Published in final edited form as:

IEEE Trans Neural Netw. 2008 March ; 19(3): 508–519. doi:10.1109/TNN.2007.914154.

An Expectation-Maximization Method for Spatio-Temporal Blind Source Separation Using an AR-MOG Source Model

Kenneth E. Hild II [Senior Member, IEEE],

Department of Radiology, The University of California at San Francisco, San Francisco, CA 94143 USA (k.hild@ieee.org)

Hagai T. Attias, and

Golden Metallic, Inc., San Francisco, CA 94147 USA (htattias@goldenmetallic.com)

Srikantan S. Nagarajan [Senior Member, IEEE]

Department of Radiology, The University of California at San Francisco, San Francisco, CA 94143 USA (Srikantan.Nagarajan@radiology.ucsf.edu)

Abstract

In this paper, we develop a maximum-likelihood (ML) spatio-temporal blind source separation (BSS) algorithm, where the temporal dependencies are explained by assuming that each source is an autoregressive (AR) process and the distribution of the associated independent identically distributed (i.i.d.) innovations process is described using a mixture of Gaussians. Unlike most ML methods, the proposed algorithm takes into account both spatial and temporal information, optimization is performed using the expectation-maximization (EM) method, the source model is adapted to maximize the likelihood, and the update equations have a simple, analytical form. The proposed method, which we refer to as autoregressive mixture of Gaussians (AR-MOG), outperforms nine other methods for artificial mixtures of real audio. We also show results for using AR-MOG to extract the fetal cardiac signal from real magnetocardiographic (MCG) data.

Keywords

Blind source separation (BSS); expectation-maximization (EM); independent components analysis (ICA); maximum likelihood (ML)

I. Introduction

BLIND SOURCE SEPARATION (BSS) methods attempt to extract a set of M unknown sources s , given only a set of observations x , which represents an unknown mixture of the sources. Herein, we assume linear, instantaneous, invertible mixtures of mutually statistically independent sources. There are several applications for which BSS, with these assumptions, has been successfully used. For example, BSS has been used to remove interference from electroencephalographic (EEG) data [1], separate the waveforms emanating from different neural generators in magnetoencephalographic (MEG) data [2], [3], and isolate the fetal cardiac signal from the maternal cardiac signal in magnetocardiographic (MCG) data [4].

There are two common types of BSS methods. The first is based on decorrelation and the second is based on independent components analysis (ICA). Decorrelation methods minimize the squared cross correlation between all pairs of source estimates at two or more lags [5]–[7]. They are useful for BSS when the sources possess sufficient spectral diversity even if all the sources are Gaussian distributed. This is not the case for ICA-based approaches. Conversely, ICA methods minimize the statistical dependence of the source estimates at lag 0

[8]-[12]. The ICA-based approaches are useful when at most one source has a Gaussian probability density function (pdf) even if all the sources have identical spectra. This is not the case for decorrelation-based approaches. Consequently, the appropriate BSS algorithm for a given application depends on the spatial and temporal structure of the sources in question. A spatio-temporal BSS method, on the other hand, would be appropriate for either of the two aforementioned cases since it combines the ICA and decorrelation criteria into one algorithm.

The maximum-likelihood (ML) criterion [13] is a popular choice for BSS algorithms [14]-[16]. For this criterion, the demixing matrix is estimated by maximizing the likelihood of the observed data assuming all the source (target) distributions are known. Part of the reason that ML methods are popular is that the ML estimator is unbiased for a large class of arbitrarily chosen score functions [17] and ML estimates are consistent and asymptotically efficient [13]. Another reason they are popular is that they perform well on real data. One particular ML BSS method [10] is a major component of EEGLAB [18], which is a widely distributed software package for processing EEG data. However, there are three possible shortcomings of most ML BSS methods. First, the ML criterion is usually formulated to exploit either the spatial or the temporal structure of the sources, but not both. Second, ML BSS methods are rarely formulated as a function of latent variables, which allows the use of the efficient expectation-maximization (EM) optimization method. Third, few ML BSS methods adapt the target distributions to optimize the criterion. Nonadapting solutions are appropriate only when the true source distributions are known (or are known approximately) and may otherwise limit performance relative to methods that can learn appropriate target distributions directly from the data.

II. Autoregressive mixture of Gaussians

A. Motivation for the Proposed Method

Any BSS algorithm that attempts to address all three of these items requires the following: 1) a structure that is capable of explaining the temporal dependencies of each source, 2) a criterion that is formulated using latent variables, and 3) the selection of a function(s) (oftentimes in parametric form) that can adequately describe each target distribution.

Structures that have been used in BSS to explain the temporal dependencies of each source include an autoregressive (AR) model [16], [19], [20], a moving average (MA) model [21], [22], and a hidden Markov model (HMM) [17], [21], [23]-[25] (an AR and an HMM are combined in [26] and [27]). We use the AR model [28] instead of the HMM because it has a lower complexity, as measured in terms of the number of adaptable parameters. The complexity of the AR (and MA) grows linearly with a linear increase in model order, whereas the complexity of an HMM grows exponentially. Likewise, we prefer the AR model over the MA model in the forward path since it requires the estimation of an finite-impulse response (FIR) filter in the backward (demixing) path, which is inherently stable.

Target distributions that have been used or suggested for BSS include the MLP [29], activation function neuron (FAN) [30], [31], histogram/lookup table (LUT) neuron [31], Parzen window estimate [32], Silverman's FFT-based kernel density estimation [33], truncated Edgeworth expansion [34], truncated Gram-Charlier expansion [35], [36], generalized hyperbolic Cauchy [37], generalized Gaussian [26], [27], [38]-[42], mixture of generalized Gaussians [15], [32], mixture of cardinal spline kernels [17], mixture of Gaussians [14], [21], [23]-[25], [43]-[47], and a mixture of (derivative of) sigmoids [16], [19], [32], [39], [43], [48], [49].

When the target pdf is a mixture of K_Q kernels [13], then the variable $Q_{m,n} \in \{1, \dots, K_Q\}$, which determines the state of the m th source at the n th time instant, can be used as the latent variable required for EM formulation. Furthermore, if the target pdf uses this type of state variable then

it can be shown (see Appendix I) that the EM optimization of the ML criterion simplifies to a weighted sum of the desired signals evaluated using the log target conditional pdf (conditioned on the latent/state variable). Hence, if the target conditional pdf is Gaussian, then the criterion reduces to a quadratic function of the desired signals. Only a few of the target distributions listed previously can take advantage of this simplified form. They include the Parzen windows estimate, Silverman's method (which involves binning the data), and the mixture of Gaussians (MOG). The main tradeoff between these methods is in the number of states, which directly affects the computational complexity of the EM optimization, and the number of parameters that must be adapted. The number of inherent/suggested states K_Q is on the order of N (the number of exemplars), 400, and 5 for Parzen's, Silverman's, and the MOG distribution, respectively. Likewise, the number of parameters that must be adapted is $M(M + K_g)$, $M(M + K_g)$, and $M(M + K_g + 3K_Q)$, respectively, where K_g is defined in the following section. We choose the MOG distribution since it has a significantly lower computational complexity with only a modest increase in the number of adapted parameters.

Combining these choices produces the proposed ML BSS method, which we refer to as autoregressive mixture of Gaussians (AR-MOG) since it assumes that each source is an autoregressive, linear process and the pdf of each innovation [the independent identically distributed (i.i.d.) input of the linear process that produces the associated source] is a mixture of Gaussians.

B. Generative Model

It is assumed that there are M mutually statistically independent sources, each of which is N samples in length. The variable s represents the $(M \times N)$ source matrix, $s_{m,1:N}$ represents the $(1 \times N)$ vector of the m th row of s , and $s_{1:M,n}$ represents the $(M \times 1)$ vector of the n th column of s . Upper case letters are used to denote constants and random variables; lower case letters are used for independent variables (i.e., space and time) and realizations of the random variables; normal font is used for scalars; and bold font is used to denote matrices and vectors. Each source $s_{mn,n}$ is assumed to be an AR process that is generated from a temporally i.i.d. innovations process $u_{mn,n}$. 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 $g_{m,0} = 1$ $m \in \{1, 2, \dots, M\}$, $g_{m,k}$ is an element of the $(M \times K_g + 1)$ matrix $\mathbf{g}$ of AR coefficients, and K_g is the order of each of the AR filters (the variable K_g can be thought of as the maximum value of $K_g(m)$ over all M sources; with this interpretation, the AR coefficients of nonmaximal length filters are padded with the appropriate number of zeros). Therefore, the sources are given by

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

The M observations at time n are assumed to be generated from the sources by means of a linear, memoryless $(M \times M)$ mixing matrix $\mathbf{A}$

$$\mathbf{x}_n = \mathbf{A} \mathbf{s}_n \quad (2)$$

for $n \in \{1, 2, \dots, N\}$. The pdf of each innovations process is assumed to be parameterized by a mixture of Gaussians (as before, the variable K_Q can be thought of as the maximum value of $K_Q(m)$ over all M sources)

$$\begin{aligned}
p_{U_{m,n}}(u_{m,n}) &= \sum_{q=1}^{K_Q} p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) \\
&\quad \cdot p_{Q_{m,n}}(Q_{m,n}=q) \\
&= \sum_{q=1}^{K_Q} \mathcal{N}(u_{m,n}|\mu_{m,q}, v_{m,q}) \pi_{m,q}
\end{aligned} \tag{3}$$

which should not be confused with $p_{U_{m,n}}(\hat{u}_{m,n})$ (the target pdf of each innovations process) or $p_{\hat{U}_{m,n}}(\hat{u}_{m,n})$ (the actual pdf of the estimated innovations $\hat{\mathbf{u}}$, and where $p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q)$ has a normal distribution denoted by $\mathcal{N}$, $\mu_{m,q}$ is the mean of the q th component (or state) of the m th source, $v_{m,q}$ is the corresponding precision, $\pi_{m,q} \equiv p_{Q_{m,n}}(Q_{m,n}=q)$ is the

corresponding prior probability (constrained such that $\sum_{q=1}^{K_Q} \pi_{m,q} = 1 \forall m$), and $Q_{m,n} \in \{1, 2, \dots, K_Q\}$ represents the state (latent variable) of the m th source at the n th time point. This model, the AR-MOG generative model, is able to describe both the temporal dependencies of the sources and the non-Gaussianity of the source distributions.

C. Criterion

Let $p_{U_{m,n}}(u_{m,n})$ denote the marginal pdf of a particular innovations process and let $p_{\mathbf{U}}(\mathbf{u})$ and $p_{U_{1:M,n}}(u_{1:M,n})$ denote the order- MN and order- M joint pdfs of the innovations, respectively. The pdfs of the sources and the observations are defined similarly. Using this notation, and the generative model from Section II-B, the data likelihood is given by

$$\begin{aligned}
p_{\mathbf{X}}(\mathbf{x}) &= \prod_{n=1}^N p_{\mathbf{X}_{1:M,n}}|\mathbf{X}_{1:M,1:n-1}(\mathbf{x}_{1:M,n}|\mathbf{x}_{1:M,1:n-1}) \\
&= \prod_{n=1}^N |\mathbf{W}| p_{\mathbf{S}_{1:M,n}}|\mathbf{S}_{1:M,1:n-1}(\mathbf{s}_{1:M,n}|\mathbf{s}_{1:M,1:n-1}) \\
&= |\mathbf{W}|^N \prod_{m=1}^M \prod_{n=1}^N p_{S_{m,n}}|s_{m,1:n-1}(s_{m,n}|s_{m,1:n-1}) \\
&= |\mathbf{W}|^N \prod_{m=1}^M \prod_{n=1}^N p_{U_{m,n}}(u_{m,n})
\end{aligned} \tag{4}$$

where $\mathbf{W} = \mathbf{A}^{-1}$, it is understood that the set of all parameters $\{\mathbf{W}, \mathbf{g}, \boldsymbol{\mu}, \mathbf{v}, \boldsymbol{\pi}, K_Q, K_g\}$ is given for each pdf, for $n = 1$ the expression $\mathbf{X}_{1:M, 1:n-1} = \mathbf{X}_{1:M, 1:0} = \mathbf{S}_{1:M, 1:0} = \boldsymbol{\theta}$ (hence, $p_{\mathbf{X}_{1:M,n}}|\mathbf{X}_{1:M, 1:0}(\mathbf{x}_{1:M,n}|\mathbf{X}_{1:M, 1:0}) = p_{\mathbf{X}_{1:M,n}}(\mathbf{x}_{1:M,n})$; this notation is used to simplify the form of the previous equations), and where $\mu_{m,n}$ can be expressed as the following function of the observations and the parameters:

$$u_{m,n} = \sum_{k=0}^{K_g} \sum_{l=1}^M W_{m,l} g_{m,k} x_{l,n-k}. \tag{5}$$

The log likelihood is given by

$$\begin{aligned}
\mathcal{L} &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \ln p_{U_{m,n}}(u_{m,n}) \\
&= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \ln \mathcal{N}(u_{m,n}|\mu_{m,q}, v_{m,q}) \pi_{m,q}.
\end{aligned} \tag{6}$$

For FM optimization, we optimize a criterion that has a similar form

$$\begin{aligned}
 \mathcal{L}' &= N \ln \mathbf{W} + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} \\
 &\quad \cdot \ln \left(\frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) \pi_{m,q}}{\gamma_{m,n,q}} \right) \\
 &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} \\
 &\quad \cdot \ln \left(\frac{N(u_{m,n} | \mu_{m,q}, \nu_{m,q}) \pi_{m,q}}{\gamma_{m,n,q}} \right)
 \end{aligned} \tag{7}$$

where $\gamma_{m,n,q}$ is a distribution over the hidden variable so that $\sum_{q=1}^{K_Q} \gamma_{m,n,q} = 1$. Notice that the inner summation in (6) is moved outside the logarithm in (7) so that the criterion becomes a quadratic function of the innovations. Maximizing $\mathcal{L}'$ is equivalent, under mild conditions, to maximizing $\mathcal{L}$ [50]. Appendix I shows that $\mathcal{L}' = \mathcal{L}$ whenever $\gamma_{m,n,q} = p_{Q_{m,n}}(Q_{m,n} = q | \mathbf{x})$ (which occurs after each E-step as can be easily shown). Appendix II contains the full derivation of the EM algorithm for learning $\{\gamma, \mathbf{W}, \mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\pi}\}$. In the derivation, the two parameters that are not learned by AR-MOG $\{K_Q, K_g\}$ are assumed to be known.

D. Parameter Initialization

In this section, we provide recommendations for initializing the parameters that are adapted by the AR-MOG algorithm and recommendations for setting the parameter values that are not adapted. The demixing matrix $\mathbf{W}$ should be initialized using principal components analysis (PCA) so that the initial source estimates are spatially uncorrelated. Also, each row should be normalized so that it has an L^2 -norm that is commensurate with the standard deviation implied by the associated target distribution $p_{U_{m,n}}(u_{m,n})$. Linear prediction coefficients can be used to initialize each $\mathbf{g}_m$, $(1:K_g+1)$ vector. The initial values for the parameter vectors $\{\boldsymbol{\mu}_m, 1:K_Q, \boldsymbol{\nu}_m, 1:K_Q, \boldsymbol{\pi}_m, 1:K_Q\}$ can be obtained by an offline ML procedure that optimally fits $p_{U_{m,n}}(\hat{u}_{m,n})$ to the initial value of $p_{U_{m,n}}(\hat{u}_{m,n})$.

The recommended setting for the parameters not adapted by AR-MOG $\{K_Q, K_g\}$ depends on several items. If K_Q is chosen to be 1 (hence, the target $p_{U_{m,n}}(\hat{u}_{m,n})$ is Gaussian), then AR-MOG reduces to the second-order statistics BSS method by Mandic and Cichocki [51]. In this case, AR-MOG uses only spectral diversity in an attempt to separate the sources. Notice that this is true regardless of the true distribution of each innovation $p_{U_{m,n}}(u_{m,n})$. If K_g is chosen to be 0, then AR-MOG uses only spatial diversity in an attempt to separate the sources. Hence, $K_Q = 1$ should not be used with $K_g = 0$. Both spectral and spatial information are used as long as $K_Q \geq 2$ and $K_g \geq 1$. The suggestion is to use a value of K_Q between 1 and 5 inclusive. Larger values of K_Q allow for a more exact description of the desired target distribution. However, ML BSS algorithms often work well even if the source model is only approximately correct [52]–[55], so small values of K_Q are sufficient to reach maximum performance. The value of 0 is sufficient for K_g when each source is i.i.d. and small values of $K_g = 0$ are preferred when the observations contain wideband noise (wideband relative to the source spectra, as discussed in Section IV-B). BSS methods that do not attempt to exploit temporal diversity of the sources essentially use $K_g = 0$. Values of $K_g \geq 1$ are appropriate for temporally correlated sources. Increasing K_g provides AR-MOG with an increased amount of information on the temporal structure of the sources, which allows it to model arbitrarily complex spectra. Due to diminishing returns on separation performance, the suggested upper bound for K_g is 5.

E. Discussion

The proposed BSS algorithm, AR-MOG, infers the hidden variable γ and learns the parameters $\{\mathbf{W}, \mathbf{g}, \boldsymbol{\mu}, \mathbf{v}, \boldsymbol{\pi}\}$ in an attempt to maximize $\mathcal{L}$. Roughly speaking, maximizing $\mathcal{L}$ consists of the following four basic objectives.

1. Adapt $\mathbf{W}$ to make the set of $\hat{\mathbf{U}}_{m,1:N}$ (spatially) statistically independent.
2. Adapt each $\mathbf{g}$, $1:K_g+1$ to remove temporal dependencies in $\hat{\mathbf{U}}_{m,1:N}$
3. Adapt $\{\mathbf{W}, \mathbf{g}\}$ such that $p_{\hat{\mathbf{U}}_{m,n}}(\hat{u}_{m,n}) = p_{U_{m,n}}^-(\hat{u}_{m,n})$.
4. Adapt $\{\boldsymbol{\mu}, \mathbf{v}, \boldsymbol{\pi}\}$ such that $p_{\hat{\mathbf{U}}_{m,n}}(\hat{u}_{m,n}) = p_{U_{m,n}}^-(\hat{u}_{m,n})$.

The first, third, and fourth of these objectives are explicit in the derivation of AR-MOG, whereas the second objective is implicit. The second objective is easily explained when $p_{U_{m,n}}^-(\hat{u}_{m,n})$ is Gaussian for a given m , in which case it is trivial to prove that the corresponding $\mathbf{g}_m$, $(1:K_g+1)$ acts as a prediction error filter [56]. It is well known that prediction error filters act to produce a temporally whitened output (in this case, $\hat{u}_{m,n}$).

The intuitive reasoning why objectives three and four lead to separation is as follows. The range of $p_{\hat{\mathbf{U}}}(\hat{\mathbf{u}})$ is limited to that achievable by adjusting $\{\mathbf{W}, \mathbf{g}\}$ and $p_{U^-}(\hat{\mathbf{u}})$ is constrained such that it always factorizes over both space and time. The best possible agreement between these two distributions, therefore, requires that $p_{\hat{\mathbf{U}}}(\hat{\mathbf{u}})$ also factorizes over both space and time. When the data is drawn from the generative model used by AR-MOG, then it is possible for all four of these objectives to be met simultaneously (for sufficiently large K_Q, K_g). In general, the final solution represents a tradeoff between these objectives. It should be kept in mind that only the first objective directly relates to separation performance, whereas the other objectives play an ancillary role.

Table I shows the pseudocode that should be followed when using AR-MOG to infer the hidden variable and learn the parameters from data. The following steps should be repeated until convergence. As with any ML method optimized using EM, convergence may be to a local maximum.

III. Literature Review

The BSS algorithms referenced in Section II-A are almost exclusively ML based. Three of these are similar to the proposed method in that they use the ML criterion, they are spatio-temporal, and they adapt the target distribution. The context-sensitive generalization of ICA (cICA) method was introduced by Pearlmutter and Parra [16], [19] and later used by Lee *et al.* for convolutive mixtures [32]. This method uses an AR process and a mixture of (derivative of) logistic pdfs for the target distributions. The hidden Markov ICA (HMICA) method by Penny *et al.* combines a two-state HMM with an AR structure for each state and assumes a generalized Gaussian distribution [26], [27]. The method by Hyvarinen uses an AR process and switches between two static target distributions based on the kurtosis of each estimated source [20] (this is a form of pseudoadaptation; several other methods also switch between predefined targets [36], [37], [40], although these methods are not spatio-temporal). AR-MOG, on the other hand, is not constrained to unimodal, symmetric target distributions (unlike HMICA), adapts the target distribution to maximize the likelihood (unlike the Hyvarinen method), and uses the efficient EM optimization method (unlike all three). Furthermore, the combination of the MOG target distribution and the AR model allows AR-MOG to have analytical solutions for all the quantities in both the E-step and the M-step (unlike all three).

Other spatio-temporal BSS algorithms that are not ML based include MRMI-SIG [12], a method by Pham [57], [58], Jade_{TD} [11], and several variants of the Bayesian Ying-Yang

(BYY) learning paradigm [39], [43], [59] (BYY learning is similar to EM under certain conditions [39]). In addition, there exists an approach that approximates full spatio-temporal separation. One can apply an ICA algorithm to the observations, determine which sources possessing spatial diversity have been recovered (using kurtosis, for example), project the observations to a subspace that lies orthogonal to the columns of the demixing matrix associated with previously extracted sources, and then apply a decorrelation BSS algorithm to the projected observations [60]. Another approach can be used that does not produce spatio-temporal separation, although it does benefit from improved performance by temporally whitening the source estimates (which is one of the advantages of using spatio-temporal separation). In this approach, which requires that all observations possess similar spectra, the observations are temporally whitened by applying the same linear filter to each observation [61].

IV. Experiments

The separation performance of the synthetic mixtures is measured using the signal-to-interference ratio (SIR), which is defined for the case of no permutations by

$$\text{SIR} = \frac{1}{M} \sum_{m=1}^M 10 \log_{10} \text{SIR}_m \text{ (dB)} \quad (8)$$

where

$$\text{SIR}_m = \frac{\sum_{n=1}^N (\mathbf{W}_{m,1:M} \mathbf{A}_{1:M,m} s_{m,n})^2}{\sum_{\substack{m'=1 \\ (m' \neq m)}}^M \sum_{n=1}^N (\mathbf{W}_{m,1:M} \mathbf{A}_{1:M,m'} s_{m',n})^2} \quad (9)$$

and $\mathbf{W}$ is the estimate of the demixing matrix. This definition is easily extended to account for any possible permutation. The value of SIR reported in the following uses the permutation that yields the maximum SIR, which is the one of interest for BSS applications. The SIR as defined previously is the mean of the ratio (in decibels) of the source power in a given source estimate to the total power of all interfering sources. If auditory sources are used then a value of SIR ≥ 20 dB is indicative of inaudible interference. The sensitivity of separation performance to additive noise is measured using the signal-to-noise ratio (SNR)

$$\text{SNR} = \frac{1}{M} \sum_{m=1}^M 10 \log_{10} \frac{\sum_{n=1}^N x_{m,n}^2}{\sum_{n=1}^N v_{m,n}^2} \text{ (dB)} \quad (10)$$

where v refers to the zero-mean white Gaussian noise that is added to the observations.

A total of ten BSS algorithms are compared. They include the proposed method, minimization/maximization of Renyi's mutual information using the stochastic information gradient (MRMI-SIG) [12], joint approximate diagonalization of eigen-matrices with temporal decorrelation (JADE_{TD}) [11], FastICA [62], extended fourth-order blind identification (EFOBI) [63], [64], temporal decorrelation separation (TDSEP) [66], AMUSE [64], [67], and three methods from Parra. The three methods from Parra assume that the sources are non-

Gaussian (Parra_{NG}) [65], nonwhite (Parra_{NW}) [65], and nonstationary (Parra_{NS}) [65]. Three of these are spatio-temporal (AR-MOG, MRMI-SIG, and JADE_{TD}), three are ICA based (FastICA, EFOBI, and Parra_{NG}), three are decorrelation based (TDSEP, AMUSE, and Parra_{NW}), and one is based on nonstationarity (Parra_{NS}). The user-defined parameters for these methods are as follows: FastICA uses the cubic nonlinearity, both AMUSE and Parra_{NW} compute cross statistics for lags 0 and 1 ($K_g = 1$), and Parra_{NS} uses two equal-length, nonoverlapping windows. Unless stated otherwise, JADE_{TD} uses $K_g = 4$ (JADE_{TD} defaults to JADE [68] when $K_g = 0$) and AR-MOG uses $K_g = 4$ and $K_Q = 3$. For JADE_{TD}, all fourth-order statistics are calculated at lag 0 and the second-order statistics are calculated for lags of 0 to K_g . Likewise, the second-order statistics of TDSEP are calculated for lags of 0 to K_g (TDSEP only uses second-order statistics). MRMI-SIG, which is only defined for $K_g = 1$, uses information from lags 0 and 1. Whenever error bars are included the results represent the mean of ten Monte Carlo trials, where each trial uses a different set of sources, and the error bars represent one standard error.

A. Noiseless Data

The first experiment measures the mean results for ten Monte Carlo trials of $M = 2$ artificially mixed audio sources, where the two audio sources are randomly chosen for each trial from nine speech sources and one music source (the sources are sampled at 16.4 kHz). The results are shown in Fig. 1. The bar plots for $N = 5 * 10^3$ and $N = 20 * 10^3$ are stacked rather than over-laid. Thus, the performance of AR-MOG for $N = 20 * 10^3$ is $97.3 - 42.5 = 54.8$ dB. The proposed method, AR-MOG, produces the best performance for both data lengths. To keep from cluttering the subsequent figures, only the first four methods used in this experiment are used henceforth.

The second experiment uses $M = 2$ i.i.d. Laplacian-distributed innovations of length $N = 10 * 10^3$ samples. Each of the innovations is filtered with an order- K_h MA filter prior to mixing, where the filters have a truncated, exponentially decaying impulse response and the decay times of the two filters are unique. Fig. 2 shows the separation performance as a function of K_h . For these data, JADE performs almost identically with JADE_{TD} when $K_h = 0$, which is expected because there is no spectral diversity for this value of K_h . For larger values of K_g , the performance of JADE_{TD} is about 5-8 dB better than JADE. The performances of MRMI-SIG, JADE, and JADE_{TD} all fall as K_h is initially increased and then flatten out for larger K_h . TDSEP, on the other hand, performs poorly for $K_h = 0$, as expected, and then improves somewhat as K_h increases.

The third experiment uses $M = 2$ artificially mixed audio sources of length $N = 35 * 10^3$ samples. As was the case for the first experiment, the two sources are randomly selected from a set of nine speech sources and one music source. Fig. 3 shows the separation performance as a function of K_g . When $K_g = 0$, JADE_{TD} reduces to JADE and TDSEP reduces to PCA, the latter of which is insufficient for separating the sources. The results for MRMI-SIG are extended to either side of $K_g = 1$ for visualization purposes. Notice that the performance of AR-MOG improves somewhat as K_g increases even though speech is not statistically stationary (the window size is slightly over 2 s in length, whereas speech is only considered stationary for periods of approximately 30 ms [69]). The JADE_{TD} method, for these data, performs only marginally better than JADE.

Notice in Fig. 3 that AR-MOG performs well for values of K_g up to 10, which is twice the largest recommended value (larger values of K_g were not attempted). This is content with the idea that the proposed method is robust with respect to K_g . An additional experiment, not shown, was performed in order to measure the separation performance as a function of K_Q . This experiment used ten Monte Carlo trials, $M = 2$ speech sources, AR filters of order $K_g = 4$, and $N = 20 * 10^3$ samples. The proposed method performed well (40 dB or slightly higher

average SIR) for values of K_Q up to 10, which is (as before) twice the largest recommended value (larger values of K_Q were not attempted). Hence, the proposed method also appears to be robust with respect to K_Q .

B. Noisy Data

The fourth experiment estimates the separation performance as a function of SNR for a mixture of $M = 2$ super-Gaussian sources where $N = 10 * 10^3$. The condition number, defined as the ratio of the largest and smallest singular values, is 2.6 for this experiment. Each source is an AR process the poles of which are located on the real axis of the complex plane at $z = 0.7, 0.9$. The consequence of the pole locations is that the sources are relatively narrowband. Also, since each AR process is filtered using the same filter, the sources possess no spectral diversity for this experiment. The results for this experiment are shown in Fig. 4. Results are given for AR-MOG using both $\{K_Q = 4, K_g = 0\}$ and $\{K_Q = 4, K_g = 4\}$.

According to these data, $\text{JADE}_{\text{TD}}(K_g = 4)$ performs worse than JADE ($K_g = 0$) for SNR values below 30 dB, although it (slightly) outperforms JADE when there is no additive noise (not shown). AR-MOG ($K_g = 0$) performs the best of all methods for SNR values between 0 and 25 dB and the MRMI-SIG algorithm ($K_g = 1$) produces the best performance for SNR values between 25 and 30 dB. If no noise is included, then AR-MOG ($K_g = 4$) outperforms all the others. When noise is present, the separation performance of AR-MOG differs from our previous experience with noiseless mixtures, where roughly speaking the performance is a nondecreasing function of K_g . These results appear to indicate that the optimum value of K_g is inversely proportional to the power of the additive noise. In the following, we explain why, and under what circumstances, the separation performance can be expected to decrease as K_g increases whenever noise is present.

Supposing that the sources are mutually statistically independent, the goal of ICA-based and spatio-temporal BSS algorithms is to make the estimated sources/innovations (spatially) statistically independent. Suppose that there is no noise and that the sources are zero-mean and non-Gaussian. In this case, the $\mathbf{W}$ that corresponds to statistically independent source estimates is $\mathbf{W} = \mathbf{DPA}^{-1}$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is a permutation matrix. Now suppose that the observations consist of only noise and that the noise is zero mean, Gaussian, and isotropic. In this case, statistically independent source estimates are possible only when $\mathbf{W} = \mathbf{R}$, where $\mathbf{R}$ is a rotation matrix (which has the property $\mathbf{RR}^T = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix).

When noise is added to a mixture of sources, the parameters that maximize are a compromise between that required by the sources and that required by the noise. Consequently, the noise has the effect of biasing the desired solution. However, the noise and the sources can agree on a solution for $\mathbf{W}$ if $\mathbf{DPA}^{-1} = \mathbf{R}$. This is true if the mixing matrix just so happens to be given by $\mathbf{A} = \mathbf{R}^T \mathbf{DP}$. The permutation matrix has no effect on separation performance, the diagonal matrix can be absorbed by the sources, and $\mathbf{R}^T$ is a rotation matrix because $\mathbf{R}$ is a rotation matrix. Consequently, we can expect good separation performance in the presence of noise whenever $\mathbf{A}$ happens to be a rotation matrix and we can expect poor performance whenever $\mathbf{A}$ is not a rotation matrix. To test these expectations, we introduce the following mixing matrix:

$$\mathbf{A} = \begin{bmatrix} \cos \theta & \sin \theta & -\sin \theta & \cos \theta \\ -\sin \theta & \cos \theta & \cos \theta & \sin \theta \end{bmatrix} \quad (11)$$

where $\theta \in (-\pi/4, \pi/4)$, which represents one (effective) period due to the gain and permutation indeterminacies of BSS. The condition number for the mixing matrix lies between 1 and 6.3 for the values of θ shown in the figure ($-\pi/4 + \pi/20 \leq \theta \leq \pi/4 - \pi/20$; the extreme values of $-\pi/4$ and $\pi/4$ are avoided since the matrix is no longer full rank at these values). With this restriction,

$\mathbf{A}$ corresponds to a rotation matrix only when $\theta = 0$ and $\mathbf{A}$ becomes increasingly less like a rotation matrix as θ moves away from 0 in either direction. The results are shown in Fig. 5 for SNR = 0 dB and SNR = 10 dB. Keep in mind that these results are measured after applying $\mathbf{W}$, the learned demixing matrix. Since the BSS algorithm does not know the values that constitute the mixing matrix, it is possible that poor separation can occur even when the observations correspond to the original unmixed sources, e.g., when $\theta = 0$. In agreement with our expectations, the separation performance improves as θ approaches 0.

The previous arguments are valid for both ICA-based and spatio-temporal BSS methods. In addition to statistical independence, many spatio-temporal methods (including AR-MOG) also attempt to whiten the spectra of the innovation estimates. Let us assume, to simplify the argument, that the sources are narrowband and the noise is white. In this case, there necessarily exists a range of frequencies for which the noise has greater energy than the sources. Furthermore, this portion of the spectra must be boosted by $\mathbf{g}$ in order to whiten the spectra of the innovation estimates. This causes the SNR measured at the output of $\mathbf{g}$ to be lower than the SNR measured at the input to $\mathbf{g}$. This phenomenon is known in the communication literature as noise enhancement [70]. The ability of the spatio-temporal method to whiten arbitrary spectra is proportional to K_g . Hence, larger values of K_g increase the potential for noise enhancement. AR-MOG with $K_g = 0$ does not contain the filters with which it could boost portions of the spectra, hence it is not susceptible to noise enhancement (although the performance is still dependent on the noise power, as expected).

To test this second claim, we performed an additional experiment (not shown) where the spectra of the additive noise matched the source spectra, thus limiting noise enhancement. In this case, the performance of AR-MOG ($K_g = 4$) is better than the performances of all the other methods. More specifically, AR-MOG ($K_g = 4$) outperformed AR-MOG ($K_g = 0$) by 7.1 dB on average (averaged over ten Monte Carlo trials per SNR value and seven equally spaced SNR values from 0 to 30 dB) and JADE_{TD} ($K_g = 4$) outperformed JADE ($K_g = 0$) by 0.9 dB. This is to be contrasted with the results where the sources are narrowband relative to the noise, as was the case for the results shown in Fig. 4, for which the mean performance of AR-MOG ($K_g = 4$) was lower than AR-MOG ($K_g = 0$) and the mean performance of JADE_{TD} ($K_g = 4$) was lower than JADE ($K_g = 0$).

C. Ten-Dimensional Data

The next experiment shows the results of separating an artificial mixture of $M = 10$ real auditory sources, where the sources are the same as used previously and the data length is $N = 35 * 10^3$. The results are shown in Fig. 6 for AR-MOG ($K_g = 4$). Fig. 6(a) represents the amplitude of the product of the mixing matrix and the estimated demixing matrix. Ideally, there should be a single light colored square in each row and in each column and the remaining squares should be dark. Fig. 6(b) shows a bar plot of the values of $\text{SIR}_m \forall m$. The dashed line in this plot represents the final (mean) SIR. In this example, only 20-25 iterations are required for convergence. The final SIR for AR-MOG ($K_g = 4$) is 41.2 dB, whereas the final SIR values for JADE, TDSEP, and MRMI-SIG are 19.2, 12.7, and 30.0 dB, respectively. No results are shown for JADE_{TD} since the time of convergence, as the algorithm was implemented, becomes unbearably long as the number of sources increases.

D. Gaussian Sources

This experiment involves the separation of $M = 2$ Gaussian-distributed sources of length $N = 10 * 10^3$ samples. The addition of the AR filters enables the proposed algorithm to use spectral diversity of the sources to perform separation. This is an especially important consideration if each $PU_{m,n}(U_{m,n})$ is Gaussian-distributed since, in this case, spectral diversity of the (assumed stationary) sources is the only possible cue for inducing separation. AR-MOG ($K_Q = 1$, $K_g =$

4) is applied to ten different data sets, each of which consists of colored Gaussian sources. For each data set, the sources are fourth-order MA linear processes (this choice does not match the mode assumed by AR-MOG). The parameters of the MA processes are randomly selected so that the spectra of the sources are random. In all cases, the SIR after convergence exceeded 30 dB.

E. Real MCG Data

In the final experiment, we attempt to extract the fetal cardiac signal from real MCG data. The observations consist of a mixture of the desired fetal cardiac source, the (much larger amplitude) maternal cardiac source, and other biological and nonbiological sources. The MCG data consists of 1 min of observations from 37 sensors at a sampling rate of 1 kHz ($N = 60 * 10^3$). Using PCA, the dimensionality of the data was reduced to $M = 5$. This dimension was chosen based on an eigenanalysis, which indicated that five eigenvalues were sufficient to explain 97% of the total energy of the observations. Separation was performed using AR-MOG with $K_g = 4$. A 1.5-s window of five randomly chosen observations is shown in Fig. 7(a). Fig. 7(b) and (c) shows, respectively, the signals that are believed to correspond to the fetal and maternal cardiac sources. This inference is based on the resemblance of the recovered signals to a series of QRS complexes (QRS complexes correspond to ventricular depolarization), the fact that the recovered signals have the expected temporal spacing of QRS complexes, and on the expectation that the fetal heart rate (147 beats/min) is larger than the maternal heart rate (92 beats/min). Notice in Fig. 7(a) that the fetal QRS complex that occurs at 10.2 s is completely masked by the larger maternal signal. However, this fetal QRS complex is clearly visible in the recovered signal. Also notice that the recovered fetal cardiac signal has no visible contamination of the maternal cardiac signal.

V. Conclusion

There are three reasons why separation performance may improve by using a spatio-temporal BSS algorithm, such as AR-MOG. The most obvious reason is that some sources may exhibit only spatial diversity and others may exhibit only spectral diversity. Another reason is that, according to the central limit theorem, the innovations are less Gaussian than the sources (unless the innovations are themselves Gaussian distributed). It is well known that, to the extent that the sources lack spectral diversity, separation performance suffers as the source distributions approach the Gaussian distribution. Finally, from estimation theory, it is known that the accuracy of the estimation of moments of possibly nonlinear functions of random variables from finite-length data is improved when the random variable in question is temporally independent (hence i.i.d. since stationarity is assumed) [56]. These last two reasons explain why the separation performance for AR-MOG ($K_Q = 4$) shown in [71, Fig. 3] does not degrade as the length of the temporal correlation of the sources increases, whereas the performance of AR-MOG ($K_Q = 0$) reduces by 30 dB, even though the sources have no spectral diversity. These potential advantages must be weighed against the one known disadvantage, which is described in Section IV-B. Namely, there are certain conditions for which the performance of spatio-temporal BSS methods is reduced because of noise enhancement. The limitations caused by noise enhancement pertain to AR-MOG (only when $K_g \neq 0$) and to the three methods most similar to AR-MOG (cICA, HMICA, and Hyvarinen's approach).

Among all the spatio-temporal BSS methods, the proposed method has several potential advantages. It uses the ML criterion (which has several very nice properties), it uses the efficient EM optimization method, the target distributions are adapted to maximize the likelihood, and the update equations for all adapted quantities have a simple, analytical form.

Acknowledgments

This work was supported by the National Institutes of Health under Grant 1 F32 NS 52048-01.

Appendix I

Equivalence of $\mathcal{L}$ AND $\mathcal{L}'$

As stated earlier, $\mathcal{L}$ and $\mathcal{L}'$ from (6) and (7), respectively, are equal when $\gamma_{m,n,q} = P_{Q_{m,n}|X}(Q_{m,n} = q|x)$. To prove the equivalence, we need to prove that $P_{Q_{m,n}|U_{m,n} = q|u_{m,n}} = P_{Q_{m,n}|X}(Q_{m,n} = q|x)$. Working backwards, we have

$$\begin{aligned}
 & p_{Q_{m,n}} |_X (Q_{m,n}=q|x) \\
 &= \frac{p_X |_{Q_{m,n}}(x|Q_{m,n}=q) p_{Q_{m,n}}(Q_{m,n}=q)}{p_X(x)} \\
 &= \frac{1}{p_X(x)} \prod_{m'=1}^M \prod_{n'=1}^N p_{U_{m',n'}} |_{Q_{m,n}} \\
 &\quad \cdot (u_{m',n'} | Q_{m,n}=q) p_{Q_{m,n}}(Q_{m,n}=q) \\
 &= \frac{1}{p_X(x)} \prod_{m'=1}^M \prod_{n'=1}^N \left(\frac{p_{U_{m',n'}}(u_{m',n'}) p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q)}{p_{U_{m,n}}(u_{m,n})} \right) \\
 &\quad \cdot p_{Q_{m,n}}(Q_{m,n}=q) \\
 &= \frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) p_{Q_{m,n}}(Q_{m,n}=q)}{p_{U_{m,n}}(u_{m,n})} \\
 &= p_{Q_{m,n}|U_{m,n}}(Q_{m,n}=q|u_{m,n}). \tag{12}
 \end{aligned}$$

The third expression is explained as follows: since $u_{m',n'}$ is independent of $Q_{m,n}$ as long as either $m' \neq m$ or $n' \neq n$, the product over m', n' of the conditional probabilities of $u_{m',n'}$ given $Q_{m,n}$ equals the product of the (marginal) probabilities of $u_{m',n'}$ multiplied by the correction term needed when $m' = m, n' = n$. Making use of this result, the proof and $\mathcal{L} = \mathcal{L}'$ (when γ is the posterior over the hidden variable) is given by

$$\begin{aligned}
 \mathcal{L} &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \ln p_{U_{m,n}}(U_{m,n}) \\
 &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \left(\sum_{q=1}^{K_Q} \gamma_{m,n,q} \right) \\
 &\quad \cdot \ln \left(\frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q') p_{Q_{m,n}}(Q_{m,n}=q')}{p_{Q_{m,n}|U_{m,n}}(Q_{m,n}=q'|u_{m,n})} \right) \\
 &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} \\
 &\quad \cdot \ln \frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) p_{Q_{m,n}}(Q_{m,n}=q)}{p_{Q_{m,n}|U_{m,n}}(Q_{m,n}=q|u_{m,n})} \\
 &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} \\
 &\quad \cdot \ln \frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) p_{Q_{m,n}}(Q_{m,n}=q)}{p_{Q_{m,n}|X}(Q_{m,n}=q|x)} \\
 &= N \ln |\mathbf{W}| + \sum_{m=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} \\
 &\quad \cdot \ln \frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q) \pi_{m,q}}{\gamma_{m,n,q}} \tag{13}
 \end{aligned}$$

The second expression makes use of the fact that $\gamma_{m,n,q}$ is a proper distribution and uses an alternative expression for the marginal distribution, which comes directly from Bayes rule and is true for any q' . Because the second expression is true for any q' , then it is also true for $q' = q$, which results in the third expression. The fourth expression uses the results of (12) and the final expression uses the definition for $\pi_{m,q}$ and our original assumption that $\gamma_{m,n,q} = P_{Q_{m,n}}|X(Q_{m,n} = q|\mathbf{x})$.

Appendix II

EM Algorithm for AR-MOG

This Appendix provides the full derivation of the EM algorithm for AR-MOG. The EM optimization method is an iterative solution to maximizing the likelihood that alternates between the E-step and the M-step.

A. E-Step

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

$$\begin{aligned}\gamma_{m,n,q} &\equiv p_{Q_{m,n}}|X(Q_{m,n}=q|\mathbf{x}) \\ &= \frac{p_{\mathbf{X}|Q_{m,n}}(\mathbf{x}|Q_{m,n}=q)p_{Q_{m,n}}(Q_{m,n}=q)}{p_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{p_{U|Q_{m,n}}(\mathbf{u}|Q_{m,n}=q)\pi_{m,q}}{p_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{p_{U_{m,n}|Q_{m,n}}(u_{m,n}|Q_{m,n}=q)\pi_{m,q}}{\xi_{m,n}}\end{aligned}\quad (14)$$

where $\xi_{m,n}$ can be determined as the value needed to ensure that $\sum_{q=1}^{K_Q} \gamma_{m,n,q} = 1 \forall_{m,n}$. This result can also be found as the solution to $\partial \mathcal{L} / \partial \gamma_{m,n,q} = 0$. For purposes of adaptation, (14) may be expressed in terms of known quantities by replacing the true conditional distribution with the associated target distribution and replacing all other quantities with their estimates (to avoid complicating the terminology the same variable name is used for the true and the estimated parameters). This produces

$$\gamma_{m,n,q} = \frac{p_{\widehat{U}_{m,n}|Q_{m,n}}(\widehat{u}_{m,n}|Q_{m,n}=q)\pi_{m,q}}{\xi_{m,n}} \quad (15)$$

where

$$p_{\widehat{U}_{m,n}|Q_{m,n}}(\widehat{u}_{m,n}|Q_{m,n}=q) = \mathcal{N}(\widehat{u}_{m,n}|\mu_{m,q}, \nu_{m,q}) \quad (16)$$

is the target (conditional) distribution.

B. M-Step

The M-step maximizes the log likelihood w.r.t. the parameters $\{\mathbf{W}, \mathbf{g}, \boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\pi}\}$ while keeping the posteriors fixed. To find the solution for $\mathbf{W}$, we start by taking the derivative of $\mathcal{L}$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W_{m,l}} = & N(\mathbf{W}^{-T})_{m,l} \\ & - \sum_{q=1}^{K_Q} \sum_{n=1}^N \left(\gamma_{m,n,q} (\widehat{u}_{m,n} - \mu_{m,q}) \cdot \nu_{m,q} \sum_{k=1}^{K_g} g_{m,k} x_{l,n-k} \right) \end{aligned} \quad (17)$$

where $(\mathbf{B})_{m,l}$ indicates the scalar entry of matrix $\mathbf{B}$ that occurs at the m th row and the l th column. Postmultiplying both sides by $\mathbf{W}^T$ produces

$$\begin{aligned} & \left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \right)_{m,l} \\ &= N I_{m,l} - \sum_{q=1}^{K_Q} \sum_{n=1}^N \left(\gamma_{m,n,q} (\widehat{u}_{m,n} - \mu_{m,q}) \right. \\ & \quad \cdot \nu_{m,q} \sum_{k=1}^{K_g} \sum_{j=1}^M g_{m,k} W_{l,j} x_{j,n-k} \left. \right) \\ &= N I_{m,l} - \sum_{j=1}^M W_{l,j} \left(\sum_{k=1}^{K_g} g_{m,k} \sum_{n=1}^N x_{j,n-k} \right. \\ & \quad \cdot \sum_{q=1}^{K_Q} \gamma_{m,n,q} (\widehat{u}_{m,n} - \mu_{m,q}) \nu_{m,q} \left. \right). \end{aligned} \quad (18)$$

Using the following variables:

$$\begin{aligned} \Xi_{m,n} &= \sum_{q=1}^{K_Q} \gamma_{m,n,q} (\widehat{u}_{m,n} - \mu_{m,q}) \nu_{m,q} \\ Y_{m,j} &= \sum_{k=1}^{K_g} g_{m,k} \sum_{n=1}^N x_{j,n-k} \Xi_{m,n} \end{aligned} \quad (19)$$

we can express (18) in matrix notation as

$$\left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \right) = N \mathbf{I} - \mathbf{W} \mathbf{Y}^T. \quad (20)$$

Equating this to an $(M \times M)$ matrix of zeros yields the following analytical solution for $\mathbf{W}$:

$$\mathbf{W} = N \mathbf{Y}^{-T}. \quad (21)$$

Experience indicates that this solution is not as numerically robust as using several iterations of gradient ascent and it oscillates between solutions that pertain to different permutations, which complicates determination of convergence for the stopping criterion. As an alternative, $\mathbf{W}$ can be updated using several iterations of the natural gradient [72] (also known as the relative gradient [52]), which is given by

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} \quad (22)$$

where η is the step size and

$$\begin{aligned}
& \left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} \right)_{m,l} \\
&= \sum_{i=1}^M \left(N I_{m,i} - \sum_{n=1}^N \sum_{q=1}^{K_Q} \sum_{k=0}^{K_g} \gamma_{m,n,q} \right. \\
&\quad \cdot \left(\widehat{u}_{m,n} - \mu_{m,q} \right) v_{m,q} g_{m,k} \widehat{s}_{i,n-k} \Big) W_{i,l}.
\end{aligned} \tag{23}$$

The update for $\mathbf{g}$ is the solution to $\partial \mathcal{L} / \partial \mathbf{g} = 0$

$$\begin{aligned}
& \frac{\partial \mathcal{L}}{\partial g_{m,k}} \\
&= - \sum_{m'=1}^M \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m',n,q} v_{m',q} \left(\widehat{u}_{m',n} - \mu_{m',q} \right) \frac{\partial \widehat{u}_{m',n}}{\partial g_{m,k}} \\
&= - \sum_{k'=0}^{K_g} g_{m,k'} \left(\sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} v_{m,q} \widehat{s}_{m,n-k} \widehat{s}_{m,n-k'} \right) \\
&\quad + \left(\sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} v_{m,q} \mu_{m,q} \widehat{s}_{m,n-k} \right).
\end{aligned} \tag{24}$$

using the following variables:

$$\begin{aligned}
\Psi_{m,k,k'} &= \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} v_{m,q} \widehat{s}_{m,n-k} \widehat{s}_{m,n-k'} \\
\Phi_{m,k} &= \sum_{n=1}^N \sum_{q=1}^{K_Q} \gamma_{m,n,q} v_{m,q} \mu_{m,q} \widehat{s}_{m,n-k}
\end{aligned} \tag{25}$$

and equating (24) with 0 using vector notation produces

$$\mathbf{g}_{m,(1:K_g+1)} \Psi_{m,(1:K_g+1),(1:K_g+1)} = \Phi_{m,(1:K_g+1)} \tag{26}$$

which yields the following analytical solution:

$$\mathbf{g}_{m,(1:K_g+1)} = \Phi_{m,(1:K_g+1)} \left(\Psi_{m,(1:K_g+1),(1:K_g+1)} \right)^{-1} \tag{27}$$

for $m \in \{1, \dots, M\}$. The update for $\mu_{m,q}$ is the solution to $\partial \mathcal{L} / \partial \mu_{m,q} = 0$

$$\frac{\partial \mathcal{L}}{\partial \mu_{m,q}} = \sum_{n=1}^N \gamma_{m,n,q} v_{m,q} \left(\widehat{u}_{m,n} - \mu_{m,q} \right). \tag{28}$$

Equating (28) with 0 yields the following analytical solution:

$$\mu_{m,q} = \frac{\sum_{n=1}^N \widehat{u}_{m,n} \gamma_{m,n,q}}{\sum_{n=1}^N \gamma_{m,n,q}}. \tag{29}$$

Notice that this result is independent of $\psi_{m,q}$ which drops out of the equation since it appears in both the numerator and the denominator. The update for $\psi_{m,q}$ is the solution to $\partial\mathcal{L}/\partial\psi_{m,q}=0$

$$\frac{\partial\mathcal{L}}{\partial\psi_{m,q}} = \frac{1}{2} \sum_{n=1}^N \gamma_{m,n,q} \frac{1}{\psi_{m,q}} - \frac{1}{2} \sum_{n=1}^N \gamma_{m,n,q} (\hat{u}_{m,n} - \mu_{m,q})^2. \quad (30)$$

Equating (30) with 0 yields the following analytical solution:

$$\psi_{m,q} = \frac{\sum_{n=1}^N \gamma_{m,n,q}}{\sum_{n=1}^N \gamma_{m,n,q} (\hat{u}_{m,n} - \mu_{m,q})^2}. \quad (31)$$

The update for π must be constrained such that $\sum_{q=1}^{K_Q} \pi_{m,q} = 1 \forall m$. This is accomplished by performing an unconstrained optimization of $\bar{\mathcal{L}} = \mathcal{L} - \sum_{m=1}^M \lambda_m \left(\sum_{q=1}^{K_Q} \pi_{m,q} - 1 \right)$, where each λ_m is a Lagrange multiplier. This produces

$$\frac{\partial \bar{\mathcal{L}}}{\partial \pi_{m,q}} = \sum_{n=1}^N \gamma_{m,n,q} \frac{1}{\pi_{m,q}} - \lambda_m. \quad (32)$$

setting (32) to 0 and solving for $\pi_{m,q}$ produces

$$\pi_{m,q} = \frac{\sum_{n=1}^N \gamma_{m,n,q}}{\lambda_m}. \quad (33)$$

The m th Lagrange multiplier can be found by summing both sides of (33) over q and applying the constraint. Substituting this result in (32) yields the following analytical solution for the prior probabilities:

$$\pi_{m,q} = \frac{\sum_{n=1}^N \gamma_{m,n,q}}{\sum_{n=1}^N \sum_{q'=1}^{K_Q} \gamma_{m,n,q'}}. \quad (34)$$

Biography

Hagai T. Attias received the Ph.D. degree in theoretical physics from Yale University, New Haven, CT. In his dissertation, he created statistical models that successfully predicted electronic transport properties of a broad class of mesoscopic systems.

He was a Sloan Postdoctoral Fellow at the University of California, San Francisco (UCSF), where he did some of the pioneering work on machine learning algorithms for audio analysis and source separation. He then became a Research Scientist at Microsoft Research, Redmond, WA, working in the Machine Learning and Applied Statistics Group. Several of his inventions

at Microsoft were incorporated into the speech recognition engine now shipping in Windows Vista. He is currently the CEO of Golden Metallic, Inc., San Francisco, CA. He has (co) authored over 60 scientific publications on machine learning theory and its applications in speech and audio processing, machine vision, and biomedical imaging. He has nine issued patents and ten pending.

Kenneth E. Hild II (M'91-SM'05) received the B.S. and M.Sc. degrees in electrical engineering from the University of Oklahoma, Norman, in 1992 and 1996, respectively, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, in 2003,

where he studied information theoretic learning and blind source separation in the Computational NeuroEngineering Laboratory.

He studied biomedical informatics at Stanford University, Palo Alto, CA. From 1995 to 1999, he was an Advisory Development Engineer in the Advanced Concepts Group, Seagate Technologies, Inc. From 2000 to 2003, he taught several graduate-level classes on adaptive filter theory and stochastic processes at the University of Florida. Since 2003, he has been with the Biomagnetic Imaging Laboratory, University of California, San Francisco, where he is applying variational Bayesian techniques for biomedical signal processing of encephalographic and cardiographic data.

Srikantan S. Nagarajan (S'90-M'91-SM'07) received the M.S. and Ph.D. degrees in biomedical engineering from the Applied Neural Control Laboratory, Case Western Reserve University (CWRU), Cleveland, OH.

He was a Postdoctoral Fellow at the Keck Center for Integrative Neuroscience, University of California, San Francisco (UCSF), under the mentorship of Drs. M. Merzenich and C. Schreiner. Subsequently, he was a tenure-track Faculty Member in the Department of Bioengineering, University of Utah. Currently, he is the Director of the Biomagnetic Imaging Laboratory, an Associate Professor in Residence in the Department of Radiology, and a member in the UCSF/UCB Joint Graduate Program in Bioengineering. His research interests are in the area of neural engineering where his goal is to better understand dynamics of brain networks involved in processing and learning of complex human behaviors such as speech, through the development of functional brain imaging technologies.

References

- [1]. Vigario RN. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalogr. Clin. Neurophysiol Sep*;1997 103(3):395–404. [PubMed: 9305288]
- [2]. Nagarajan SS, Attias HT, Hild KE II, Sekihara K. A graphical model for estimating stimulus-evoked brain responses from MEG data with large background brain activity. *NeuroImage Apr*;2006 30(2):400–416. [PubMed: 16360320]
- [3]. Cao J, Murata N, Amari S, Cichocki A, Takeda T. A robust approach to independent component analysis of signals with high-level noise measurements. *IEEE Trans. Neural Netw May*;2003 14(3):631–645. [PubMed: 18238044]
- [4]. Comani S, Mantini D, Allea G, Di Luzio S, Romani GL. Fetal magnetocardiographic mapping using independent component analysis. *Physiol. Meas Dec*;2004 25(6):1459–1472. [PubMed: 15712724]
- [5]. Van Gerven S, Van Compernelle D. Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness. *IEEE Trans. Signal Process Jul*;1995 43(7):1602–1612.
- [6]. Weinstein E, Feder M, Oppenheim A. Multi-channel signal separation by decorrelation. *IEEE Trans. Speech Audio Process Oct*;1993 1(4):405–413.
- [7]. Wu H-C, Principe JC. A unifying criterion for blind source separation and decorrelation: Simultaneous diagonalization of correlation matrices. *Neural Netw. Signal Process Sep*;1997 :496–505.

- [8]. Koldovsky Z, Tichavsky P, Oja E. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramer-Rao lower bound. *IEEE Trans. Neural Netw* Sep;2006 17(5):1265–1277. [PubMed: 17001986]
- [9]. Boscolo R, Pan H, Roychowdhury VP. Independent component analysis based on nonparametric density estimation. *IEEE Trans. Neural Netw* Jan;2004 15(1):55–65. [PubMed: 15387247]
- [10]. Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* Nov;1995 7(6):1129–1159. [PubMed: 7584893]
- [11]. Muller, K-R.; Philips, P.; Ziehe, A. Jade: Combining higherorder statistics and temporal information for blind source separation (with noise); *Proc. Int. Workshop Independent Component Anal Signal Separat.*; Jan. 1999; p. 87-92.
- [12]. Hild KE II, Erdogmus D, Principe JC. An analysis of entropy estimators for blind source separation. *Signal Process* Jan;2006 86(1):182–194.
- [13]. Bishop, C. *Neural Networks for Pattern Recognition*. Oxford Univ. Press; Oxford, U.K.: 1995.
- [14]. Moulines, E.; Cardoso, J-F.; Gassiat, E. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models; *Proc. Int. Conf. Acoust. Speech Signal Process.*; Apr. 1997; p. 3617-3620.
- [15]. Pham DT, Garat P. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Signal Process* Jul;1997 45(7):1712–1725.
- [16]. Pearlmutter BA, Parra LC. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. *Adv. Neural Inf. Process. Syst* Dec;1996 :613–619.
- [17]. Hosseini S, Jutten C, Pham DT. Markovian source separation. *IEEE Trans. Signal Process* Dec; 2003 51(12):3009–3019.
- [18]. Delorme A, Makeig S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* Mar;2004 134(1):9–21. [PubMed: 15102499]
- [19]. Pearlmutter, BA.; Parra, LC. A context-sensitive generalization of ICA; *Proc. Int. Conf. Neural Inf. Process.*; Sep. 1996; p. 151-157.
- [20]. Hyvarinen A. A unifying model for blind separation of independent sources. *Signal Process* Jul; 2005 85(7):1419–1427.
- [21]. Attias H. Independent factor analysis with temporally structured sources. *Adv. Neural Inf. Process. Syst* Nov;2000 :386–392.
- [22]. Dyrholm, M.; Hansen, LK. CICAAR: Convolutional ICA with an auto-regressive inverse model; *Proc. Int. Workshop Independent Component Anal. Signal Separat.*; Sep. 2004; p. 594-601.
- [23]. Attias, H. Blind source separation by dynamical graphical models; *Proc. Int. Conf. Artif. Neural Netw.*; Sep. 1999; p. 126-131.
- [24]. Choudrey, RA.; Roberts, SJ. Bayesian ICA with Hidden Markov model sources; *Proc. Int. Workshop Independent Component Anal. Signal Separat.*; Apr. 2003; p. 809-814.
- [25]. Snoussi H, Mohammad-Djafari A. Bayesian unsupervised learning for source separation with mixture of Gaussians prior. *J. VLSI Signal Process. Syst* Jun./Jul;2004 37(23):263–279.
- [26]. Penny, WD.; Everson, RM.; Roberts, SJ.; Girolami, M., editors. *Advances in Independent Component Analysis*. Springer-Verlag; Berlin, Germany: 2000. Hidden Markov independent components analysis; p. 3-22.
- [27]. Penny, WD.; Roberts, SJ.; Everson, RM. Hidden Markov independent components for biosignal analysis; *Proc. Int. Conf. Med. Signal Inf. Process.*; Sep. 2000; p. 244-250.
- [28]. Jackson, LB. *Signals and Systems*. Addison-Wesley; Boston, MA: 1991.
- [29]. Roth Z, Baram Y. Multidimensional density shaping by sigmoids. *IEEE Trans. Neural Netw* Sep; 1996 7(5):1291–1298. [PubMed: 18263522]
- [30]. Fiori S, Bucciarelli P. Probability density estimation using adaptive activation function neurons. *Neural Process. Lett* Feb;2001 13(1):31–42.
- [31]. Fiori S. Hybrid independent component analysis by adaptive LUT activation function neurons. *Neural Netw* Jan;2002 15(1):85–94. [PubMed: 11958492]
- [32]. Lee, TW.; Orglmeister, R. A contextual blind separation of delayed and convolved sources; *Proc. Int. Conf. Acoust. Speech Signal Process.*; Apr. 1997; p. 1199-1202.

- [33]. Vlassis N, Motomura Y. Efficient source adaptivity in independent component analysis. *IEEE Trans. Neural Netw* May;2001 12(3):559–566. [PubMed: 18249888]
- [34]. Comon P. Independent component analysis, a new concept? *Signal Process* Apr;1994 36(3):287–314.
- [35]. Yang HH, Amari S-I. Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Comput* Oct;1997 9(7):1457–1482.
- [36]. Girolami, M.; Fyfe, C. Generalised independent component analysis through unsupervised learning with Bussgang properties; *Proc. Int. Conf. Neural Netw.*; Jun. 1997; p. 1788-1791.
- [37]. Lee T-W, Girolami M, Sejnowski TJ. Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Comput* Feb;1999 11(2): 417–441. [PubMed: 9950738]
- [38]. Liang, H.; Logothetis, NK. Adaptive independent component analysis through a maximum likelihood approach; *Proc. Int. Conf. Neural Netw. Brain*; Oct. 1998; p. 353-356.
- [39]. Xu L, Cheung CC, Amari S-I. Learned parametric mixture based ICA algorithm. *Neurocomputing* Jul;1998 22(13):69–80.
- [40]. Choi S, Cichocki A, Amari S-I. Flexible independent component analysis. *J. VLSI Signal Process. Syst* Aug;2000 26(12):25–38.
- [41]. Liang H. Adaptive independent component analysis of multichannel electrogastrograms. *Med. Eng. Phys* Mar;2001 23(2):91–97. [PubMed: 11413061]
- [42]. Ichir MM, Mohammad-Djafari A. A mean field approximation approach to blind source separation with Lp priors. *Eusipco*. Sep;2005
- [43]. Xu L. Independent component analysis and extensions with noise and time: A Bayesian Ying-Yang learning perspective. *Neural Inf. Process.—Lett. Rev* Oct;2003 1(1):1–52.
- [44]. Attias H. Independent factor analysis. *Neural Comput* May;1999 11(4):803–851. [PubMed: 10226184]
- [45]. Choudrey, RA.; Roberts, SJ. Flexible Bayesian independent component analysis for blind source separation; *Proc. Int. Workshop Independent Component Anal. Signal Separat.*; Dec. 2001; p. 90-95.
- [46]. Deligne S, Gopinath R. An EM algorithm for convolutive independent component analysis. *Neurocomputing* Dec;2002 49(14):187–211.
- [47]. Snoussi, H.; Mohammad-Djafari, A. *Neural Netw. Signal Process. Sep. 2001* Unsupervised learning for source separation with mixture of Gaussians prior for sources and Gaussian prior for mixture coefficients; p. 293-302.
- [48]. Xu, L.; Cheung, CC.; Yang, HH.; Amari, S-I. Independent Component Analysis by the information-theoretic approach with mixture of densities; *Proc. Int. Conf. Neural Netw.*; Jun. 1997; p. 1821-1826.
- [49]. Xu, L. Bayesian Ying-Yang system and theory as a unified statistical learning approach: (III) Models and algorithms for dependence reduction, data dimension reduction, ICA, and supervised learning; *Proc. Int. Workshop Theor. Aspects Neural Comput.: Multidiscip. Persp.*; May 1997; p. 43-60.
- [50]. Neal, RM.; Hinton, GE.; Jordan, MI., editors. *Learning in Graphical Models*. MIT Press; Cambridge, MA: 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants; p. 355-368.
- [51]. Mandic, DP.; Cichocki, A. An online algorithm for blind extraction of sources with different dynamical structures; *Proc. Int. Workshop Independent Component Anal. Signal Separat.*; Apr. 2003; p. 645-650.
- [52]. Cardoso J-F, Laheld BH. Equivariant adaptive source separation. *IEEE Trans. Signal Process* Dec; 1996 44(12):3017–3030.
- [53]. Amari S-I, Cardoso J-F. Blind source separation-semiparametric statistical approach. *IEEE Trans. Signal Process* Nov;1997 45(11):2692–2700.
- [54]. Cruces-Alvarez SA, Cichocki A, Amari S-I. On a new blind signal extraction algorithm: Different criteria and stability analysis. *IEEE Signal Process. Lett* Aug;2002 9(8):233–236.
- [55]. Cardoso J-F. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process. Lett* Apr;1997 4(4):112–114.

- [56]. Haykin, S. Adaptive Filter Theory. Vol. 4th ed.. Prentice-Hall; Englewood Cliffs, NJ: 2001.
- [57]. Pham, D-T. Mutual information approach to blind separation of stationary sources; Proc. Int. Workshop Independent Component Anal. Signal Separat.; Jan. 1999; p. 215-220.
- [58]. Pham DT. Mutual information approach to blind separation of stationary sources. IEEE Trans. Inf. Theory Jul;2002 48(7):1935–1946.
- [59]. Xu L. BYY harmony learning, independent state space, and generalized APT financial analysis. IEEE Trans. Neural Netw Jul;2001 12(4):822–849. [PubMed: 18249916]
- [60]. Castells F, Rieta JJ, Millet J, Zarzoso V. Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias. IEEE Trans. Biomed. Eng Feb;2005 52(2):258–267. [PubMed: 15709663]
- [61]. Hyvarinen, A. Independent component analysis for time-dependent stochastic processes; Proc. Int. Conf. Artif. Neural Netw.; Sep. 1998; p. 541-546.
- [62]. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw May;1999 10(3):626–634. [PubMed: 18252563]
- [63]. Soon, VC.; Tong, L.; Huang, YF.; Liu, R. An extended fourth order blind identification algorithm in spatially correlated noise; Proc. Int. Conf. Acoust. Speech Signal Process.; Apr. 1990; p. 1365-1368.
- [64]. Tong L, Liu R-W, Soon VC, Huang Y-F. Indeterminacy and identifiability of blind identification. IEEE Trans. Circuits Syst May;1991 38(5):499–509.
- [65]. Parra L, Sajda P. Blind source separation via generalized eigenvalue decomposition. J. Mach. Learn. Res Dec;2003 4:1261–1269.
- [66]. Ziehe, A.; Muller, K-R. TDSEP—An efficient algorithm for blind separation using time structure; Proc. Int. Conf. Artif. Neural Netw.; Sep. 1998; p. 675-680.
- [67]. Tong, L.; Soon, VC.; Huang, YF.; Liu, R. Amuse: A new blind identification algorithm; Proc. IEEE Int. Symp. Circuits Syst.; May 1990; p. 1784-1787.
- [68]. Cardoso JF, Souloumiac A. Blind beamforming for non-Gaussian signals. Inst. Electr. Eng. Proc. —F Dec;1993 140(6):362–370.
- [69]. Rabiner, LR.; Schafer, RW. Digital Processing of Speech Signals. Prentice-Hall; Upper Saddle River, NJ: 1978.
- [70]. Proakis, JG. Digital Communications. Vol. 3rd ed.. McGraw-Hill; Boston, MA: 1995.
- [71]. Hild, KE., II; Attias, HT.; Nagarajan, SS. An EM method for spatio-temporal blind source separation using an AR-MOG source model; Proc. Int. Workshop Independent Component Anal. Signal Separat.; Mar. 2006; p. 98-105.
- [72]. Amari S. Neural learning in structured parameter spaces—Natural Riemannian gradient. Adv. Neural Inf. Process. Syst Dec;1996 :127–133.

Fig. 1.

Separation performance for two data lengths averaged over ten Monte Carlo trials. The results for $N = 20 \times 10^3$ are stacked on the results for $N = 5 \times 10^3$.

Fig. 2.

Separation performance as a function of K_h . The error bars represent one standard error, which is based on ten Monte Carlo trials.

Fig. 3. Separation performance as a function of K_g . The error bars represent one standard error, which is based on ten Monte Carlo trials.

Fig. 4.

Separation performance as a function of SNR. The error bars represent one standard error, which is based on ten Monte Carlo trials.

Fig. 5. Separation performance as a function of the angle parameter θ , where the mixing matrix is given by (11) and white Gaussian noise is added to the sensors.

Fig. 6. Results of AR-MOG adaptation for demixing ten auditory sources. (a) Absolute value of the product of the mixing and demixing matrix. (b) SIR_m for all ten source estimates (the dotted line represents the mean SIR). (c) SIR as a function of iteration number. (d) Likelihood as a function of iteration number.

Fig. 7.

Results of the extraction of the fetal cardiac source from real MCG data. (a) Representative 1.5-s window of MCG recordings. (b) Recovered fetal cardiac signal using AR-MOG.

TABLE I

PSEUDOCODE FOR AR-MOG

E-step:	update γ using Eq. (15)
Convergence:	compute $\mathbf{L}$ using Eq. (7)
M-step:	update $\mathbf{W}$ using Eq. (21), or Eqs. (22), (23)
	update $\mathbf{g}$ using Eq. (27)
	update $\boldsymbol{\mu}$ using Eq. (29)
	update $\mathbf{v}$ using Eq. (31)
	update $\boldsymbol{\pi}$ using Eq. (34)