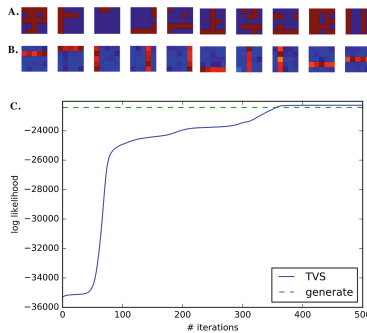


Fig. 3. Application of a shallow SBN with TVS to artificial data (bars test). **A** Ten examples of the training data points. **B** Visualization of the learned weight matrix W of the shallow SBN. All bars are discovered, one for each hidden unit. **C** Learning curve of the free energy. The dashed green line shows the true log likelihood of the SBN with the parameters used for generating the training data.

For Algorithm 2 we used $|\mathcal{K}^{(n)}| = S = 50$ and very few samples for variation were found sufficient ($M_q = M_p = 5$). Results are shown in Fig. 3. The free energy (3) converges to even somewhat higher values than the (here still computable) ground-truth likelihood because of the limited size of training data.

Table 1. Comparison of different models with two layers and different numbers of latents H on binarized MNIST. (*) taken from [2], (\diamond) from [3], (\dagger) from [26], (\ddagger) from [27]. For TVS the final free energy on the test set can directly be computed by iterating the TV-E-step. The right column reports these values as the estimated test log-likelihood (log-LL). Hence, for TVS the log-LL values are a lower bound estimation while results by [26, 27] are from Monte Carlo estimations (and not necessarily lower bounds). For SBNs with 200 hidden units, we observed that TVS slightly out-performs NVIL, and its performance is comparable to RWS. The results of [2] can not directly serve as a comparison of variational approaches (additional knowledge in the form of sparse priors were used).

Model	H	Approx. log-LL
SBN (TVS)	100	-121.91
SBN (TVS)	200	-111.23
SBN (Gibbs)*	200	-94.3
SBN (VB)*	200	-117.0
SBN (NVIL) \diamond	200	-113.1
SBN (WS) \dagger	200	-120.7
SBN (RWS) \dagger	200	-103.1
SBN (AIR) \ddagger	200	-100.9

Binarized MNIST. Finally we apply SBNs to the Binarized MNIST dataset (downloaded from [28], converted as in [29]). We used Algorithm 2 with $|\mathcal{K}^{(n)}| = S = 50$, $M_p = 10$, $M_q = 20$ (no sampler adjustment) and truncated marginal distributions approximated using an MLP with one hidden layer of 500 hidden units and \tanh activation (2nd Approximation). Table 1 compares SBNs optimized by TVS with other models and optimization approaches.

5 Conclusion

The TVS approach studied here is different from previous approaches [6, 9–12] as it does *not* rely on a parametric form of a variational distribution which is then, e.g., sampled from to estimate parameter updates. In contrast, for TVS, the drawn samples *themselves* define the variational distribution and act as its variational parameters. Changing the used samples changes the variational distribution. TVS is thus by definition directly coupling sampling and variational EM which in conjunction with its ‘black box’ applicability is the main contribution of this study. One benefit of the tight coupling seems to be that none of

the diverse variance reduction techniques (which were central to BBVI or NVIL) are required. TVS can thus be considered as the most directly applicable ‘black box’ approach. We have here shown a proof of concept. More advanced models and further algorithmic improvements will be the subject of future studies.

Acknowledgement. We acknowledge funding by the MWK (Low.Saxony) VWZN3189, DFG grant EXC 1077/1, and support for GE through ERC InvariantClass grant 320959.

References

1. Goodfellow, I., Courville, A.C., Bengio, Y.: Large-scale feature learning with spike-and-slab sparse coding. In: ICML (2012)
2. Gan, Z., Heno, R., Carlson, D., Carin, L.: Learning deep sigmoid belief networks with data augmentation. In: AISTATS (2015)
3. Mnih, A., Gregor, K.: Neural var. inf. and learning in BNs. In: ICML (2014)
4. Sheikh, A.S., Lücke, J.: Select-and-sample for spike-and-slab sparse coding. In: NIPS, vol. 29, pp. 3927–3935 (2016)
5. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233 (1999)
6. Hoffman, M.D., Blei, D.: Structured var. inf. In: AISTATS (2015)
7. Salimans, T., Kingma, D.P., Welling, M., et al.: Markov chain Monte Carlo and variational inference: bridging the gap. In: ICML, pp. 1218–1226 (2015)
8. Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D., Turner, R.: Black-box alpha divergence minimization. In: ICML, pp. 1511–1520 (2016)
9. Ranganath, R., Gerrish, S., Blei, D.M.: Black box var. inf. In: AISTATS (2014)
10. Tran, D., Blei, D., Airoldi, E.M.: Copula var. inf. In: NIPS, pp. 3564–3572 (2015)
11. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning (2015)
12. Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M.: Automatic differentiation variational inference. *CoRR* abs/1603.00788 (2016)
13. Lücke, J., Eggert, J.: Expectation truncation and the benefits of preselection in training generative models. *JMLR* **11**, 2855–900 (2010)
14. Forster, D., Lücke, J.: Can clustering scale sublinearly with its clusters? In: AISTATS 2018 (2018, to appear). [arXiv:1711.03431](https://arxiv.org/abs/1711.03431)
15. Lücke, J.: Truncated variational expectation maximization (2016). *arXiv preprint: arXiv:1610.03113*
16. Shelton, J.A., Gasthaus, J., Dai, Z., Lücke, J., Gretton, A.: GP-select. *Neural Comput.* **29**, 2177–2202 (2017)
17. Gu, S., Ghahramani, Z., Turner, R.E.: Neural adaptive sequential Monte Carlo. In: NIPS (2015)
18. Mnih, A., Rezende, D.J.: Var. inf. for Monte Carlo objectives. In: ICML (2016)
19. Haft, M., Hofman, R., Tresp, V.: Gen. binary codes. *Pat. An. Appl.*, pp. 269–84 (2004)
20. 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). https://doi.org/10.1007/978-3-642-15995-4_56
21. van Hateren, J.H., van der Schaaf, A.: *Proc. Royal Soc. Lon. B* **265**, 359–366 (1998)

22. Exarchakis, G., Lücke, J.: Discrete sparse coding. *Neur. Comp.* **29**, 2979–3013 (2017)
23. Saul, L.K., Jaakkola, T., Jordan, M.I.: Mean field theory for sigmoid belief networks. *J. Artif. Intell. Res.* **4**(1), 61–76 (1996)
24. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: *ICLR* (2014)
25. Rezende, D.J., Mohamed, S., Wierstra, D.: *ICML* (2014)
26. Bornschein, J., Bengio, Y.: Reweighted wake-sleep. In: *ICLR* (2015)
27. Hjelm, R.D., et al.: Iterative refinement of the approximate posterior for directed belief networks. In: *NIPS* (2016)
28. Larochelle, H.: Binarized MNIST (2011). cs.toronto.edu/~larocheh/
29. Murray, I., Salakhutdinov, R.: Evaluating probabilities under high-dimensional latent variable models. In: *NIPS* (2008)



Using Hankel Structured Low-Rank Approximation for Sparse Signal Recovery

Ivan Markovsky¹(✉) and Pier Luigi Dragotti²

¹ Department ELEC, Vrije Universiteit Brussel (VUB), Pleinlaan 2, Building K, 1050 Brussels, Belgium

imarkovs@vub.ac.be

² EEE Department, Imperial College London, Exhibition Road, London SW7-2AZ, UK

p.dragotti@imperial.ac.uk

Abstract. Structured low-rank approximation is used in model reduction, system identification, and signal processing to find low-complexity models from data. The rank constraint imposes the condition that the approximation has bounded complexity and the optimization criterion aims to find the best match between the data—a trajectory of the system—and the approximation. In some applications, however, the data is sub-sampled from a trajectory, which poses the problem of sparse approximation using the low-rank prior. This paper considers a modified Hankel structured low-rank approximation problem where the observed data is a linear transformation of a system’s trajectory with reduced dimension. We reformulate this problem as a Hankel structured low-rank approximation with missing data and propose a solution methods based on the variable projections principle. We compare the Hankel structured low-rank approximation approach with the classical sparsity inducing method of ℓ_1 -norm regularization. The ℓ_1 -norm regularization method is effective for sum-of-exponentials modeling with a large number of samples, however, it is not suitable for damped system identification.

Keywords: Low-rank approximation · Hankel structure
Sparse approximation · Missing data estimation
Sum-of-exponentials modeling · ℓ_1 -norm regularization

1 Introduction

The problem considered is defined as follows: Given

- full row rank matrix $A \in \mathbb{R}^{n_g \times n_p}$ with $n_g < n_p$,
- vector of measurements b ,
- structure specification $\mathcal{S} : \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{m \times n}$, and
- rank constraint r ,

$$\begin{aligned} & \text{minimize over } \hat{p} \quad \|b - A\hat{p}\|_2 \\ & \text{subject to} \quad \text{rank}(\mathcal{S}(\hat{p})) \leq r. \end{aligned} \tag{1}$$

The measurements b are obtained as

$$b = A\bar{p} + \tilde{b},$$

where \bar{p} is a vector that we aim to estimate (the “true value”) and \tilde{b} is zero mean Gaussian measurement noise with covariance matrix that is a multiple of the identity. The prior knowledge that makes the estimation of \bar{p} a well posed problem is that it is sparse in the sense that the matrix $\mathcal{S}(\bar{p})$ has low rank:

$$\text{rank}(\mathcal{S}(\bar{p})) \leq r. \tag{2}$$

Therefore, we impose the low-rank prior knowledge on the estimate \hat{p} in problem (1).

Problem (1) is a structured low-rank approximation problem. Its novel element with respect to related problems considered in the literature [1–4, 8, 10] is that a subset of n_g samples are observed. In structured low-rank approximation problem formulations considered in the literature, all n_p samples are available for the estimation of \bar{p} . Our main result, presented in Sect. 2, is a reformulation of problem (1) as an equivalent structured low-rank approximation problem with missing data [8]. Section 3 presents a solution method based on the variable projection principle [5].

Section 4 considers the special case of (1) when the structure \mathcal{S} is Hankel. Hankel structured low-rank approximation has applications in computer algebra, system theory, and signal processing. In the case of a Hankel matrix structure, the rank constraint (2) is equivalent to the constraint that the to-be-estimated vector \bar{p} satisfies a recursive relation [6, 7]

$$a_0 p_t + a_1 p_{t+1} + \dots + a_r p_{t+r} = 0, \quad \text{for } t = 1, \dots, n_p - r.$$

Equivalently, $(\bar{p}_1, \dots, \bar{p}_{n_p})$ is a sum-of-polynomials-times-damped-exponentials discrete-time signal [9]. In system theoretic terms, $(\bar{p}_1, \dots, \bar{p}_{n_p})$ is the output of a discrete-time autonomous linear time-invariant system of order at most r . Therefore, (1) can be viewed as the problem of identifying an autonomous linear time-invariant system from partial noisy measurements that are a linear transformation of a system’s output.

We compare the approach of solving the autonomous linear time-invariant system identification problem via (1) with method based on ℓ_1 -norm regularization. This latter approach imposes sparsity on the frequency domain representation of the signal. Indeed, an r -sparse frequency domain signal is a sum of r complex exponentials in the time-domain. However, the frequencies are constrained to belong to the grid $\{k\omega_0 \mid k \in \mathbb{Z}\}$, where $\omega_0 := 2\pi/n_p$. Therefore, the accuracy of the ℓ_1 -norm regularization method for autonomous linear time-invariant system identification is limited. Another essential difference between (1) and the ℓ_1 -norm approach is that the ℓ_1 -norm approach can not deal with

damped exponentials and polynomials. Indeed, damping gives rise to “skirts” in the frequency domain, so that the signal is no longer k -sparse in the frequency domain, however, it is sparse in the sense of (2). Section 5 shows numerical examples.

2 Link to Missing Data Estimation

We use the notation $p_{1:n_g}$ for the subvector $[p_1 \cdots p_{n_g}]^\top$ consisting of the first n_g elements of p .

Theorem 1. *Problem (1) is equivalent to the structured low-rank approximation problem with missing values*

$$\begin{aligned} & \text{minimize over } \hat{p}' && \|b - \hat{p}'_{1:n_g}\|_2 \\ & \text{subject to} && \text{rank}(\mathcal{S}'(\hat{p}')), \end{aligned} \quad (3)$$

where

$$\mathcal{S}'(\cdot) := \mathcal{S}(V \cdot) \quad \text{and} \quad \hat{p}' = V^{-1} \hat{p},$$

with a nonsingular matrix V , such that $AV = [I_{n_g} \ 0]$.

Proof. Using the change of variables $\hat{p}' = V^{-1} \hat{p}$, where V is a nonsingular matrix, problem (1) becomes

$$\begin{aligned} & \text{minimize over } \hat{p}' && \|b - A' \hat{p}'\|_2 \\ & \text{subject to} && \text{rank}(\mathcal{S}'(\hat{p}')) \leq r, \end{aligned} \quad (4)$$

where $A' = AV$ and $\mathcal{S}'(\cdot) := \mathcal{S}(V \cdot)$. By the full row rank assumption, we can choose V , so that

$$A' = AV = [I_{n_g} \ 0]. \quad (5)$$

With this choice of V , problem (4) becomes (3).

Note that if the original structure \mathcal{S} is affine, the new structure \mathcal{S}' is also affine.

Example 1. Let A consists of the first n_g rows of the $n_p \times n_p$ discrete cosine transform matrix C . Since C is orthonormal, we have that $V = C^\top$ satisfies condition (5). The change of variables $\hat{p}' = V^\top \hat{p}$ then transforms the problem into the frequency domain.

3 Solution Method

Next, we present a local optimization method for solving problem (3). First, we represent the rank constraints in the kernel form

$$\text{rank}(\mathcal{S}(\hat{p})) \leq r \quad \iff \quad \text{there is } R \in \mathbb{R}^{(m-r) \times m}, \text{ such that} \\ R \mathcal{S}(\hat{p}) = 0 \text{ and } R \text{ is full row rank.} \quad (6)$$

Then, we use the variable projection principle to eliminate \hat{p} , which results in a nonlinear least-squares in R .

Representing the constraint of (3) in the kernel form (6), leads to the double minimization problem

$$\text{minimize over } R \in \mathbb{R}^{(m-r) \times m} \quad f(R) \quad \text{subject to } R \text{ is full row rank,} \quad (7)$$

where

$$f(R) := \min_{\hat{p}} \|p - \hat{p}\|_2 \quad \text{subject to } R\mathcal{S}(\hat{p}) = 0. \quad (8)$$

The computation of $f(R)$, called “inner” minimization, is over the estimate \hat{p} of p . The minimization over the kernel parameter $R \in \mathbb{R}^{(m-r) \times m}$ is called “outer”. The inner minimization problem is a projection of the columns of $\mathcal{S}(p)$ onto the model $\mathcal{B} := \ker(R)$. Note that, the projection depends on the parameter R , which is the variable in the outer minimization problem. Thus, the name “variable projection”.

The general linear structure

$$\mathcal{S} : \mathbb{R}^{n_p} \rightarrow \mathbb{R}^{m \times n}, \quad \mathcal{S}(\hat{p}) = \sum_{k=1}^{n_p} S_k \hat{p}_k \quad (9)$$

is specified by the n_p matrices $S_1, \dots, S_{n_p} \in \mathbb{R}^{m \times n}$. Let

$$\mathbf{S} := [\text{vec}(S_1) \cdots \text{vec}(S_{n_p})] \in \mathbb{R}^{mn \times n_p},$$

so that

$$\text{vec}(\mathcal{S}(\hat{p})) = \mathbf{S}\hat{p}, \quad \text{or} \quad \mathcal{S}(\hat{p}) = \text{vec}^{-1}(\mathbf{S}\hat{p}). \quad (10)$$

Define the change of variables

$$\hat{p} \mapsto \Delta p = p - \hat{p}.$$

Then, the constraint of the optimization problem becomes

$$\begin{aligned} R\mathcal{S}(\hat{p}) = 0 &\iff R\mathcal{S}(p - \Delta p) = 0 \\ &\iff R\mathcal{S}(p) - R\mathcal{S}(\Delta p) = 0 \\ &\iff \text{vec}(R\mathcal{S}(\Delta p)) = \text{vec}(R\mathcal{S}(p)) \\ &\iff \underbrace{[\text{vec}(RS_1) \cdots \text{vec}(RS_{n_p})]}_{G(R)} \Delta p = \underbrace{\text{vec}(R\mathcal{S}(p))}_{h(R)} \\ &\iff G(R)\Delta p = h(R). \end{aligned}$$

Assuming that

$$n_p \leq (m - r)n \quad (11)$$

the inner minimization problem (8) with respect to the new variable Δp is a generalized linear least norm problem

$$f(R) = \min_{\Delta p} \|\Delta p_{1:n_g}\|_2 \quad \text{subject to } G(R)\Delta p = h(R). \quad (12)$$

(12) is not a standard least norm problem due to the presence of missing data (or equivalently singularity of the cost function), however, it has an analytic solution [8, Theorem 2.1].

For the outer minimization problem in (7), *i. e.*, the minimization of M over R , subject to the constraint that R is full row rank, we use general purpose constrained local optimization methods [11], representing the full row rank constraint as $RR^T = I_{m-r}$. This is a nonconvex optimization problem, so that there is no guarantee that a globally optimal solution is found.

4 Hankel Structured Sparse Approximation Problems and ℓ_1 -norm Regularization

In this section, we consider the special case of problem (1) when the structure \mathcal{S} is Hankel

$$\mathcal{H}_m(p) := \begin{bmatrix} p_1 & p_2 & p_3 & \cdots & p_{n_p-m+1} \\ p_2 & p_3 & \ddots & & p_{n_p-m+2} \\ p_3 & \ddots & & & p_{n_p-m+3} \\ \vdots & & & & \vdots \\ p_m & p_{m+1} & \cdots & \cdots & p_{n_p} \end{bmatrix}. \quad (13)$$

By the result of Theorem 1, problem (1) is a Hankel structured low-rank approximation with missing data. In turn, Hankel structured low-rank approximation is a linear time-invariant system identification problem with missing data. Therefore, equivalently, we consider a problem of system identification with missing data.

An alternative approach for missing data estimation with sparsity prior is ℓ_1 -norm regularization. Sparsity of a signal in the frequency domain means that the signal is a sum of a few exponentials. In the paper, we consider real-valued time-domain signals, so that the frequency domain signal has an additional symmetry property.

A signal that is a sum of n -complex exponentials can be represented as an output of an autonomous linear time-invariant system of order n . Alternatively, such a signal can be represented as the impulse response of a n -th order linear time-invariant system. Representing exactly or approximately a given signal as an output of an autonomous linear time-invariant system or as the impulse response of an input/output linear time-invariant system are fundamental problems in system theory and system identification.

Next, we explain the similarities and differences between sparse approximation by ℓ_1 -norm minimization in the frequency domain and sparse approximation by Hankel structured low-rank approximation. The underlying assumption for the ℓ_1 -norm minimization problem is that the data b is generated as

$$b = Dx + \tilde{b}, \quad (14)$$

where D is a $n_g \times n_p$ matrix with $n_g < n_p$, x is k -sparse with $k \ll n_p$, and \tilde{b} is a zero mean Gaussian random vector with covariance matrix $\sigma^2 I$. Moreover, it is assumed that D consists of the first n_g rows of the inverse discrete cosine transform matrix C^\top . Due to the properties of D (submatrix of the inverse discrete cosine transform) and x (k -sparse vector), $\bar{b} := Dx$ is a sum of k cosines with frequencies on the grid

$$0\frac{2\pi}{n_p}, 1\frac{2\pi}{n_p}, 2\frac{2\pi}{n_p}, \dots, (n_p - 1)\frac{2\pi}{n_p}. \tag{15}$$

Assuming that enough observations are available, namely

$$n_g \geq 2r + 1, \quad \text{where } r := 2k,$$

the Hankel matrix $\mathcal{H}_{r+1}(\bar{b})$ with $r + 1$ rows and $n_g - r$ columns, constructed from \bar{b} has rank r . Vice versa,

$$\text{rank}(\mathcal{H}_{r+1}(\bar{b})) \leq r \tag{16}$$

implies that \bar{b} is a sum of at most $2n$ polynomials-times-damped-exponentials.

Note that (16) does not impose a constraint that the frequencies are on (15); they can be any real numbers in the interval $[0, 2\pi)$. Also (16) allows damped cosines while the model $\bar{b} = Dx$ does not allow damping. Therefore, (16) is not equivalent to $\bar{b} = Dx$ with x k -sparse. For large values of n_p , (15) approximates “well” the interval $[0, 2\pi)$.

5 Numerical Examples

In this section we consider the Hankel structured low-rank approximation problem

$$\begin{aligned} &\text{minimize} \quad \text{over } \hat{b} \quad \|b - A\hat{p}\|_2 \\ &\text{subject to} \quad \text{rank}(\mathcal{H}_{r+1}(\hat{p})) \leq r \end{aligned} \tag{17}$$

First, we specialize the variable projections method described in Sect. 3 to the Hankel structured case and demonstrate on a simulation example that the resulting algorithm allows us to separate signal from noise. Then, we compare numerically the variable projections method with the ℓ_1 -norm regularization method in setup of (14).

Autonomous System Identification from Data With Missing Values

In the case of Hankel structure (13), the matrices S_k in (9) are

$$S_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

and the $G(R)$ matrix in (12) is

$$G(R) = \begin{bmatrix} R_0 & R_1 & \cdots & R_r & & \\ & R_0 & R_1 & \cdots & R_r & \\ & & \ddots & \ddots & & \ddots \\ & & & & R_0 & R_1 & \cdots & R_r \end{bmatrix},$$

where all missing elements are zeros. Fast ($O(n_p)$) implementation of the variable projection algorithm, taking into account the structure of $G(R)$ for the cost function and Jacobian evaluation is presented in [12].

Example 2. A random second order ($r = 2$) autonomous linear time-invariant system is generated in Matlab by the function `drss` and a random trajectory \bar{p} of the system with $n_p = 50$ samples is then generated. The sampling matrix A is $[I_{n_g} \ 0]$, where $n_g = 20$, *i.e.*, only the first 40% of the samples of \bar{p} are observed. Finally, zero mean, white, Gaussian noise with standard deviation s is added to the true samples.

Figure 1 shows the relative estimation error

$$e := \|\bar{b} - \hat{p}_{1:n_g}\|_2 / \|\bar{b}\|_2$$

from a Monte Carlo experiment with standard deviations varying in the interval $[0, 0.1]$ (signal-to-noise ratio varying from 46 dB to infinity). The result shows that the low-rank prior allows us to filter noise from the data. Indeed, the error e in using the noisy data (solid black line) is higher and increases faster than the error e in using the estimate \hat{p} (the solution of problem (3) obtained with the variable projections algorithm).

Moreover, it can be shown that in the simulation setup of the example, the solution of problem (1) gives the maximum likelihood estimator, so that it is statistically optimal.

Comparison with the ℓ_1 -norm Regularization Method

In this section, we consider data generated from the compressive sensing model (14) with $n_p = 100$, $n_g = 20$, $k = 2$, and noise standard deviation $s = 0.1$. The true data \bar{p} is a sum of two sines with frequencies on the grid (15). With this simulation setup, the ℓ_1 -norm regularization method recovers the correct frequencies with 100% success rate.

The low-rank constraint (2) with Hankel structured matrix and rank $n = 4$ imposes the weaker prior that the signal is a sum-of-damped exponentials, *i.e.*, damping is allowed and the frequencies are not assumed to be on the grid (15). Nevertheless, in the above simulation example the estimator defined by problem (17) also recovers the correct frequencies with 100% success rate.

Both the ℓ_1 -norm regularization method and (17) fail when the number of given samples n_g is decreased and/or the noise standard deviation s is increased. The ℓ_1 -norm regularization method fails for a smaller number of samples and at a higher noise standard deviation s . The reader can reproduce the reported results by downloading the SLRA package (<http://slra.github.io/>) and <http://homepages.vub.ac.be/~imarkovs/software/ica18.tar>.

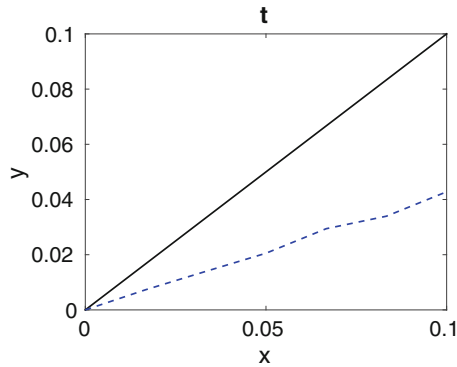


Fig. 1. The maximum likelihood (ML) estimator obtained by solving problem (1) with the variable projections algorithm (blue dashed line) improves the relative estimation error in comparison with the use of the raw noisy data (black solid line). (Color figure online)

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC Grant 258581 “Structured low-rank approximation: Theory, algorithms, and applications”; Fund for Scientific Research (FWO-Vlaanderen), FWO projects G028015N “Decoupling multivariate polynomials in nonlinear system identification”; G090117N “Block-oriented nonlinear identification using Volterra series”; and FWO/FNRS Excellence of Science project 30468160 “Structured low-rank matrix/tensor approximation: numerical optimization-based algorithms and applications”.

References

1. Bresler, Y., Macovski, A.: Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Trans. Acoust. Speech Sign. Process.* **34**, 1081–1089 (1986)
2. Cadzow, J.: Signal enhancement—a composite property mapping algorithm. *IEEE Trans. Sign. Proc.* **36**, 49–62 (1988)
3. Chu, M., Funderlic, R., Plemmons, R.: Structured low rank approximation. *Linear Algebra Appl.* **366**, 157–172 (2003)
4. De Moor, B.: Structured total least squares and L_2 approximation problems. *Linear Algebra Appl.* **188–189**, 163–207 (1993)
5. Golub, G., Pereyra, V.: Separable nonlinear least squares: the variable projection method and its applications. *Inst. Phys. Inverse Prob.* **19**, 1–26 (2003)
6. Markovsky, I.: Structured low-rank approximation and its applications. *Automatica* **44**(4), 891–909 (2008)
7. Markovsky, I.: *Low-Rank Approximation: Algorithms, Implementation. Applications.* Springer, London (2018). <https://doi.org/10.1007/978-1-4471-2227-2>
8. Markovsky, I., Usevich, K.: Structured low-rank approximation with missing data. *SIAM J. Matrix Anal. Appl.* **34**(2), 814–830 (2013)

9. Polderman, J., Willems, J.C.: Introduction to Mathematical Systems Theory. Springer-Verlag, New York (1998)
10. Rosen, J., Park, H., Glick, J.: Total least norm formulation and solution of structured problems. *SIAM J. Matrix Anal. Appl.* **17**, 110–126 (1996)
11. Usevich, K., Markovsky, I.: Optimization on a Grassmann manifold with application to system identification. *Automatica* **50**, 1656–1662 (2014)
12. Usevich, K., Markovsky, I.: Variable projection for affinely structured low-rank approximation in weighted 2-norms. *J. Comput. Appl. Math.* **272**, 430–448 (2014)



Probabilistic Sparse Non-negative Matrix Factorization

Jesper Løve Hinrich^(✉) and Morten Mørup

Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Richard Petersens Plads Bld. 321,
2800 Kongens Lyngby, Denmark
jehi@dtu.dk

Abstract. In this paper, we propose a probabilistic sparse non-negative matrix factorization model that extends a recently proposed variational Bayesian non-negative matrix factorization model to explicitly account for sparsity. We assess the influence of imposing sparsity within a probabilistic framework on either the loading matrix, score matrix, or both and further contrast the influence of imposing an exponential or truncated normal distribution as prior. The probabilistic methods are compared to conventional maximum likelihood based NMF and sparse NMF on three image datasets; (1) A (synthetic) swimmer dataset, (2) The CBCL face dataset, and (3) The MNIST handwritten digits dataset. We find that the probabilistic sparse NMF is able to automatically learn the level of sparsity and find that the existing probabilistic NMF as well as the proposed probabilistic sparse NMF prunes inactive components and thereby automatically learns a suitable number of components. We further find that accounting for sparsity can provide more part based representations but for the probabilistic modeling the choice of priors and how sparsity is imposed can have a strong influence on the extracted representations.

Keywords: Non-negative matrix factorization · Sparsity
Bayesian modeling · Sparse non-negative matrix factorization

1 Introduction

Non-negative matrix factorization (NMF) also denoted positive matrix factorization [1] has become a popular feature extraction tool due to its easy interpretable part based representations [2]. NMF is based on the decomposition $\mathbf{X}^{\mathbf{I} \times \mathbf{J}} \approx \mathbf{W}^{\mathbf{I} \times \mathbf{D}} \mathbf{H}^{\mathbf{D} \times \mathbf{J}}$ in which the elements of both \mathbf{W} and \mathbf{H} are non-negative, i.e. $w_{id} \geq 0 \forall i, d$ and $h_{dj} \geq 0 \forall d, j$. Several procedures for fitting the parameters in NMF have been proposed. For the least squares NMF objective $\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$ this includes multiplicative update [3], active sets [4], projected gradient [5], and component-wise updating [6, 7].

Unfortunately, NMF is not in general unique [8, 9]. In particular, if the data does not sufficiently span the positive orthant multiple non-negative representations may equally well represent the data and the representation is not guaranteed to be part based. In order to provide part based representation and alleviate

issues of non-unique representations sparsity in the NMF model has been proposed. This includes fixing each of the extracted feature representation to a given sparsity level (quantified by $\text{sparseness}(\mathbf{w}_d) = \frac{\sqrt{I} - \|\mathbf{w}_d\|_1 / \|\mathbf{w}_d\|_2}{\sqrt{I-1}}$) [10] or regularizing the NMF objective by a sparsity promoting penalty such as the ℓ_1 -norm, i.e. minimizing the objective $\|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda\|\mathbf{W}\|_1$, keeping either the norm of each component of \mathbf{H} fixed or regularizing by the ℓ_2 -norm squared corresponding to imposing a Gaussian prior (c.f. [11]). Unfortunately, these sparse approaches require a user defined tuning of the sparsity level or degree of regularization λ .

Whereas initial work on NMF was based on maximum-likelihood estimation the NMF model has been advanced to probabilistic modeling. This includes for the least squares objective (i.e., Gaussian noise assumption) probabilistic inference based on Markov-Chain Monte-Carlo [12, 13] and variational Bayesian inference [14]. Benefits of probabilistic NMF includes quantifying the number of components [12, 15, 16] and accounting for parameter uncertainty and noise.

Within the variational Bayesian framework we propose several formulations of probabilistic sparse non-negative matrix factorization (psNMF) and investigate their ability to promote part based representations. Neither sparsity nor priors based on truncated normal factors were investigated in the context of variational inference by [14]. The approach to enforcing sparsity is inspired by [17] where sparsity was considered on one factor in probabilistic PCA. In contrast, we investigate non-negative factors and the effect of modeling sparsity on one or both factors and two ways of constraining the dense factor (infinity norm or unit variance prior). On three image datasets we contrast the proposed psNMF to conventional NMF (c.f. [18]) and the recently proposed variational Bayesian NMF [14] as well as to sparse NMF tuned through fixed sparsity degrees [10] and conventional user defined levels of regularization [11].

2 Methods

Bayesian inference aim to identify the posterior distribution of parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}, \dots\}$ given the data \mathbf{X} , i.e. $P(\boldsymbol{\theta}|\mathbf{X})$. The exact posterior is in general intractable and this paper relies on variational Bayesian inference (c.f. [19]) to approximate $P(\boldsymbol{\theta}|\mathbf{X})$. VB uses a set of simpler distributions $Q(\boldsymbol{\theta})$ to obtain a lowerbound on the evidence (ELBO), i.e. $\log P(\mathbf{X}) \geq \int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) \log \frac{P(\mathbf{X}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} = \mathcal{L}(Q)$ where $Q(\boldsymbol{\theta})$ are chosen such that $\mathcal{L}(Q)$ is tractable. VB identifies a local optimum of the (variational) posterior distribution by maximizing the ELBO.

2.1 Probabilistic Non-negative Matrix Factorization

Recently, variational inference for probabilistic NMF assuming Gaussian noise was proposed in [14] assuming exponential distributions on the factors. We presently consider this model as well as the corresponding model assuming truncated normal distribution as priors. The generative model of the considered probabilistic non-negative matrix factorization is specified as:

$$\lambda_d \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda), \tag{1}$$

$$w_{id} | \lambda_d \sim \mathcal{TN}(0, \lambda_d^{-1}, 0, \infty) \quad \text{or} \quad w_{id} | \lambda_d \sim \text{Exponential}(\lambda_d^{-1}), \tag{2}$$

$$h_{dj} | \lambda_d \sim \mathcal{TN}(0, \lambda_d^{-1}, 0, \infty) \quad \text{or} \quad h_{dj} | \lambda_d \sim \text{Exponential}(\lambda_d^{-1}), \tag{3}$$

$$\tau \sim \text{Gamma}(\alpha_\tau, \beta_\tau), \quad \mathbf{x}_j | \mathbf{W}, \mathbf{h}_j, \tau \sim \mathcal{N}(\mathbf{x}_j | \mathbf{W}\mathbf{h}_j, \tau^{-1} \mathbf{I}_I). \tag{4}$$

Where the reconstruction error $\mathbf{x}_j - \mathbf{W}\mathbf{h}_j$ follows a normal distribution with noise precision (inverse variance) τ . Each component of \mathbf{W} and \mathbf{H} (i.e. $\mathbf{w}_d, \mathbf{h}_d$) shares a common prior λ_d used to infer the scale of a component, see also [14]. Note λ_d represents the precision or rate for the truncated normal and exponential distribution, respectively. This prior can be used for model order selection and is commonly called an automatic relevance determination (ARD) prior. The posterior distribution of $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}, \tau, \{\lambda_d\}_{d=1,2,\dots,D}\}$ is inferred from the data \mathbf{X} while the shape α_* and rate β_* of the Gamma distributions are fixed. In absence of a priori information, weak priors are specified (c.f. $\alpha_* = \beta_* = 10^{-4}$), thereby the distribution of τ and λ_d are determined primarily from the data.

2.2 Probabilistic Sparse Non-negative Matrix Factorization

Probabilistic modeling can be used to automatically infer the sparsity pattern and level, as supported by the data. The Bayesian framework facilitates sparse modeling through the prior distribution of the factor matrices (\mathbf{W}, \mathbf{H}), along with a (hyper-)prior distribution on its parameters. In the context of probabilistic sparse principal component analysis, [17] proposed using either a Laplace, Gamma or Jeffrey’s distribution as an element-wise ARD prior for obtaining a sparse representation. Inspired by [17], we propose the probabilistic sparse NMF (psNMF) by placing a sparsity enforcing prior on each element of the associated factor (\mathbf{W}, \mathbf{H}). For conciseness, this is presently shown for the truncated normal formulation. A sparse representation may be desired in either \mathbf{W}, \mathbf{H} or both. Below we specify the generation of \mathbf{W}, \mathbf{H} when both factors are sparse,

$$\begin{aligned} \lambda_{id}^{(\mathbf{W})} &\sim \text{Gamma}(\alpha_\lambda^{(\mathbf{W})}, \beta_\lambda^{(\mathbf{W})}), & w_{id} | \lambda_{id}^{(\mathbf{W})} &\sim \mathcal{TN}\left(0, \left(\lambda_{id}^{(\mathbf{W})}\right)^{-1}, 0, \infty\right), \\ \lambda_{dj}^{(\mathbf{H})} &\sim \text{Gamma}(\alpha_\lambda^{(\mathbf{H})}, \beta_\lambda^{(\mathbf{H})}), & h_{dj} | \lambda_{dj}^{(\mathbf{H})} &\sim \mathcal{TN}\left(0, \left(\lambda_{dj}^{(\mathbf{H})}\right)^{-1}, 0, \infty\right). \end{aligned} \tag{5}$$

When only one factor is sparse (e.g. \mathbf{W}), the scale of the other factor (e.g. \mathbf{H}) should be fixed to avoid scale ambiguity in the solution. We consider two approaches; (1) Using a truncated normal distribution with unit variance as a regularizer. (2) Restricting the maximum value of the factor to 1 (e.g., constraining according to the infinity-norm $\|\mathbf{H}\|_\infty \leq 1$) using a Uniform(0,1) prior distribution. The generative model for psNMF with sparsity on one factor is,

$$\begin{aligned} \lambda_{id}^{(\mathbf{W})} &\sim \text{Gamma}(\alpha_\lambda^{(\mathbf{W})}, \beta_\lambda^{(\mathbf{W})}) \quad , \quad w_{id} | \lambda_{id}^{(\mathbf{W})} \sim \mathcal{TN}\left(0, \left(\lambda_{id}^{(\mathbf{W})}\right)^{-1}, 0, \infty\right), \\ h_{dj} &\sim \mathcal{TN}(0, 1, 0, \infty) \quad \text{or} \quad h_{dj} \sim \text{Uniform}(0, 1). \end{aligned} \tag{6}$$

If \mathbf{H} is sparse and \mathbf{W} is dense, the distributions are switched. The psNMF models with exponential factors are found by changing $\mathcal{TN}(0, \lambda_*, 0, \infty)$ to $\text{Exponential}(\lambda_*)$.

2.3 Variational Distributions and Update Rules

The exact posterior distribution $P(\boldsymbol{\theta}|\mathbf{X})$ is intractable, thus its variational approximation $Q(\boldsymbol{\theta}|\mathbf{X})$ is used instead. The Q -distributions are chosen to factorize over parameters, i.e. $Q(\boldsymbol{\theta}) = \prod_i Q(\theta_i)$, called a mean-field approximation. For psNMF with sparsity on both factors,

$$Q(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^I \prod_{d=1}^D \mathcal{TN}(\mathbf{w}_{id} | \mu_{\mathbf{w}_{id}}, \sigma_{\mathbf{w}_{id}}^2, 0, \infty) \text{Gamma}(\lambda_{id}^{(\mathbf{W})} | \tilde{\alpha}_{\lambda_{id}^{(\mathbf{W})}}, \tilde{\beta}_{\lambda_{id}^{(\mathbf{W})}}) \quad (7)$$

$$\prod_{j=1}^J \prod_{d=1}^D \mathcal{TN}(\mathbf{h}_{dj} | \mu_{\mathbf{h}_{dj}}, \sigma_{\mathbf{h}_{dj}}^2, 0, \infty) \text{Gamma}(\lambda_{dj}^{(\mathbf{H})} | \tilde{\alpha}_{\lambda_{dj}^{(\mathbf{H})}}, \tilde{\beta}_{\lambda_{dj}^{(\mathbf{H})}}) \text{Gamma}(\tau | \tilde{\alpha}_\tau, \tilde{\beta}_\tau).$$

Due to the Gaussian noise assumption, $Q(\mathbf{W})$ and $Q(\mathbf{H})$ follow a truncated normal distribution, regardless of whether $P(\mathbf{W})$ or $P(\mathbf{H})$ is truncated normal or exponential distributed. Variational update rules are determined using Eq. (13) from [19] and results are shown below when both \mathbf{W} and \mathbf{H} are sparse,

$$\sigma_{\mathbf{w}_{id}}^2 = \left(\langle \tau \rangle \langle \mathbf{h}_d \mathbf{h}_d^\top \rangle + \langle \lambda_{id}^{(\mathbf{W})} \rangle \right)^{-1}, \quad (8)$$

$$\mu_{\mathbf{w}_{id}} = \sigma_{\mathbf{w}_{id}}^2 \langle \tau \rangle \left(\langle \mathbf{h}_d \rangle \langle \mathbf{x}_i^\top \rangle - \sum_{d' \neq d} \langle \mathbf{w}_{id'} \rangle \langle \mathbf{h}_d \mathbf{h}_{d'}^\top \rangle \right), \quad (9)$$

$$\sigma_{\mathbf{h}_{dj}}^2 = \left(\langle \tau \rangle \langle \mathbf{w}_d^\top \mathbf{w}_d \rangle + \langle \lambda_{dj}^{(\mathbf{H})} \rangle \right)^{-1}, \quad (10)$$

$$\mu_{\mathbf{h}_{dj}} = \sigma_{\mathbf{h}_{dj}}^2 \langle \tau \rangle \left(\langle \mathbf{w}_d^\top \rangle \langle \mathbf{x}_j \rangle - \sum_{d' \neq d} \langle \mathbf{h}_{d'j} \rangle \langle \mathbf{w}_d^\top \mathbf{w}_{d'} \rangle \right), \quad (11)$$

$$\tilde{\beta}_\tau = \beta_\tau + \frac{1}{2} \cdot \left(\text{trace}(\mathbf{X}^\top \mathbf{X}) + \text{trace}(\langle \mathbf{W}^\top \mathbf{W} \rangle \langle \mathbf{H} \mathbf{H}^\top \rangle) - 2 \cdot \text{trace}(\mathbf{X}^\top \langle \mathbf{W} \rangle \langle \mathbf{H} \rangle) \right), \quad \tilde{\alpha}_\tau = \alpha_\tau + \frac{I \cdot J}{2}, \quad (12)$$

$$\tilde{\alpha}_{\lambda_{id}^{(\mathbf{W})}} = \alpha_{\lambda_{id}^{(\mathbf{W})}} + \frac{1}{2}, \quad \tilde{\beta}_{\lambda_{id}^{(\mathbf{W})}} = \beta_{\lambda_{id}^{(\mathbf{W})}} + \frac{1}{2} \langle w_{id}^2 \rangle, \quad (13)$$

$$\tilde{\alpha}_{\lambda_{dj}^{(\mathbf{H})}} = \alpha_{\lambda_{dj}^{(\mathbf{H})}} + \frac{1}{2}, \quad \tilde{\beta}_{\lambda_{dj}^{(\mathbf{H})}} = \beta_{\lambda_{dj}^{(\mathbf{H})}} + \frac{1}{2} \langle h_{dj}^2 \rangle, \quad (14)$$

where $\langle \cdot \rangle$ is the expected value under the variational distribution, i.e. $\mathbb{E}[\cdot]_{Q(\boldsymbol{\theta})}$. The psNMF model reduces to pNMF [14] with truncated normal priors when λ is placed on the columns and shared between the factors, i.e.

$$\tilde{\alpha}_{\lambda_d} = \alpha_{\lambda_d} + \frac{I+J}{2}, \quad \tilde{\beta}_{\lambda_d} = \beta_{\lambda_d} + \frac{1}{2} \sum_{i=1}^I \langle w_{id}^2 \rangle + \frac{1}{2} \sum_{j=1}^J \langle h_{dj}^2 \rangle. \quad (15)$$

Changing $P(\mathbf{W})$ from a truncated normal to an exponential distribution determines whether the prior affects μ or σ of $Q(\mathbf{W})$. The exponential distribution affects the former (see Eq. (16)) while the truncated normal distribution affects the latter (see Eqs. (8) and (9)).

$$\begin{aligned}\mu_{\mathbf{w}_{id}} &= \sigma_{\mathbf{w}_{id}}^2 \left(\langle \tau \rangle \left(\langle \mathbf{h}_d \rangle \langle \mathbf{x}_i^\top \rangle - \sum_{d' \neq d} \langle \mathbf{w}_{id'} \rangle \langle \mathbf{h}_d \mathbf{h}_{d'}^\top \rangle \right) - \langle \lambda_{i,d}^{(\mathbf{W})} \rangle \right) \\ \sigma_{\mathbf{w}_{id}}^2 &= (\langle \tau \rangle \langle \mathbf{h}_d \mathbf{h}_d^\top \rangle)^{-1}.\end{aligned}\quad (16)$$

Matlab implementations of the above models can be found at: <https://github.com/JesperLH/psNMF-LVA2018>.

3 Results and Discussion

We investigate the influence on the estimated components with respect to conventional NMF using the implementation of [18] restricted to exclusively use HALS updating, sparse NMF (sparseness imposed on \mathbf{W}) optimizing for an explicit sparsity level γ (see [10]), denoted PH-sNMF(γ) or penalizing model complexity by $\lambda \|\mathbf{w}_d\|_1$ and $\eta \|\mathbf{H}\|_F^2$. Sparsity is here implicitly controlled by varying $\lambda = 10^{-2}, 10^0, 10^2$ and fixing $\eta = 1$. We used the implementation `sparsenmfnnls` from [11], denoted YA-sNMF(λ). For probabilistic modeling, we considered sparsity imposed on either \mathbf{W} , \mathbf{H} , or both, denoted psNMF(S, \cdot), psNMF(\cdot , S) and psNMF(S, S), respectively. The dense factor is modeled using either a uniform Uniform(0, 1) or by fixing the rate of the exponential or precision of truncated normal distribution to 1, denoted as psNMF(\cdot , Inf) and psNMF(\cdot , I), respectively. When non-negativity is modeled by the exponential distribution the suffix (Exp) is added, e.g. psNMF(Exp). All models were run for 200 iterations and restarted 20 times to mitigate the effects of local optima. For each model, the restart achieving the lowest root mean square error (RMSE) or highest ELBO (probabilistic models) is shown.

3.1 Noisy Swimmer Dataset

The swimmer dataset [8] consists of $J = 256$ images (32×32 pixels, $I = 1024$), each constructed with a static torso region and four limbs. There are four possible articulations for each limb and all possible combinations ($4^4 = 256$) are constructed. To investigate the benefit of accounting for noise in probabilistic modeling we consider a noisy variant of the swimmer data with additive white noise. The noise variance is set such that the signal-to-noise ratio (SNR) is 5 decibels (SNR = 5 dB), i.e. more signal than noise. The SNR_{dB} is defined as $SNR_{dB} = 10 \log_{10}(\frac{Power_{signal}}{Power_{noise}})$, where $Power_{\star}$ is the sum of squared elements.

The non-probabilistic NMF and PH-sNMF methods are unable to find the underlying parts based representation due to an incorrect model order and the presence of noise, see Fig. 1. The best non-probabilistic method is YA-sNMF($\lambda = 10^2$)

which mostly separates the articulations including the (invariant) torso as a separate component. The size of the subspace $D = 16$ is correctly identified, but the right leg is not separated correctly. Setting λ correctly is extremely important as $\lambda = 1$ identifies too many and $\lambda = 10^2$ too few components.

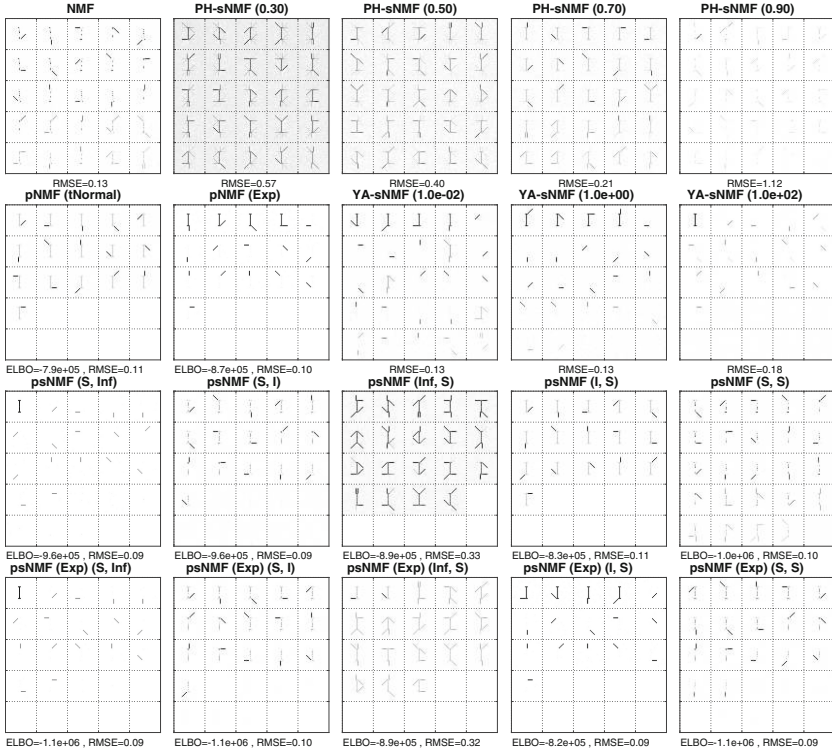


Fig. 1. Noisy Swimmer Dataset (SNR = 5 dB): The RMSE error of $\mathbf{X}_{\text{noiseless}} \approx \mathbf{W}\mathbf{H}$ is given for all methods, where \mathbf{W} and \mathbf{H} have been estimated from $\mathbf{X}_{\text{noisy}}$.

The probabilistic methods pNMF, psNMF(I,S) and psNMF(S,I) based both on the truncated normal distribution and exponential prior are barely affected by the noise and correctly identify a representation using $D^* = 16$ components with the torso either placed across the different articulations or along with the articulation of one of the limbs (pNMF(Exp) and psNMF(Exp)(I,S)). When using the uniform distribution and sparsity on \mathbf{W} the torso is separated out as a 17th component. The representation estimated by psNMF(Inf, S) stand out, as the estimated components resemble actual data points. We attribute this to the uniform distribution apriori having equal support on the entire $[0; 1]$ interval whereas the truncated normal and exponential priors favor inactive pixels making these differences of the priors when facing limited observations, i.e. $I \gg J$ highly influential on the extracted components. For sparsity on both \mathbf{W} and \mathbf{H} ,

the representation of psNMF is similar to NMF and YA-NMF(0.01). The evidence lowerbound (ELBO) identifies pNMF as the most likely model, followed by psNMF(Exp)(I, S) and when inspecting the RMSE in Fig. 1 we observe that the probabilistic models in general provide a better reconstruction of the noise-free data than the conventional NMF and sparse NMF procedures while automatically identifying a suitable number of components.

3.2 MIT CBCL Face Dataset

Lee and Seung [2] used a variant of the CBCL Face Database¹ for illustrating how NMF finds a part based representation. The dataset consists of 2429 aligned facial images of size 19×19 pixels. Each image is normalized to have 0.25 mean and standard deviation, then clipped to $[0, 1]$. The images are then vectorized and stacked into a data matrix $\mathbf{X}^{361 \times 2429}$.

In Fig. 2, a parts based representation comparable to [2] is found by all models, except for PH-sNMF and YA-sNMF(10^2). For PH-sNMF the sparsity level clearly affects the identified components. In contrast, implicit sparsity (YA-sNMF) results in an NMF like parts based representation. Increasing penalization prunes components, but setting $\lambda = 10^2$ removes valuable information. While similar solutions are found, the probabilistic NMF finds a model order of $D = 46$ whereas ordinary NMF uses all the available components ($D = 49$).

A probabilistic approach to sparsity automatically identifies the sparsity pattern, as shown in Fig. 2. There is little difference between using unit variance (I) or the infinity norm constraint or switching between the exponential and truncated normal formulation. The main differences arise from which factors sparsity is imposed upon. If sparsity is enforced on the pixel mode, i.e. \mathbf{W} , the extracted component images become more sparse. In contrast, sparsity on \mathbf{H} results in denser component images as the model seek to use as few components as possible in \mathbf{H} thereby providing a sparse reconstruction. The ELBO identifies pNMF as the most likely model, followed by psNMF with sparsity on \mathbf{W} . The least likely model is psNMF with sparsity on both \mathbf{W} and \mathbf{H} . The lack of support for sparsity is unsurprising, as a part based representation is already achieved by NMF and pruning by pNMF.

For $D = 49$, psNMF, psNMF(S,Inf), psNMF(Exp)(S,Inf) and YA-sNMF(10^2) prune the number of components to be 46, 48, 47 and 5, respectively. If $D = 100$ (results not shown), then pNMF(Exp), psNMF(S, I), psNMF(Exp)(S, I) and YA-sNMF(1) also prunes the number of components (to be 88, 62, 83 and 89). This shows the truncated normal formulation identifies a smaller basis (but with denser components) than its exponential counterpart. Neither NMF, PH-sNMF or psNMF with sparsity on \mathbf{H} or both factors prune any components.

¹ CBCL Face Database #1 MIT Center For Biological and Computation Learning <http://www.ai.mit.edu/projects/cbcl>. A copy is available at <http://www.ai.mit.edu/projects/cbcl.old/software-datasets/faces.tar.gz>.

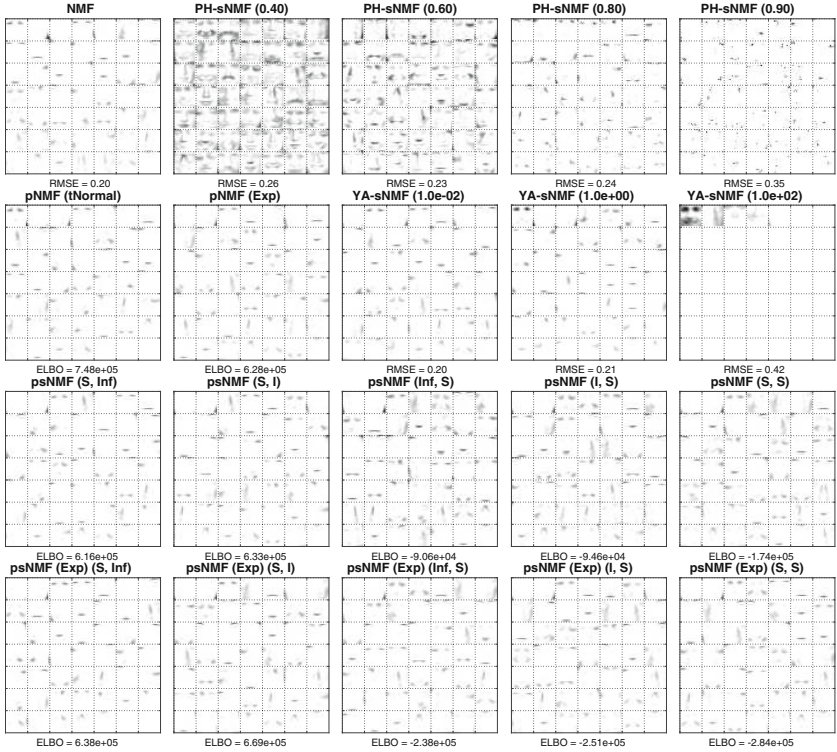


Fig. 2. The CBCL face database.

3.3 Handwritten Digits

The MNIST dataset [20] contains a training set of 60,000 gray-scale images (28×28 pixels) of handwritten digits (0–9). The images are individually vectorized and then stacked into a data matrix $\mathbf{X}^{784 \times 60,000}$. Similar solutions are identified by NMF, PH-sNMF(0.90) and YA-sNMF (with varying sparsity), where small strokes represented the underlying subspace, see Fig. 3. The strokes of PH-sNMF(0.90) and YA-sNMF are more sharply defined than those of NMF due to the enforced sparsity. In contrast to YA-sNMF, the sparsity setting of PH-sNMF greatly affect the estimated subspace.

The probabilistic models find representations similar to the non-probabilistic methods. The methods without sparsity or sparsity only on \mathbf{W} prunes inactive components. Again the truncated normal formulation identifies a smaller basis (but with denser components) than the exponential formulation. Similarly, using Uniform(0, 1) over unit precision/rate identifies a smaller but denser basis. The ELBO points to psNMF(Exp)(S,I) as the most likely model followed by psNMF(Exp). When sparsity is enforced on \mathbf{H} or both factors, the identified subspace is similar to that of NMF and YA-sNMF. No components, even if $D = 100$ (not shown), are pruned using these formulations.

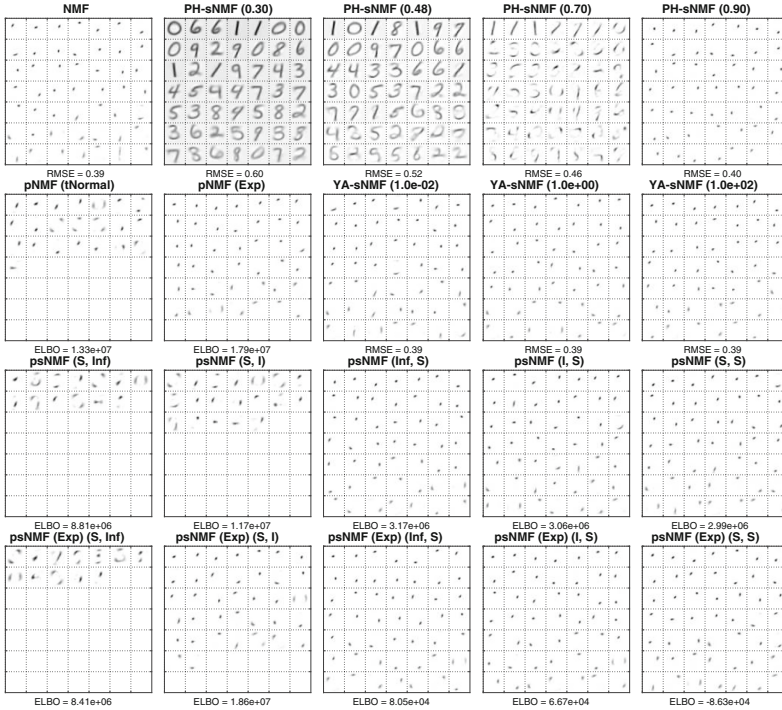


Fig. 3. Handwritten digits (MNIST) dataset.

4 Conclusion

We introduced the probabilistic sparse non-negative matrix factorization (psNMF) model based on variational Bayesian inference which in contrast to conventional sparse NMF is able to automatically tune the level of sparsity. The influence of priors and specification of sparsity on either \mathbf{W} , \mathbf{H} or both was investigated on three datasets and compared to NMF, sparse NMF with explicit (PH-sNMF) or implicit sparsity (YA-sNMF) and the recently proposed probabilistic NMF based on variational inference [14].

For the noisy Swimmer data, the probabilistic methods were able to determine a suitable model order identifying all 16 articulations and potentially the torso as a separate additional component. Facing a limited number of observations compared to image pixels we further observed that the specification of priors heavily influenced the results. For the CBCL face data we found the conventional and probabilistic sparse NMF resulted in similar part based representations regardless of how sparsity was imposed and priors specified, whereas the result of the MNIST data was heavily influenced by how the priors and sparsity was specified. Here imposing sparsity on \mathbf{W} resulted in fewer yet less part based but denser components than imposing sparsity on \mathbf{H} . We further observed differences in the extracted components using the truncated normal

distribution and the exponential prior where the truncated normal distribution in general extracted fewer more dense components. In contrast to existing sparse NMF relying on a user defined a priori specification of sparsity level or tuning of the regularization, we found that the probabilistic sparse NMF admitted automatic tuning of regularization. However, as the extracted components in some cases were heavily influenced by the specification of priors and how sparsity was imposed the ELBO may here provide an important quantitative tool for selecting between the different probabilistic model specifications to determine a suitable representation of data in terms of its constituting parts.

References






1. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**(2), 111–126 (1994)
2. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999)
3. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems*, pp. 556–562 (2001)
4. Bro, R., De Jong, S.: A fast non-negativity-constrained least squares algorithm. *J. Chemom.* **11**(5), 393–401 (1997)
5. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **19**(10), 2756–2779 (2007)
6. Bro, R.: Multi-way analysis in the food industry: models, algorithms, and applications. Ph.D. thesis, Amsterdam: Universiteit van Amsterdam (1998)
7. Cichocki, A., Zdunek, R., Amari, S.: Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) *ICA 2007*. LNCS, vol. 4666, pp. 169–176. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74494-8_22
8. Donoho, D., Stodden, V.: When does non-negative matrix factorization give a correct decomposition into parts? In: *Advances in Neural Information Processing Systems*, pp. 1141–1148 (2004)
9. Laurberg, H., Christensen, M.G., Plumbley, M.D., Hansen, L.K., Jensen, S.H.: Theorems on positive data: On the uniqueness of NMF. *Comput. Intell. Neurosci.* **2008**, 10 (2008)
10. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **5**, 1457–1469 (2004)
11. Li, Y., Ngom, A.: The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol. Med.* **8**(1), 10 (2013)
12. Schmidt, M.N., Winther, O., Hansen, L.K.: Bayesian non-negative matrix factorization. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) *ICA 2009*. LNCS, vol. 5441, pp. 540–547. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00599-2_68
13. Schmidt, M.N., Mohamed, S.: Probabilistic non-negative tensor factorization using Markov chain Monte Carlo. In: *2009 17th European Signal Processing Conference*, pp. 1918–1922. IEEE (2009)

14. Brouwer, T., Frellsen, J., Lió, P.: Comparative study of inference methods for bayesian nonnegative matrix factorisation. In: Ceci, M., Hollmén, J., Todorovski, L., Vens, C., Dvzeroski, S. (eds.) ECML PKDD 2017. LNCS (LNAI), vol. 10534, pp. 513–529. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71249-9_31
15. Zhong, M., Girolami, M.: Reversible jump mcmc for non-negative matrix factorization. In: International Conference on Artificial Intelligence and Statistics (2009)
16. Schmidt, M.N., Mørup, M.: Infinite non-negative matrix factorization. In: 2010 18th European Signal Processing Conference, pp. 905–909. IEEE (2010)
17. Guan, Y., Dy, J.: Sparse probabilistic principal component analysis. In: Artificial Intelligence and Statistics, pp. 185–192 (2009)
18. Nielsen, S.F.V., Mørup, M.: Non-negative tensor factorization with missing data for the modeling of gene expressions in the human brain. In: 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). IEEE (2014)
19. Bishop, C.M.: Variational principal components (1999)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)

Biomedical Data and Methods



Application of Independent Component Analysis to Tumor Transcriptomes Reveals Specific and Reproducible Immune-Related Signals

Urszula Czerwinska^{1,3}(✉) , Laura Cantini¹ , Ulykbek Kairov² ,
Emmanuel Barillot¹ , and Andrei Zinovyev¹ 

¹ Institut Curie, INSERM U900, PSL Research University,
Mines ParisTech, 26 rue d'Ulm, Paris, France
urszula.czerwinska@curie.fr

² Laboratory of Bioinformatics and Computational Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Astana, Kazakhstan

³ Center for Interdisciplinary Research, Paris Descartes University, Paris, France
<https://sysbio.curie.fr/>

Abstract. Independent Component Analysis (ICA) can be used to model gene expression data as an action of a set of statistically independent hidden factors. The ICA analysis with a downstream component analysis was successfully applied to transcriptomic data previously in order to decompose bulk transcriptomic data into interpretable hidden factors. Some of these factors reflect the presence of an immune infiltrate in the tumor environment. However, no foremost studies focused on reproducibility of the ICA-based immune-related signal in the tumor transcriptome. In this work, we use ICA to detect immune signals in six independent transcriptomic datasets. We observe several strongly reproducible immune-related signals when ICA is applied in sufficiently high-dimensional space (close to one hundred). Interestingly, we can interpret these signals as cell-type specific signals reflecting a presence of T-cells, B-cells and myeloid cells, which are of high interest in the field of oncoimmunology. Further quantification of these signals in tumoral transcriptomes has a therapeutic potential.

Keywords: Blind source separation · Unsupervised learning
Genomic data analysis · Cancer · Immunology

1 Introduction

In many fields of science (biology, technology, sociology) observations on a studied system represent complex mixtures of signals of various origins. It is known that tumors are engulfed in a complex microenvironment (TME) that critically impacts progression and response to therapy. In the light of recent findings [1],

many cancer biologists believe that the state of tumor microenvironment (in particular, the composition of immune system-related cells) defines the long-term effect of the cancer treatment.

In biological systems information is coded in a form of DNA that do not vary a lot between different individuals of the same species. In order to trigger a function in an organism, a part of the DNA is transcribed to RNA, depending on the intrinsic and extrinsic factors, and after additional modification messenger RNA (mRNA) is translated into a protein (i.e. digestive enzyme) that fulfill a role in the organism. The mRNA information (also called transcriptome) can be captured with experimental methods at high throughput (transcriptomics) and provides an approximation of the state of the studied system (i.e. a tissue).

Given the way transcriptomic data is collected, in the resulting dataset, for each observation or sample, the measured transcripts' expression (a putative gene expression that is transcribed to mRNA, and before it is translated to a protein) level is affected by a mixture of signals coming from various sources. Thus, we adopt a hypothesis that a transcriptome is a mixture of different signals (that can be biological or technical), including cell-type specific signals.

Recent works [2–4] showed that expression data from complex tissues (such as tumor microenvironment) can be used to estimate the cell-specific expression profiles of the main cellular components present in a tumor sample. This methodology is based on a linear model of a mixture of signals and their interaction and termed cell-type deconvolution. The mentioned methods take advantage of the prior knowledge (and, at the same time, heavily depend) on the specific transcriptomic signatures (characteristic genes and their weights) of cell types composing TME; therefore, they fall into supervised learning category.

A methodology using an unsupervised data decomposition was applied, so far, in the context of tumor clonality deconvolution by Roman et al. [5]. Some attempts were made to apply Non-negative Matrix factorization to transcriptomic data as well. However, they were either applied in very simplified context of *in vitro* cell mixtures [6] or without a specific focus on the immune signals [7].

In our work, we propose to apply an unsupervised method that will decompose mixture into hidden sources, which will be as independent as possible, based uniquely on data structure and without any prior knowledge. For this purpose, we apply Independent Component Analysis (ICA) [8] that solves blind source separation problem. ICA defines a new coordinate system in the multi-dimensional space such that the distributions of the data point projections on the new axes become as mutually independent as possible. To achieve this, the standard approach is maximizing the non-gaussianity of the data point projection distributions.

As a result of ICA, conventionally, data matrix X can be approximated: $X \approx AS$, where X is a matrix of data of size $m \times n$, A is a $m \times k$ matrix, $k < m$ and S is $k \times n$ matrix [9]. In our pipeline, input data matrix $n \times m$ (n genes/probes in rows and m samples in columns) is first transposed before applying ICA to $m \times n$. Thus columns of A ($m \times k$) can be named components (m -dimensional vectors) of mixing proportions for each sample m . The S matrix

$(k \times n)$ is transposed to $n \times k$ where rows are projections of data vectors onto the components (a k -dimensional vector for each of n data points).

ICA has been applied for the analysis of transcriptomic data for blind separation of biological, environmental and technical factors affecting gene expression [9–13].

The interpretation of the results of any matrix factorization-based method applied to transcriptomics data is done by the analysis of the resulting pairs of metagenes and metasamples, associated to each component and represented by sets of weights for all genes and all samples, respectively [7,9]. Standard statistical tests applied to these vectors can then relate a component to a reference gene set (e.g., cell cycle genes), or to clinical annotations accompanying the transcriptomic study (e.g., tumor grade). The application of ICA to multiple expression datasets has been shown to uncover insightful knowledge about cancer biology [11,14]. In [11] a large multi-cancer ICA-based metaanalysis of transcriptomic data defined a set of metagenes associated with factors that are universal for many cancer types. Metagenes associated with cell cycle, inflammation, mitochondria function, GC-content, gender, basal-like cancer types reflected the intrinsic cancer cell properties.

In our previous work, we introduced a ranking of independent components based on their stability in multiple independent components computation runs and define a distinguished number of components (Most Stable Transcriptome Dimension, MSTD) corresponding to the point of the qualitative change of the stability profile [15].

However, an interesting observation can be made employing a number of components going far beyond the MSTD ($M \gg \text{MSTD}$), that we call here *overdecomposition*. Applying this approach, one can discover more specific components that remain reproducible between independent datasets. In this work, we present results of overdecomposition with focus on the fine decomposition of the immune signal into cell-type specific signals.

In this analysis, we used a set of six independent breast cancer transcriptomic datasets (BRCA TCGA [16], METABRIC [17], BRCACIT [18], BRCA BEK [19], BRCA WAN [20] and BRCA BCR [21]) to evaluate a detectability and a reproducibility of the immune cell-type related signal. Each dataset contains gene expression measured in breast tumor biopsy for a number of patients. Therefore each measured gene expression here can be a mix of expression from different cells: tumor cells, stroma cells (fibroblasts), immune cells or normal connective tissue.

Throughout this publication we employ terms: *stability*, *conservation* and *reproducibility* that we define as follows. Stability of an independent component, in terms of varying the initial starts of the ICA algorithm, is a measure of internal compactness of a cluster of matched independent components produced in multiple ICA runs for the same dataset and with the same parameter set but with random initialization. Conservation of an independent component in terms of choosing various orders of the ICA decomposition is a correlation between matched components computed in two ICA decompositions of different orders (reduced data dimensions) for the same dataset. Reproducibility of an independent

component is an (average) correlation between the components that can be matched after applying the ICA method using the same parameter set but for different datasets. We claim that if a component is reproduced between the datasets of the same cancer type, then it can be considered a reliable signal less affected by technical dataset peculiarities. If the component is reproduced in datasets from many cancer types, then it can be assumed to represent a universal cancerogenesis mechanism, such as cell cycle or infiltration by immune cells.

2 Methods

2.1 ICA Overdecomposition Procedure

Our pipeline can be described as follows. Started with six public transcriptomic data of breast cancer, we apply the fastICA algorithm [8] accompanied by the icasso package [22] to improve the components estimation and to rank the components based on their stability. In order to run the analysis we used open source BIODICA tool (ICA applied to BIOlogical Data), available from <https://github.com/LabBandSB/BIODICA>. It provides both a command line and a user-friendly Graphical User Interface (GUI) for high-performance ICA analysis, including bootstrapping and further stability analysis. It also allows the computation of MSTD index, introduced in [15]. BIODICA software links to downstream analysis enabling the interpretation of components, such as standard statistical methods, i.e. enrichment test, and non-standard methods, such as using projection on top of molecular maps (InfoSigMap, [23]). The downstream analysis was not exhaustively employed in this publication as we focused on specific immune signals.

ICA was applied to each transcriptomic dataset separately. For each analyzed transcriptomic dataset, we computed M independent components (ICs), using *pow3* nonlinearity and symmetrical approach to the decomposition. The number of dimensions was set to 100 ($M = 100$) as it is significantly greater than MSTD for these datasets (that is in the order of $M = 30$). Each component of the resulting S matrix was oriented in the direction of its heavy tail, being defined as the tail with the maximum sum of absolute weight values, so that it always has the positive sign.

2.2 Interpretation of Components

In order to confirm that we can recover expected known signals performing the overdecomposition procedure, we correlate reference metagenes with the S matrix. Correlations are performed on common genes for each component and metagene. The result was graphically represented using R package *ggplot2* [24]. An interpretation is assigned to a component only if its assignment is reciprocal. In our analysis reciprocity is defined as follows. Given correlations between the set of metagenes $M = \{M_1, \dots, M_m\}$ and S matrix $S = \{IC_1, \dots, IC_N\}$, if $S_i = \operatorname{argmax}_k(\operatorname{corr}(M_j, S_k))$ and $M_j = \operatorname{argmax}_k(\operatorname{corr}(S_i, M_k))$, then S_i

and M_j are reciprocal. In this way, the breast cancer metagenes were matched against the following set of previously defined metagenes [11] - reference metagenes: MYOFIBROBLASTS, BLCAPATHWAYS, STRESS, GC CONTENT, SMOOTH MUSCLE, MITOCHONDRIAL TRANSLATION, INTERFERON, BASALLIKE, CELLCYCLE, UROTHERIALDIFF. Details about construction of reference metagenes and their interpretation can be found in Biton et al. 2014 [11]. The correlation plot was visualized in Cytoscape 2.8 [25].

2.3 Selecting Immune-Related Components

In order to preselect immune-related signals, we focused on all Independent Components (ICs) with Pearson correlation > 0.1 between IMMUNE metagene and ICs (columns of the S matrix). The interpretation was given using Fisher exact test on 100 top-ranked genes of each of the preselected components and Immgen [26] signatures containing in total 6467 genes of six immune cell types: $\alpha\beta$ T-cells, $\gamma\delta$ T-cells, B-cells, CD+, Myeloid cells, NK cells and four non-immune cell types: Fetal-Liver, Stem cells, Stromal cells and Pasmocytoid, 241241 signatures in total, each of 480 genes in average.

2.4 Comparing Independent Components from Different Datasets

Following the methodology developed previously in [11], the metagenes computed in two independent datasets were compared by computing a Pearson correlation coefficient between their corresponding gene weights. Since each dataset can contain a different set of genes, the correlation is computed on the genes which are common for a pair of datasets. Note that this common set of genes can be different for different pairs of datasets. The same correlation-based comparison was done with previously defined and annotated metagenes. In all correlation-based comparisons, the absolute value of the correlation coefficient was used.

3 Results

3.1 Most of Known Metagenes Can Be Found in Overdecomposed Datasets

In all six overdecomposed datasets of breast cancer, we could find major reference metagenes [11]. As an example, we present results for METABRIC dataset [17] (Fig. 1) where we can observe correlations between metagenes and all 100 ICs. For some metagenes (MYOFIBROBLASTS, INTERFERON, MITOCHONDRIAL TRANSLATION, CELL CYCLE), there is only one reciprocal and strongly (>0.3) correlated component, which can be understood as a good signal reproducibility. Some other as STRESS, BASALLIKE and SMOOTH MUSCLE can have two similarly correlated components. This is probably due to component split in higher-order decomposition. Importantly, reference metagenes were

defined in significantly lower dimensional space ($M = 25$) and as a result of high-dimensional decomposition, these signals are decomposed to more specific sources that can still be interpreted in biological terms. For few components, no strong correlations with metagenes were found (UROTHELIALDIFFERENTIATION and BLCPATHWAYS). As these metagenes are more specific to Bladder cancer, we can consider them as negative control here. Also, GC Content and IMMUNE metagenes have several corresponding components. The IMMUNE metagene is considered here as a special case as we can find several components correlated to it and, in addition, their interpretation can be interesting for biological applications. We investigate more about the immune-related components in the Subsect. 3.3.

3.2 Reproducibility of the Signals in Breast Cancer Datasets

It would be reasonable to expect that the main biological signals are characteristic for a given cancer type. Thus, they should be the same when one studies molecular profiles of different independent cohorts of patients. For this reason, we expect that for multiple datasets related to the same cancer type, the ICA decompositions should be somewhat similar; hence, reciprocally matching each other.

We correlated the ICA overdecompositions of all six datasets with each other and with the forementioned metagenes [11]. One can notice from the correlation graph (Fig. 2A), that some pseudo-cliques characterized with strong correlation coefficient (thick edges) and reciprocal (green) edges are present in the mass of low correlation coefficients edges. If the edges with correlation coefficient < 0.4 are filtered out, we can better visualize a collection of pseudo-cliques (Fig. 2B). Some of those pseudo-cliques are connected to a metagene and can be given an interpretation directly, some others would need a further investigation of the gene signature in order to attribute a meaning to them. We can see that in some pseudo-cliques not all datasets are represented. It may suggest that some signals, still reproducible, are not representative for all datasets. In order to explain, why a signal is missing, one should first interpret the signal, then try to understand the similarities or differences of samples based on provided metadata. From our previous analysis [11], the components that do not find reciprocity (absent from the pseudo-cliques) are either dataset specific or they correspond to unknown batch effects that cannot be guessed without an additional knowledge. It is remarkable that despite overdecomposition, the metagenes conceived in lower-dimensional space are highly conserved and reproducible, which suggests the overdecomposition does not diminish strong signals conceived in “optimal” dimensional space (i.e. MSTD). Of note, these datasets were produced using various technologies of transcriptomic profiling.

3.3 Three Pseudo-cliques Related to Three Immune Cell Types

To better understand the reproductibility of the immune-related signal, we extracted only components correlated with IMMUNE > 0.1 . Hence, we obtain

three strongly connected cliques (Fig. 3) and some disconnected components. We interpreted each of the ICs with an enrichment test. The results of Fisher exact test indicate mainly three cell types T-cell, B-cell and Myeloid cells with a p-value < 0.05 as indicated in the Fig. 3. While T-cell and Myeloid cell are indicated with very high certainty, the B-cell signal seems to be more complex. The results of the enrichment test for the B-cell component are less explicit as among the most enriched pathways, different cell types (T-cells and Natural Killers) are listed together with dominating B-cell signal. However, this can be explained by functional and phenotypic similarities between NK and B cells [27]. Also, T cell and B cell as they are both lymphocytes, they share common features. It is worth highlighting that definition of cell type signature is a part of ongoing debate [28] and here we use them as an indicator of possible signal definitions. Also, some ICs belonging to one pseudo-clique are correlated (with lower coefficients) with ICs from another pseudo-clique (i.e. BRCABCR IC2). It may suggest an inclination of the signal towards the other phenotype. As far as components not included in pseudo-cliques are concerned, through interpretation BRCACIT IC42 can be associated with B cells, METABRIC IC28 with Myeloid cells, BRCAWAN IC68 and BRCABEK IC27 with T-cells. Thus, the correlations of the disconnected components, even though they are low, they are most probably not spurious. Some other components not included in the pseudo-cliques like BRCAWAN IC28 and BRCABCR IC19 seem to contain stroma elements. It would be worth understanding more deeply the nature of each signal and interpret in terms of biological functions or sub-phenotypes.

4 Discussion

The overdecomposition of six breast cancer datasets, where different normalization methods and different transcriptome profiling platforms were used, showed that even in high order blind source separation, the ICA-based analysis can be reproducible between datasets. Moreover, the most stable signals are conserved and not affected by the number of dimensions. Interestingly, for some signals we can observe a split into more specific signals that can still be interpreted in biological terms. In the case of the immune-related signals, it allows robust reproduction of three main signals that form pseudo-cliques on the correlations graph in the Fig. 3. This result let us believe that ICA allows separation of signals in cancer transcriptomes in an unsupervised manner and detect the most represented immune cell-types. We found highly interesting that technically non-stable signal is found reproducible and interpretable in the six breast cancer datasets.

The question about the choice of ICA over other available blind source separation methods can be asked. We address this question more extensively in a publication in preparation comparing NMF, ICA and PCA for transcriptome BSS. From our expertise (unpublished data) NMF applied to transcriptomes can effectively separate sources and their proportions (proven in controlled mixtures of different cell types or tissues). However, when NMF was applied to noisy tumor

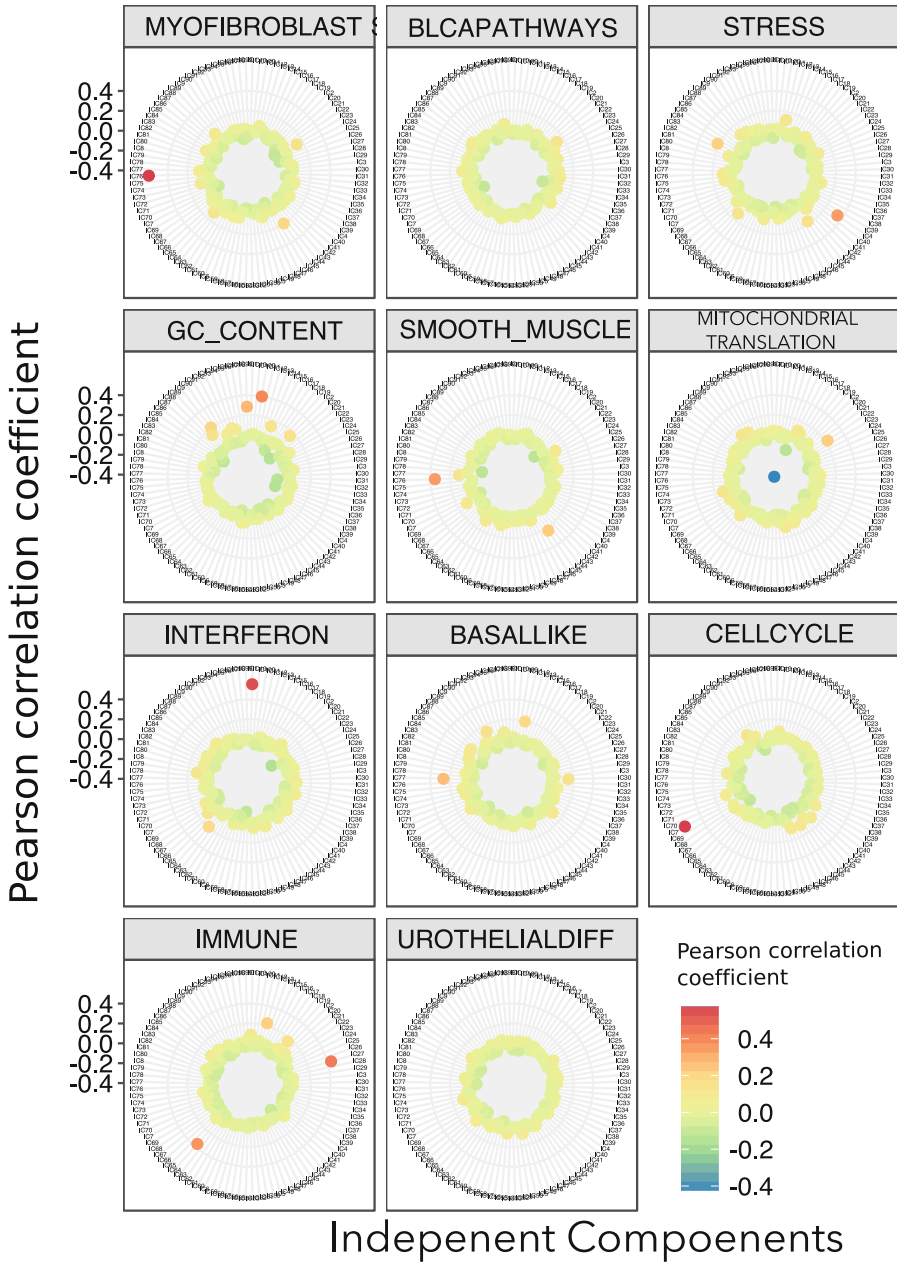


Fig. 1. Correlations between 11 metagenes [11] and 100 independent components of METABRIC dataset [17]. Each panel shows correlation coefficients between a given metagene and 100 ICs of METABRIC, the components are ordered in the same manner for all panels from 1 to 100 in a circle. For a high correlation coefficient, the point is red, for low, it is blue (see legend). (Color figure online)

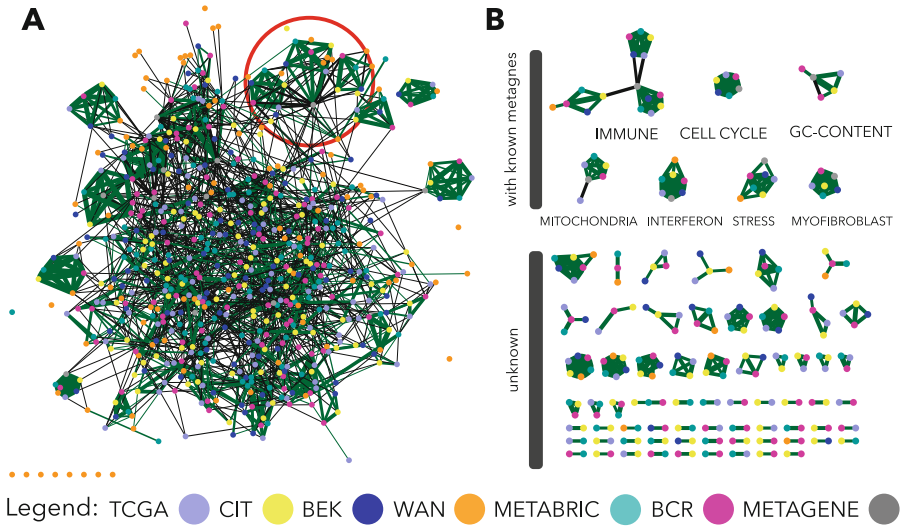


Fig. 2. Correlation plot of six tumor datasets and the reference metagenes [11] A- Correlation graph between decompositions into 100 ICs of the six transcriptomic datasets and the 11 reference metagenes. The IMMUNE metagene and related ICs in encircled; B - collection of pseudo-cliques extracted from the correlation graph A through filtering out edges of the Pearson correlation coefficient < 0.4 . They were split in two groups, the ones that are directly interpretable via their correlation with a metagene and cliques that are not related to any known metagene; The thickness of edges is proportional to the Pearson correlation coefficients, green color indicates reciprocity of edges, colors of nodes indicate dataset (see legend). (Color figure online)

transcriptomes, obtained source profiles were not highly reproducible between different datasets. Our unpublished research showed that NMF profiles are highly affected by mean gene expression. Therefore, NMF decomposition applied to breast cancer transcriptomes followed by correlation of obtained profiles did not reveal meaningful pseudo-cliques as the ICA-based analysis discussed in this article.

In order to translate our findings into real biomedical application, more time should be dedicated to analyze ICA signatures in details, to report their similarities and differences. As well as, this analysis could be applied in a pan-cancer manner to observe the reproducibility of the signal among different tumor types. Such an analysis would possibly identify components and/or genes linked with patients' survival or response to treatment and eventually, use them to compose a predictive score for tumor immune therapy outcome.

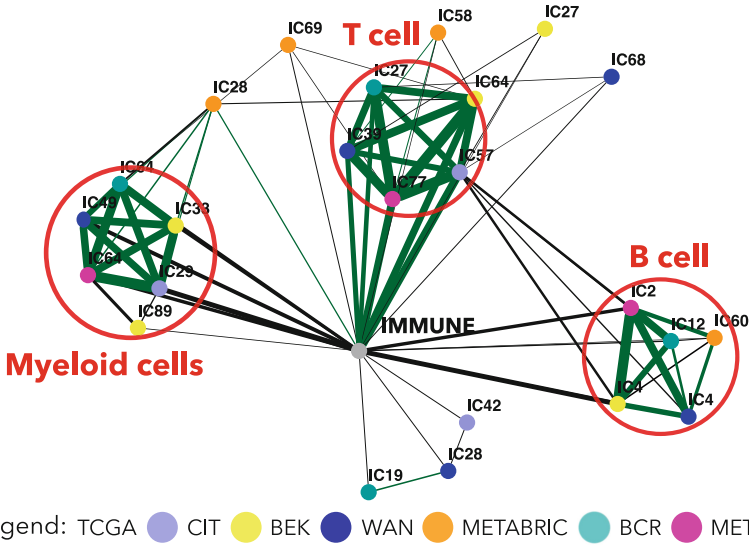


Fig. 3. Correlation graph of ICs correlated with IMMUNE metagene > 0.1 . Three pseudo-cliques are encircled and labeled according to the results of Fisher exact test. The thickness of edges is proportional to the Pearson correlation coefficients, green color indicates reciprocal edges, colors of nodes indicate dataset (see legend). (Color figure online)

5 Conclusions

We applied overcomposition into one hundred components of six transcriptomic datasets using Independent Components Analysis, a blind source separation algorithm. We used a known collection of ranked ICA-derived genetic signatures (that we call reference metagenes) to conclude that most of the signals are conserved in the higher dimensions. We noticed that some of the components split into more specific signals. Our correlation analysis of the ICA overdecompositions of the transcriptomes stated that majority of components are reproducible between datasets. Our more focused investigation of immune-related ICs demonstrated that three cell types can be named: T-cell, B-cell and myeloid cells as a reproducible source signal in the breast cancer datasets. Further interpretation of those cell-type related genomic signatures can find application in immunology therapeutics as predictive biomarkers for immunotherapies.

Acknowledgments. We thank Vassili Soumelis for discussions on multidimensionality of biological systems. This work has been funded by INSERM Plan Cancer N BIO2014-08 COMET grant under ITMO Cancer BioSys program and by ITMO Cancer (AVIESAN) who provided 3-year PhD grant. We would like to acknowledge as well foundation Bettencourt Schueller and Center for Interdisciplinary Research funding for the training of the PhD student.

References

1. Swartz, M.A., Iida, N., Roberts, E.W., Sangaletti, S., Wong, M.H., Yull, F.E., Coussens, L.M., DeClerck, Y.A.: Tumor microenvironment complexity: emerging roles in cancer therapy (2012)
2. Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., et al.: Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**(1), 218 (2016)
3. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., Alizadeh, A.A.: Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**(5), 453–457 (2015)
4. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E., Gfeller, D.: Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017)
5. Roman, T., Xie, L., Schwartz, R.: Automated deconvolution of structured mixtures from heterogeneous tumor genomic data. *PLoS Comput. Biol.* **13**(10), e1005815 (2017)
6. Gaujoux, R., Seoighe, C.: Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.* **12**(5), 913–921 (2012)
7. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**(12), 4164–4169 (2004)
8. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Netw.* **13**(45), 411–430 (2000)
9. Zinovyev, A., Kairov, U., Karpenyuk, T., Ramanculov, E.: Blind source separation methods for deconvolution of complex signals in cancer biology. *Biochem. Biophys. Res. Commun.* **430**(3), 1182–1187 (2013)
10. 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 Comput. Biol.* **3**(8), 1539–1554 (2007)
11. Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouisson, S., DeReyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., Radvanyi, F.: Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* **9**(4), 1235–1245 (2014)
12. Gorban, A., Kegl, B., Wunch, D., Zinovyev, A.: *Principal Manifolds for Data Visualisation and Dimension Reduction*. Lecture notes in Computational Science and Engineering, vol. 58, p. 340. Springer, Heidelberg (2008)
13. Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J.C., Charnock-Jones, D.S., Print, C.G., Smith, S.K.: Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* **23**(39), 6677–6683 (2004)
14. Bang-Berthelsen, C.H., Pedersen, L., Fløyel, T., Hagedorn, P.H., Gylvin, T., Pociot, F.: Independent component and pathway-based analysis of miRNA-regulated gene expression in a model of type 1 diabetes. *BMC Genomics* **12**, 97 (2011)
15. Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., Zinovyev, A.: Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics* **18**(1), 712 (2017)

16. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al.: The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**(10), 1113 (2013)
17. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Aparicio, S., Brenton, J.D., Ellis, I., Huntsman, D., Pinder, S., Murphy, L., Bardwell, H., Ding, Z., Jones, L., Liu, B., Papatheodorou, I., Sammut, S.J., Wishart, G., Chia, S., Gelmon, K., Speers, C., Watson, P., Blamey, R., Green, A., MacMillan, D., Rakha, E., Gillett, C., Grigoriadis, A., De Rinaldis, E., Tutt, A., Parisien, M., Troup, S., Chan, D., Fielding, C., Maia, A.T., McGuire, S., Osborne, M., Sayalero, S.M., Spiteri, I., Hadfield, J., Bell, L., Chow, K., Gale, N., Kovalik, M., Ng, Y., Prentice, L., Tavaré, S., Markowitz, F., Langerød, A., Provenzano, E., Purushotham, A., Børresen-Dale, A.L., Caldas, C.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403), 346–352 (2012)
18. Guedj, M., Marisa, L., De Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A.L., Feugeas, J.P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., De Thé, H., Theillet, C.: A refined molecular taxonomy of breast cancer. *Oncogene* **31**(9), 1196–1206 (2012)
19. Bekhouche, I., Finetti, P., Adelaïde, J., Ferrari, A., Tarpin, C., Charafe-Jauffret, E., Charpin, C., Houvenaeghel, G., Jacquemier, J., Bidaut, G., Birnbaum, D., Viens, P., Chaffanet, M., Bertucci, F.: High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS ONE* **6**(2), e16950 (2011)
20. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-Van Gelder, M.E., Yu, J., Jatkoa, T., Berns, E.M., Atkins, D., Foekens, J.A.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**(9460), 671–679 (2005)
21. Reyat, F., Rouzier, R., Depont-Hazelzet, B., Bollet, M.A., Pierga, J.Y., Alran, S., Salmon, R.J., Fourchotte, V., Vincent-Salomon, A., Sastre-Garau, X., Antoine, M., Uzan, S., Sigal-Zafrani, B., de Rycke, Y.: The molecular subtype classification is a determinant of sentinel node positivity in early breast carcinoma. *PLoS ONE* **6**(5), e20297 (2011)
22. Himberg, J., Hyvärinen, A.: ICASSO: software for investigating the reliability of ICA estimates by clustering and visualization. In: *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, vol. 2003, pp. 259–268, January 2003
23. Cantini, L., Calzone, L., Martignetti, L., Rydenfelt, M., Blüthgen, N., Barillot, E., Zinoviyev, A.: Classification of gene signatures for their information value and functional redundancy. *npj Syst. Biol. Appl.* **4**(1), 2 (2018)
24. Wickham, H.: *ggplot2 Elegant Graphics for Data Analysis*, vol. 35 (2009)
25. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003)
26. Shay, T., Kang, J.: *Immunological Genome Project and systems immunology* (2013)

27. Kerdiles, Y.M., Almeida, F.F., Thompson, T., Chopin, M., Vienne, M., Bruhns, P., Huntington, N.D., Raulet, D.H., Nutt, S.L., Belz, G.T., Vivier, E.: Natural-Killer-like B cells display the phenotypic and functional characteristics of conventional B cells. *Immunity* **47**(2), 199–200 (2017)
28. Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B., Raue, A.: Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Commun.* **8**(1), 2032 (2017)



Probit Latent Variables Estimation for a Gaussian Process Classifier: Application to the Detection of High-Voltage Spindles

Rémi Souriau¹, Vincent Vigneron^{1(✉)}, Jean Lerbet¹, and Hsin Chen²

¹ IBISC EA 4526, Univ. Evry, Université Paris-Saclay, Paris, France
{remi.souriau,vincent.vigneron,jean.lerbet}@ibisc.univ-evry.fr

² Electrical Engineering Department,
National Tsing Hua University, Hsinchu, Taiwan
hchen@ee.nthu.edu.tw

Abstract. The Deep Brain Stimulation (DBS) is a surgical procedure efficient to relieve symptoms of some neurodegenerative disease like the Parkinson's disease (PD). However, apply permanently the deep brain stimulation due to the lack of possible control lead to several side effects. Recent studies shown the detection of High-Voltage Spindles (HVS) in local field potentials is an interesting way to predict the arrival of symptoms in PD people. The complexity of signals and the short time lag between the apparition of HVS and the arrival of symptoms make it necessary to have a fast and robust model to classify the presence of HVS ($Y = 1$) or not ($Y = -1$) and to apply the DBS only when needed. In this paper, we focus on a Gaussian process model. It consists to estimate the latent variable f of the probit model: $\Pr(Y = 1|input) = \Phi(f(input))$ with Φ the distribution function of the standard normal distribution.

Keywords: Deep learning · Gaussian processes · Autoencoder Classification · High-Voltage Spindle · Parkinson diseases

1 Introduction

The Parkinson's disease (PD) is a progressive *neurodegenerative* disease. The depletion of the dopamine in the basal ganglia network leads to several symptoms like rigidity, posture instability, slow motion or pain for example. The expectation of the number of PD victims in Asian countries is 6.17 millions in 2030 [2]. The deep brain stimulation (DBS) is a surgical procedure used to relieve disabling neurological symptoms for diseases like PD [9]. A high-frequency stimulation signal (around 130 Hz) is continuously applied to a deep-brain region called the

R. Souriau—This work was partly supported by the National Tsing Hua University (Hsinchu, Taiwan) and Ministry of Science and Technology, R.O.C. (Taiwan).

subthalamic nucleus (STN) to relieve the symptoms. The main drawback of the DBS is the absence of any control on stimulation to minimize side effects. In addition, contemporary DBS implant requires another surgery to replace battery every 6 or 7 years.

Recent studies show we can predict the arrival of PD symptoms by the detection of high-voltages spindles (HVS) in recorded signals in local field potentials (LFPs) [1]. The HVS signals as *e.g.* in Fig. 1 are synchronous spike-and-wave patterns in LFPs oscillating in the 5–13 Hz frequency band. Suppressing HVS signals is found useful for delaying the progress of PD and deleting symptoms. Being able to detect HVS make possible the realization of a *closed-loop system* to control the DBS. However the diffusion of signals in the brain is *nonlinear* and there is only few milliseconds between the HVS wave and the apparition of PD symptoms. Hence a fast and robust model is needed for real time HVS detection and to apply the high frequency signal only when it is needed.

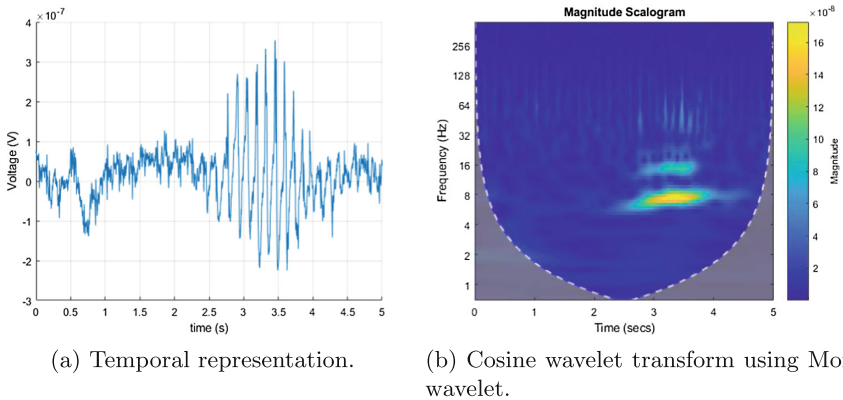


Fig. 1. Signals recorded in LFPs in two different representations. HVS are located between 2.5 and 5 s. HVS are characterized by a fundamental frequency between 5 and 13 Hz.

In this paper, the PD rat model is used. Data are collected from eight intracortical channels from different cortical regions. In this paper, we investigate performance of the Gaussian Process (GP) [3] for the detection of HVS. The GP model is a *Bayesian network* with continuous variables. Bayesian network model relations of *causality* between variables and in our study, data collected are the result of a diffusion of signals between neurons in the brain. Moreover, relations between variables model by a GP are nonlinear. Section 2 presents how data are collected and the preprocessing of data. Details of the model are developed in Sect. 3. The two last sections give main results and discuss some future improvement and other possible approaches.

2 Data Collection

2.1 Data Acquisition and Data Preparation

The PD rat model has been used to develop and evaluate the results. The description of the procedure to extract data is given in [8, sect. 2].

Table 1. List of brain region where LFPs signals were recorded.

NOTATION	REGION NAME
M1D	Layer 5b of the primary motor cortex
M1U	Layer 2/3 of the primary motor cortex
M2D	Layer 5b of the secondary motor cortex
M2U	Layer 2/3 of the secondary motor cortex
SD	Layer 5b of the primary somatosensory cortex
SU	Layer 2/3 of the primary somatosensory cortex
STRI	Dorsal region of striatum
THAL	Ventrolateral thalamus

The LFPs were recorded from eight different brain regions listed in Table 1. The frequency sampling of signals was 1 kHz and the recording duration of one session was 60 s (60,000 samples). Several sessions have been recorded on PD rats.

GP classifier is a model which requires a *supervised* learning.

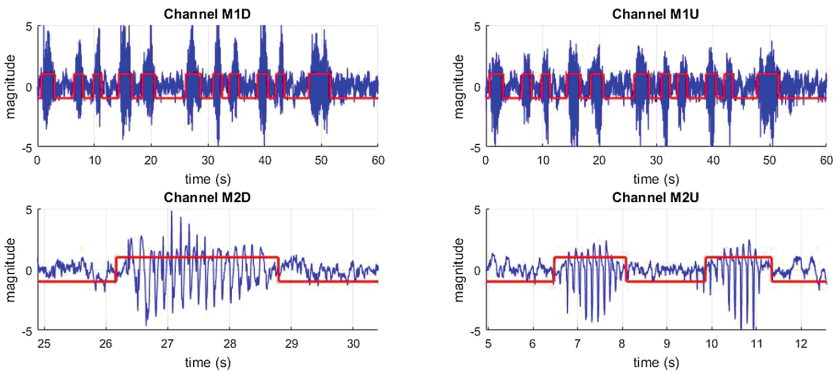


Fig. 2. PSD mean as a function of time for channels M1D, M2D, M1U and M2U. The red line represent the ground truth: if $\frac{3}{4}$ of signal magnitude is above the threshold, then we consider we detect the presence of the HVS. (Color figure online)

The data preparation step consists to construct *feature* signals from collected data to *train* the model and feature signals for *testing* the model.

The data preparation step consists to construct *feature* signals from collected data to *train* the model and feature signals for *testing* the model.

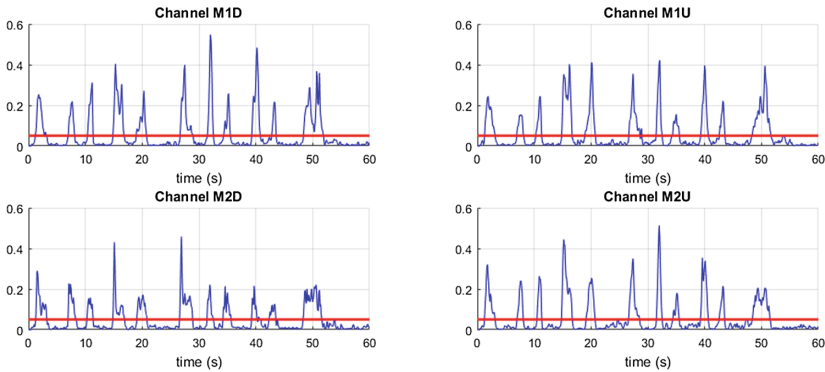


Fig. 3. Result of the classification on the four same channels. Channels MD2 and MU2 are zoomed to observe with more precision the switch of state HVS/no HVS and reverse

One 60s session has been used for the training step and the other session has been used for the testing step. The preprocessing step is the same for the two data sets. HVS wave's fundamental frequency is between 5 and 13 Hz and, according to observations in Fig. 1b, harmonics of HVSs are visible around 30 Hz. High frequencies noises and continuous components of signals were deleted with a second order Butterworth filter between 1 and 200 Hz while preserving much of the HVS frequency content. Each channel was normalized independently from each other (zero mean and unit variance).

Then we define the prediction class vector $Y_n \forall n$ for the training set and the testing set. The presence of HVS is characterized by the apparition of *spike-and-wave* patterns with a fundamental frequency between 5 and 13 Hz. We estimate the Power Spectral Density (PSD) with a periodogram using the Hanning window with a 500 ms time window each 100 ms. Then by computing the PSD mean between 5 and 13 Hz and performing an interpolation, we plot the PSD mean between 5–13 Hz as a function of time (see Fig. 2). We then defined our *ground truth* by thresholding our observations above at least one quarter of the signal magnitude.

Result are given in Figs. 2 and 3.

3 Model

3.1 Gaussian Process Classifier

A closed-loop DBS system delivers the stimulation only when needed.

Mathematically this consists to build a two-class classifier \mathcal{C} capable to identify the presence or the absence of HVS. The available information for the classification are the values of the R channels and p previous signal values for each channel. Let $X_n \in \mathbb{R}^{R \times (p+1)}$ denotes the concatenated feature vector recorded between times n and $n - p$ and $Y_n \in \{-1, 1\}$ be the associated output of the *supervised classifier*. Suppose also the database is shared in a training set for which Y_n is known and a test set for which Y_n is *unknown*. The aim of the model is to estimate Y_n from new observations X_n .

We note in the following $\mathcal{D} = \{X_n, Y_n\}_{n \in [1, N]}$ the training set with $X = \{X_n\}_{n \in [1, N]}$ being independent randomly selected input observations and $Y = \{Y_n\}_{n \in [1, N]}$ the associated output decision respectively. The GP classifier focus on modeling the *posterior* probabilities by defining the *latent* variables $f_n = f(X_n)$.

The model used here is the *probit* model: $\Pr(Y = 1|X) = \Phi(f(X))$ where Φ denotes the cumulative density function of the standard normal distribution. The likelihood of the probit model with independent observations and given $f = \{f(X(n))\}_{n \in [1, N]}$ is:

$$p(Y|f) = \prod_{n=1}^N p(Y_n|f_n) = \prod_{n=1}^N \Phi(Y_n f_n). \tag{1}$$

In a GP, f is a stochastic process which associates a zero mean normal random value for an input $X(n)$. For the training set \mathcal{D} we have $p(f|X, \Theta) \sim \mathcal{N}(0, \mathbf{C}_N)$ where Θ is a set of hyper-parameters and \mathbf{C}_N is a covariance matrix modeled with a *squared exponential* and a Gaussian noise [7]:

$$\mathbf{C}_N(X_i, X_j) = \theta_0^2 \exp \left(-\frac{1}{2} \sum_{n=1}^{\dim(X_i)} \frac{(X_i^{(n)} - X_j^{(n)})^2}{\lambda_n^2} \right) + \theta_1^2 \delta_{(X_i, X_j)}. \tag{2}$$

$X_i^{(n)}$ is the n th component of X_i and $\delta_{(\cdot)}$ is the Kronecker delta. The set of hyper-parameters Θ is composed of $\{\theta_1, \theta_2, \{\lambda_n\}_{n \in [1, N]}\}$. Baye's posterior probability rule of the latent variable f with Θ known can be written:

$$p(f|\mathcal{D}, \Theta) = \frac{p(Y|f)p(f|X, \Theta)}{p(\mathcal{D}|\Theta)} = \frac{\mathcal{N}(f|0, \mathbf{C}_N)}{p(\mathcal{D}|\Theta)} \prod_{n=1}^N \Phi(Y_n f_n). \tag{3}$$

With the marginalization of Eq. (3) for a new observation X_{N+1} we obtain:

$$\Pr(f_{N+1}|\mathcal{D}, \Theta, X_{N+1}) = \int \Pr(f_{N+1}|f, X, \Theta, X_{N+1}) \Pr(f|\mathcal{D}, \Theta) df, \tag{4}$$

and the expectation of the Eq. 4 gives:

$$\Pr(Y_{N+1}|\mathcal{D}, \Theta, X_{N+1}) = \int \Pr(Y_{N+1}|f_{N+1}) \Pr(f_{N+1}|\mathcal{D}, \Theta, X_{N+1}) df_{N+1} \tag{5}$$

We model the posterior probability $q(f|\mathcal{D}, \Theta) \sim \mathcal{N}(m, A)$ to compute $\Pr(Y_{N+1} = 1|\mathcal{D}, \Theta, X_{N+1})$. And then, for a new observation $N + 1$, we can show that the posterior probability of f_{N+1} is $q(f_{N+1}|\mathcal{D}, \Theta, X_{N+1}) \sim \mathcal{N}(\mu, \sigma)$ with:

$$\begin{cases} \mu &= k^T \mathbf{C}_N^{-1} m, \\ \sigma^2 &= \kappa - k^T (\mathbf{C}_N^{-1} - \mathbf{C}_N^{-1} A \mathbf{C}_N^{-1}) k. \end{cases} \quad (6)$$

where $k = (\mathbf{C}_N(X_1, X_{N+1}), \dots, \mathbf{C}_N(X_N, X_{N+1}))^T$ is the covariance function vector between each observation of the training set and the new observation X_{N+1} and $\kappa = \mathbf{C}_N(X_{N+1}, X_{N+1}) = \theta_0^2 + \theta_1^2$ is the variance of X_{N+1} .

With the approximation of $\Pr(f|\mathcal{D}, \Theta)$, Eq. (5) becomes:

$$\Pr(Y_{N+1} = 1|D, \Theta, X_{N+1}) = \Phi\left(\frac{\mu}{\sqrt{1 + \sigma^2}}\right) \quad (7)$$

Training a GP consist to find Θ , m and A . We can learn Θ by computing the log-likelihood of the log posterior probability $\log q(Y|X, \Theta)$ Eq. 8 (see [7, Chapter 5]) and his gradient in function of Θ .

$$\log q(Y|X, \Theta) = -\frac{1}{2} f^T \mathbf{C}_N^{-1} f + \log p(Y|f) - \frac{1}{2} \log \det \left(I + W^{\frac{1}{2}} \mathbf{C}_N W^{\frac{1}{2}} \right) \quad (8)$$

With $W = -\Delta_f \log p(Y|f)$ (Hessian) and f such the unnormalized log likelihood $\log p(f|\mathcal{D}, \Theta)$ is maximized:

$$\begin{aligned} \log p(f|\mathcal{D}, \Theta) &= \log p(Y|f) + \log p(f|X) \\ &= \log p(Y|f) - \frac{1}{2} f^T \mathbf{C}_N^{-1} f - \frac{1}{2} \log \det (\mathbf{C}_N) - \frac{N}{2} \log 2\pi. \end{aligned} \quad (9)$$

For a given Θ , we can find $m = \arg \max_f \log p(f|\mathcal{D}, \Theta)$ by using the Newton’s method. Equations 10 and 11 give the gradient and the hessian of $\log p(f|\mathcal{D}, \Theta)$, respectively.

$$\nabla_f \log p(f|\mathcal{D}, \Theta) = \nabla_f \log p(Y|f) - \mathbf{C}_N^{-1} f. \quad (10)$$

$$\Delta_f \log p(f|\mathcal{D}, \Theta) = \Delta_f \log p(Y|f) - \mathbf{C}_N^{-1}. \quad (11)$$

The maximization of $\log p(f|\mathcal{D}, \Theta)$ makes use of the first and second order partial derivation of $\log p(Y|F)$ in function of f_i .

$$\frac{\partial}{\partial f_i} \log p(Y|f) = \frac{Y_i \phi(f_i)}{\Phi(Y_i f_i)}. \quad (12)$$

$$\frac{\partial^2}{\partial f_i^2} \log p(Y|f) = -\frac{\phi(f_i)^2}{\Phi(Y_i f_i)^2} - \frac{Y_i f_i \phi(f_i)}{\Phi(Y_i f_i)}. \quad (13)$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. Learning m allow to compute Eq. 8 and their gradient in function of Θ . We implement a

gradient descent search of the optimum Θ^* that leads to the following iterative algorithm:

$$\Theta^{(k+1)} = \Theta^{(k)} - \alpha_k \nabla_{\Theta} \log q(Y|X, \Theta). \quad (14)$$

But Eq. (14) requires to inverse the $N \times N$ matrix \mathbf{C}_N at each iteration which can be time consuming for a large number of observations. Once m and Θ found, we can compute $A = (\mathbf{C}_N^{-1} + W)^{-1}$.

Finally, the GP classifier is learned by identifying the covariance matrix between observations \mathbf{C}_N as a function of the hyper-parameters Θ , the mean vector m is learning for each iteration of Θ then and the covariance matrix A is deduced.

Once the learning is done, the prediction step consists to compute the covariance vector k between the new observation X_{N+1} and the training set X and then estimates the probability $\Pr(Y_{N+1} = 1|\mathcal{D}, \Theta, X_{N+1})$. If $\Pr(Y_{N+1} = 1|\mathcal{D}, \Theta, X_{N+1}) > 0.5$ then $Y_{N+1} = 1$ and $Y_{N+1} = -1$ else.

3.2 Input *autoencoding*

Learning the model consist in two step: learning the hyper-parameters Θ and learning the parameters of $q(f_{N+1}|\mathcal{D}, \Theta, X_{N+1})$. HVS have a fundamental frequency between 5 and 13 Hz. With a the maximal period of 200 ms. Choosing $p = 199$ to have at least one period of the signal leads to a model with high dimensions: the input size of X_n is then $(p + 1) \times R = 1600$ and Θ has 1602 parameters. To reduce the dimensionality of the input vector X_n (which makes it difficult to use for real time applications) we use an autoencoder (see Fig. 4). This autoencoder consists in a 3 layers neural network that compresses input data onto the hidden layer. We present to the input and the output layers the same input vector X_n . The activation function of the hidden layer is sigmoidal function $s(\cdot)$ that permits nonlinear combination of the inputs:

$$s(x_j) = \frac{1}{1 + \exp(-b_j - \sum_i w_{ij} x_i)} \quad (15)$$

where $(x_1, \dots, x_p)^T$ is the input vector. Learning this autoencoder consists to find biases b_j and weights w_{ij} of the input neurons. The output layer has to be the closest possible to the input layer. The training algorithm is the *scaled conjugate gradient* [5] using the mean square error with L_2 sparsity regularized loss function [6].

Figure 4 gives an example of result for a number of observations $N = 500$ and the size of the autoencoder $H = 10$. The sensitivity and the specificity (see Sect. 4) are, respectively, 82.92% and 99.31%.

4 Experimental Results

Detection of HVS has been applied on different rats with various set of parameters for the learning stage such as the number of observations N or the size of the hidden layer of the autoencoder H .

The smaller the parameters, the lower the number of parameters: learning the model and use it become very fast by reducing the dimensionality. Choosing N small means taking the risk to have not enough or representative observations.

For each rat, one signal session record has been used for the learning step and another session has been used for the testing step. The criteria of performance for models are the sensitivity $Se = TP / (TP + FN)$ and the specificity $Sp = TN / (TN + FP)$, with TP is the number of true positive, TN is the number of true negative, FN is the number of false negative and FP is the number of false positive. The sensitivity gives the true positive rate: the number of correct detection under the number of correct detection and miss. The specificity give the true negative rate: the number of correct non detection under the number of correct non detection and false detection. We reproduce 5 times for each set of parameters the learning and the testing stages and compute the mean and the variance of Se and Sp to verify the performance and the robustness, regarding random sampling.

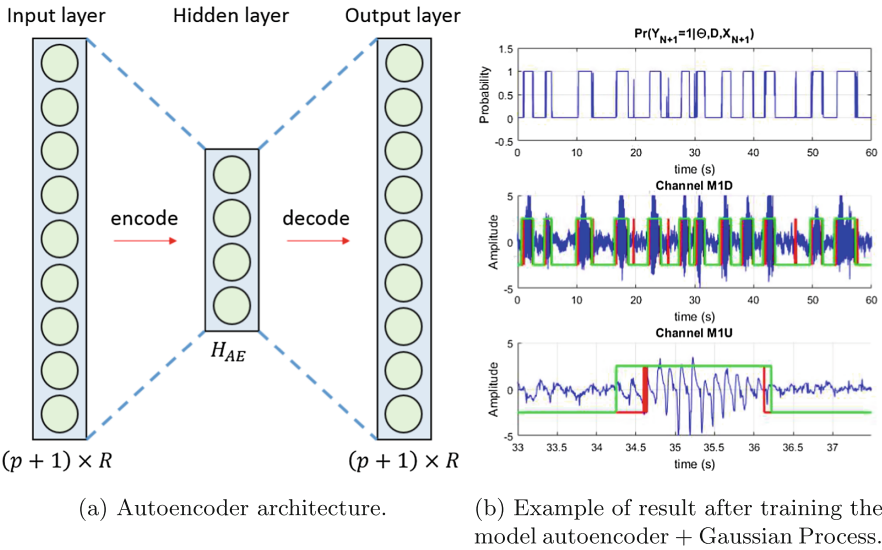


Fig. 4. (4a) autoencoder neural network: input and output layer dimensions equally set to $(p + 1) \times R$. H_{AE} is the size of the hidden layer (parameters). (4b): *upper plot* is $\Pr(Y_{N+1} = 1 | \mathcal{D}, \Theta, X_{N+1})$ as a function of time. Figures below are two among five channels of the testing set. Green line marks the *ground-truth* defined in the preprocessing step. The red line is the decision made by the GP classifier. The lower figure zoomed on a HVS. (Color figure online)

Results are summarized in Table 2. Data collection for each rat is different: the rat 2 provides data from channels M1U, M1D, SU and SD; rat 3 provides data from channels M1U, M1D, STRI and SD; rat 1 provides data from all channels¹.

¹ See Table 1 as a reminder.

Table 2. Result of the experience. N is the number of observations used from the training set. H is the size of the hidden layer of the encoder. Variance equal to .0000 in the table mean the value is less than 10^{-4} . Bold numbers highlight most relevant results.

N	H	RAT NUMBER 1				RAT NUMBER 2				RAT NUMBER 3			
		Sensitivity		Specificity		Sensitivity		Specificity		Sensitivity		Specificity	
		mean	var	mean	var	mean	var	mean	var	mean	var	mean	var
50	10	.58	.0192	.88	.0161	.63	.0370	.47	.0596	.23	.0159	.83	.0082
50	30	.63	.0095	.96	.0031	.69	.0099	.57	.0212	.04	.0039	.97	.0019
50	50	.62	.0058	.97	.0008	.68	.0672	.52	.1624	.12	.0053	.91	.0056
50	100	.62	.0050	.96	.0013	.46	.0384	.78	.0531	.10	.0168	.93	.0047
200	10	.82	.0021	.95	.0016	.61	.0082	.56	.0044	.24	.0195	.86	.0061
200	30	.81	.0016	.90	.0032	.70	.0088	.71	.0087	.25	.0274	.82	.0128
200	50	.81	.0006	.91	.0097	.71	.0016	.73	.0057	.13	.0257	.90	.0173
200	100	.60	.0126	.96	.0097	.61	.0367	.82	.0104	.01	.0001	.99	.0000
500	10	.85	.0010	.98	.0001	.67	.0003	.65	.0036	.14	.0242	.94	.0043
500	30	.82	.0019	.95	.0006	.70	.0044	.67	.0043	.25	.0083	.86	.0028
500	50	.83	.0012	.89	.0005	.70	.0026	.68	.0017	.21	.0220	.85	.0141
500	100	.74	.0226	.94	.0054	.76	.0029	.70	.0005	.85	.0032	.85	.0027

Lavielle and Teysseires [4] design an unsupervised approach for changepoint detection which models data within a segment as multivariate Gaussian with known covariance. But the mean can change from segment to segment. See Fig. as an illustration. Besides the fact the change point algorithm is not *real time* and needs to set the number of points in advance, the changepoint detection can detect HVS with an advance of 62 ± 6.55 samples but of 82 ± 32.3 samples for the Gaussian Process.

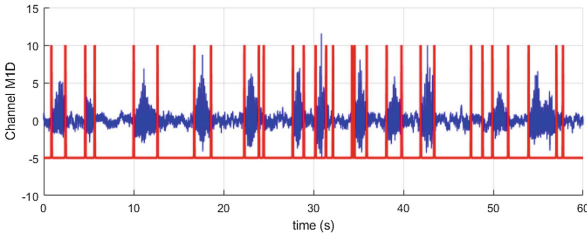


Fig. 5. Unsupervised change point detection: 30 points only.

5 Discussion and Conclusion

Table 2 highlights some tendencies in the parameters. First, the number of observations is critical for fine sensitivity and specificity. Taking too little observation can alter the overall knowledge of the system : in this case the variance is often more important for $N = 50$ than for bigger N . But too much data leads to big with covariance matrices too long to calculate. Increasing H (hidden layer number of neurons) increases the sensitivity but decreases the specificity: the model

tends to detect HVS all the time. On the opposite for small H we compress a lot of data by taking the risk to loose information. A large hidden layer better preserves the information but (i) the problem becomes hard to optimize (too much parameters) (ii) learning the model become time-consuming.

Results for rat 2 and 3 are not as fine as the first rat. By looking closely step by step signals of the two rats it appear that the intensity of the noise is much more important than in signals of the first rat. Means over channels of signal-to-ratio of the three rats are respectively, 35 decibels, 14 decibels and 10 decibels. Moreover, in rat 2 and 3, appearance of signals differs according to the various channels: some HVS do not appear in all channel which make the preprocessing step not relevant for those two rats. This is why results of rat 2 and 3 are not reliable to conclude with a high confidence level about the robustness of the model.

In a future work, we will develop an approach based on *unsupervised learning* model because by defining ourselves the groundtruth we may have missed some complex features in the signal which could have helped for predicting HVS. *Restricted Boltzmann Machines* is a promising stochastic model which, by exploring latent variables could find such hidden features.

References

1. Dejean, C., Gross, C.E., Bioulac, B., Boraud, T.: Dynamic changes in the cortex-basal ganglia network after dopamine depletion in the rat. *J. Neurophysiol.* **100**(1), 385–396 (2008)
2. Dorsey, E.R., Constantinescu, R., Thompson, J.P., Biglan, K.M., Holloway, R.G., Kieburtz, K., Marshall, F.J., Ravina, B.M., Schifitto, G., Siderowf, A., et al.: Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **68**(5), 384–386 (2007)
3. Kuss, M., Rasmussen, C.E.: Assessing approximate inference for binary Gaussian process classification. *J. Mach. Learn. Res.* **6**, 1679–1704 (2005)
4. Lavielle, M., Teyssiere, G.: Detection of multiple change-points in multivariate time series. *Lith. Math. J.* **46**(3), 287–306 (2006)
5. Møller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw.* **6**(4), 525–533 (1993)
6. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vis. Res.* **37**(23), 3311–3325 (1997)
7. Rasmussen, C.E.: Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) *ML -2003. LNCS (LNAI)*, vol. 3176, pp. 63–71. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4
8. Vigneron, V., Syed, T.Q., Chen, H.: Automatic detection of high-voltage spindles for Parkinson’s disease. In: *BIOSIGNALS*, pp. 372–378 (2015)
9. Vitek, J.L.: Mechanisms of deep brain stimulation: excitation or inhibition. *Mov. Disord.* **17**(S3) (2002)



Spatial Filtering of EEG Signals to Identify Periodic Brain Activity Patterns

Dounia Mulders^{1,2(✉)}, Cyril de Bodt¹, Nicolas Lejeune², André Mouraux²,
and Michel Verleysen¹

¹ ICTEAM Institute, Université catholique de Louvain,
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
{dounia.mulders,cyril.debodt,michel.verleysen}@uclouvain.be

² IONS Institute, Université catholique de Louvain,
Avenue Mounier 53, 1200 Woluwe-Saint-Lambert, Belgium
{nicolas.lejeune,andre.mouraux}@uclouvain.be

Abstract. Long-lasting periodic sensory stimulation is increasingly used in neuroscience to study, using electroencephalography (EEG), the cortical processes underlying perception in different modalities. This kind of stimulation can elicit synchronized periodic activity at the stimulation frequency in neuronal populations responding to the stimulus, referred to as a steady-state response (SSR). While the frequency analysis of EEG recordings is particularly well suited to capture this activity, it is limited by the intrinsic noisy nature of EEG signals and the low signal-to-noise ratio (SNR) of some responses. This paper compares and adapts spatial filtering methods for periodicity maximization to enhance the SNR of periodic EEG responses, a key condition to generalize their use as a research or clinical tool. This approach uncovers both temporal dynamics and spatial topographic patterns of SSRs, and is validated using EEG data from 15 healthy subjects exposed to periodic cool and warm stimuli.

Keywords: Periodic Component Analysis · Spatial filtering
Generalized Rayleigh quotient · EEG · Thermal stimulation
Steady-states

1 Introduction

Understanding the neural mechanisms underlying human perception of stimuli from different modalities, such as visual, auditory, tactile or nociceptive, is a challenging issue in neuroscience. In this context, scalp electroencephalography (EEG) is particularly suited to record brain activity, as it is non-invasive and directly measures neuronal activity with a high temporal resolution of the millisecond order. Meanwhile, most studies consider brief sensory stimuli, lasting less than one second and eliciting well-known event-related potentials (ERPs) [10]. However, the recording of neural responses to long-lasting periodic

stimulation is increasingly proposed in several works as an alternative to probe sensory perception [2], since it reveals new aspects of the sensory information processing [8]. Such long-lasting stimuli indeed induce a periodic activity at the stimulation frequency in some neuronal populations. This so-called steady-state response (SSR) can usually be recorded with multichannel EEG.

One of the main advantages of the aforementioned periodic stimuli is that the signal-to-noise ratio (SNR) of the SSR is usually higher in the frequency domain compared to the time domain. Therefore, most neuroscience works up to now use frequency analyses to highlight the SSR features. The EEG time course and frequency transform at some specific electrodes can be studied, as well as the distribution across the scalp of the signal amplitude at the stimulation frequency. Although these approaches provide direct measures of the signal periodicity, their efficiency is limited when the SNR is low. For instance, this is the case with nociceptive SSRs generated from infrared laser stimulation [3]. In the meantime, whereas linear filtering methods have been successfully developed in the context of brain-computer interfaces (BCI) to classify steady-state visually evoked-potentials (SSVEP) [9, 15], they were surprisingly not yet adapted to extract, interpret and characterize the EEG activity related to different periodic stimuli. Indeed, while the optimized filters can lead to high classification accuracies, their spatial patterns may also refine the analysis of SSRs.

In this context, this paper compares filtering methods maximizing the periodicity of the extracted components to study and, more importantly, interpret the cortical processing of periodic stimuli. The filters are constrained to be linear, thereby defining meaningful topographic patterns of the associated components. We propose to adapt four spatial filtering methods to enhance the SNR of SSRs. The two first methods are derived from a measure of periodicity in the time domain, first introduced by Saul and Allen [14]. This approach, called Periodic Component Analysis (π CA), was initially used to extract periodic components from speech signals. Variants have been developed, handling for instance non-strictly periodic signals such as the electrocardiogram [12]. A third method is based on Canonical Correlation Analysis (CCA) between the multichannel EEG signals and a relevant reference periodic signal [9]. Finally, the last method directly optimizes the spectral concentration of the filtered signals at the fundamental stimulation frequency and its harmonics.

This paper is organized as follows. Section 2 defines the compared methods in the context of our application. Section 3 presents empirical results validating the proposed methods on an EEG data set collected on 15 healthy subjects. Finally, Sect. 4 concludes and presents further perspectives.

2 Methods

This section introduces four methods aiming at extracting periodic components by filtering a noisy multidimensional signal $\mathbf{x}(t) \in \mathbb{R}^C$, assumed to have a zero-mean (i.e. $\sum_t \mathbf{x}(t) = 0$). Since the ultimate goal is to interpret the links between these components and the original signals, the spatial filters are constrained

to be linear and will be denoted by vectors $\mathbf{w} \in \mathbb{R}^C$. These filters are optimized to define a maximally periodic component $s(t) := \mathbf{w}^T \mathbf{x}(t)$ of fundamental frequency f_1 and corresponding fundamental period $T_1 = 1/f_1$. Once an optimal filter is found by optimizing a cost function F , a second optimal filter can be found, leading to a component in the orthogonal subspace of the first one. A matrix $W \in \mathbb{R}^{C \times d}$ is hence recursively built, whose columns are the filters ranked in decreasing order of periodicity of the filtered components as measured by F , which determines $d \leq C$. Pseudo-inverting the matrix $W^T \in \mathbb{R}^{d \times C}$ then estimates patterns of activity of the extracted components. Each filtered signal indeed has a fixed projection across the components of \mathbf{x} : writing the linear forward model as $\mathbf{x}(t) = \mathbf{A}\mathbf{u}(t)$, with $\mathbf{u}(t) \in \mathbb{R}^{d \times 1}$ the periodic *sources*, estimates $W^T \approx A^{-1}$ are derived; the first column of $(W^T)^{-1}$ approximates the spatial pattern of the first estimated source signal $\mathbf{u}_1(t)$ [11].

2.1 Periodic Component Analysis

Periodic Component Analysis (π CA) [14] defines an optimal linear filter by minimizing a scale-invariant periodicity measure of the filtered signal $s(t) = \mathbf{w}^T \mathbf{x}(t)$:

$$\mathbf{w}_{\pi 1} = \arg \min_{\mathbf{w}} \left\{ F_{\pi 1}(\mathbf{w}) = \frac{\sum_t |s(t + T_1) - s(t)|^2}{\sum_t |s(t)|^2} = \frac{\mathbf{w}^T A_{\mathbf{x}}(T_1) \mathbf{w}}{\mathbf{w}^T C_{\mathbf{x}}(0) \mathbf{w}} \right\}, \quad (1)$$

where $A_{\mathbf{x}}(T_1) = \mathbb{E}_t \{ (\mathbf{x}(t + T_1) - \mathbf{x}(t)) (\mathbf{x}(t + T_1) - \mathbf{x}(t))^T \}$ and $C_{\mathbf{x}}(\tau) = \mathbb{E}_t \{ \mathbf{x}(t + \tau) \mathbf{x}(t)^T \}$. The minimization of this generalized Rayleigh quotient is solved by the generalized eigenvalue decomposition (GEVD) of the matrix pair $(A_{\mathbf{x}}(T_1), C_{\mathbf{x}}(0))$. These two matrices being symmetric, the matrix of generalized eigenvectors W sorted in decreasing order of magnitude of the associated generalized eigenvalues gives the components $W^T \mathbf{x}(t)$ ranked in decreasing order of periodicity [4].

2.2 Periodic Component Analysis Variant

Another periodicity measure can alternatively be optimized as:

$$\mathbf{w}_{\pi 2} = \arg \max_{\mathbf{w}} \left\{ F_{\pi 2}(\mathbf{w}) = \frac{\sum_t |s(t + T_1) \cdot s(t)|}{\sum_t |s(t)|^2} = \frac{\mathbf{w}^T C_{\mathbf{x}}(T_1) \mathbf{w}}{\mathbf{w}^T C_{\mathbf{x}}(0) \mathbf{w}} \right\}. \quad (2)$$

This defines a variant [12] of π CA, denoted here by π CA₂. This problem can be similarly solved by a GEVD of the matrices $(C_{\mathbf{x}}(T_1), C_{\mathbf{x}}(0))$. It is noteworthy that whenever $C_{\mathbf{x}}(T_1)$ is symmetric, $A_{\mathbf{x}}(T_1) = 2 \cdot (C_{\mathbf{x}}(0) - C_{\mathbf{x}}(T_1))$ and (2) is hence equivalent to (1). For any real-world signal \mathbf{x} however, $C_{\mathbf{x}}(T_1)$ is unlikely to be symmetric. Therefore (1) and (2) will typically not define the same filters. Since (1) seems more generally suited for periodicity maximization, π CA is expected to outperform π CA₂ in the current application.

2.3 Canonical Correlation Analysis

Another approach to extract periodic components from a multidimensional signal is based on Canonical Correlation Analysis (CCA) [5]. The principle is to maximize the correlation between a filtered signal and a linear combination of the components of a reference signal $\mathbf{y}(t)$, with the same length as \mathbf{x} . In our setting, the components of \mathbf{y} are defined from the Fourier series of a periodic signal of fundamental frequency f_1 : $\mathbf{y}(t) = (\sin(2\pi f_1 t) \cos(2\pi f_1 t) \sin(2\pi 2f_1 t) \dots \sin(2\pi N_h f_1 t) \cos(2\pi N_h f_1 t))^T$, where N_h is a parameter indicating the number of accounted harmonics [9]. The CCA optimization problem is

$$(\mathbf{w}_{\pi 3}, \mathbf{w}_{\pi 3-y}) = \arg \max_{\mathbf{w}, \mathbf{w}_y} \left\{ F_{\pi 3}(\mathbf{w}, \mathbf{w}_y) = \frac{\mathbf{w}^T C_{\mathbf{x};\mathbf{y}} \mathbf{w}_y}{\sqrt{\mathbf{w}^T C_{\mathbf{x}}(0) \mathbf{w} \cdot \mathbf{w}_y^T C_{\mathbf{y}}(0) \mathbf{w}_y}} \right\}, \quad (3)$$

where $C_{\mathbf{x};\mathbf{y}} = \mathbb{E}_t\{\mathbf{x}(t)\mathbf{y}(t)^T\}$. Only the optimal filter $\mathbf{w}_{\pi 3}$ is interesting in our context. The filter $\mathbf{w}_{\pi 3-y}$ being also optimized, (3) amounts finding the filtered signal $\mathbf{w}_{\pi 3}^T \mathbf{x}(t)$ which is maximally correlated with an arbitrary periodic signal whose frequency content is limited to $N_h f_1$. The solution to (3) is obtained by diagonalizing $C_{\mathbf{x};\mathbf{y}}$, $C_{\mathbf{x}}(0)$ and $C_{\mathbf{y}}(0)$ using only two matrices W and W_y with the filters in their columns [7].

2.4 Spectral Contrast Maximization

Spectral contrast maximization (SCM) consists in maximizing the magnitude of some frequency components of the filtered signal, with respect to the whole spectrum energy [13]. It is recommended when the searched components are more easily separable in the frequency domain, i.e. when the frequency band to amplify is known a priori. Let $S(f) := \mathcal{F}_f\{s(t)\} = \mathbf{w}^T \mathcal{F}_f\{\mathbf{x}(t)\} = \mathbf{w}^T X(f)$ denotes the Fourier transform of the filtered signal at frequency f . The optimal SCM filter is then defined as

$$\mathbf{w}_{\pi 4} = \arg \max_{\mathbf{w}} \left\{ F_{\pi 4}(\mathbf{w}) = \frac{\mathbb{E}_{f \in \nu} \{|S(f)|^2\}}{\mathbb{E}_{f \in \mu} \{|S(f)|^2\}} = \frac{\mathbf{w}^T S_{\mathbf{x}} \mathbf{w}}{\mathbf{w}^T C_{\mathbf{x}}(0) \mathbf{w}} \right\}, \quad (4)$$

with $\nu := \{\pm f_1, \pm 2f_1, \dots, \pm N_h f_1\}$ the set of considered frequencies, μ the whole frequency range (the Nyquist band for discrete signals), $S_{\mathbf{x}} := \mathbb{E}_{f \in \nu} \{X(f)X(f)^*\}$ and using the Parseval's identity at the denominator. The set ν contains negative frequencies to ensure the realness of the cross-spectrum matrix $S_{\mathbf{x}}$. Again, N_h is the number of harmonics to consider, including the fundamental frequency. This problem is solved using the GEVD of $(S_{\mathbf{x}}, C_{\mathbf{x}}(0))$.

Although the two last methods are formulated differently, they are intrinsically related. Indeed, CCA maximizes the correlation of the filtered signal with an arbitrary sum of sines and cosines at the harmonic frequencies, while SCM maximizes the Fourier amplitudes of the filtered signal at the same frequencies. In both cases, normalization ensures a scale-invariance of the solutions.

3 Processing EEG Signals

The comparison of the performances of the methods introduced in Sect. 2 is conducted on an EEG data set, which is first described in Sect. 3.1. Section 3.2 defines the quality criterion employed to quantitatively compare the methods, and Sect. 3.3 finally summarizes the results.

3.1 Experimental Setting

We recorded scalp EEG on 15 healthy subjects to whom we applied sinusoidal stimulations with a thermal cutaneous stimulator (TCS) in 4 different conditions, as shown in Fig. 1: warm and cool with either a fixed or a variable active surface along the stimulation cycles. Five distinct zones of the TCS stimulation surface could indeed be controlled independently. These 4 conditions were chosen in order to determine, for both warm and cool stimuli, whether alternating the position of the active surface along the stimulation cycles could improve the SNR of the induced SSR. Varying the active surface is indeed likely to limit the response habituation, which can for instance be due to skin receptor fatigue. Each stimulus consisted in a 0.2 Hz sinusoidal waveform lasting 15 periods (i.e. 75 s) and was applied to the right forearm. Each subject received 12 trials from each condition, leading to 48 trials in total presented in a randomized order. To reduce artifacts, these 12 trials are averaged for each condition. The EEG was sampled at 1000 Hz and recorded using 64 electrodes placed on the scalp according to the international 10/10 system. All signals were high-pass filtered above 0.05 Hz to remove slow drifts (4th order zero-phase Butterworth filter).

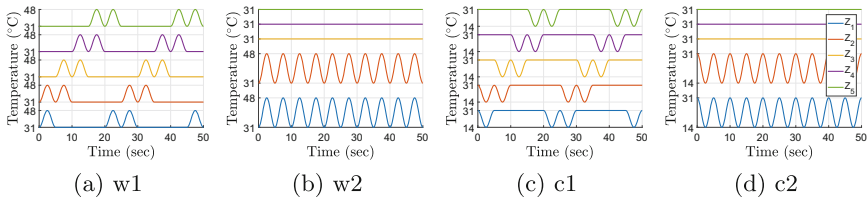


Fig. 1. Stimulation temperature profiles (only two thirds of the whole stimulus length of 75 s are shown for readability). Z_i means ‘zone i ’ of the stimulator. w1 and w2 (resp. c1 and c2) indicate the two warm (resp. cool) conditions with a variable and fixed stimulation surface.

3.2 Quality Measure

In order to assess the quality of an extracted component, we define a periodicity measure for a given unidimensional signal $y(t)$ and fundamental frequency f_1 . First, since each frequency amplitude $|Y(f)|$ is affected by some background noise, the average amplitude at 10 neighboring frequencies (5 higher and 5 lower) is removed from each frequency amplitude [8], resulting in a *noise-subtracted spectrum* $Y_{NS}(f) \in \mathbb{R}$. Then, the periodicity measure is defined as

$$M_\pi(y) = 100 \cdot \frac{\sum_{k=1}^{\lfloor f_s/(2 \cdot f_1) \rfloor} Y_{NS}(k \cdot f_1)}{\sum_f |Y(f)|}, \quad (5)$$

with f_s the sampling frequency. In addition to accounting for the background noise, this measure is normalized with respect to the total signal amplitude. A positive (resp. negative) M_π suggests that the spectrum amplitude of y contains local maxima (resp. minima) on average at the harmonics $k \cdot f_1$. It is noteworthy that the components extracted using the methods from Sect. 2 should maximize this measure. Meanwhile, these methods are constrained to produce a linear filtering of the original signals, thereby providing topographical patterns of the extracted components which can be interpreted.

3.3 Periodic Components Extraction

Before analyzing the results obtained with the methods described in Sect. 2, we can observe whether the periodic components are visible in the raw EEG signals. The topographies of the Fourier transforms at $f_1 = 0.2$ Hz, shown in Fig. 2, suggest that the periodic components seem to be most prominent at centro-frontal electrodes, and in particular at FCZ (see Fig. 2b). The periodicity of the EEG signal at this location will hence be compared to the periodicity of filtered signals. The EEG time courses averaged over the stimulation periods as well as the spectra at this electrode are shown in Figs. 3 and 4 for all subjects. In both figures, the periodicity is visible for the warm conditions, while it is less clear for the cool ones, especially when the stimulation surface is fixed (condition c2). The average over the periods highlights more easily the periodic structure in the EEG, and in particular gives an estimate of the latency between the temperature and EEG peaks.

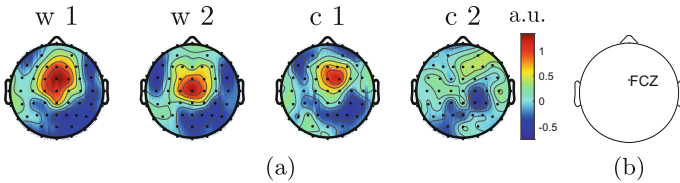


Fig. 2. Scalp topographies of (a) the group-level average noise-subtracted spectrum amplitudes at $f_1 = 0.2$ Hz and (b) the position of the FCZ electrode.

Performances of the periodicity-maximization methods are given in Table 1. First, the filter obtained from π CA, $\mathbf{w}_{\pi 1}$, outperforms the performances reached by $\mathbf{w}_{\pi 2}$. The poor results of $\mathbf{w}_{\pi 2}$ can partly be explained by the cross-channel symmetry hypothesis used to derive $F_{\pi 2}$ from $F_{\pi 1}$. Importantly, all filtering methods except π CA₂ lead to a filtered signal with an improved periodicity compared to the raw EEG signal at FCZ, even when this raw signal is hardly

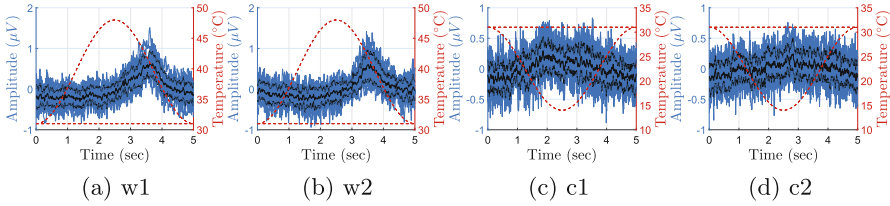


Fig. 3. EEG time courses averaged over the stimulation periods, at electrode FCZ for all 4 conditions. There is one curve per subject and the group-level average is in bold, with intervals of \pm one standard deviation around the mean delimited with dotted lines. Dashed lines indicate the stimulation temperature.

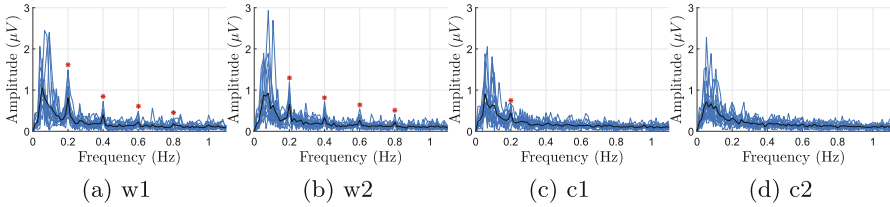


Fig. 4. EEG frequency spectrum at electrode FCZ for each subject and the 4 conditions. The group-level average is in bold. A star indicates significance of the noise-subtracted peaks (Sect. 3.2) at $k \cdot f_1 = k \cdot 0.2$ Hz (paired t-test vs 0).

periodic, such as for condition c2 for instance, which corresponds to cool stimulations with a fixed stimulation surface. Two values of the parameter N_h are shown for CCA and SCM, chosen according to an analysis of the method performances as a function of N_h , not depicted here for space limitations. This analysis revealed a saturating increase of the performances which was consistent for all the conditions: as long as N_h is chosen higher than approximately 8, M_π has almost reached a plateau. The results of this table also indicate that:

- the SS responses obtained when stimulating the forearm with a variable surface (conditions w1 and c1) exhibit a more pronounced periodicity compared to the stimuli applied with a fixed surface (w2 and c2). The periodicity measure indeed shows this difference for almost all the filtered and raw signals.
- CCA and SCM extract the same periodic components (for all signals). This is not very surprising regarding the similarity between these two methods that was discussed in Sect. 2; a deeper algorithmic comparison is left for future works.
- performances of CCA and SCM are improved when the number of harmonics is increased, for all conditions. This further motivates the idea of extracting periodic non sinusoidal components instead of analyzing raw EEG frequency spectra: a single spatial pattern can regroup information from several harmonics.

The spatial patterns and filters obtained with the best filtering method (SCM with $N_h = 10$) are shown in Fig. 5. These spatial patterns indicate the distribution of the most periodic component across the scalp. The spatial filters on the other hand have more intricate scalp topographies than their associated patterns as they need to cancel the other interfering (noise) components [1]. In addition, the associated component time courses, averaged over the stimulation periods, and their Fourier transforms are given in Figs. 6 and 7, the y-scales of the latter differing from Fig. 4. These two last figures show the high periodicity of the filtered signals. This is in striking contrast with the raw signals at FCZ, especially for condition c2 depicted in Fig. 3d (with its spectrum in Fig. 4d), where no clear periodicity was visible. However, all methods rank, according to M_π , the components for each condition (i.e. the entries in each column of Table 1) in almost the same order as the FCZ signals (i.e. first column of Table 1), which encourages the valid interpretation of the filtered signal. Further validation will be conducted to ensure that the periodic amplified activity indeed reflects stimulation-related patterns. The average curves from Fig. 6 are also interesting as they show the time lag between the temperature peaks and the extracted part of the SSR. In particular, we observe a longer time lag for both warm conditions compared to the cool ones. This is in accordance with the fact that cool stimuli activate thinly-myelinated A δ fibers [6], while the employed warm periodic stimuli most probably activate unmyelinated C fibers with slower conduction velocities [3].

Table 1. Mean(std) for the 15 subjects of the periodicity measure M_π of the components extracted with the periodicity-maximization methods of Sect. 2. For each stimulation type (row), the best performances are in bold. Italic characters indicate that the corresponding signal is not significantly less periodic than the best one of the same row. Significativity is computed with paired t-tests and is adjusted for multiple comparisons using the Holm-Bonferroni correction.

	FCZ signal	π CA	π CA ₂	CCA		SCM	
				$N_h = 1$	$N_h = 10$	$N_h = 1$	$N_h = 10$
w1	1.02(0.49)	1.60(1.72)	-0.12(0.19)	2.84(1.20)	3.04(1.15)	2.84(1.20)	3.04(1.15)
w2	0.80(0.51)	1.50(1.65)	-0.12(0.18)	2.70(0.96)	2.93(1.01)	2.70(0.96)	2.93(1.01)
c1	0.19(0.24)	0.39(0.40)	-0.12(0.11)	1.99(0.56)	2.17(0.57)	1.99(0.56)	2.17(0.57)
c2	0.06(0.31)	0.44(0.51)	-0.11(0.14)	1.88(0.41)	2.03(0.47)	1.88(0.41)	2.03(0.47)

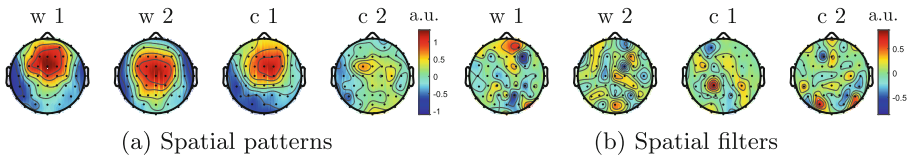


Fig. 5. Group-level average spatial patterns and spatial filters (defined at the beginning of Sect. 2) of the first component extracted with SCM ($N_h = 10$).

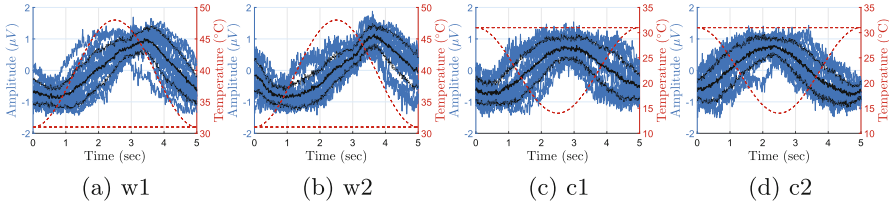


Fig. 6. Average over the stimulation periods of the optimal component extracted with SCM ($N_h = 10$). There is one curve per subject and the group-level average is in bold, with intervals of \pm one standard deviation around the mean delimited with dotted lines. Dashed lines indicate the stimulation temperature.

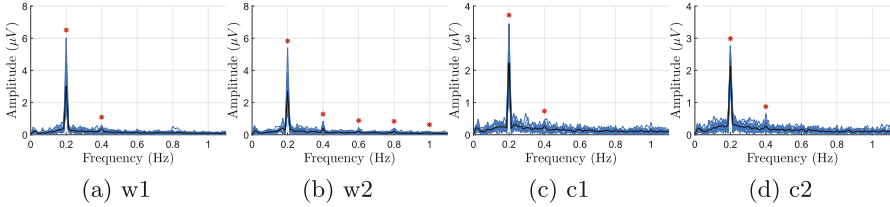


Fig. 7. Frequency spectrum of the components extracted with SCM ($N_h = 10$) for each subject. The group-level average is in bold. A star indicates significance of the noise-subtracted peaks (Sect. 3.2) at $k \cdot 0.2$ Hz (paired t-test vs 0).

4 Conclusions and Perspectives

This paper suggests employing spatial filters to enhance the SNR of EEG responses elicited by periodic sensory stimulation. Four approaches are detailed and compared on an EEG data set recorded on 15 healthy subjects exposed to four different kinds of long lasting sinusoidal thermal stimuli. We show that these methods are able to extract periodic components from signals which do not necessarily exhibit a pronounced temporal periodicity. The estimated spatial activity patterns as well as the component time courses can hence characterize the steady-state responses.

As to further perspectives, the filtering methods considered in this work have been applied to the EEG signals averaged over the trials, enhancing their phase-locked components. Studying the periodic responses on a trial-basis, possibly using tensor methods, would enable determining whether and to which extent phase variability across trials affects the observed SSR. Another line of work is related to the analysis of the multiple suboptimal components, in terms of periodicity, extracted by the linear filters defined in Sect. 2. Whereas this paper focuses on the periodicity and spatial patterns of the optimal component identified by each method, it is very likely that the linear space spanned by the spatial patterns reflecting stimulation-related activity is more than one-dimensional in the studied EEG data. The considered linear filters moreover cannot compensate some phase changes across channels reflecting the propagation of the SSR

within brain regions. This hence suggests analyzing the links between the time dynamics of different filtered components and their spatial localization on the scalp.

Acknowledgments. DM and CdB are Research Fellows of the Fonds de la Recherche Scientifique - FNRS. The authors gratefully thank Prof. Christian Jutten for insightful discussions.

References

1. Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.R.: Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage* **56**(2), 814–825 (2011)
2. Colon, E., Legrain, V., Mouraux, A.: Steady-state evoked potentials to study the processing of tactile and nociceptive somatosensory input in the human brain. *Neurophysiol. Clin./Clin. Neurophysiol.* **42**(5), 315–323 (2012)
3. Colon, E., Liberati, G., Mouraux, A.: EEG frequency tagging using ultra-slow periodic heat stimulation of the skin reveals cortical activity specifically related to C fiber thermoreceptors. *NeuroImage* **146**, 266–274 (2017)
4. Golub, G.H., Van Loan, C.F.: *Matrix Computations*, vol. 3. JHU Press, Baltimore (2012)
5. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**(12), 2639–2664 (2004)
6. Hüllemann, P., Nerdal, A., Binder, A., Helfert, S., Reimer, M., Baron, R.: Cold-evoked potentials—ready for clinical use? *Eur. J. Pain* **20**(10), 1730–1740 (2016)
7. Krzanowski, W.: *Principles of Multivariate Analysis*, vol. 23. OUP, Oxford (2000)
8. Mouraux, A., Iannetti, G.D., Colon, E., Nozaradan, S., Legrain, V., Plaghki, L.: Nociceptive steady-state evoked potentials elicited by rapid periodic thermal stimulation of cutaneous nociceptors. *J. Neurosci.* **31**(16), 6079–6087 (2011)
9. Nakanishi, M., Wang, Y., Wang, Y.T., Mitsukura, Y., Jung, T.P.: A high-speed brain speller using steady-state visual evoked potentials. *Int. J. Neural Syst.* **24**(06), 1450019 (2014)
10. Pfurtscheller, G., Da Silva, F.L.: Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin. Neurophysiol.* **110**(11), 1842–1857 (1999)
11. Samadi, S., Amini, L., Cosandier-Rimélé, D., Soltanian-Zadeh, H., Jutten, C.: Reference-based source separation method for identification of brain regions involved in a reference state from intracerebral EEG. *IEEE Trans. Biomed. Eng.* **60**(7), 1983–1992 (2013)
12. Sameni, R., Jutten, C., Shamsollahi, M.B.: Multichannel electrocardiogram decomposition using periodic component analysis. *IEEE Trans. Biomed. Eng.* **55**(8), 1935–1940 (2008)
13. Sameni, R., Jutten, C., Shamsollahi, M.B.: A deflation procedure for subspace decomposition. *IEEE Trans. Sig. Process.* **58**(4), 2363–2374 (2010)
14. Saul, L.K., Allen, J.B.: Periodic component analysis: an eigenvalue method for representing periodic structure in speech. In: *Advances in Neural Information Processing Systems*, pp. 807–813 (2001)
15. Wittevrongel, B., Van Hulle, M.M.: Frequency- and phase encoded SSVEP using spatiotemporal beamforming. *PLoS One* **11**(8), e0159988 (2016)



Static and Dynamic Modeling of Absence Epileptic Seizures Using Depth Recordings

Saeed Akhavan^{1,2}(✉), Ronald Phlypo¹, Hamid Soltanian-Zadeh^{2,3},
Mahmoud Kamarei², and Christian Jutten^{1,4}

¹ Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab., Grenoble, France
{saeed.akhavan-bahabadi,ronald.phlypo,christian.jutten}@gipsa-lab.fr

² School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
{hszadeh,kamarei}@ut.ac.ir

³ Medical Image Analysis Lab., Henry Ford Health System, Detroit, MI, USA

⁴ Institut Universitaire de France, Paris, France

Abstract. This research temporally explores absence epileptic seizures using depth cortical data recorded from different layers of the somatosensory cortex of Genetic Absence Epilepsy Rats from Strasbourg (GAERS). We characterize the recorded absence seizures by a linear combination of a few static and dynamic sources. Retrieving these sources from the recorded absence seizures is the main target of this study which helps us uncover the temporal evolution of absence seizures. The method used in this study provides an interesting and original solution to the classical data denoising consisting in removing the background activity and cleaning the data. The obtained results show the presence of a static source and a few specific dynamic sources during the recorded absence seizures. It is also shown that the sources have similar origins in different GAERS.

Keywords: Absence seizure · Static and dynamic source
Static and dynamic structure · Temporal evolution

1 Introduction

Absence epilepsy is a form of epilepsy which is accompanied by appearance of sudden absence seizures in different regions of the brain [12]. Recent studies about the origin of absence seizures in the brain show that one region of the somatosensory cortex starts absence seizures, and after a few cycles, a circuit between the somatosensory cortex and the thalamus continues absence seizures [4, 10]. In order to locally investigate the starting region of absence seizures,

The data used in this study were acquired at Grenoble institute of Neurosciences (GIN) in the team Synchronisation and Modulation of Neural Networks in Epilepsy (SyMoNNE) supervised by Dr. A. Depaulis. Also, this work has been partly supported by the European project 2012-ERC-AdG-320684 CHESS.

i.e., the somatosensory cortex, a data set was acquired at Grenoble Institute of Neurosciences (GIN) from different layers of the somatosensory cortex of Genetic Absence Epilepsy Rats from Strasbourg (GAERS) [12]. In this research, we investigate the temporal evolution of absence seizures using the recorded data.

Temporal analysis of absence seizures has attracted a lot of attention during the last decade [2,3,9]. As one of the major works, [2] studies the temporal behavior of the sources generating the absence seizures using the intracranial EEG (iEEG) data recorded from different regions of GAERS brain. At first, blind source separation methods [7] are applied on sliding time windows of the data. Then, the obtained sources are compared in different time windows using cross correlation criterion. It is shown that the sources become more stationary after a short delay from absence seizures onset [2]. The temporal analysis of absence seizures has also been performed in humans suffering from absence epilepsy, and it is shown that the cortical activations occur earlier than thalamic activations during absence seizures [11].

In [1], a sequence of Markovian states is assigned to the time windows of an absence seizure, and a few substates are considered for each state. Then, hidden Markov model (HMM) is applied on the data and the states and their substates are extracted. The obtained results show that there are a few common substates among the states indicating the presence of some background activities during absence seizures.

Based on the results obtained in [1], we model the absence seizures by a linear combination of the static and dynamic sources. The static sources show the background epileptic activities, and the dynamic sources are supplementary to the static sources in producing the absence seizures. The static sources have a fixed structure with respect to the recording electrode, while the dynamic sources may have a variable structure. We propose a method to extract these sources and their structures from the data, and then, we analyze the temporal evolution of the absence seizures based on the extracted results. The proposed method also provides an interesting solution to the classical electroencephalography (EEG) denoising consisting in removing the EEG background activity and cleaning the EEG epileptic data.

2 Materials

2.1 Data

An electrode with $n = 16$ sensors was vertically implanted in the somatosensory cortex of four GAERS rats, and extracellular field potentials were recorded. The distance between each pair of adjacent sensors and the sampling rate are $150 \mu\text{m}$ and 20KHz , respectively.

Spikes are the most important epileptic events during absence seizures. Since the data were acquired locally, the spikes approximately appear in different channels simultaneously. In fact, when a spike appears, we have a spike time window consisting of 16 spikes as shown in Fig. 1. Therefore, each absence seizure can be considered as a train of spike time windows. Our model for the absence seizures

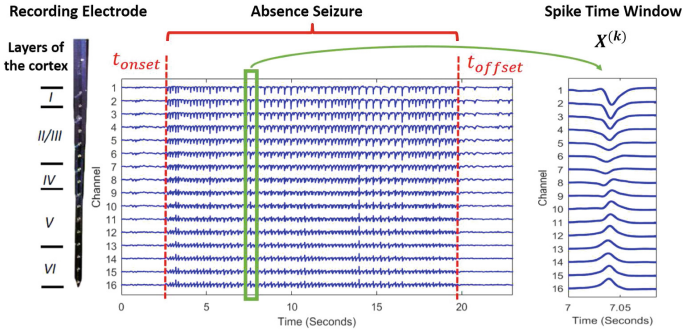


Fig. 1. Recording electrode (left), an absence seizure (center) and a spike time window (right).

is defined based on the spike time windows, which can be detected and aligned following the proposed method in [13] and [6], respectively.

2.2 Model for Absence Seizures

Due to the quasi-static assumption of Maxwell’s laws, we assume that the sensors on the electrode record the instantaneous linear mixture of the sources located around the electrode. We categorize the sources into the static and dynamic sources [1]. The static sources contribute in the generation of the data in all of the spike time windows and have a fixed structure, while the dynamic sources participate in some of the spike time windows and do not have a fixed structure. Figure 2 shows the schematic diagrams of the considered model for four spike time windows.

2.3 Problem Formulation

Consider an absence seizure with K spike time windows. Each spike time window $\mathbf{X}^{(k)} \in \mathbb{R}^{n \times L}$ consists the data of $n = 16$ channels for $L = 1750$ samples (87.5 ms) [1, 12]. According to the considered model, $\mathbf{X}^{(k)}$ is expressed as follows:

$$\mathbf{X}^{(k)} = \mathbf{A}\mathbf{S}^{(k)} + \mathbf{B}^{(k)}\mathbf{U}^{(k)} + \mathbf{N}^{(k)} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n_s}$ and $\mathbf{S}^{(k)} \in \mathbb{R}^{n_s \times L}$ show the static structure and the static sources, respectively. n_s denotes the number of static sources. If we assume $n_d^{(k)}$ dynamic sources are active in the k^{th} spike time window, $\mathbf{B}^{(k)} \in \mathbb{R}^{n \times n_d^{(k)}}$ and $\mathbf{U}^{(k)} \in \mathbb{R}^{n_d^{(k)} \times L}$ represent the dynamic structure and the dynamic sources, respectively. Finally, $\mathbf{N}^{(k)} \in \mathbb{R}^{n \times L}$ is considered the noise matrix with independent and identically distributed (i.i.d.) and zero-mean Gaussian entries. Therefore, the entries of the noise matrix are independent, and each entry has $\mathcal{N}(0, \sigma_0^2)$ distribution.

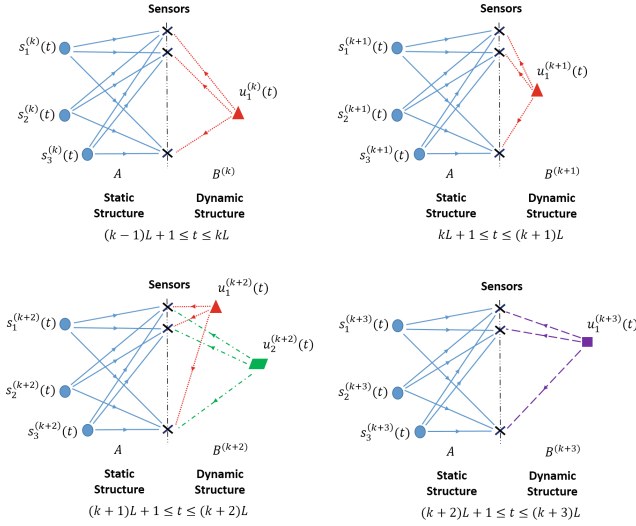


Fig. 2. The static sources and the dynamic sources are respectively shown in the left and right sides of the sensors for four consecutive spike time windows. The static sources have a static structure while the dynamic sources can move over time windows.

We also consider the following assumptions:

- (\mathcal{A}_1) Due to the scaling ambiguity [7], we must put some constraints on \mathbf{A} or $\mathbf{S}^{(k)}$, and also $\mathbf{B}^{(k)}$ or $\mathbf{U}^{(k)}$. We assume that the columns of \mathbf{A} are unit norm, and the dynamic sources are unit power in each spike time window.
- (\mathcal{A}_2) All of the sources are considered uncorrelated in each spike time window. In fact, we assume that there is no linear relationship among the sources.
- (\mathcal{A}_3) Since the dynamic sources may only appear in one spike time window, we need more information to extract them [7]. Therefore, we assume the dynamic sources are also statistically independent in each spike time window.
- (\mathcal{A}_4) Due to identifiability issues [7], we assume that the number of sources is less than the number of sensors, and the concatenation of the static and the dynamic structure $[\mathbf{A} \ \mathbf{B}^{(k)}]$ leads to a full rank matrix for all of the spike time windows.

Now, the problem definition is complete. The set of unknown parameters is as follows:

$$\Theta = \{\mathbf{A}, n_s, \bigcup_{k=1}^K \{\mathbf{S}^{(k)}, n_d^{(k)}, \mathbf{B}^{(k)}, \mathbf{U}^{(k)}\}\} \quad (2)$$

Based on the observations $\mathbf{X}^{(k)}$ ($k = 1, 2, \dots, K$) and the considered assumptions, Θ must be estimated.

3 Proposed Method

At first, we estimate the number of sources. Then, the static structure and the dynamic sources are obtained. Finally, the static sources and the dynamic structures are extracted.

3.1 Number of Sources

Since there is no prior information about the number of static sources (n_s), we obtain the results for different n_s . Then, we select the one which has the best biophysiological interpretation (e.g., the shape of the sources must be smooth). Therefore, we assume n_s is fixed and known.

For estimating $n_d^{(k)}$, based on the considered assumptions, we express the autocorrelation matrix of the observations $\mathbf{R}_x^{(k)} \in \mathbb{R}^{n \times n}$ in each spike time window as follows:

$$\mathbf{R}_x^{(k)} = \mathbf{A}\mathbf{R}_s^{(k)}\mathbf{A}^T + \mathbf{B}^{(k)}\mathbf{R}_u^{(k)}\mathbf{B}^{(k)T} + \mathbf{R}_n^{(k)} \tag{3}$$

where $\mathbf{R}_s^{(k)} \in \mathbb{R}^{n_s \times n_s}$ shows the autocorrelation matrix of the static sources. According to (\mathcal{A}_2) , it is a diagonal matrix, and its diagonal entries may change in different spike time windows because the static sources can be non-stationary. $\mathbf{R}_u^{(k)} \in \mathbb{R}^{n_d^{(k)} \times n_d^{(k)}}$ represents the autocorrelation matrix of the dynamic sources, and it is equal to \mathbf{I} (the identity matrix) according to (\mathcal{A}_1) and (\mathcal{A}_2) . Finally, $\mathbf{R}_n^{(k)} \in \mathbb{R}^{n \times n}$ is the autocorrelation matrix of noise which is equal to $\sigma_0^2\mathbf{I}$.

The total number of sources in each spike time window, and hence $n_d^{(k)}$, are obtained by calculating the eigen decomposition of $\mathbf{R}_x^{(k)}$ and thresholding the eigenvalues following the method proposed in [8].

3.2 Static Structure

We minimize the following objective function to find the static structure [7]:

$$f(\Theta) = \sum_{k=1}^K \|\mathbf{R}_x^{(k)} - \mathbf{A}\mathbf{R}_s^{(k)}\mathbf{A}^T - \underbrace{\mathbf{B}^{(k)}\mathbf{R}_u^{(k)}\mathbf{B}^{(k)T}}_{\mathbf{R}_d^{(k)}}\|_F^2 \tag{4}$$

Here, we just focus on estimating $\{\mathbf{A}, \bigcup_{k=1}^K \{\mathbf{R}_s^{(k)}, \mathbf{R}_d^{(k)}\}\}$. Although $\mathbf{R}_s^{(k)}$ and $\mathbf{R}_d^{(k)}$ are not important parameters, but they must be obtained during the optimization. According to the explained assumptions, the following constraints must also be considered in the optimization:

$$\begin{aligned} c_1 &: \text{diag}(\mathbf{A}^T\mathbf{A}) = \mathbf{I}, \\ c_2 &: \text{rank}(\mathbf{R}_s^{(k)}) = n_s, \quad \mathbf{R}_s^{(k)} = \text{diag}(\mathbf{R}_s^{(k)}), \quad \mathbf{R}_s^{(k)} \succeq 0, \\ c_3 &: \text{rank}(\mathbf{R}_d^{(k)}) = n_d^{(k)}, \quad \mathbf{R}_d^{(k)} \succeq 0, \quad k = 1, 2, \dots, K \end{aligned} \tag{5}$$

where $diag(\cdot)$ makes the non-diagonal entries of a matrix equal to zero, $rank(\cdot)$ represents the rank of a matrix, and $\succeq 0$ shows that a matrix is positive semidefinite. We use alternating minimization to solve the proposed constrained optimization problem. The minimization with respect to \mathbf{A} , $\mathbf{R}_s^{(k)}$ and $\mathbf{R}_d^{(k)}$ can respectively be performed using gradient-projection (GP), non-negative least square (NNLS) and eigen decomposition methods [5].

3.3 Dynamic Sources

Once the static structure is obtained, we can use the concept of singular value decomposition (SVD) to find the null space of \mathbf{A} and omit the static sources in each spike time window. Assuming $\mathbf{V} \in \mathbb{R}^{n \times (n-n_s)}$ spans the null space of \mathbf{A} , we left multiply both sides of (1) by \mathbf{V}^T :

$$\underbrace{\mathbf{V}^T \mathbf{X}^{(k)}}_{\mathbf{X}'^{(k)}} = \underbrace{\mathbf{V}^T \mathbf{A} \mathbf{S}^{(k)}}_0 + \underbrace{\mathbf{V}^T \mathbf{B}^{(k)}}_{\mathbf{B}'^{(k)}} \mathbf{U}^{(k)} + \underbrace{\mathbf{V}^T \mathbf{N}^{(k)}}_{\mathbf{N}'^{(k)}} \tag{6}$$

where $\mathbf{X}'^{(k)} \in \mathbb{R}^{(n-n_s) \times L}$, $\mathbf{B}'^{(k)} \in \mathbb{R}^{(n-n_s) \times n_d^{(k)}}$ and $\mathbf{N}'^{(k)} \in \mathbb{R}^{(n-n_s) \times L}$ can be considered the new data, the new dynamic structure and new noise, respectively. Since the dynamic sources are statistically independent according to (\mathcal{A}_3) , we can use the proposed independent component analysis (ICA) method in [8] to estimate the dynamic sources. It should be mentioned that the number of dynamic sources $n_d^{(k)}$ has been already estimated in each spike time window, $\mathbf{B}'^{(k)}$ is not equal to zero according to (\mathcal{A}_4) , and each column of $\mathbf{N}'^{(k)}$ has $\mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$ distribution.

3.4 Static Sources and Dynamic Structures

According to (1), since the entries of noise are Gaussian and independent, minimizing the following objective function leads to finding the maximum likelihood estimation of the static sources and the dynamic structure in each spike time window:

$$g(\mathbf{S}^{(k)}, \mathbf{B}^{(k)}) = \|\mathbf{X}^{(k)} - \mathbf{A} \mathbf{S}^{(k)} - \mathbf{B}^{(k)} \mathbf{U}^{(k)}\|_F^2 \tag{7}$$

We can simply minimize this objective function using alternation minimization.

By extracting the static sources and the dynamic structures in all of the spike time windows, all of the parameters are determined.

4 Experimental Results

We apply the proposed method on the absence seizures for different number of static sources. The best results are extracted by considering $n_s = 1$ in all of the absence seizures. In fact, when we consider $n_s > 1$, the sources become non-smooth and incomprehensible. The estimated number of dynamic sources is also

equal to one ($n_d^{(k)} = 1$) for all of the spike time windows. An absence seizure with $K = 390$ spike time windows from the first rat is considered to show the results. The obtained results for all of the spike time windows are shown in Fig. 3. The static sources and the dynamic structures are normalized to better show the results. The obtained static sources are similar and they can be considered as one cluster. The center of this cluster, which is the average of the static sources, is shown in red in Fig. 3.

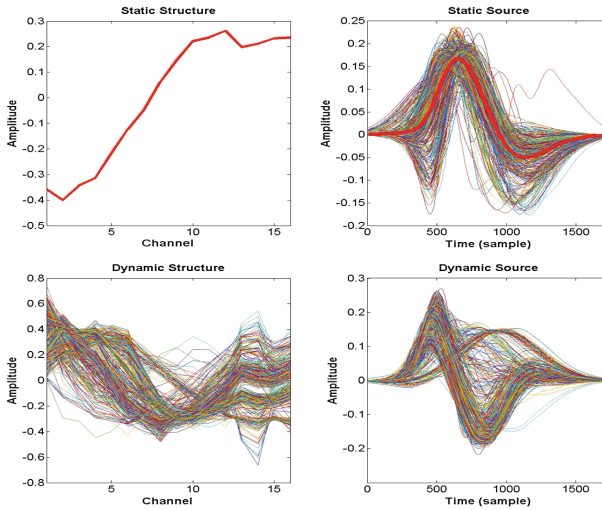


Fig. 3. The obtained results for an absence seizure with $K = 390$ spike time windows (Color figure online).

It seems that the dynamic sources can also be partitioned into a few clusters. If we apply *k-means* clustering on the extracted dynamic sources, three clusters are obtained as shown in Fig. 4. The corresponding dynamic structures of each cluster are also shown in Fig. 4.

Based on the obtained results, we can conclude that the linear superposition of a static source with one of three dynamic sources generates the spike time windows of the absence seizure as shown in Fig. 5.

Since there is a specific dynamic source in each spike time window, a sequence of clusters can be assigned to the absence seizure as shown in Fig. 6. It can be observed that the second dynamic source disappears in the end of the absence seizure, thus, it is an unstable source.

The same results are obtained when we apply the proposed method on other absence seizures of the first rat. For other rats, the obtained static structure and centers of clusters for the dynamic structures are similar to the results of the first rat, and there is also an unstable dynamic source, but the centers of clusters for the static and dynamic sources are different. For instance, the obtained centers of clusters for the absence seizures of the second rat are shown in Fig. 7.

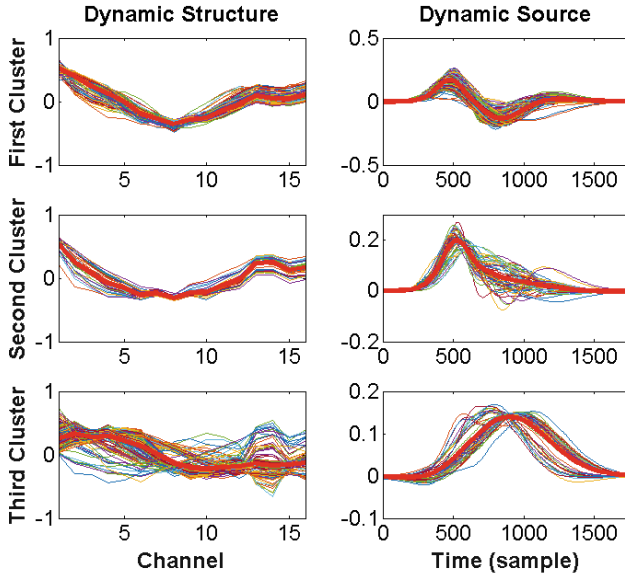


Fig. 4. There are three kinds of dynamic sources during absence seizures.

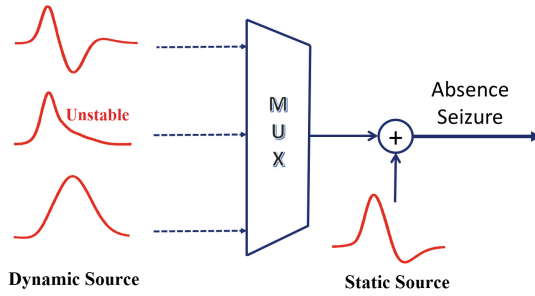


Fig. 5. The linear combination of a static source with one of three kinds of dynamic sources generates the spike time windows of absence seizures. MUX stands for multiplexer, that just allows one source to pass.

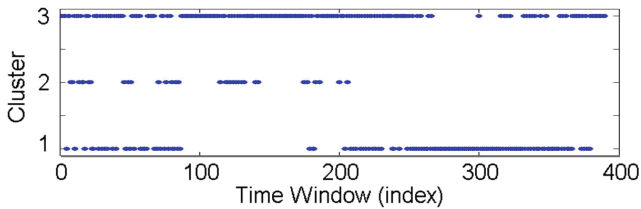


Fig. 6. Sequence of clusters for an absence seizure with $K = 390$ spike time windows.

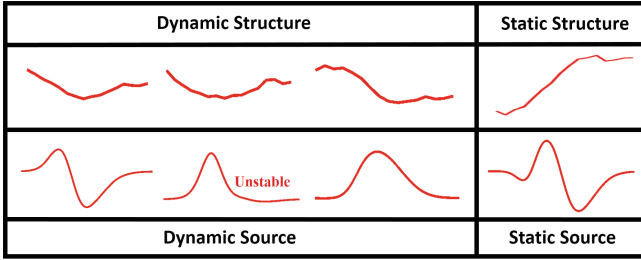


Fig. 7. The obtained centers of clusters for the second rat.

By comparing Fig. 7 with Figs. 3 and 4, we observe that the corresponding structures are the same in the first and second rats. The correlation coefficient for the corresponding static structures and centers of clusters for the dynamic structures in the first and second rats are respectively equal to 0.99, 0.97, 0.98 and 0.95. Therefore, we can conclude that the origins of the sources are similar in the rats, however, the propagated signals from the origins are different.

We also cross validate the obtained results by computing the reconstruction error. The spike time windows of each absence seizure ($\mathbf{X}^{(k)}$) are reconstructed ($\hat{\mathbf{X}}^{(k)}$) using the obtained results in other absence seizures (i.e., the static structure + the centers of cluster for dynamic structures, static sources, and dynamic sources). Then, the relative reconstruction error is calculated as follows:

$$Er = \frac{1}{K} \sum_{k=1}^K \frac{\|\hat{\mathbf{X}}^{(k)} - \mathbf{X}^{(k)}\|_F^2}{\|\mathbf{X}^{(k)}\|_F^2} \tag{8}$$

The obtained relative reconstruction errors for five absence seizures of the first rat are reported in Table 1. The errors in other rats have the same order of magnitude as the first rat which show the accuracy and generality of the proposed model and the obtained results for the recorded absence seizures.

Table 1. Relative reconstruction error for five absence seizures of the first rat. The absence seizures respectively consist of $K_1 = 87$, $K_2 = 94$, $K_3 = 95$, $K_4 = 88$ and $K_5 = 390$ spike time windows.

Training on	Testing on				
Seizure	1	2	3	4	5
1	0.05	0.11	0.13	0.12	0.09
2	0.07	0.06	0.10	0.09	0.08
3	0.08	0.11	0.06	0.10	0.09
4	0.10	0.09	0.10	0.08	0.10
5	0.09	0.10	0.11	0.12	0.03

5 Conclusion and Future Work

We analyzed the temporal evolution of absence epileptic seizures using the data recorded from different layers of the somatosensory cortex of GAERS rats. We showed that the linear combination of a static source and one of three (two stable + one unstable) dynamic sources generates the spike time windows of recorded absence seizures. It was also shown that the origin of the sources are similar in the rats, but the propagated signals are different. The concentration of this study was on the temporal evolution of the recorded absence seizures. In the future work, we will investigate the spatial characterization of the data, and provide a comprehensible spatio-temporal analysis of absence seizures.

References

1. Akhavan, S., Phlypo, R., Soltanian-Zadeh, H., Studer, F., Depaulis, A., Jutten, C.: Characterizing absence epileptic seizures from depth cortical measurements. In: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 444–448, August 2017
2. Amini, L., Jutten, C., Pouyatos, B., Depaulis, A., Roucard, C.: Dynamical analysis of brain seizure activity from EEG signals. In: 2014 22nd European Signal Processing Conference (EUSIPCO), pp. 36–40, September 2014
3. Amor, F., Baillet, S., Navarro, V., Adam, C., Martinerie, J., Le Van Quyen, M.: Cortical local and long-range synchronization interplay in human absence seizure initiation. *Neuroimage* **45**(3), 950–962 (2009)
4. Avoli, M.: A brief history on the oscillating roles of thalamus and cortex in absence seizures. *Epilepsia* **53**(5), 779–789 (2012)
5. Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge (2004)
6. Cabasson, A., Meste, O.: Time delay estimation: a new insight into the woody’s method. *IEEE Sig. Process. Lett.* **15**, 573–576 (2008)
7. Comon, P., Jutten, C.: *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Elsevier (2010)
8. Joho, M., Mathis, H., Lambert, R.H.: Overdetermined blind source separation: using more sensors than source signals in a noisy mixture. In: *Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 81–86 (2000)
9. Marten, F., Rodrigues, S., Suffczynski, P., Richardson, M.P., Terry, J.R.: Derivation and analysis of an ordinary differential equation mean-field model for studying clinically recorded epilepsy dynamics. *Phys. Rev. E* **79**, 021911 (2009)
10. Meeren, H.K.M., Pijn, J.P.M., Coenen, A.M.L., Lopes Da Silva, O.H.: Cortical focus drives widespread corticothalamic networks during spontaneous absence seizures in rats. *J. Neurosci.* **22**, 1480–1495 (2002)
11. Moeller, F., LeVan, P., Muhle, H., Stephani, U., Dubeau, F., Siniatchkin, M., Gotman, J.: Dynamic analysis of absence seizures in humans: all the same but all different. *Neuropediatrics* **41**(02), V1287 (2010)

12. Polack, P.O., Guillemain, I., Hu, E., Deransart, C., Depaulis, A., Charpier, S.: Deep layer somatosensory cortical neurons initiate spike-and-wave discharges in a genetic model of absence seizures. *J. Neurosci.* **27**(24), 6590–6599 (2007)
13. Quiroga, R.Q., Nadasdy, Z., Ben-Shaul, Y.: Unsupervised Spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16**(8), 1661–1687 (2004)

Applications of LVA/ICA



Multichannel Audio Source Separation Exploiting NMF-Based Generic Source Spectral Model in Gaussian Modeling Framework

Thanh Thi Hien Duong^{1,2(✉)}, Ngoc Q. K. Duong³, Cong-Phuong Nguyen^{1,4},
and Quoc-Cuong Nguyen⁴

¹ International Research Institute MICA,
Hanoi University of Science and Technology, Hanoi, Vietnam

² Information Technology Faculty,
Hanoi University of Mining and Geology, Hanoi, Vietnam
duongthihienthanh@hmg.edu.vn

³ Imaging Science Lab, Technicolor, Cesson-Sévigné, France
quang-khanh-ngoc.duong@technicolor.com

⁴ Department of Instrumentation and Industrial Informatic,
Hanoi University of Science and Technology, Hanoi, Vietnam
{phuong.nguyencong, cuong.nguyenquoc}@hust.edu.vn

Abstract. Nonnegative matrix factorization (NMF) has been well-known as a powerful spectral model for audio signals. Existing work, including ours, has investigated the use of generic source spectral models (GSSM) based on NMF for single-channel audio source separation and shown its efficiency in different settings. This paper extends the work to multichannel case where the GSSM is combined with the source spatial covariance model within a unified Gaussian modeling framework. Especially, unlike a conventional combination where the estimated variances of each source are further constrained by NMF separately, we propose to constrain the total variances of all sources altogether and found a better separation performance. We present the expectation-maximization (EM) algorithm for the parameter estimation. We demonstrate the effectiveness of the proposed approach by using a benchmark dataset provided within the 2016 Signal Separation Evaluation Campaign.

Keywords: Multichannel audio source separation
Generic spectral model · Nonnegative matrix factorization
Spatial covariance model · Gaussian modeling

1 Introduction

Audio source separation, which aims at separating individual sound sources from their mixture, is crucial in many practical applications such as speech enhancement, sound post-production, and robotics. Despite numerous efforts in the past

decades, its performance in real-world conditions is still far from perfect [1]. To improve the separation performance, depending on specific scenario where certain side information can be known, a range of *informed* source separation algorithms has been proposed in the literature [2]. Such side information can be *e.g.*, score associated with musical sources [3], text associated with spoken speeches [4], motion associated with audio-visual objects in a video [5], or deformed references [6]. Following this trend, very abstract semantic information just about the type of audio source (*e.g.*, if a source in the mixture is speech, musical instrument, or environmental sound) has been used to create a universal speech model in [7] or the universal sound class model in [8]. Exploiting this idea, we have investigated the use of generic speech and noise model for single-channel speech separation in [9] and shown its promising result. Further more, we have proposed to combine the block sparsity constraint investigated in [7] with the component sparsity constraint presented in [8] in a common formulation so as to take into account the advantage of both of them.

It is interesting to note that most cited work above [3–5, 7–9] considered only a single channel case, where the mixtures are mono, and exploited non-negative matrix factorization (NMF) [10, 11] to model the spectral characteristics of audio sources. When more recording channels are available, multichannel source separation algorithm should be considered as it allows to exploit important information about the spatial locations of the sources. Such additional information has been shown to greatly improve the separation performance. To date, the spatial cues can be modeled by *e.g.*, the interchannel time difference and interchannel intensity difference [12], the rank-1 mixing vector in the frequency domain [13, 14], or the full-rank spatial covariance matrix in Gaussian modeling framework [15, 16]. In this paper, we present an extension of our previous work [9] to multichannel case where the NMF-based GSSM is combined with the powerful full-rank spatial covariance model in a Gaussian modeling paradigm [15]. Note that the combination of NMF with such spatial covariance model has been investigated in several works [16–18]. However, our work is different from [17, 18] in the sense that we use the pre-trained GSSM so as the intermediate source variances are better constrained. As consequence, the overall algorithm is much less sensitive to the parameter initialization and it does not suffer from the well-known permutation problem. Our work is also different from [16] as we exploit the mixed group sparsity constraint in the optimization algorithm in order to automatically select the most representative spectral components in the GSSM. Especially, unlike all existing approaches [16–18] where the estimated variances of each source are independently constrained by NMF, we propose to constrain the total variances of all sources altogether so as the parameters are estimated in a more global consistent way.

The structure of the rest of the paper is as follows. We introduce the problem formulation and modeling in Sect. 2. We then present the proposed multichannel algorithm with the details of parameter estimation in Sect. 3. The effectiveness of the proposed approach are validated in Sect. 4. Finally we conclude in Sect. 5.

2 Problem Formulation and Modeling

Let us denote by $\mathbf{c}_j(t) \in \mathbb{R}^{I \times 1}$ the contribution of j -th source, $j = 1, 2, \dots, J$, to an array of I microphones, and $\mathbf{x}(t) = \sum_{j=1}^J \mathbf{c}_j(t)$ the mixture signal. The objective of source separation is to estimate $\mathbf{c}_j(t)$ given $\mathbf{x}(t)$. As most source separation algorithms operate in the frequency domain, we denote by $\mathbf{c}_j(n, f)$ and $\mathbf{x}(n, f)$ the short-term Fourier transform (STFT) of $\mathbf{c}_j(t)$ and $\mathbf{x}(t)$, respectively, where $n = 1, 2, \dots, N$ presents time frame index and $f = 1, 2, \dots, F$ the frequency bin index. The mixing model in the frequency domain writes:

$$\mathbf{x}(n, f) = \sum_{j=1}^J \mathbf{c}_j(n, f). \quad (1)$$

2.1 General Gaussian Modeling Framework

We consider the nonstationary Gaussian modeling framework [15], where $\mathbf{c}_j(n, f)$ is modeled as a zero-mean complex Gaussian random vector $\mathbf{c}_j(n, f) \sim \mathcal{N}_c(\mathbf{0}, \boldsymbol{\Sigma}_j(n, f))$. Here $\mathbf{0}$ denotes a $I \times 1$ vector of zeros, and the covariance matrix $\boldsymbol{\Sigma}_j(n, f)$ is factorized as

$$\boldsymbol{\Sigma}_j(n, f) = v_j(n, f) \mathbf{R}_j(f), \quad (2)$$

where $v_j(n, f)$ are scalar time-dependent *variances* encoding the spectro-temporal power of the sources and $\mathbf{R}_j(f)$ are time-independent $I \times I$ *spatial covariance matrices* encoding their spatial characteristics. Under the assumption that the source images are statistically independent, the mixture vector $\mathbf{x}(n, f)$ also follows a zero-mean multivariate complex Gaussian distribution with the covariance matrix computed as

$$\boldsymbol{\Sigma}_{\mathbf{x}}(n, f) = \sum_{j=1}^J v_j(n, f) \mathbf{R}_j(f). \quad (3)$$

Denoting by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n, f) = \mathbb{E}(\mathbf{x}(n, f)\mathbf{x}^H(n, f))$ the empirical covariance matrix, which can be numerically computed by local averaging over neighborhoods of (n, f) [15, 16] the negative log-likelihood is computed as

$$\mathcal{L}(\theta) = \sum_{n, f} \text{tr}(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n, f)\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n, f)) + \log \det(\pi\boldsymbol{\Sigma}_{\mathbf{x}}(n, f)), \quad (4)$$

where $\det()$ presents the matrix determinant. Under this model, the parameters $\{v_j(n, f), \mathbf{R}_j(f)\}_{j, n, f}$ can be estimated in the Maximum likelihood (ML) sense by minimizing $\mathcal{L}(\theta)$. Then the STFT coefficients of the source images are obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering as

$$\hat{\mathbf{c}}_j(n, f) = v_j(n, f) \mathbf{R}_j(f) \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n, f) \mathbf{x}(n, f). \quad (5)$$

Finally, the estimated time-domain source images $\hat{\mathbf{c}}_j(t)$ can be obtained by performing the inverse STFT of $\hat{\mathbf{c}}_j(n, f)$.

2.2 NMF-Based Spectral Model

As mentioned earlier, NMF has been widely applied to single channel audio source separation where the spectrogram of the mixture is factorized by two smaller matrices known as the spectral dictionary and the activation [11]. When adapting NMF to the considered Gaussian modeling framework, the nonnegative source variance $v_j(n, f)$ can be approximated as

$$v_j(n, f) \approx \hat{v}_j(n, f) = \sum_{k=1}^{K_j} w_{jfk} h_{jkn}, \quad (6)$$

where w_{jfk} is an entry of the spectral basis matrix $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$, h_{jkn} is an entry of the activation matrix $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$, and K_j the number of latent components in the NMF model to model the j -th source. Given the matrix of the source variances $\mathbf{V}_j = \{v_j(n, f)\}_{n,f} \in \mathbb{R}_+^{F \times N}$, the corresponding NMF parameters can be estimated by minimizing the Itakura-Saito divergence, which offers scale invariant property, as

$$\min_{\mathbf{H}_j \geq 0, \mathbf{W}_j \geq 0} D(\mathbf{V}_j \| \mathbf{W}_j \mathbf{H}_j), \quad (7)$$

where $D(\mathbf{V}_j \| \mathbf{W}_j \mathbf{H}_j) = \sum_{n=1}^N \sum_{f=1}^F d_{IS}(v_j(n, f) \| \hat{v}_j(n, f))$, and $d_{IS}(x \| y) = \frac{x}{y} - \log(\frac{x}{y}) - 1$.

The parameters $\{\mathbf{W}_j, \mathbf{H}_j\}$ are usually initialized with random non-negative values and are iteratively updated via the well-known multiplicative update (MU) rules [10, 11]. To our best knowledge, this NMF formulation for the source variances within the presenting Gaussian modeling framework was first presented in [17], and then further discussed in [18].

3 Proposed Approach

We will first introduce the GSSM construction in Sect. 3.1. We then discuss the novel GSSM fitting with a sparsity constraint in Sect. 3.2. Finally, we present the derived EM algorithm in Sect. 3.3. *Note that we focus on NMF as spectral model in this paper, however the whole idea of the proposed approach can actually be used for other spectral models than NMF.*

3.1 GSSM Construction

We assume that the types of sources in the mixture are known and some examples of them are available. This is actually feasible in practice as we often know at least what type of target signal to extract from a recording, *e.g.*, in the speech enhancement usecase, one target source is speech and another is noise. Let the spectrogram of p -th example of the j -th source $s_j^p(t)$ be denoted by \mathbf{V}_j^p . First,

\mathbf{V}_j^p is used to learn a corresponding NMF spectral dictionary, denoted by \mathbf{W}_j^p , by optimizing the criterion similarly to (7):

$$\min_{\mathbf{H}_j^p \geq 0, \mathbf{W}_j^p \geq 0} D(\mathbf{V}_j^p \| \mathbf{W}_j^p \mathbf{H}_j^p) \quad (8)$$

where \mathbf{H}_j^p is the time activation matrix. Given \mathbf{W}_j^p for all examples $p = 1, \dots, P_j$ of the j -th source, the GSSM for the j -th source is constructed as

$$\mathbf{U}_j = [\mathbf{W}_j^1, \dots, \mathbf{W}_j^{P_j}], \quad (9)$$

then the GSSM for all the sources is computed by

$$\mathbf{U} = [\mathbf{U}_1, \dots, \mathbf{U}_J]. \quad (10)$$

As an example for speech and noise separation, in practical implementation we may need several speech examples from different male voices and female voices (*e.g.*, $P_1 = 4$), and several examples of different types of noise such as those from outdoor environment, cafeteria, waterfall, street (*e.g.*, $P_2 = 5$).

3.2 GSSM Fitting with Mixed Group Sparsity Constraint

The GSSM for all sources \mathbf{U} constructed in (10) becomes a large matrix when the number of examples P_j for each source increases, and it is actually a redundant dictionary since different examples may share similar spectral patterns. Thus in the NMF model fitting, sparsity constraint is naturally needed so as to automatically select only a subset of \mathbf{U} which represents the sources in the mixture [19]. In other words, the model-based spectrogram of the mixture $\tilde{\mathbf{V}} = \sum_{j=1}^J \mathbf{V}_j$ is decomposed by solving the following optimization problem

$$\min_{\mathbf{H} \geq 0} D(\tilde{\mathbf{V}} \| \mathbf{U}\mathbf{H}) + \lambda \Omega(\mathbf{H}) \quad (11)$$

where $\Omega(\mathbf{H})$ presents a penalty function imposing sparsity on the activation matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, and λ is a trade-off parameter determining the contribution of the penalty. Note that unlike existing approaches [16–18] where the matrix of the estimated variances of each source \mathbf{V}_j was constrained by NMF independently as (7), we propose here to constrain the matrix of the total variances of all sources $\tilde{\mathbf{V}}$ altogether by (11). This can be seen as an additional NMF-based separation step applied on the source variances, while the existing works does not perform any additional separation of the variances, but more like denoising of the already separated variances. In our recent work [9] we investigated a general form for the penalty function as

$$\Omega(\mathbf{H}) = \alpha \sum_{g=1}^G \log(\epsilon + \|\mathbf{H}_g\|_1) + (1 - \alpha) \sum_{k=1}^K \log(\epsilon + \|\mathbf{h}_k\|_1). \quad (12)$$

The first term on the right-hand side of Eq.(12) presents the *block* sparsity-inducing penalty (which enforces the activation of relevant examples only while

omitting irrelevant ones since their corresponding activation block in \mathbf{H} will likely converge to zero), the second term presents the *component* sparsity-inducing penalty (which enforces the activation of relevant components in \mathbf{U} only), $\alpha \in [0, 1]$ weights the contribution of each term. In (12), $\mathbf{h}_k \in \mathbb{R}_+^{1 \times N}$ is a row (or component) of \mathbf{H} , \mathbf{H}_g is a subset of \mathbf{H} representing the activation coefficients for g -th block, G is the total number of blocks, ϵ is a non-zero constant (*i.e.*, set by $3 * 10^{-6}$ in our experiment), and $\|\cdot\|_1$ denotes ℓ_1 -norm operator (*i.e.*, the maximum absolute column sum of the matrix). In the considered setting, a block represents one training example for a source and G is the total number of used examples (*i.e.*, $G = \sum_{j=1}^J P_j$).

By putting (12) into (11), we have a complete criterion for estimating \mathbf{H} given $\tilde{\mathbf{V}}$ and the pre-trained spectral model \mathbf{U} . The derived MU rule for updating \mathbf{H} is presented in [9] and summarized in the Algorithm 1, where \mathbf{Y}_g is a uniform matrix of the same size as \mathbf{H}_g and \mathbf{z}_k a uniform row vector of the same size as \mathbf{h}_k .

3.3 Proposed Multichannel Algorithm

Within the presenting Gaussian modeling framework, EM algorithm has been derived to estimate the parameters $\{v_j(n, f), \mathbf{R}_j(f)\}_{j,n,f}$ by considering the set of hidden STFT coefficients of all the source images $\{\mathbf{c}_j(n, f)\}_{n,f}$ as the *complete data*. In the E-step, the Wiener filter $\mathbf{Q}_j(n, f)$ and the expected covatiance $\tilde{\Sigma}_j(n, f)$ of the spatial images of the j -th source are computed. Then in the M-step, $\mathbf{R}_j(f)$ and $v_j(n, f)$ are updated by minimizing (4), which gives close-form solution. The detail of this EM derivation can be found in [15, 18]. For the proposed approach as far as the GSSM concerned, the E-step of EM algorithm remains the same. In the M-step, we additionally perform the optimization defined in (11) by MU rules so as the estimated intermediate source variance $v_j(n, f)$ is further updated with the supervision of the GSSM. The detail of EM algorithm for the parameter estimation is summarized in Algorithm 1.

4 Experiments

4.1 Dataset and Settings

We validated the performance of the proposed algorithm in a popular but very important speech enhancement usecase where we knew already two types of sources in the mixture: speech and noise. For a better comparison with the state of the art, we used the benchmark development dataset of the ‘‘Two-channel mixtures of speech and real-world background noise’’ (BGN) task¹ within the SiSEC 2016 [1]. This devset contained stereo mixtures of 10 second duration and 16 KHz sampling rate. They were mixed from male/female speeches and noises recorded from six different public environments: cafeteria (Ca), square (Sq), and subway (Su). Overall there were nine mixtures of two sources: three with Ca

¹ <https://sisec.inria.fr/sisec-2016/bgn-2016/>.

Algorithm 1. EM algorithm for the parameter update

```

// E-step (perform calculation for all  $j, n, f$ ):
 $\Sigma_j(n, f) = v_j(n, f) \mathbf{R}_j(f)$  // equation (2)
 $\Sigma_{\mathbf{x}}(n, f) = \sum_{j=1}^J v_j(n, f) \mathbf{R}_j(f)$  // equation (3)
 $\mathbf{Q}_j(n, f) = \Sigma_j(n, f) \Sigma_{\mathbf{x}}^{-1}(n, f)$ 
 $\hat{\Sigma}_j(n, f) = \mathbf{Q}_j(n, f) \hat{\Sigma}_{\mathbf{x}}(n, f) \mathbf{Q}_j^H(n, f) + (\mathbf{I} - \mathbf{Q}_j(n, f)) \Sigma_j(n, f)$ 

// M-step (perform calculation for all  $j, n, f$ )
 $\mathbf{R}_j(f) = \frac{1}{N} \sum_{n=1}^N \frac{1}{v_j(n, f)} \hat{\Sigma}_j(n, f)$  // update  $\mathbf{R}_j(f)$ 
 $v_j(n, f) = \frac{1}{I} \text{tr}(\mathbf{R}_j^{-1}(f) \hat{\Sigma}_j(n, f))$  // update  $v_j(n, f)$ 

 $\mathbf{V}_j = \{v_j(n, f)\}_{n, f}$ 
 $\tilde{\mathbf{V}} = \sum_{j=1}^J \mathbf{V}_j$ 
// Perform NMF in the M-step to further constrain source spectra by the GSSM
for  $iter = 1, \dots, \text{MU-iteration}$  do
  for  $g = 1, \dots, G$  do
     $\mathbf{Y}_g \leftarrow \frac{1}{\epsilon + \|\mathbf{H}_g\|_1}$ 
  end for
   $\mathbf{Y} = [\mathbf{Y}_1^T, \dots, \mathbf{Y}_G^T]^T$ 
  for  $k = 1, \dots, K$  do
     $\mathbf{z}_k \leftarrow \frac{1}{\epsilon + \|\mathbf{h}_k\|_1}$ 
  end for
   $\mathbf{Z} = [\mathbf{z}_1^T, \dots, \mathbf{z}_K^T]^T$ 
   $\hat{\mathbf{V}} = \mathbf{U}\mathbf{H}$ 
   $\mathbf{H} \leftarrow \mathbf{H} \odot \left( \frac{\mathbf{U}^T (\tilde{\mathbf{V}} \odot \hat{\mathbf{V}}^{-2})}{\mathbf{U}^T (\tilde{\mathbf{V}}^{-1}) + \lambda (\alpha \mathbf{Y} + (1-\alpha) \mathbf{Z})} \right)^{\frac{1}{2}}$  // MU rule
end for

 $v_j(n, f) = [\mathbf{U}_j \mathbf{H}_j]_{n, f}$  // updating constrained spectra

```

noise, four with Sq noise, and two with Su noise. The signal-to-noise ratio was drawn randomly per mixture between -17 and $+12$ dB by the dataset creators.

For training the GSSM for speech and noise, we took one male voice and two female voices from the SiSEC 2015². These three speech examples were also 10-s long. Five noise training examples were extracted from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND)³. Again they were 10-s long and contained three types of environmental noise: cafeteria, square, metro. We made sure that these examples used for GSSM training are different from those in the devset, which were used for testing. The number of NMF components in \mathbf{W}_j^p for each speech example was set to 32, while that for noise example was 16, and each \mathbf{W}_j^p was obtained after 20 MU iterations. Other parameter settings were as follows. The STFT window length of 50% overlapping was 1024. The spatial covariance matrix $\mathbf{R}_j(f)$ for noise was initialized following the diffuse

² <https://sisec.inria.fr/sisec-2015/2015-underdetermined-speech-and-music-mixtures/>.

³ <http://parole.loria.fr/DEMAND/>.

model, while $\mathbf{R}_j(f)$ for speech was initialized following the direct+diffuse model [15] assuming the direction-of-arrival (DoA) for speech source is 90° . For testing, we firstly varied the number of EM and MU iterations and found that generally the convergence obtained after about 10 iterations. Specifically, the best result was obtained by 15 EM iterations and 10 MU iterations. The trade-off parameter λ determining the contribution of the sparsity-inducing penalty in (11) and the factor α weighting the contribution of each penalty term in (12) were tested with different values: $\lambda = \{1, 10, 25, 50, 100, 200, 500\}$, $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ and we found that the algorithm is less sensitive to the choice of α , while more sensitive to the choice of λ and $\lambda > 10$ decreases the separation performance. The best choice for these parameters are $\lambda = 10, \alpha = 0.2$.

4.2 Comparison Results

We compare the speech separation performance of the proposed approach with several state of the art and baseline algorithms as follows:

- Liu’s method: the algorithm performed Time Difference of Arrival (TDOA) clustering based on GCC-PHAT and participated to the same SiSEC 2016 campaign [1]. The separation results were submitted by the authors and evaluated by the SiSEC organizers.
- Wood’s method [20]: this algorithm firstly applied NMF to the magnitude spectrograms of the mixtures with channels concatenated in time. Each dictionary atom was then clustered to either the speech or the noise according to its spatial origin. Again the separation results for devset were submitted to the SiSEC 2016 campaign and evaluated by the SiSEC organizers.
- Arberet’s method [17]: using the similar Gaussian modeling framework, the algorithm further constrained the estimated source variances by unsupervised NMF where the parameters were obtained by optimizing the criterion (7) in the M-step of EM algorithm instead of (11) like us. Such optimization criterion was implemented by Ozerov *et al.* in [18].
- Baseline 1: the presenting GSSM + full-rank spatial covariance approach but there is no sparsity constraint in (11) (*i.e.*, $\lambda = 0$). This is to investigate the importance of the sparsity constraint (12) in the GSSM fitting.
- Baseline 2: the presenting GSSM + full-rank spatial covariance approach but the estimated variances of each source \mathbf{V}_j are further constrained by NMF where the corresponding activation matrix \mathbf{H}_j obtained by optimizing the following criterion:

$$\min_{\mathbf{H}_j \geq 0} D(\mathbf{V}_j \| \mathbf{U}_j \mathbf{H}_j) + \lambda \Omega(\mathbf{H}_j) \quad (13)$$

We submitted results obtained by this method to the SiSEC 2016 BGN task and obtained the best performance among other submitting methods in term of the overall signal-to-distortion (SDR) ratio [1].

- Proposed method: the presenting GSSM + full-rank spatial covariance approach where the matrix of the total variances of all sources $\tilde{\mathbf{V}}$ is constrained by NMF and the activation matrix is obtained by optimizing (11). EM algorithm for the corresponding parameter updates is present in Algorithm 1.

The separation performance (for speech source only) for all approaches was evaluated by the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), the signal-to-artifacts ratio (SAR), and the source image-to-spatial distortion ratio (ISR), measured in dB [21]. These values are shown in Table 1 where the higher the better.

Table 1. Average speech separation performance obtained on the devset of the BGN task of SiSEC 2016. Results for Liu’s method and Wood’s method were submitted by the authors [1].

Methods	SDR	SIR	SAR	ISR
Liu’s method	−7.0	−1.4	15.0	3.1
Wood’s method [20]	1.9	3.6	3.7	5.1
Arberet’s method [17, 18]	4.4	4.6	12.1	15.9
Baseline 1 (No sparsity constraint)	0.4	−1.1	9.5	8.3
Baseline 2 ($\lambda = 10, \alpha = 0.2$)	7.4	8.9	12.7	11.3
Proposed method ($\lambda = 10, \alpha = 0.2$)	7.7	10.7	11.6	13.9

It is interesting to see that the result obtained by the Baseline 1 is lower than that of Arberet’s method, even the former used the pre-trained GSSM while the later was completely unsupervised. It reveals that the GSSM itself is redundant and contains some irrelevant spectral patterns with the actual sources in the mixture. Thus constraining the source variances by the GSSM without a relevant spectral pattern selection guided by the sparsity penalty is even worse than unsupervised NMF case where the spectral patterns were randomly initialized and then updated by MU rules. The importance of such sparsity penalty is explicitly confirmed by the fact that the result obtained by the Baseline 2 was far more better than that of the Baseline 1. It is also not surprising to see that the Baseline 2 clearly outperforms Arberet’s method as the former exploited additional information about the types of sources in the mixtures so as to learn the GSSM in advance. We also tested the case where the small size dictionary obtained by jointly decomposing all training examples for the target signal, but the performance was lower than the Baseline 2. Finally, the proposed method offers the best separation performance in terms of SDR and SIR, the two important criteria. This confirms the effectiveness of the proposed approach where the GSSM is successfully combined with the spatial covariance model in a unified Gaussian modeling framework. Furthermore, the benefit of the new criterion (11) compared to the conventional one (13) for the NMF parameter estimation is supported. *Our further analysis, which is not described here due to the lack*

of space, shows in addition that with such new criterion, the algorithm is less sensitive to the parameter initialization and the choice of hyper-parameters λ and α as compared to the Baseline 2.

5 Conclusion

We have presented a novel multichannel audio source separation algorithm, which exploits the use of generic source spectral model within the well-established Gaussian modeling framework. Such redundant GSSM can be easily learned from source examples by NMF and shown to be very useful in guiding the source separation. Especially, we have proposed a new optimization criterion in order to better constrain the intermediate source variances estimated in each EM iteration. Experiment with a benchmark dataset from the SiSEC 2016 campaign has confirmed the effectiveness of the proposed approach compared to both the state of the art and the baselines. Motivated by the GSSM, future work can be devoted to extending the current approach so as to exploit in addition the use of a *generic spatial covariance model*, which remains to be defined.

References

1. Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., Fontecave, J.: The 2016 signal separation evaluation campaign. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) LVA/ICA 2017. LNCS, vol. 10169, pp. 323–332. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_31
2. Liutkus, A., Durrieu, J.L., Daudet, L., Richard, G.: An overview of informed audio source separation. In: International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pp. 1–4. IEEE (2013)
3. Ewert, S., Pardo, B., Mueller, M., Plumbley, M.D.: Score-informed source separation for musical audio recordings: an overview. *IEEE Sig. Process. Mag.* **31**(3), 116–124 (2014)
4. Magoarou, L.L., Ozerov, A., Duong, N.Q.K.: Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization. *J. Sig. Process. Syst.* **79**(2), 117–131 (2015)
5. Parekh, S., Essid, S., Ozerov, A., Duong, N.Q.K., Perez, P., Richard, G.: Motion informed audio source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)
6. Souviraà-Labastie, N., Olivero, A., Vincent, E., Bimbot, F.: Multi-channel audio source separation using multiple deformed references. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**, 1775–1787 (2015)
7. Sun, D.L., Mysore, G.J.: Universal speech models for speaker independent single channel source separation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 141–145 (2013)
8. Badawy, D.E., Duong, N.Q.K., Ozerov, A.: On-the-fly audio source separation - a novel user-friendly framework. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(2), 261–272 (2017)

9. Duong, H.T.T., Nguyen, Q.C., Nguyen, C.P., Tran, T.H., Duong, N.Q.K.: Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint. In: Proceedings of the ACM SoICT, pp. 247–251 (2015)
10. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural and Information Processing Systems 13, pp. 556–562 (2001)
11. Févotte, C., Bertin, N., Durrieu, J.L.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Comput.* **21**(3), 793–830 (2009)
12. Mandel, M., Ellis, D.: EM localization and separation using interaural level and phase cues. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 275–278 (2007)
13. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE Trans. Audio Speech Lang. Process.* **19**(3), 516–527 (2011)
14. Kitamura, D., Ono, N., Sawada, H., Kameoka, H., Saruwatari, H.: Efficient multi-channel nonnegative matrix factorization exploiting rank-1 spatial model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 276–280 (2015)
15. Duong, N.Q.K., Vincent, E., Gribonval, R.: Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Audio Speech Lang. Process.* **18**(7), 1830–1840 (2010)
16. Fakhry, M., Svaizer, P., Omologo, M.: Audio source separation in reverberant environments using beta-divergence based nonnegative factorization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(7), 1462–1476 (2017)
17. Arberet, S., Ozerov, A., Duong, N.Q.K., Vincent, E., Gribonval, R., Vanderghenst, P.: Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In: Proceedings of the IEEE ISSPA, pp. 1–4 (2010)
18. Ozerov, A., Vincent, E., Bimbot, F.: A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **20**(4), 1118–1133 (2012)
19. Lefèvre, A., Bach, F., Févotte, C.: Itakura-Saito non-negative matrix factorization with group sparsity. In: Proceedings of the IEEE ICASSP, pp. 21–24 (2011)
20. Wood, S., Rouat, J.: Blind speech separation with GCC-NMF. In: Proceedings of the Interspeech, pp. 3329–3333 (2016)
21. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)



Perceptual Evaluation of Blind Source Separation in Object-Based Audio Production

Philip Coleman^{1,2(✉)}, Qingju Liu², Jon Francombe^{1,3},
and Philip J. B. Jackson²

¹ Institute of Sound Recording, University of Surrey, Guildford, UK
p.d.coleman@surrey.ac.uk

² Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, UK

³ BBC Research & Development, Salford, UK

Abstract. Object-based audio has the potential to enable multimedia content to be tailored to individual listeners and their reproduction equipment. In general, object-based production assumes that the objects—the assets comprising the scene—are free of noise and interference. However, there are many applications in which signal separation could be useful to an object-based audio workflow, e.g., extracting individual objects from channel-based recordings or legacy content, or recording a sound scene with a single microphone array. This paper describes the application and evaluation of blind source separation (BSS) for sound recording in a hybrid channel-based and object-based workflow, in which BSS-estimated objects are mixed with the original stereo recording. A subjective experiment was conducted using simultaneously spoken speech recorded with omnidirectional microphones in a reverberant room. Listeners mixed a BSS-extracted speech object into the scene to make the quieter talker clearer, while retaining acceptable audio quality, compared to the raw stereo recording. Objective evaluations show that the relative short-term objective intelligibility and speech quality scores increase using BSS. Further objective evaluations are used to discuss the influence of the BSS method on the remixing scenario; the scenario shown by human listeners to be useful in object-based audio is the worst-case scenario among those tested.

1 Introduction

Research into blind source separation (BSS), where an estimate of a clean audio source can be obtained knowing only a mixture of sounds, has been active for many years. Generally, the performance of such approaches has been evaluated in terms of the quality of the estimated audio signal after processing, suppression of interference, and absence of artefacts, using tools such as BSS Eval [21] and PEASS [3]. However, in *remixing*, where the estimated audio signal is combined with other audio before being presented to the listener, some separation artefacts

may be masked, increasing the utility for source separation techniques in contexts such as broadcast where high-quality content is required.

Opportunities for source separation are also emerging in the context of object-based audio. Here, instead of content being mastered for a particular production format, a set of *audio objects* is transmitted. Audio objects usually comprise a single ‘clean’ audio channel and corresponding metadata that describe how the audio should ideally be rendered for the end user. Object-based audio allows for customisation of content to each listener’s reproduction setup, and personalisation of content to their personal preferences. However, clean audio objects may not always be available. In this paper, we investigate applications of BSS for clean object estimation in the context of an object-based workflow.

Other recent work has also sought to exploit the potential for using BSS as part of a remix. In [12], perceptual model results were used to show that the speech quality achieved by remixing estimated sources was higher than the quality of the estimated sources in isolation. In [20], subjective tests were conducted to investigate the extent to which users were satisfied by personalising object-based content, with a source separation scenario considered. The MARuSS (Musical Audio Repurposing using Source Separation) project has worked on the problem of musical remix and upmix using deep learning-based BSS [16], including separation of vocals from the remainder of the mix [18], and perceptual evaluation of BSS in the context of remixing [17, 22].

The work described in this paper extends the work by Coleman *et al.* [2, Sect. VII.C] in three ways. First, we investigate two additional BSS algorithms; second, we extend the presentation and discussion of the objective metrics; third, we evaluate the effects of remixing both talkers instead of just the quieter talker as in [2]. The paper is organised as follows. In Sect. 2, we motivate the use of source separation in the context of an object-based production workflow and present the background theory for the BSS approaches implemented. In Sect. 3, we present the results of subjective and objective experiments for speech stimuli. Finally, in Sect. 4 we conclude.

2 Background

In this section, the application scenario for BSS in object-based audio is described, and the BSS methods under test are briefly introduced.

2.1 Object-Based Production Workflow

An object-based scene is composed of a number of audio signals, together with metadata describing how they should be rendered for the end user. Traditionally, it is assumed that the audio signals are clean, that is, not contaminated with artefacts or interference from other sources. Then, in a standard object-based workflow, metadata would be manually authored by the sound designer in post-production. This process is time consuming, both in terms of capturing the required source signals and authoring the metadata. Consequently, a new

workflow stage of *objectification* has recently been proposed [2], wherein audio objects and their metadata are estimated from audio and video signals that may form part of the final audible production or may serve purely as production aids.

Although a strictly object-based production would encode each individual sound source as an object, a commonly used pragmatic approach is to mix close microphone object signals with a channel-based capture of the entire scene. This enables opportunities for editing, remixing or personalising content (compared to a traditional channel-based broadcast of the same mix) and is supported in current standards [4, 7]. Furthermore, if close microphone signals are not available (for example, if there is limited time to set up equipment), BSS can potentially be used to estimate the object signals. In this case, remixing can still take place in post-production. Coleman *et al.* [2] explored two use-cases for audio separation algorithms in object-based production (BSS for speech; beamforming for music). The analysis of the results from the speech use case is extended here.

2.2 Blind Source Separation Methods

Three BSS methods are considered in this paper. The first is a traditional time-frequency (TF) masking method statistically-characterised with a Gaussian mixture model (GMM), where binaural features of inter-aural level difference (ILD) and inter-aural phase difference (IPD) are exploited to iteratively refine the GMM parameters for the separation mask generation [13]. We denote this method as “Mandel”. The second uses similar principles, yet takes into account ILD and IPD as well as mixing vector features [1], and is denoted as “Alinaghi”. Unlike the above methods, with unsupervised learning processes, the third method is based on deep neural networks (DNNs), where the commonly used spectral features and non-linearly-transformed binaural spatial features are fed into a hybrid DNN structure, consisting of convolutional layers and fully-connected layers [11]. The spatial features are iteratively refined using the DNN output. This method is denoted as “Liu”. The training process for Liu was performed on a simulated data set lasting around 12 h in a reverberant room (RT60 640 ms). It is noteworthy that the mixing scenario for training the DNN used in Liu does not correspond to the conditions of the data recorded for the experiments reported in this paper: the talker positions, microphones, and balance between dominant and interfering speakers were all different.

3 Experiments

To investigate the utility of BSS to enable object-based remix of stereo speech content, listening tests were conducted, and objective scores were obtained using predictive perceptual models and signal-based metrics. In this section, the setup for each experiment is described and the results are presented and discussed.

3.1 Speech Stimuli

Performances were recorded in a large recording studio (dimensions $14.55 \times 17.08 \times 6.50$ m; RT60 1.1 s) using a number of microphone techniques [5]. TIMIT sentences [6] spoken simultaneously by two talkers were recorded with a pair of high-quality omnidirectional microphones, 18 cm apart, approximately 4 m from the talkers. Lapel microphone signals were also recorded, to provide close reference signals for the objective evaluation. In the stereo recording, one talker was 4.6 dB louder than the other, according to the relative estimated signal-to-interference ratios (SIRs) calculated by BSS Eval [21]. Therefore, for the subjective tests, the application scenario was to estimate the speech uttered by the *quieter* talker, i.e., to allow the talker at -4.6 dB SIR to be better level-balanced in post-production.

3.2 Subjective Evaluation

Listening tests were conducted using a standardized “0+5+0” surround setup [8] with Genelec 8020B loudspeakers in an acoustically-treated listening room (RT60 conforming to ITU recommendation BS.1116-3 [10] above 400 Hz). In the subjective experiment (also reported by Coleman *et al.* [2]), listeners were presented with the stereo recording (left and right signals rendered directly to $\pm 30^\circ$) and a BSS-estimated object extracted by Mandel’s method. They were asked to “*adjust the slider [controlling the extracted object level] until the target talker is as clear and easy to understand as possible, whilst ensuring that the overall audio quality remains at an acceptable level (compared to the reference).*” The BSS object was rendered at azimuths $\{0, 15, 30^\circ\}$, with three repeats, giving nine ratings per listener. Additionally, a threshold of audibility was determined: listeners were presented with the same stimulus (object at 0°) and asked to “*adjust the [object] level to the point immediately before the mix is different to the reference.*” This part also included three repeats. Ten experienced listeners completed the tests, of whom seven were native English speakers. The results are shown as boxplots for each participant (showing the range of the data, the quartiles, and medians with 95% confidence notches) in Fig. 1. It can be seen that for most participants, the thresholds of audibility and acceptability are significantly different. The results of participant 5 were removed from further analysis due to the large variance in threshold judgements. The results from the remaining participants were normally distributed, both for audibility (Lilliefors test, $p = 0.08$) and acceptability (Lilliefors test, $p > 0.50$). The mean mixing level averaged over azimuth (0.2 dB relative to the reference) differed significantly from the threshold of audibility (-14.9 dB) according to a two-sample t -test ($t = 9.73$, $p < 0.01$). There is therefore a region (15 dB range) in which the BSS-extracted object is audible and makes the target talker clearer, while maintaining acceptable quality. An analysis of variance (ANOVA) showed no significant effects of azimuth ($F = 0.85$, $p = 0.43$) or repeat ($F = 0.98$, $p = 0.38$) on the acceptability threshold.

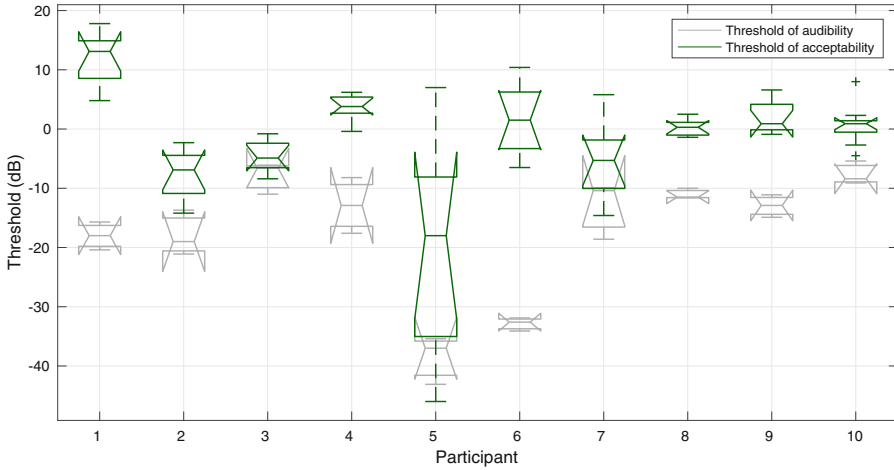


Fig. 1. Box plots of perceptual thresholds of audibility and acceptability, for remixing Talker 2 (estimated by Mandel’s method). Notches show 95% confidence intervals around the median [14].

3.3 Objective Evaluation

Objective evaluation was conducted to support the listening test analysis. The objective evaluation employed two metrics: short-time objective intelligibility (STOI) [19], in the range $[0,1]$, which predicts speech intelligibility; and perceptual evaluation of speech quality (PESQ) [15], in the range $[-0.5, 4.5]$, which predicts speech quality. The mono sum of the stereo reference, mixed with the extracted speech object at relative levels in the range ± 20 dB, was presented to the models. Prior to processing, all signals were downsampled to 16 kHz and each test mixture was loudness-matched to the reference lapel microphone signal using a Matlab implementation of [9]. Objective scores were calculated as the average of scores obtained individually for each sentence in the recording (four clips with average duration 3.2s for Talker 2 as target; five clips with average duration 2.7s for Talker 1 as target). Relative STOI and PESQ scores (target talker score – interfering talker score) were calculated for Mandel (corresponding to the subjective experiment described above), Alinaghi, and Liu.

The STOI scores are plotted in Fig. 2 for Talker 1 (left) and Talker 2 (right). The -0.1 relative STOI score for the target talker in the original stereo recording (relative SIR -4.6 dB) confirms that the interfering talker is more intelligible than the target talker before mixing the extracted object into the scene. By increasing the object’s level in the mixture, the relative STOI scores increase. At the mean mixing level determined in the subjective tests using Mandel’s method, the relative scores are both positive, implying that introducing the separated speech into the mix has resulted in an enhancement in speech intelligibility. Moreover, Mandel’s method, as tested subjectively, performed worst among the three methods tested. For both talkers, Alinaghi was predicted to

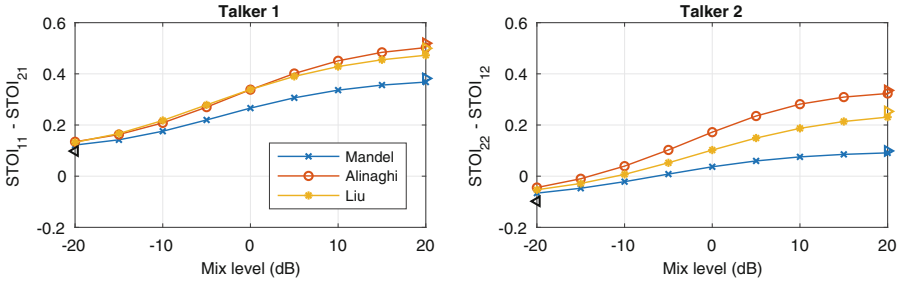


Fig. 2. Relative STOI scores, where $STOI_{AB}$ denotes the score for target Talker A when adjusting the level of Talker B. The mixture score (\triangleleft) and object-only scores for each method (\triangleright) are also marked.

give the greatest improvement in speech intelligibility; Liu was ranked second while improving upon Mandel.

The PESQ scores are plotted in Fig. 3 for Talker 1 (left) and Talker 2 (right). For all methods, and both Talkers, the relative PESQ scores increase with mixing level, implying that the separated speech is closer to the reference lapel microphone signal than the mixture. However, the subjective results indicate that the relative PESQ score does not fully convey the listening experience of the remixed speech, because the listeners identified a threshold of acceptability above which the target quality was not acceptable. For the PESQ scores, Mandel also performs worst among the methods tested. Alinaghi performs best for Talker 2, and well for Talker 1, although the relative scores for Liu are best for Talker 1 above a mix level of 0 dB. This performance is analysed further in terms of the signal-based metrics discussed below.

The objective evaluation was extended by obtaining the signal-to-interference, -artefact, and -distortion ratios (SIR, SAR, and SDR respectively) for each method, at each remix level, for both talkers. These results are plotted

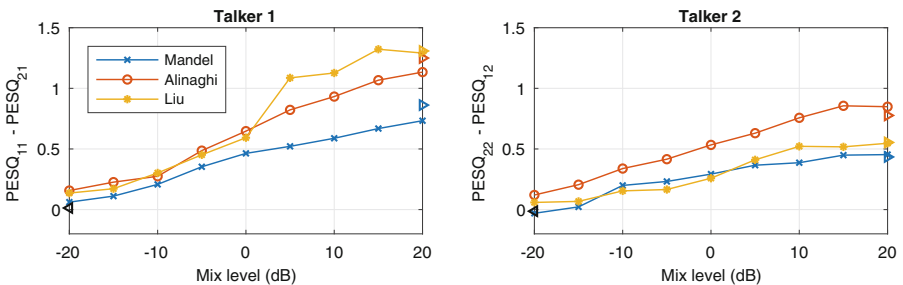


Fig. 3. Relative PESQ scores, where $PESQ_{AB}$ denotes the score for target Talker A when adjusting the level of Talker B. The mixture score (\triangleleft) and object-only scores for each method (\triangleright) are also marked.

in Fig. 4 for Talker 1 (left) and Talker 2 (right). Scores are absolute (i.e., only the target talker is taken into account). The SIR scores show that Mandel performs worst among the methods tested. For Talker 2, Liu is close to Mandel but slightly better, while Alinaghi performs over twice as well. The scores for Talker 1 are higher overall. Liu and Alinaghi give similar performance, but Liu exceeds Alinaghi for mix levels above 0 dB. These trends closely mirror the relative PESQ scores shown in Fig. 3, suggesting that SIR is the dominant signal quality property contributing to the relative PESQ scores.

The SAR scores have different trends for each talker. For Talker 2 (quieter in original mix), the SAR decreases with mix level, which may explain why listeners found there to a trade off between speech intelligibility and target quality. On the other hand, SAR actually increases with mix level for Talker 1 (apart from Liu, which remains approximately stable with mix level). Thus, if Mandel or Alinaghi were applied to remix Talker 1, the thresholds of acceptable quality would likely be higher than those reported in the subjective tests described above. Finally, the SDR scores for each method and talker increase with mix level, with Alinaghi outperforming Mandel and Liu in each case.

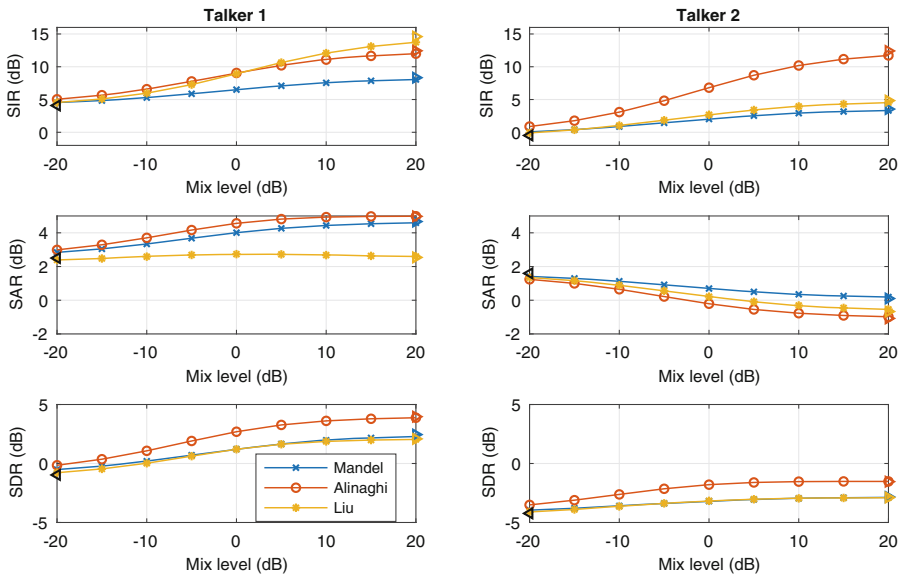


Fig. 4. Signal-based evaluations of SIR (top row), SAR (middle), and SDR (middle), adjusting the level for Talker 1 (left) and Talker 2 (right). The mixture score (\triangleleft) and object-only scores for each method (\triangleright) are also marked.

4 Conclusions

Subjective and objective results were presented to evaluate the performance of speech remixing, enabled by BSS. Such remixing has applications in object-based audio, where a producer may wish to make adjustments to a mix not facilitated by the available microphone signals, or an object-based renderer may adjust a mix based on a listener's personal preference or accessibility settings. The subjective scores showed that, in a challenging scenario with two interfering talkers, the quieter talker could be made clearer by mixing in an object estimated by BSS, while retaining acceptable audio quality. STOI, an objective perceptual model, was used to verify that the relative speech intelligibility increased with mix level. The SAR for Talker 2 for Mandel's method (corresponding to the subjective test scenario) reduced with mix level, which could explain why listeners felt that the quality degraded after the mean acceptability threshold at a mix level of 0.2 dB.

Further predictions of speech intelligibility, quality, and signal-based metrics of SIR, SAR and SDR suggested that the scenario considered for the subjective tests was the worst case among the two talkers and the three tested BSS algorithms (Mandel, Alinaghi, and Liu). In particular, the objective metrics suggested that Alinaghi may perform well compared to Mandel. Furthermore, as the DNN in Liu was trained on binaural features (including ILD), yet omnidirectional microphones were used here, the method would likely perform better if the training conditions were closer to the application example studied.

Further work should investigate whether the perceptual acceptability thresholds increase for the other methods tested. Other aspects not tested here that could be developed in future include respatalisation of BSS-estimated sources, and the applications to other sound sources, e.g. musical instruments. Finally, the possibility of creating an object-based scene with only BSS-extracted sources (i.e., no underlying channel-based recording) could be investigated. In [2, Sect. III.C], we made some informal comments about this scenario; in general the BSS-extraction allows for respatalization and some level control of the mixed sources, but degradations in the target quality due to the BSS are more exposed.

Acknowledgements. This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). Relevant data can be accessed via <https://doi.org/10.15126/surreydata.00845514>.

References

1. Alinaghi, A., Jackson, P.J., Liu, Q., Wang, W.: Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(9), 1434–1448 (2014)
2. Coleman, P., Franck, A., Francombe, J., Liu, Q., de Campos, T., Hughes, R., Menzies, D., Galvez, S., Tang, Y., Woodcock, J., et al.: An audio-visual system for object-based audio: from recording to listening. *IEEE Trans. Multimedia* (2018, in press). <https://doi.org/10.1109/TMM.2018.2794780>

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. European Telecommunications Standards Institute: Digital audio compression (AC-4) standard part 2: immersive and personalized audio, ETSI-TS-103-190-2. European Telecommunications Standards Institute (2015)
5. Francombe, J., Brookes, T., Mason, R., Flindt, R., Coleman, P., Liu, Q., Jackson, P.: Production and reproduction of program material for a variety of spatial audio formats. In: 138 Convention Audio Engineering Society, Warsaw, Poland (2015)
6. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Linguistic Data Consortium, Philadelphia (1993)
7. Herre, J., Hilpert, J., Kuntz, A., Plogsties, J.: MPEG-H 3D audio – the new standard for coding of immersive spatial audio. *IEEE J. Sel. Top. Sig. Process.* **9**(5), 770–779 (2015)
8. ITU-R: Recommendation ITU-R BS.2051-0: Advanced sound system for programme reproduction. International Telecommunication Union (2014)
9. ITU-R: Recommendation BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunication Union (2015)
10. ITU-R: Recommendation ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems. International Telecommunication Union (2015)
11. Liu, Q., Xu, Y., Coleman, P., Jackson, P.J.B., Wang, W.: Iterative deep neural networks for speaker-independent binaural blind speech separation. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Accepted 2018
12. Liu, Q., Wang, W., Jackson, P.J., Cox, T.J.: A source separation evaluation method in object-based spatial audio. In: *Signal Processing Conference (EUSIPCO)*, 2015 23rd European, pp. 1088–1092. IEEE (2015)
13. Mandel, M.I., Weiss, R.J., Ellis, D.P.W.: Model-based expectation-maximization source separation and localization. *IEEE Trans. Audio Speech Lang. Process.* **18**(2), 382–394 (2010)
14. McGill, R., Tukey, J.W., Larsen, W.A.: Variations of box plots. *Am. Stat.* **32**(1), 12–16 (1978)
15. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752. IEEE, Salt Lake City, May 2001
16. Roma, G., Grais, E.M., Simpson, A.J., Plumbley, M.D.: Music remixing and upmixing using source separation. In: *Proceedings of the 2nd AES Workshop on Intelligent Music Production* (2016)
17. Simpson, A.J.R., Roma, G., Grais, E.M., Mason, R.D., Hummersone, C., Plumbley, M.D.: Psychophysical evaluation of audio source separation methods. In: Tichavský, P., Babaie-Zadeh, M., Michel, O.J.J., Thirion-Moreau, N. (eds.) *LVA/ICA 2017. LNCS*, vol. 10169, pp. 211–221. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-53547-0_21
18. Simpson, A.J.R., Roma, G., Plumbley, M.D.: Deep karaoke: extracting vocals from musical mixtures using a convolutional deep neural network. In: Vincent, E., Yeredor, A., Koldovský, Z., Tichavský, P. (eds.) *LVA/ICA 2015. LNCS*, vol. 9237, pp. 429–436. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-22482-4_50

19. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* **19**(7), 2125–2136 (2011)
20. Torcoli, M., Herre, J., Paulus, J., Uhle, C., Fuchs, H., Hellmuth, O.: The adjustment/satisfaction test (a/st) for the subjective evaluation of dialogue enhancement. In: *Audio Engineering Society Convention 143*. Audio Engineering Society (2017)
21. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
22. Wierstorf, H., Ward, D., Mason, R., Grais, E.M., Hummersone, C., Plumbley, M.D.: Perceptual evaluation of source separation for remixing music. In: *Audio Engineering Society Convention 143* (2017)



Muticriteria Decision Making Based on Independent Component Analysis: A Preliminary Investigation Considering the TOPSIS Approach

Guilherme Dean Pelegrina¹(✉), Leonardo Tomazeli Duarte²,
and João Marcos Travassos Romano¹

¹ School of Electrical and Computer Engineering (FEEC),
University of Campinas (UNICAMP), Campinas, Brazil
pelegrina@decom.fee.unicamp.br, romano@dmo.fee.unicamp.br

² School of Applied Sciences (FCA),
University of Campinas (UNICAMP), Limeira, Brazil
leonardo.duarte@fca.unicamp.br

Abstract. This work proposes the application of independent component analysis to the problem of ranking different alternatives by considering criteria that are not necessarily statistically independent. In this case, the observed data (the criteria values for all alternatives) can be modeled as mixtures of latent variables. Therefore, in the proposed approach, we perform ranking by means of the TOPSIS approach and based on the independent components extracted from the collected decision data. Numerical experiments attest the usefulness of the proposed approach, as they show that working with latent variables leads to better results compared to already existing methods.

Keywords: Multi-criteria decision making · Dependent criteria
Independent component analysis · Latent variables · TOPSIS

1 Introduction

Many practical situations in multicriteria decision making (MCDM) consist in obtaining a ranking of a set of alternatives based on their evaluation according to a set of criteria [1, 2]. The main difference between the existing methods that perform ranking in MCDM is related to the criteria aggregation procedure. For instance, a natural way to perform aggregation is to consider a simple weighted sum [2] for all criteria and for a given alternative. Another strategy can be found

The authors would like to thank the São Paulo Research Foundation (FAPESP - process n. 2016/21571-4 and 2017/23879-9) for the financial support. The authors also thank the National Council for Scientific and Technological Development (CNPq, Brazil) for funding their research.

in TOPSIS method (TOPSIS stands for Technique for Order Preferences by Similarity to an Ideal Solution) [3]. In this method, one firstly defines a positive and a negative ideal alternative. Then, aggregation for a given alternative is done by calculating the Euclidean distances between the alternative under evaluation and the (positive and negative) ideal alternatives.

The original versions of the aforementioned approaches do not take into account any relation among criteria, which may lead to biased results in the aggregation step. Indeed, if, for instance, there are two criteria strongly correlated which are governed by a latent factor, then such a latent factor will have a strong influence on the aggregation step. In view of this inconvenient, there are some methods that try to deal with possible relations among the observed criteria [4–8]. Among them, an interesting approach is an extended version of TOPSIS [5, 7, 8]. In this version, instead of considering the Euclidean distance in the aggregation step, one applies the Mahalanobis distance. Therefore, the calculation of the distance measure takes into account the covariance matrix among criteria.

However, a question that arises is whether the information about the covariance among criteria is sufficient to mitigate the biased effect of dependent criteria. Motivated by this question, this paper proposes a novel three-step procedure to deal with correlated criteria in decision making problems. In the first step of our proposal, we formulate the problem as a Blind Source Separation (BSS) [9] problem and apply an Independent Component Analysis (ICA) method to estimate the latent variables. The second step comprises the elimination of permutation and/or scale ambiguities provided by ICA. In the third step, we perform the TOPSIS approach based on the Euclidean distance on the estimated latent variables in order to obtain a global evaluation of the alternatives, thus allowing a final ranking. Aiming at verifying the proposed ICA-TOPSIS approach, we performed numerical experiments on synthetic data and compared the results obtained by our approach and the TOPSIS based on Mahalanobis distance.

The rest of this paper is organized as follows. Section 2 discusses the main theoretical aspects about multicriteria decision making and blind source separation problems. Then, in Sect. 3, we present the proposed ICA-TOPSIS approach. The numerical experiments are described in Sect. 4. Finally, in Sect. 5, we present our conclusions and future perspectives.

2 Theoretical Background

This section presents the theoretical aspects involved in multicriteria decision making and blind source separations problems.

2.1 Multicriteria Decision Making Problems and TOPSIS Method

The most relevant problems in MCDM consist in ranking a set of K alternatives ($A = [A_1, A_2, \dots, A_K]$) based on a set of M criteria ($C = [C_1, C_2, \dots, C_M]$). For each alternative A_i , $v_{i,j}$ represents its evaluation with respect to the criterion C_j .

Therefore, in a MCDM problem, we often face with the following decision matrix (or decision data):

$$\mathbf{V} = \begin{matrix} & C_1 & C_2 & \dots & C_M \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_K \end{matrix} & \begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,M} \\ v_{2,1} & v_{2,2} & \dots & v_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{K,1} & v_{K,2} & \dots & v_{K,M} \end{bmatrix} \end{matrix} \quad (1)$$

Based on the decision matrix \mathbf{V} and the set of weights $\mathbf{w} = [w_1, w_2, \dots, w_M]$, which represent the ‘‘importance’’ of criterion C_j in the decision problem, the goal is to aggregate $v_{i,j}$, $j = 1, \dots, M$ in order to obtain a global evaluation for each alternative A_i and, then, to establish a ranking.

Several methods have been developed to deal with MCDM problems. Among them, a widely used one is the TOPSIS, developed by Hwang and Yoon [3]. The main idea of this method is to determine the ranking based on the distances between each alternative and the (positive and negative) ideal solutions, as will be described in the sequel. The following steps describe the algorithm¹:

1. The first step comprises the normalization of each evaluation $v_{i,j}$, given by

$$r_{i,j} = \frac{v_{i,j}}{\sqrt{\sum_{i=1}^K v_{i,j}^2}}, \quad i = 1, \dots, K, \quad j = 1, \dots, M. \quad (2)$$

2. Based on $r_{i,j}$, we calculate the weighted normalized evaluation, given by

$$p_{i,j} = w_j r_{i,j}, \quad i = 1, \dots, K, \quad j = 1, \dots, M. \quad (3)$$

3. In this step, we determine the positive ideal solution (PIS) and the negative ideal solution (NIS), given by

$$PIS = \mathbf{p}^+ = \{p_1^+, p_2^+, \dots, p_M^+\}, \quad (4)$$

where $p_j^+ = \max\{p_{i,j} | 1 \leq i \leq K\}$, $j = 1, \dots, M$, and

$$NIS = \mathbf{p}^- = \{p_1^-, p_2^-, \dots, p_M^-\}, \quad (5)$$

where $p_j^- = \min\{p_{i,j} | 1 \leq i \leq K\}$, $j = 1, \dots, M$.

4. Given PIS and NIS derived in the last step, we calculate the distances (using Euclidean distance) from each evaluation vector $\mathbf{p}_i = [p_{i,1}, p_{i,2}, \dots, p_{i,M}]$ representing alternative A_i and both ideal solutions, described as follows:

$$D_i^+ = \sqrt{(\mathbf{p}_i - \mathbf{p}^+)^T (\mathbf{p}_i - \mathbf{p}^+)}, \quad i = 1, \dots, K \quad (6)$$

and

$$D_i^- = \sqrt{(\mathbf{p}_i - \mathbf{p}^-)^T (\mathbf{p}_i - \mathbf{p}^-)}, \quad i = 1, \dots, K. \quad (7)$$

¹ We considered in this paper that all the criteria are to be maximized, i.e. the larger the better. However, if there are criteria to be minimized in the problem, some simple adaptations must be incorporated in the algorithm steps. For further details, please see [3].

5. In the last step, we determine the similarity measure of each alternative A_i to the ideal solutions, given by

$$u_i = \frac{D_i^-}{D_i^+ + D_i^-}, \quad i = 1, \dots, K, \tag{8}$$

and derive the ranking according to u_i in descending order.

In this approach, one may note that the criteria are aggregated without taking into account any interaction between them. For example, in scenarios in which the criteria are correlated, i.e. they are composed by a combination of latent variables, disregarding the interaction may lead to biased results. In this context, an extended version of TOPSIS was proposed [7,8], which takes into account the Mahalanobis distance [10] (instead of Euclidean distance) and, therefore, exploit the covariance among criteria. In this version, the distances calculated in step 4 are given by

$$DM_i^+ = \sqrt{(\mathbf{r}_i - \mathbf{r}^+)^T \mathbf{\Delta}^T \mathbf{\Sigma}^{-1} \mathbf{\Delta} (\mathbf{r}_i - \mathbf{r}^+)}, \quad i = 1, \dots, K \tag{9}$$

and

$$DM_i^- = \sqrt{(\mathbf{r}_i - \mathbf{r}^-)^T \mathbf{\Delta}^T \mathbf{\Sigma}^{-1} \mathbf{\Delta} (\mathbf{r}_i - \mathbf{r}^-)}, \quad i = 1, \dots, K, \tag{10}$$

where $\mathbf{r}_i = [r_{i,1}, r_{i,2}, \dots, r_{i,M}]$, \mathbf{r}^+ and \mathbf{r}^- are, respectively, the positive and the negative ideal solutions derived from the normalized data $\mathbf{R} = (r_{i,j})_{K \times M}$, $\mathbf{\Delta} = \text{diag}(w_1, w_2, \dots, w_M)$ is the diagonal matrix whose elements are composed by the weights \mathbf{w} and $\mathbf{\Sigma} \in \mathbb{R}^{M \times M}$ is the covariance matrix of \mathbf{R} . The similarity measure is calculated as described in step 5.

2.2 Blind Source Separation Problems and Independent Component Analysis

Let us suppose a set of signal sources $\mathbf{s}(k) = [s_1(k), s_2(k), \dots, s_N(k)]$ that were linearly mixed according to

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k) + \mathbf{g}(k), \tag{11}$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the mixing matrix, $\mathbf{x}(k) = [x_1(k), x_2(k), \dots, x_M(k)]$ is the set of mixed signals and $\mathbf{g}(k) = [g_1(k), g_2(k), \dots, g_M(k)]$ is an additive white Gaussian noise (AWGN). In this linear case, BSS problems consist in retrieving the signal sources $\mathbf{s}(k)$ based only on the observed mixed data $\mathbf{x}(k)$, i.e. without the knowledge of both $\mathbf{s}(k)$ and mixing matrix \mathbf{A} [9]. This can be achieved by adjusting a separating matrix $\mathbf{B} \in \mathbb{R}^{N \times M}$ that provides a set of estimates $\mathbf{y}(k) = [y_1(k), y_2(k), \dots, y_N(k)]$, given by

$$\mathbf{y}(k) = \mathbf{B}\mathbf{x}(k), \tag{12}$$

which should be as close as possible from $\mathbf{s}(k)$. In this scenario, the separating matrix \mathbf{B} should converge to the inverse of the unknown mixing matrix \mathbf{A} .

However, given the permutation and scaling ambiguities inherent in BSS methods [9], \mathbf{B} may not be exactly the inverse of \mathbf{A} . As will be discussed later on this paper, we made some assumptions on the problem in order to avoid these inconveniences.

There are several approaches used to deal with BSS problems. A common one, called ICA, is based on the assumption that the sources are i.i.d. (independent and identically distributed) and non-Gaussian. Given the mixing process expressed in (11), the observed sources are not independent anymore but close to Gaussian. Therefore, a simplified strategy to recover signal sources that are statistically independent is to formulate an optimization problem in which the cost function leads to the minimization of a Gaussian measure (e.g. kurtosis or negentropy) of the retrieved signals. An algorithm that is based on these assumptions is known as FastICA [11]. Another method that is used in BSS problems is the Infomax, proposed by Bell and Sejnowski [12]. This method, as demonstrated by Cardoso [13], is closed-related to the maximum likelihood approach, which estimate the separating matrix \mathbf{B} from the distribution of $\mathbf{x}(k)$. Both strategies will be used in our experiments.

3 The Proposed ICA-TOPSIS Approach

In several problems in MCDM the criteria are dependent. For example, consider the case of determining a ranking of K students evaluated according to their grades in sociology, mathematics and physics². It is possible that both grades in mathematics and physics are correlated criteria, since they usually measure similar competences. Therefore, the aggregation based on the collected data may lead to biased results. In this case, one may think that a proper analysis should be made in the latent variables $\mathbf{l}(k) = [l_1(k), l_2(k), \dots, l_N(k)]^T$ associated with the collected data \mathbf{V} through the mixing process

$$\mathbf{V}^T = \mathbf{A}\mathbf{l}(k) + \mathbf{g}(k), \quad (13)$$

where $\mathbf{A} \in \mathbb{R}^{M \times N}$ represents the mixing process acting on the latent variables $\mathbf{l}(k)$ and $\mathbf{g}(k) = [g_1(k), g_2(k), \dots, g_M(k)]$ is an additive white Gaussian noise (AWGN). One may note that Eq. (13) is similar to (11), with $\mathbf{l}(k)$ and \mathbf{V}^T representing, respectively, the set of signal sources and the mixed signals. Therefore, aiming at performing the MCDM analysis on the latent variables, as mentioned in Sect. 1, the application of Mahalanobis distance in TOPSIS approach may not be sufficient to deal with dependent criteria, since only the information of covariance among criteria is taken into account.

In this context, this paper proposes to deal with the problem of dependent criteria in MCDM applying an ICA-TOPSIS approach, which comprises three steps. In the first one, we formulate a BSS problem whose aim is to recover the latent variables based on the mixed decision data \mathbf{V} . In this formulation,

² It is worth mentioning that this MCDM problem is addressed by other works in the literature [4, 14].

we consider that the number of criteria is equal to the number of latent variables, which leads to the determined case $M = N$ in BSS. Therefore, after estimating the separating matrix \mathbf{B} , we obtain the estimated latent variables $\hat{\mathbf{l}}(k) = [\hat{l}_1(k), \hat{l}_2(k), \dots, \hat{l}_N(k)]^T$, given by

$$\hat{\mathbf{l}}(k) = \mathbf{B}\mathbf{V}^T, \tag{14}$$

similarly as described in (12).

The second step comprises the adjustment of the estimated latent variables in order to avoid permutation and/or scale ambiguities. In this procedure, we made the assumption that the diagonal elements in the mixing matrix \mathbf{A} is positive and greater, in absolute value, than all the off-diagonal elements in the same row, i.e. each latent variable has a positive majority influence in each mixed criterion. Therefore, based on the separating matrix \mathbf{B} and, consequently, on the estimated mixing matrix $\hat{\mathbf{A}} = \mathbf{B}^{-1}$, we perform the following adjustment³:

- For the first row in $\hat{\mathbf{A}}$, we find the column q in which the greater absolute value is located. Therefore, we permute the first and the q columns of $\hat{\mathbf{A}}$. In order to correctly resetting the estimated latent variables, we also permute the first and the q estimates. After repeating this procedure for all rows in $\hat{\mathbf{A}}$, we obtain the estimated mixing matrix partially adjusted $\hat{\mathbf{A}}^{Adj_p}$ and avoid the permutation ambiguity provided by the BSS method.
- Based on the assumption that the diagonal elements in the mixing matrix \mathbf{A} is positive, if a diagonal element q' of $\hat{\mathbf{A}}^{Adj_p}$ is negative, we multiply all the elements in the same column of q' by -1 . This leads to the signal inversion of the estimated latent variable $\hat{l}_{q'}$, since Eq. (13) needs to be valid. After verifying all the diagonal elements of $\hat{\mathbf{A}}^{Adj_p}$ and performing the signal changes, we obtain the final adjusted estimated mixing matrix $\hat{\mathbf{A}}^{Adj_f}$ and avoid the scale ambiguity provided by the -1 factor.

In order to illustrated these adjustments, suppose that we achieve the estimated mixing matrix

$$\hat{\mathbf{A}} = \begin{bmatrix} 1.52 & -2, 95 \\ 2.01 & 0.85 \end{bmatrix}$$

associated with the retrieved sources $\hat{\mathbf{l}}(k) = [\hat{l}_1(k), \hat{l}_2(k)]^T$. Based on our assumptions, the first adjustment leads to

$$\hat{\mathbf{A}}^{Adj_p} = \begin{bmatrix} -2.95 & 1.52 \\ 0.85 & 2.01 \end{bmatrix},$$

and to the retrieved sources partially adjusted $\hat{\mathbf{l}}^{Adj_p}(k) = [\hat{l}_2(k), \hat{l}_1(k)]$. One may note the permutation of both columns. In the second adjustment, we obtain

$$\hat{\mathbf{A}}^{Adj_f} = \begin{bmatrix} 2.95 & 1.52 \\ -0.85 & 2.01 \end{bmatrix}$$

³ It is worth mentioning that the scale ambiguity provided by a positive factor or a negative factor different from -1 is automatically mitigated in the normalization step of TOPSIS.

and $\hat{\mathbf{I}}^{Adj_f}(k) = [-\hat{l}_2(k), \hat{l}_1(k)]$, which corrects the signal of the retrieved sources.

After performing the ICA and eliminating the ambiguities, the third step of the proposed approach comprises the application of TOPSIS based on Euclidean distance in $\hat{\mathbf{I}}^{Adj_f}(k)$ and the ranking determination.

4 Numerical Experiments

Aiming at verifying the application of the proposed ICA-TOPSIS approach to deal with dependent criteria in MCDM problems, we performed numerical experiments based on synthetic data and compared the results with the ones provided by existing methods. The next section describes the considered data and the obtained results.

4.1 Data Generation

In this paper, we performed the experiments based on a decision data comprised by 100 alternatives and 2 criteria, both with the same importance ($w_1 = w_2 = 0.5$). The latent variables were randomly generated according to a uniform distribution in the range $[0, 1]$. In order to derive the “collected” observed data \mathbf{V} , we considered the mixing matrix

$$\mathbf{A} = \begin{bmatrix} 1.00 & -0.15 \\ 0.30 & 1.00 \end{bmatrix}$$

and the mixing process described in (11), in which $\mathbf{s}(k)$ and $\mathbf{x}(k)$ represent the latent variables and the observed data \mathbf{V} , respectively. Moreover, the additive noise was applied considering a Signal-to-Noise Ratio (SNR), given by

$$SNR = 10 \log_{10} \frac{\sigma_{signal}^2}{\sigma_{noise}^2}, \quad (15)$$

where σ_{signal}^2 and σ_{noise}^2 are, respectively, the signal power and the noise power, in the range $(0, 50]$.

4.2 Comparison Between the Considered Approaches

In order to verify the application of the proposal, we first generate the latent variables and derive the ranking according to the original TOPSIS method (based on Euclidean distance). This ranking is considered as the correct one, since it is obtained directly from the (unknown) latent variables. Therefore, we perform the mixing process and, given the mixed observed data, we apply the proposed ICA-TOPSIS approach (based on FastICA and Infomax algorithms), the original TOPSIS and the TOPSIS based on Mahalanobis distance. The obtained results are compared according to a performance index called normalized Kendall tau

distance [15], which calculates the percentage of pairwise disagreements between two rankings. This measure is defined by

$$\tau = \frac{N_D}{K(K-1)/2}, \tag{16}$$

where N_D is the number of pairwise disagreements between the rankings and K is the number of alternatives. Therefore, τ close to zero indicates that there is no disagreement between the two rankings, i.e. the obtained ranking is the same that the correct one provided by the original TOPSIS method applied on the latent variables.

Figure 1 presents the Kendall tau distance for each considered method and SNR value (averaged over 1000 realizations). One may note that the TOPSIS based on Mahalanobis distance improves the original version of this method, leading to lower values of τ . However, the best results were obtained applying the ICA-TOPSIS, specially for SNR values greater than 25 dB. In terms of the FastICA and Infomax algorithms, the former achieved a better performance.

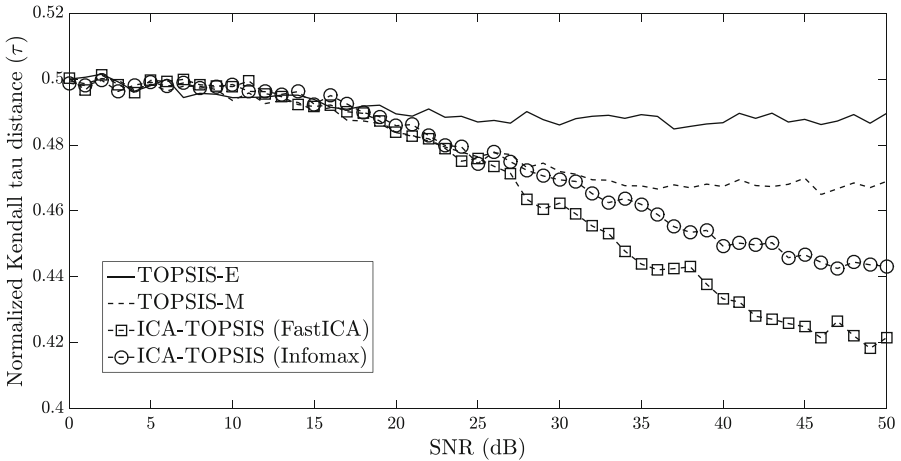


Fig. 1. Comparison of the Kendall tau distances for the original TOPSIS based on Euclidean distance (TOPSIS-E), the TOPSIS based on Mahalanobis distance (TOPSIS-M) and the proposed approach (with FastICA and Infomax).

5 Conclusions and Perspectives

Dependent criteria is an important issue in multicriteria decision making. In order to deal with this problem, several methods has been developed, such as the TOPSIS based on Mahalanobis distance. In this work, we presented preliminaries discussions on a novel approach used to mitigate biased results provided by dependent criteria. This approach, called ICA-TOPSIS, comprises the application of independent component analysis in order to extract the latent variables

from the observed decision data and, then, the use of the original TOPSIS to derive the ranking based on the retrieved independent data.

Based on the MCDM scenario considered in this work and the obtained results, one may remark that the proposed ICA-TOPSIS approach leads to better results compared to the methods found in the literature. For instance, our proposal achieved lower Kendall tau values compared to the TOPSIS based on Mahalanobis distance, which is used in several works in the literature. A possible explanation for this result is that the ICA methods exploit the independence among criteria, which is stronger than the covariance information used in TOPSIS based on Mahalanobis distance. Since we consider a MCDM problem comprised by a mixture of latent variables, our proposal can better mitigate the biased effect of the criteria dependence.

It is worth mentioning that this work presented initial results on the application of ICA-TOPSIS approach to deal with MCDM problems. Future works comprise a further understanding on this proposal, especially on the latent variable estimation step. Different numbers of criteria and alternatives will also be considered in new experiments. Moreover, we aim at verifying the performance of the proposed approach on decision problems based on real data.

References

1. Figueira, J., Greco, S., Ehrgott, M. (eds): Multiple criteria decision analysis: State of the art surveys. Springer's International Series in Operations Research & Management Science, 2nd edn. Springer, New York (2016)
2. Tzeng, G., Huang, J.: Multiple Attribute Decision Making: Methods and Applications. CRC Press, New York (2011)
3. Hwang, C.-L., Yoon, K.: Multiple Attribute Decision Making: Methods and Applications. Springer, Heidelberg (1981)
4. Grabisch, M.: The application of fuzzy integrals in multicriteria decision making. *Eur. J. Oper. Res.* **89**, 445–456 (1996)
5. Antuchevičienė, J., Zavadskas, E.K., Zakarevičius, A.: Multiple criteria construction management decisions considering relations between criteria. *Technol. Econ. Dev. Econ.* **16**(1), 109–125 (2010)
6. Bondor, C.I., Museşan, A.: Correlated criteria in decision models: Recurrent application of TOPSIS method. *Appl. Med. Inf.* **30**(1), 55–63 (2012)
7. Vega, A., Aguarón, J., García-Alcaraz, J., Moreno-Jiménez, J.M.: Notes on dependent attributes in TOPSIS. *Procedia Comput. Sci.* **31**, 308–317 (2014)
8. Wang, Z.-X., Wang, Y.-Y.: Evaluation of the provincial competitiveness of the Chinese high-tech industry using an improved TOPSIS method. *Expert Syst. Appl.* **41**, 2824–2831 (2014)
9. Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, Oxford (2010)
10. Mahalanobis, P.C.: On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936)
11. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
12. Bell, A., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**(6), 1129–1159 (1995)

13. Cardoso, J.F.: Infomax and maximum likelihood for blind source separation. *IEEE Sig. Process. Lett.* **4**(4), 112–114 (1997)
14. Kojadinovic, I.: Unsupervised aggregation of commensurate correlated attributes by means of the Choquet integral and entropy functionals. *Int. J. Intell. Syst.* **23**, 128–154 (2008)
15. Kendall, M.G.: A new measure of rank correlation. *Biometrika* **30**(1/2), 81–93 (1938)

Author Index

- Ablin, Pierre 151
Absil, P.-A. 69, 139
Akbayrak, Semih 24
Akhavan, Saeed 534
Albera, Laurent 3
Attux, Romis 193
- Bach, Francis 446
Badeau, Roland 13
Baraniuk, Richard G. 395
Barillot, Emmanuel 501
Beaumont, Guillaume 385
Bro, Rasmus 89
- Cabral Farias, Rodrigo 36
Cantini, Laura 501
Çapan, Gökhan 24
Cardoso, Jean-François 151
Cemgil, Ali Taylan 24
Ceritli, Taha Yusuf 24
Chaux, Caroline 417
Chazan, Shlomo E. 319
Chen, Hsin 514
Chrétien, Stéphane 127
Čmejla, Jaroslav 270
Cohen, Jeremy Emile 36, 89, 456
Coleman, Philip 558
Comon, Pierre 57
Corey, Ryan M. 238
Czerwinska, Urszula 501
- Dai, Zhenwen 467
Dantas, Cássio F. 456
de Bodt, Cyril 524
De Geeter, Jeroen 79
De Lathauwer, Lieven 139
de Oliveira, Pedro Marinho R. 46
Deville, Alain 204
Deville, Yannick 183, 193, 204
Dixon, Simon 329
Dong, Xuan 351
- Dragotti, Pier Luigi 479
Dreesen, Philippe 79
Drémeau, Angélique 385
Duan, Zhiyao 372
Duarte, Leonardo Tomazeli 193, 568
Duong, Ngoc Q. K. 547
Duong, Thanh Thi Hien 547
- Emiya, Valentin 417
Eskimez, Sefik Emre 372
Ewert, Sebastian 329
Exarchakis, Georgios 467
- Fablet, Ronan 385
Fano Yela, Delia 280
Fantinato, Denis G. 193
Fazzi, Antonio 99
Fontaine, Mathieu 13
Francombe, Jon 558
- Gallivan, Kyle A. 69
Gannot, Sharon 319
Gerkmann, Timo 407
Goldberger, Jacob 319
Grais, Emad M. 340
Gramfort, Alexandre 151
Green, Owen 306
Gribonval, Rémi 456
Guerrero, Andréa 183
Guglielmi, Nicola 99
Guo, Peng 228
- Han, Xu 3
Hinrich, Jesper Løve 488
Ho, Olivier 127
Hosseini, Shahram 183
- Ishteva, Mariya 79
Ito, Nobutaka 293
- Jackson, Philip J. B. 558
Jouni, Mohamad 57
Jutten, Christian 171, 193, 534

- Kachenoura, Amar 3
 Kairov, Ulykbek 501
 Kamarei, Mahmoud 534
 Koldovský, Zbyněk 161, 270
 Kolossa, Dorothea 228
 Kong, Qiuqiang 361
 Kopriva, Ivica 107
 Kounovský, Tomáš 270
 Krawczyk-Becker, Martin 407
 Krémé, A. Marina 417
- Lahat, Dana 171
 Lejeune, Nicolas 524
 Lerbet, Jean 514
 Liu, Qingju 558
 Liutkus, Antoine 13, 293
 Lücke, Jörg 467
- Maddox, Ross K. 372
 Málek, Jiří 270
 Markovsky, Ivan 99, 479
 Matilainen, Markus 248
 Maymon, Yanir 228
 Metzler, Christopher A. 395
 Mørup, Morten 488
 Mouraux, André 524
 Mulders, Dounia 524
 Mura, Mauro Dalla 57
- Neves, Aline 193
 Nguyen, Cong-Phuong 547
 Nguyen, Quoc-Cuong 547
 Nordhausen, Klaus 248
- Olikier, Guillaume 139
 Ono, Nobutaka 161
- Pelegrina, Guilherme Dean 568
 Phlypo, Ronald 534
 Plumbley, Mark D. 340, 361, 446
 Průša, Zdeněk 429
- Rafaely, Boaz 228
 Rajmic, Pavel 429
 Reiss, Joshua D. 259
- Renard, Emilie 69
 Rencker, Lucas 446
 Rivet, Bertrand 36
 Roma, Gerard 306
 Romano, João Marcos Travassos 568
- Sadowski, Tomasz 116
 Sandler, Mark 280
 Sawada, Hiroshi 217
 Schniter, Philip 395
 Schymura, Christopher 228
 Senhadji, Lotfi 3
 Serizel, Romain 13
 Shu, Huazhong 3
 Şimşekli, Umut 13
 Singer, Andrew C. 238
 Soltanian-Zadeh, Hamid 534
 Souriau, Rémi 514
 Stoller, Daniel 329
 Stöter, Fabian-Robert 13, 293
 Stowell, Dan 259, 280
- Tichavský, Petr 161
 Tremblay, Pierre Alexandre 306
- Verleysen, Michel 524
 Veselý, Vítězslav 429
 Vigneron, Vincent 514
 Virta, Joni 248
- Wang, Wenwu 361, 446
 Ward, Dominic 340
 Wierstorf, Hagen 340
 Wilkinson, William J. 259
 Williamson, Donald S. 351
- Xu, Chenliang 372
 Xu, Yong 361
- Zarzoso, Vicente 46
 Závaška, Pavel 429
 Zdunek, Rafał 116
 Zermini, Alfredo 361
 Zinovyev, Andrei 501