An EM Method for Spatio-temporal Blind Source Separation Using an AR-MOG Source Model

Kenneth E. Hild II[1], Hagai T. Attias[2], and Srikantan S. Nagarajan[1]

[1] Dept. of Radiology, University of California at San Francisco, CA, USA 94122
k.hild@ieee.org, Srikantan.Nagarajan@radiology.ucsf.edu
[2] Golden Metallic, San Francisco CA, USA 94147
htattias@goldenmetallic.com

Abstract. A maximum likelihood blind source separation algorithm is developed. The temporal dependencies are explained by assuming that each source is an AR process and the distribution of the associated i.i.d. innovations process is described using a Mixture of Gaussians (MOG). Unlike most maximum likelihood methods the proposed algorithm takes into account both spatial and temporal information, optimization is performed using the Expectation-Maximization method, and the source model is learned along with the demixing parameters.

1 Introduction

Blind source separation (BSS) involves the application of a linear transformation to an observed set of M mixtures, $\boldsymbol{x}$, in an attempt to extract the original M (unmixed) sources, $\boldsymbol{s}$. Two of the main types of BSS methods for stationary data include decorrelation approaches and approaches based on Independent Components Analysis (ICA). Methods based on decorrelation minimize the squared cross-correlation between all possible pairs of source estimates at two or more lags [1], [2], [3]. Methods based on ICA attempt to make the source estimates statistically independent at lag 0 [4], [5], [6]. Herein it is assumed that the sources are mutually statistically independent, the mixing matrix is invertible, and there are as many sensors as there are sources. If, in addition, at most one source has a Gaussian probability density function (pdf) then ICA methods are appropriate for BSS even if all the sources have identical spectra, whereas this is not the case for decorrelation methods. Similarly, if the M sources possess sufficient spectral diversity then decorrelation methods are appropriate for BSS even if all the sources are Gaussian-distributed, whereas this is not the case for ICA methods. Consequently, the appropriate BSS algorithm for a given application depends on the spatial and temporal structure of the sources in question.

The approach presented here, AR-MOG, differs from most ML methods [7], [8], [9] in three important ways. First, the proposed criterion makes use of both the spatial and temporal structure of the sources. Consequently, AR-MOG may be used in situations for which either of the above two types of BSS algorithms

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 98–105, 2006.

are appropriate. Second, AR-MOG is formulated in terms of latent variables so that it can be optimized using the Expectation-Maximization (EM) method. Third, instead of assuming the target distributions are known, the proposed method learns the target distributions directly from the observations.

2 Generative Model

It is assumed that there are M mutually statistically independent sources, each of which are N samples in length. The variable $\boldsymbol{s}$ represents the $(M \times N)$ source matrix, $\boldsymbol{s}_{m,1:\mathrm{N}}$ represents the $(1 \times N)$ vector of the m^{th} row of $\boldsymbol{s}$, and $\boldsymbol{s}_{1:\mathrm{M},n}$ represents the $(M \times 1)$ vector of the n^{th} column of $\boldsymbol{s}$. Each source, $s_{m,n}$, is assumed to be an autoregressive (AR) process that is generated from a temporally i.i.d. innovations process, $u_{m,n}$. The relationship between a given source and the associated innovations process is assumed to be $u_{m,n} = \sum_{k=0}^{K_g} g_{m,k} s_{m,n-k}$, where $g_{m,0} = 1$ $m \in \{1, 2, \ldots, M\}$, $g_{m,k}$ is an element of the $(M \times K_g + 1)$ matrix $\boldsymbol{g}$ of AR coefficients, and K_g is the order of each of the AR filters. The sources are therefore given by

$$s_{m,n} = -\sum_{k=1}^{K_g} g_{m,k} s_{m,n-k} + u_{m,n} \quad . \tag{1}$$

The M observations at time n are assumed to be generated from the sources by means of a linear, memory-less $(M \times M)$ mixing matrix, i.e., $\boldsymbol{x}_n = \boldsymbol{A}\boldsymbol{s}_n$.

The pdf of each innovations process is assumed to be parameterized by a Mixture of Gaussians (MOG),

$$\begin{aligned} p_{\mathrm{U}_{m,n}}(u_{m,n}) &= \sum_{q=1}^{K_Q} p_{\mathrm{U}_{m,n}|\mathrm{Q}_{m,n}}(u_{m,n}|Q_{m,n}=q) p_{\mathrm{Q}_{m,n}}(Q_{m,n}=q) \\ &= \sum_{q=1}^{K_Q} \mathcal{N}(u_{m,n}|\mu_{m,q}, \nu_{m,q})\, \pi_{m,q} \quad , \end{aligned} \tag{2}$$

which should not be confused with $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ (the target pdf of each innovations process) or $p_{\hat{\mathrm{U}}_{m,n}}(\hat{u}_{m,n})$ (the actual pdf of the estimate of the innovations), and where $p_{\mathrm{U}_{m,n}|\mathrm{Q}_{m,n}}(u_{m,n}|Q_{m,n} = q)$ has a normal distribution, $\mu_{m,q}$ is the mean of the q^{th} component (or state) of the m^{th} source, $\nu_{m,q}$ is the corresponding precision, $\pi_{m,q} \equiv p_{\mathrm{Q}_{m,n}}(Q_{m,n} = q)$ is the corresponding prior probability (constrained such that $\sum_{q=1}^{K_Q} \pi_{m,q} = 1 \ \forall m$), and $Q_{m,n} \in \{1, 2, \ldots, K_Q\}$ represents the state (latent variable) of the m^{th} source at the n^{th} time point. This particular generative model is able to describe both the non-Gaussianity and the temporal dependencies of the sources.

3 Criterion

Let $p_{\mathrm{U}_{m,n}}(u_{m,n})$ denote the marginal pdf of a particular innovations process and let $p_{\mathbf{U}}(\boldsymbol{u})$ and $p_{\mathbf{U}_{1:\mathrm{M},n}}(\boldsymbol{u}_{1:\mathrm{M},n})$ denote the order-MN and order-M joint pdf's

of the innovations, respectively. It is assumed that all variables are identically distributed in time (although this is not valid for the outputs of the IIR filter $s_{m,n}$ until after the transients have died out). Using this notation and the preceding generative model the data likelihood is given by

$$p_{\mathbf{x}}(\boldsymbol{x}) = \prod_{n=1}^{N} p_{\mathbf{x}_{1:\mathrm{M},n}|\mathbf{x}_{1:\mathrm{M},1:n-1}}(\boldsymbol{x}_{1:\mathrm{M},n}|\boldsymbol{x}_{1:\mathrm{M},1:n-1}) = |\boldsymbol{W}|^N \prod_{m=1}^{M}\prod_{n=1}^{N} p_{\mathrm{U}_{m,n}}(u_{m,n}) \ , \tag{3}$$

where $\boldsymbol{W} = \boldsymbol{A}^{-1}$ (hence, $\boldsymbol{s}_n = \boldsymbol{W}\boldsymbol{x}_n$), it is understood that the set of all parameters, $\{\boldsymbol{W}, \boldsymbol{g}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\pi},\ K_Q, K_g\}$, is given for each pdf, and where $u_{m,n} = \sum_{k=0}^{K_g}\sum_{l=1}^{M} W_{m,l} g_{m,k} x_{l,n-k}$. Hence, the log likelihood is given by

$$\begin{aligned} \mathcal{L} &= N \ln |\boldsymbol{W}| + \sum_{m=1}^{M}\sum_{n=1}^{N} \ln p_{\mathrm{U}_{m,n}}(u_{m,n}) \\ &= N \ln|\boldsymbol{W}| + \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{q=1}^{K_Q} \gamma_{m,n,q} \ln \frac{p_{\mathrm{U}_{m,n}\mathrm{Q}_{m,n}}(u_{m,n}, Q_{m,n}=q)}{\gamma_{m,n,q}} \ , \end{aligned} \tag{4}$$

where the latter expression is given as a function of the posterior state probabilities, $\gamma_{m,n,q} \equiv p_{\mathrm{Q}_{m,n}|\mathbf{x}}(Q_{m,n}=q|\boldsymbol{x})$. Adaptation using EM, which is guaranteed to converge (possibly to a local maximum), involves maximizing (4) by alternating between the E-step and the M-step.

4 EM Algorithm for AR-MOG

In this section we present an EM algorithm for inferring the model from the data and extracting independent sources.

4.1 E-Step

The E-step maximizes the log likelihood w.r.t. the posteriors, $\gamma_{m,n,q}$, while keeping the parameters fixed. The estimates of the posteriors are given by

$$\hat{\gamma}_{m,n,q} = \frac{p_{\bar{\mathrm{U}}_{m,n}|\hat{\mathrm{Q}}_{m,n}}(\hat{u}_{m,n}|\hat{Q}_{m,n}=q)\hat{\pi}_{m,q}}{\xi_{m,n}} \ , \tag{5}$$

where $\xi_{m,n}$ ensures that $\sum_{q=1}^{K_Q} \hat{\gamma}_{m,n,q} = 1\ \forall\, m,n$, the true pdf's (conditioned on the state) have been replaced with the target pdf's, and all other quantities have been replaced with their estimates (denoted using the hat symbol).

4.2 M-Step

The M-step maximizes the log likelihood w.r.t. the parameter estimates $\{\hat{\boldsymbol{W}}, \hat{\boldsymbol{g}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\pi}}\}$ while keeping the posteriors fixed. The two parameters that are not learned by AR-MOG, $\{\hat{K}_Q, \hat{K}_g\}$, are assumed to be known. The update of $\hat{\boldsymbol{W}}$ is performed using multiple iterations of gradient ascent where

$$\frac{\partial \mathcal{L}}{\partial \hat{W}_{m,l}} = \sum_{i=1}^{M} \left(NI_{m,i} - \sum_{n=1}^{N} \sum_{q=1}^{\hat{K}_Q} \sum_{k=0}^{\hat{K}_g} \hat{\gamma}_{m,n,q} (\hat{u}_{m,n} - \hat{\mu}_{m,q}) \hat{\nu}_{m,q} \hat{g}_{m,k} \hat{s}_{i,n-k} \right) \hat{W}_{i,l} \quad , \tag{6}$$

which makes use of the natural gradient [10] (also known as the relative gradient [11]). The solution for the matrix of AR coefficients is

$$\begin{aligned} \boldsymbol{\Phi}_{m,k} &= \sum_{n=1}^{N} \sum_{q=1}^{K_Q} \hat{\gamma}_{m,n,q} \hat{\nu}_{m,q} (\hat{\mu}_{m,q} - \hat{s}_{m,n}) \hat{s}_{m,n-k} \\ \boldsymbol{\Psi}_{m,k,k'} &= \sum_{n=1}^{N} \sum_{q=1}^{K_Q} \hat{\gamma}_{m,n,q} \hat{\nu}_{m,q} \hat{s}_{m,n-k} \hat{s}_{m,n-k'} \\ \hat{\boldsymbol{g}}_{m,1:\mathrm{N}} &= \boldsymbol{\Phi}_{m,1:\mathrm{M}} (\boldsymbol{\Psi}_{m,1:\mathrm{M},1:\mathrm{M}})^{-1} \end{aligned} \tag{7}$$

for $m \in \{1, \ldots, M\}$. The solutions for the parameters that constitute the target distributions are

$$\begin{aligned} \hat{\mu}_{m,q} &= \frac{\sum_{n=1}^{N} \hat{u}_{m,n} \hat{\gamma}_{m,n,q}}{\sum_{n=1}^{N} \hat{\gamma}_{m,n,q}} \\ \hat{\nu}_{m,q} &= \frac{\sum_{n=1}^{N} \hat{\gamma}_{m,n,q}}{\sum_{n=1}^{N} (\hat{u}_{m,n} - \hat{\mu}_{m,q})^2 \hat{\gamma}_{m,n,q}} \\ \hat{\pi}_{m,q} &= \frac{\sum_{n=1}^{N} \hat{\gamma}_{m,n,q}}{\sum_{n=1}^{N} \sum_{q'=1}^{\hat{K}_Q} \hat{\gamma}_{m,n,q'}} \quad . \end{aligned} \tag{8}$$

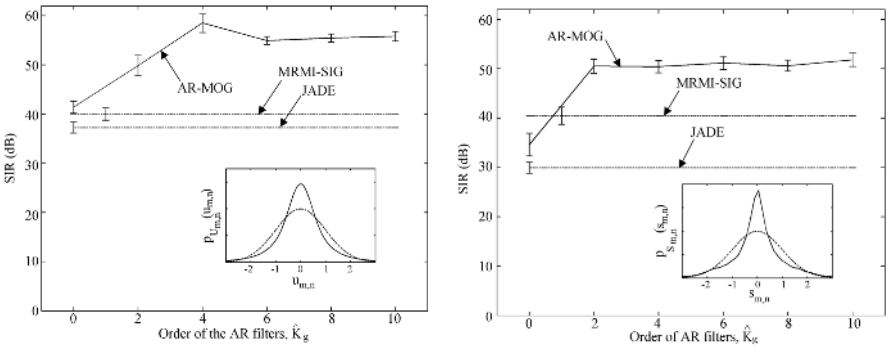
5 Experiments

Several different experiments are performed in order to assess the separation performance of AR-MOG. Separation performance is gauged using the signal-to-interference ratio (SIR), which is defined by

$$\mathrm{SIR} = \frac{1}{M} \sum_{m=1}^{M} 10 \log_{10} \frac{\sum_{n=1}^{N} (\hat{\boldsymbol{W}}_{m,1:\mathrm{M}} \boldsymbol{A}_{1:\mathrm{M},m} s_{m,n})^2}{\sum_{\substack{m'=1 \\ (m' \neq m)}}^{M} \sum_{n=1}^{N} (\hat{\boldsymbol{W}}_{m,1:\mathrm{M}} \boldsymbol{A}_{1:\mathrm{M},m'} s_{m',n})^2} \quad (\mathrm{dB}) \quad .$$

Unless otherwise specified the data is drawn from the same model that is used by AR-MOG, the innovations are assumed to have the same distribution, $M = 2$, and $N = 10^4$. The error bars represent one standard error. When they are included the mean results represent the average of 10 Monte Carlo trials. Results from JADE [12], which does not use temporal dependencies ($\hat{K}_g = 0$), and MRMI-SIG [6], which essentially uses $\hat{K}_g = 1$, are also included as benchmarks.

Figure 1a shows the mean separation performance of AR-MOG as a function of $\hat{K}_g$, where $K_g = 10$ and $\hat{K}_Q = K_Q = 4$. The means, precisions, and priors are not adapted in this experiment or the next experiment so that the change in performance due to the addition of the AR filters may be better quantified. For $\hat{K}_g = 0$



(a) Experiment 1 (b) Experiment 2

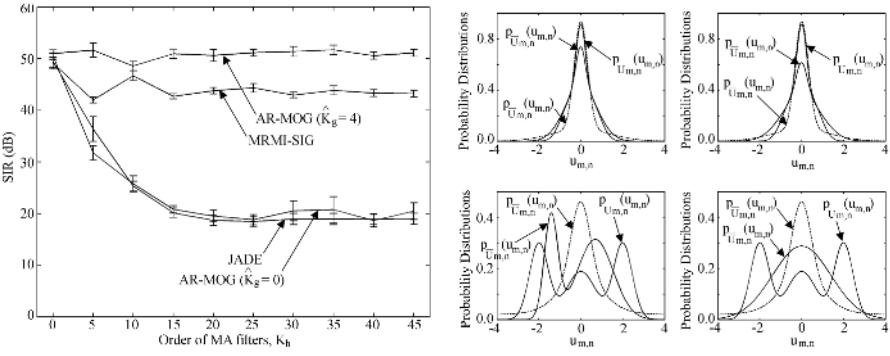
Fig. 1. Separation performance as a function of $\hat{K}_g$. (a) Experiment 1. The inset shows $p_{\mathrm{U}_{m,n}}(u_{m,n})$ and a Gaussian distribution (dashed line) having the same mean and variance. (b) Experiment 2. The inset shows $p_{\mathrm{S}_{m,n}}(s_{m,n})$ and a Gaussian distribution (dashed line) having the same mean and variance.

AR-MOG defaults to the case where no AR model is used, i.e., $\hat{\boldsymbol{u}} = \hat{\boldsymbol{s}}$. When no temporal dependencies are used the AR-MOG method performs similarly, but slightly better, than JADE and MRMI-SIG. For $\hat{K}_g \geq 4$ the performance improvement of AR-MOG is approximately 15-20 dB.

The separation results in Fig. 1a represent a best case scenario for AR-MOG since the data are drawn from the model. The results shown in Fig. 1b use (artificially-mixed) real speech data not drawn from the AR-MOG model. Performance is shown as a function of $\hat{K}_g$ where $\hat{K}_Q = 3$. The target distribution $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ is chosen to be unimodal and super-Gaussian since speech is known to be approximately Laplacian. For the speech data the performance of both AR-MOG and JADE are reduced by approximately 5-8 dB with respect to the first experiment. Figure 1b shows that it is not strictly necessary for the sources to be stationary processes for AR-MOG to perform well (speech is commonly assumed to be stationary, but only for very short segments [13]).

The third experiment shows the sensitivity of the three BSS algorithms to an increase in the temporal correlation of the sources. For this experiment $N = 3 * 10^4$, $\hat{K}_Q = K_Q = 3$, $\hat{K}_g = 4$, and each $\boldsymbol{s}_{m,1:\mathrm{N}}$ is related to the associated $\boldsymbol{u}_{m,1:\mathrm{N}}$ by means of a moving average (MA) filter, $\boldsymbol{h}_{m,1:K_h+1}$. Performance is shown in Fig. 2a as the order of this filter, K_h, is varied (increasing K_h increases the overall correlation at an exponentially decreasing rate). Unlike the previous experiments the means and variances are adapted. For this dataset increasing the temporal correlation (i.e., K_h) causes the separation performance of JADE and MRMI-SIG to decrease by roughly 30 dB and 6 dB, respectively. The performance of AR-MOG is not affected by the change in temporal correlation.

The fourth experiment attempts to measure the separation performance as a function of the initialization of $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$. For each case considered the separation performance is given when the parameters that constitute $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ are



(a) Experiment 3 (b) Experiment 4

Fig. 2. (a) Separation performance as a function of K_h (the length of each h_m) for Experiment 3. (b) Initial (dashed line) and final $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ distributions and $p_{\mathrm{U}_{m,n}}(u_{m,n})$ for Experiment 4. Upper-left: Case 1. Upper-right: Case 2. Lower-left: Case 3. Lower-right: Case 4.

adapted and when they are fixed. The resulting SIR values are shown in Table I, where the left column corresponds to when $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ is adapted, whereas the right column keeps $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ fixed at the distribution used for initialization of the left column results. The initial and final $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ distributions and the true distribution, $p_{\mathrm{U}_{m,n}}(u_{m,n})$, are shown in Fig. 2b. For Cases 1 & 2 the initial innovations distribution and the true distribution are similar and for Cases 3 & 4 the assumed (initial) innovations distribution is far from correct. Likewise, for Cases 1 & 3 $\hat{K}_g = K_g = 0$ and for Cases 2 & 4 $\hat{K}_g = K_g = 4$. When the initial innovations distribution is similar to the true distribution the separation performance is excellent independent of whether or not $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ is adapted. When they are not similar, based on these results, it is advantageous to adapt $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$. Notice that $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$ gets trapped in a local maximum for Cases 3 and 4. This is indicated by the fact that the target distribution converges to a bimodal solution for Case 3 and a unimodal solution for Case 4. If AR-MOG is initialized with the true distribution the final SIR is 62.9 and 67.7 dB, respectively, and the target distributions for both cases converge to a trimodal solution. The fact that the final target distribution is incorrect does not necessarily preclude the possibility of

Table 1. Final SIR separation performance for Experiment 4

Case	Adapt $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$	Fixed $p_{\bar{\mathrm{U}}_{m,n}}(u_{m,n})$
1	45.6 dB	46.3 dB
2	59.5 dB	47.1 dB
3	44.2 dB	0.0 dB
4	51.0 dB	39.5 dB

achieving good separation performance, as indicated in Table I, because it is neither sufficient nor necessary for good separation performance that the final $p_{\bar{\mathbb{U}}_{m,n}}(u_{m,n})$ approximates $p_{\mathbb{U}_{m,n}}(u_{m,n})$. What is necessarily required (but is not sufficient, e.g., for Gaussian distributions) is that $p_{\hat{S}_{m,n}}(s_{m,n})$ approximates $p_{\mathbb{S}_{m,n}}(s_{m,n})$ for each m (and allowing for possible permutations). The ability of AR-MOG to separate sources even if $p_{\bar{\mathbb{U}}_{m,n}}(u_{m,n})$ is incorrect is identical to the well-known fact that ML methods that assume the cumulative density function (cdf) is sigmoidal are oftentimes able to separate sources even if the cdf of each source is not sigmoidal [4], [11], [14], [15], [16], [17]. There is no assurance that AR-MOG will be able to find a solution for $p_{\bar{\mathbb{U}}_{m,n}}(u_{m,n})$ that allows for good separation, but Table I indicates that it may be advantageous to try to improve on the original assumptions.

6 Conclusions

This paper develops a BSS algorithm that is based on maximizing the data likelihood where each source is assumed to be an AR process and the innovations are described using a MOG distribution. It differs from most ML methods in that it uses both spatial and temporal information, the EM algorithm is used as the optimization method, and the parameters that constitute the source model are adapted to maximize the criterion. Due to the combination of the AR process and the MOG model, the update equations for each parameter has a very simple form. The separation performance was compared to several other methods, one that does not take into account temporal information and one that does. The proposed method outperforms both. Future work will focus on incorporating noise directly into the model in a manner similar to that used for the Independent Factor Analysis method [18].

Acknowledgments

This work was supported by National Institutes of Health grants 1 F32 NS 52048-01 and RO1 DC 4855.

References

1. Van Gerven, S., Van Compernolle, D.: Signal separation by symmetric adaptive decorrelation: stability, convergence, and uniqueness. IEEE Trans. on Signal Proc. **43** (1995) 1602–1612
2. Weinstein, E., Feder, M., Oppenheim, A.: Multi-channel signal separation by decorrelation. IEEE Trans. on Speech and Audio Proc. **1** (1993) 405–413
3. Wu, H.C., Principe, J.C.: A unifying criterion for blind source separation and decorrelation: simultaneous diagonalization of correlation matrices. Neural Networks for Signal Proc. (1997) 496–505
4. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation **7** (1995) 1129–1159

5. Muller, K.R., Philips, P., Ziehe A.: Jade$_{\mathrm{TD}}$: Combining higher-order statistics and temporal information for blind source separation (with noise). Intl. Workshop on Independent Component Analysis and Signal Separation (1999) 87–92
6. Hild II, K.E., Erdogmus, D., Principe, J.C.: An Analysis of Entropy Estimators for Blind Source Separation. Signal Processing **86** (2006) 182–194
7. Moulines, E., Cardoso, J.F., Gassiat, E.: Maximum Likelihood for blind source separation and deconvolution of noisy signals using mixture models. Intl. Conf. on Acoustics, Speech, and Signal Processing **5** (1997) 3617–3620
8. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans. on Signal Proc. **45** (1997) 1712–1725
9. Pearlmutter, B.A., Parra, L.C.: Maximum likelihood blind source separation: A context-sensitive generalization of ICA. Advances in Neural Information Proc. Systems **9** (1996) 613–619
10. Amari, S.: Neural learning in structured parameter spaces - Natural Riemannian gradient. Advances in Neural Information Proc. Systems **9** (1996) 127–133
11. Cardoso, J.F., Laheld, B.H.: Equivariant adaptive source separation. IEEE Trans. on Signal Proc. **44** (1996) 3017–3030
12. Cardoso, J.F., Souloumiac, A.: Blind beamforming for non-Gaussian signals. IEE Proceedings-F **140** (1993) 362–370
13. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals (1978)
14. Hosseini, S., Jutten, C., Pham, D.T.: Markovian source separation. IEEE Trans. on Signal Proc. **51** (2003) 3009–3019
15. Amari, S.I., Cardoso, J.F.: Blind source separation-Semiparameteric statistical approach. IEEE Trans. on Signal Proc. **45** (1997) 2692–2700
16. Cruces-Alvarez, S.A., Cichoki, A., Amari, S.I.: On a new blind signal extraction algorithm: Different criteria and stability analysis. IEEE Signal Proc. Letters **9** (2002) 233–236
17. Cardoso, J.F.: Infomax and maximum likelihood for blind source separation. IEEE Signal Proc. Letters **4** (1997) 112–114
18. Attias, H.: Independent factor analysis. Neural Computation **11** (1999) 803–851