Mixture Convolutive Independent Component Analysis

Jason A. Palmer[1,3], Kenneth Kreutz-Delgado[1], Qin Wang[2], and Scott Makeig[3]

[1] Department of Electrical and Computer Engineering
University of California San Diego, La Jolla, CA 92093
{japalmer, kreutz}@ece.ucsd.edu
[2] Department of Computing Science
University of Alberta, Edmonton, Alberta, Canada, T6G 2E8
wqin@cs.ualberta.ca
[3] Swartz Center for Computational Neuroscience
University of California San Diego, La Jolla, CA 92093
scott@sccn.ucsd.edu

Abstract. We propose a mixture model for blind source separation and deconvolution with adaptive source densities. Data is modelled as a multivariate locally linear random process. We derive an expression for the asymptotic likelihood of a linear process segment, which allows us to formulate and optimize a mixture model via the EM algorithm. The mixture model is able to represent nonstationary (locally, or piecewise stationary) signals. We exploit a convexity-based inequality to ensure monotonic increase of the likelihood with respect to the source density parameters. The model is applied to analysis of EEG signals.

1 Introduction

The maximum likelihood/minimum mutual information approach to independent component analysis [1] has been applied successfully to the analysis of a variety of signals and data types. This framework has also been extended to multichannel deconvolution. In the convolutive case, algorithms are usually not developed probabilistically since the inverse z-transform is used in the definition of the likelihood, or cost function. In this paper we develop a probabilistic framework for the analysis of convolutive signals, allowing us to adapt a mixture model to deal with nonstationarity, and mixture source densities to adapt to arbitrary source densities.

Since multichannel deconvolution is a linear operation, the output can be expressed as the product of a (possibly infinite sized) matrix and an input vector, just as in the instantaneous linear mixing case. The essential difference between the convolutive case and the instantaneous case is that the matrix in the convolutive linear operation has a particular structure, specifically a block Toeplitz structure, and thus resides in a particular subspace. Thus derivatives of functions of the block Toeplitz matrix $\overline{\mathbf{W}}$, in particular the derivative of $\log|\det\overline{\mathbf{W}}|$ in the likelihood, will differ from derivatives of unconstrained demixing linear operators. However, the block Toeplitz structure allows us to calculate the determinant in terms of the blocks in a single row, using Szegö's limit formula concerning Toeplitz matrices. We can then calculate derivatives with respect to the individual blocks rather than the entire matrix. The block generalization of the Szegö

also allows us to efficiently calculate the likelihood, which we then use in an EM algorithm to adapt a mixture model involving multiple multichannel deconvolving filters models. The linear formulation of the convolutive case is particularly useful in deriving the natural gradient of the likelihood, since it can be formulated as the product of the ordinary block Toeplitz gradient with the Hermitian block Toeplitz matrix $\overline{\mathbf{W}}^T\overline{\mathbf{W}}$. The approach may be more intuitive for some than the z-transform approach used in [1].

2 Asymptotic Likelihood of Convolutive Component Model

We define a multivariate linear process [7, 14] to be a real-valued discrete-time multivariate random process of the form,

$$\mathbf{x}(t) = \sum_{k=-\infty}^{\infty} \mathbf{A}_k \, \mathbf{s}(t-k) \tag{1}$$

where $\mathbf{A}_k \in \mathbb{R}^{n\times n}$ with $\sum_k |[\mathbf{A}_k]_{ij}|^p < \infty$, and $\mathbf{s}(t)$ is i.i.d. for all t, and the components $s_i(t)$ of $\mathbf{s}(t)$ are independent (but not necessarily identically distributed). Note that the independent time series $\mathbf{s}(t)$ is distinct from the innovations representation of a second-order process. The innovation sequence is merely uncorrelated, and is only unique if it is causal. The independent series $\mathbf{s}(t)$ does not exist for every process, and is unique for all non-Gaussian linear processes [5, 15].

Given a multivariate discrete-time time series $\mathbf{x}(t)$, observed at $t = 1, 2, \ldots, T$, we divide the length T series into a set of time series segments of length $2N + 1 \ll T$. In the model, each segment is generated independently by one of M linear process models. Dependency among the segments (for example among overlapping segments) can be accounted for by imposing a Markov dependence structure on the segments, but we shall assume here for simplicity that the segments are generated independently of one another, with model prior probabilities γ_h, $h = 1, \ldots, M$.

Given a finite dimensional random vector $\mathbf{s}$ with density $p_{\mathbf{s}}(\mathbf{s})$ and an invertible linear transformation $\mathbf{A} = \mathbf{W}^{-1}$, the density of the random vector $\mathbf{x} = \mathbf{A}\mathbf{s}$ is given by,

$$p_{\mathbf{x}}(\mathbf{x}) \;=\; \frac{1}{|\det \mathbf{A}|}\, p_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x}) \;=\; |\det \mathbf{W}|\, p_{\mathbf{s}}(\mathbf{W}\mathbf{x})$$

The convolutive model (1) is a linear transformation of the process $\mathbf{s}(t)$. We suppose that the matrix filter is of finite duration, i.e. $\mathbf{A}_k = \mathbf{0}$ for $|k| > L$, where $L \ll N$. Let $N_0 = 2N + 1$. If we form the $nN_0 \times nN_0$ block Toeplitz matrix,

$$\overline{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_{-1} & \cdots & \\ \mathbf{A}_1 & \mathbf{A}_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{A}_{-1} \\ & \cdots & \mathbf{A}_1 & \mathbf{A}_0 \end{bmatrix}$$

and define the matrix $\mathbf{X}_t \equiv [\mathbf{x}(t-N) \;\cdots\; \mathbf{x}(t) \;\cdots\; \mathbf{x}(t+N)]$ and define $\mathbf{S}_t$ similarly, then we have,

$$\text{vec}(\mathbf{X}_t) \approx \overline{\mathbf{A}}\text{vec}(\mathbf{S}_t)$$

where the equation is only approximate for the first L and last L vectors in $\mathbf{X}_t$ since $\overline{\mathbf{A}}$ is block banded. We suppose that the (two-sided) inverse matrix filter exists,

$$\mathbf{W}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{l=-\infty}^{\infty}\mathbf{A}_l e^{-i\omega l}\right)^{-1} e^{i\omega k}\,d\omega$$

and can be approximated by the truncated filter with $\mathbf{W}_k = \mathbf{0}$ for $|k| > L$. Then we have,

$$\text{vec}(\mathbf{S}_t) \approx \bar{\mathbf{W}}_N \text{vec}(\mathbf{X}_t)$$

where we define,

$$\overline{\mathbf{W}}_N \equiv \overbrace{\begin{bmatrix} \mathbf{W}_0 & \mathbf{W}_{-1} & \cdots & \mathbf{W}_{-L} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{W}_1 & \mathbf{W}_0 & \mathbf{W}_{-1} & \ddots & \mathbf{W}_{-L} & \mathbf{0} & \ddots \\ \vdots & \mathbf{W}_1 & \mathbf{W}_0 & \ddots & & \ddots & \ddots \\ \mathbf{W}_L & \ddots & \ddots & \ddots & \mathbf{W}_{-1} & & \\ \mathbf{0} & \mathbf{W}_L & & \mathbf{W}_1 & \mathbf{W}_0 & \ddots & \\ \mathbf{0} & \mathbf{0} & \ddots & & \ddots & \ddots & \\ \vdots & \ddots & \ddots & & & & \end{bmatrix}}^{2N+1 \text{ blocks}}$$

Then we have,

$$p_{\mathbf{x}}\big(\text{vec}(\mathbf{X}_t)\big) \;\approx\; \big|\det\big(\overline{\mathbf{W}}_N \overline{\mathbf{W}}_N^T\big)\big|^{\frac{1}{2}}\, p_{\mathbf{s}}\big(\overline{\mathbf{W}}_N \text{vec}(\mathbf{X}_t)\big)$$

For large N, the matrix $\overline{\mathbf{W}}_N \overline{\mathbf{W}}_N{}^T$ tends to the symmetric block Toeplitz matrix $\overline{\mathbf{R}}_N$ with blocks $\mathbf{R}_k$ given by $\mathbf{R}_k = \sum_l \mathbf{W}_l \mathbf{W}_{k+l}^T$ for $k = -2L, \ldots, 2L$. To evaluate the determinant of $\overline{\mathbf{R}}_N$ asymptotically, we use the following extension of the classical Szegö limit theorem for Toeplitz matrices [6, 10].

Theorem 1 *For a set of block Toeplitz matrices $\{T_n\}$, generated by f, we have,*

$$\lim_{n\to\infty}\sum_{\lambda\in\sigma(T_n)} F(\lambda) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{k}\sum_{\lambda\in\sigma(f(x))} F(\lambda)dx$$

Taking $F = \log$, we have the following asymptotic form for the Toeplitz determinant:

$$\lim_{N\to\infty}\left(\det\overline{\mathbf{R}}_N\right)^{1/N_0} = \exp\left(\frac{1}{4\pi}\int_{-\pi}^{\pi}\log\det S_{\mathbf{W}}(\omega)\,d\omega\right)$$

where $N_0 = 2N+1$, and

$$S_{\mathbf{W}}(\omega) = \sum_k \mathbf{R}_k e^{-i\omega k} = \left(\sum_k \mathbf{W}_k e^{-i\omega k}\right)\left(\sum_k \mathbf{W}_k^T e^{i\omega k}\right)$$

Thus for the asymptotic approximation to the likelihood of the $n \times N_0$ sample $\mathbf{X}_t$ we have,

$$\frac{1}{N_0}\log p_{\mathbf{x}}(\mathbf{X}_t) \;\approx\; \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\left|\det\Big(\sum_k \mathbf{W}_k e^{-i\omega k}\Big)\right| d\omega + \frac{1}{N_0}\log p_{\mathbf{s}}\big(\overline{\mathbf{W}}_N \text{vec}(\mathbf{X}_t)\big)$$

In the implementation, we define the source estimates,

$$\mathbf{y}_\tau = \sum_{k=-L}^{L} \mathbf{W}_k \mathbf{x}_{t+\tau-k}, \quad \tau = -N+L, \ldots, N-L$$

The number of $\mathbf{y}$ vectors is less than the number of $\mathbf{x}$ vectors since we discard the edges. Thus, given the data segment $\mathbf{X}_t = [\mathbf{x}_{t-N} \cdots \mathbf{x}_{t+N}]$, we define the approximate likelihood $q_{\mathbf{x}}$ by,

$$\log q_{\mathbf{x}}(\mathbf{X}_t) \;=\; \frac{N_1}{2\pi}\int_{-\pi}^{\pi}\log\left|\det\Big(\sum_k \mathbf{W}_k e^{-i\omega k}\Big)\right| d\omega \;+ \sum_{\tau=-N+L}^{N-L}\sum_{i=1}^{n}\log q_i(y_{i\tau}) \quad (2)$$

where $N_1 = 2(N-L)+1$ and $q_i(y)$ is the approximating density of the ith source.

3 Maximizing the Likelihood

Let $\mathbf{C} \in \mathbb{C}^{n\times n}$ be square and non-singular. We use the complex derivative defined by

$$\frac{\partial g(\mathbf{C})}{\partial \mathbf{C}} = \frac{1}{2}\left(\frac{\partial g(\mathbf{C})}{\partial \text{Re}\mathbf{C}} - i\frac{\partial g(\mathbf{C})}{\partial \text{Im}\mathbf{C}}\right)$$

$\partial g/\partial \mathbf{C}^*$ is defined similarly but as a sum rather than a difference. If $g : \mathbb{C}^{n\times n} \to \mathbb{R}$ is a real valued function of a complex matrix, and $\mathbf{H} : \mathbb{R}^{n\times n} \to \mathbb{C}^{n\times n}$ is a complex matrix valued function of a real matrix, then we have the following chain rule

$$\frac{\partial}{\partial \mathbf{B}_{ij}} g\big(\mathbf{H}(\mathbf{B})\big) \;=\; \text{tr}\left(\frac{\partial g}{\partial \mathbf{H}}\frac{\partial \mathbf{H}^T}{\partial \mathbf{B}_{ij}} + \frac{\partial g}{\partial \mathbf{H}^*}\frac{\partial \mathbf{H}^{*T}}{\partial \mathbf{B}_{ij}}\right) \;=\; 2\,\text{Re tr}\left(\frac{\partial g}{\partial \mathbf{H}}\frac{\partial \mathbf{H}^T}{\partial \mathbf{B}_{ij}}\right)$$

since $\frac{\partial g}{\partial \mathbf{H}^*} = \left(\frac{\partial g}{\partial \mathbf{H}}\right)^*$ and $\frac{\partial \mathbf{H}^*}{\partial \mathbf{B}_{ij}} = \left(\frac{\partial \mathbf{H}}{\partial \mathbf{B}_{ij}}\right)^*$ according to our assumptions.

Now, using the fact that $(\partial/\partial \mathbf{C}) \log\det \mathbf{C}\mathbf{C}^H = \mathbf{C}^{-T}$, taking the derivative of the Toeplitz determinant term in (2) with respect to $\mathbf{W}_k$, we get,

$$\frac{\partial(1^{\text{st}}\,\text{term})}{\partial[\mathbf{W}_k]_{ij}} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\text{Re tr}\left[\Big(\sum_l \mathbf{W}_l e^{-i\omega l}\Big)^{-T} E_{ij}^T\, e^{-i\omega k}\right] d\omega$$

so that for the matrix derivative, we have,

$$\frac{\partial(1^{\text{st}}\,\text{term})}{\partial \mathbf{W}_k} \;=\; \text{Re}\,\frac{1}{2\pi}\int_{-\pi}^{\pi}\Big(\sum_l \mathbf{W}_l^T e^{-i\omega l}\Big)^{-1} e^{-i\omega k} d\omega \;=\; \mathbf{A}_{-k}^T$$

where E_{ij} is the matrix with 1 in the (i,j)th element and 0 elsewhere, and $\mathbf{A}_k$, $k = \ldots,-1,0,1,\ldots$ is kth element in the inverse filter of $\mathbf{W}_k$, $k = \ldots,-1,0,1,\ldots$. Thus the gradient of the determinant term is the block Teoplitz matrix $\overline{\mathbf{A}}^T$ yielding the natural gradient with respect to the determinant term,

$$\overline{\mathbf{A}}^T \overline{\mathbf{W}}^T \overline{\mathbf{W}} = \overline{\mathbf{W}} \tag{3}$$

For the derivative of the second term in (2) with respect to $\mathbf{W}_k$, we have,

$$\frac{\partial(2^{\text{nd}}\,\text{term})}{\partial \mathbf{W}_k} = \sum_\tau \mathbf{g}_\tau \mathbf{x}_{t+\tau-k}^T \tag{4}$$

where $\mathbf{g}_\tau \equiv -\nabla_{\mathbf{y}} \log q(\mathbf{y}_\tau)$. Multiplying the block Toeplitz matrix with blocks given by (4) by $\overline{\mathbf{W}}^T$ on the right, we get the block Toeplitz matrix with blocks

$$\sum_\tau \sum_l \mathbf{g}_\tau \mathbf{x}_{t+\tau-l}^T \mathbf{W}_{l-k}^T \;=\; \sum_\tau \mathbf{g}_\tau \mathbf{y}_{\tau-k}^T$$

Then multiplying this on the right by $\overline{\mathbf{W}}$, we get the block Toeplitz matrix with blocks,

$$\sum_\tau \sum_l \mathbf{g}_\tau \mathbf{y}_{\tau-l}^T \mathbf{W}_{k-l} \;=\; \sum_\tau \mathbf{g}_\tau \mathbf{u}_{\tau-k}^T \tag{5}$$

where we define,

$$\mathbf{u}_\tau \equiv \sum_{l=-L}^{L} \mathbf{W}_l^T \mathbf{y}_{\tau+l}, \;\; \tau = -N+2L, \ldots, N-2L$$

Again the number of $\mathbf{u}$ vectors is smaller than the number of $\mathbf{y}$ vectors since we discard the edges. The natural gradient is then,

$$\Delta \mathbf{W}_k = \mathbf{W}_k - \frac{1}{TN_2} \sum_{t=1}^{T} \sum_{\tau=-N+2L}^{N-2L} \mathbf{g}_{\tau+k} \mathbf{u}_\tau^T$$

4 Mixtures of Strong Super-Gaussians

Definition 1 *A symmetric univariate density $p(s)$ is **strongly super-gaussian** if $g(s) \equiv -\log p(\sqrt{s})$ is concave on $(0,\infty)$, and **strongly sub-gaussian** if $g(s)$ is convex.*

In [13] we discuss these densities in some detail, and derive relationships between them and the hyperprior representation used in the evidence framework [9] and the Variational Bayes framework [2]. Here we limit consideration to strongly super-gaussian mixture densities. If $p(s)$ is strongly super-gaussian, we have $f(s) \equiv g(s^2)$, with g concave on $(0,\infty)$. This implies that, $\forall\, s,t$,

$$f(t) - f(s) \;=\; g\big(t^2\big) - g\big(s^2\big) \;\leq\; g'\big(s^2\big)\big(t^2 - s^2\big) \;=\; \tfrac{1}{2}\tfrac{f'(s)}{s}\big(t^2 - s^2\big) \tag{6}$$

Table 1. Some common strongly super-gaussian densities.

Density Name	Density Form $\propto$	$f'(s)/s$
Gen. Gauss., $0 < \rho \leq 2$	$\exp(-\lvert s\rvert^{\rho})$	$\lvert s\rvert^{\rho-2}$
Student's t, $\nu > 0$	$(1 + s^2/\nu)^{-(\nu+1)/2}$	$(\nu+1)/(\nu+s^2)$
Jeffrey's prior	$1/s$	$1/s^2$
Logistic	$1/\cosh^2(s/2)$	$\tanh(s/2)/s$
Symmetric α-stable	no closed form	no closed form

In the inequality (6), the term $f'(s)/s$ is a fixed weight in a quadratic function of t and plays the role of a variance or scale parameter. Table 1 gives the form of this variational weight parameter for some common strongly super-gaussian densities. The algorithm is developed for the class of densities that are mixtures of strongly super-gaussian densities,

$$p(s) = \sum_{j=1}^{m} \alpha_j \sqrt{\beta_j}\, p_j\Big(\sqrt{\beta_j}\big(s - \mu_j\big)\Big)$$

where $\sum_j \alpha_j = 1$, $\alpha_j \geq 0$, $\beta_j > 0$, and the p_j are strongly super-gaussian. Note that $p(s)$ is not strongly super-gaussian in general. The use of different location parameters μ_j allows the representation of sub-gaussian densities for example.

5 The EM Algorithm

We follow the framework of [11] and [16] in deriving the EM algorithm. The log likelihood of the data decomposes as follows,

$$\begin{aligned} \log p(\mathbf{x};\theta) &= \int q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}\, d\mathbf{z} \;+\; D\big(q(\mathbf{z}) \big\| p(\mathbf{z}|\mathbf{x};\theta)\big) \\ &\equiv -F(q;\theta) \;+\; D(q\,\|\,p_\theta) \end{aligned} \tag{7}$$

where q is an arbitrary density and D is the Kullback-Leibler divergence. The term $F(q;\theta)$ is commonly called the *variational free energy* [16, 11]. This representation is useful if $F(q;\theta)$ can be easily minimized with respect to θ.

Since the KL divergence is non-negative, and equal to 0 if $q = p_\theta$, and the left hand side of (7), it follows that,

$$-\log p(\mathbf{x};\theta) = \min_q F(q;\theta)$$

where equality is obtained if and only if $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x};\theta)$ almost everywhere. The EM algorithm, then, at the lth iteration, given θ^l, proceeds as follows,

$$q^l = p\big(\mathbf{z}|\mathbf{x};\theta^l\big)\,, \quad \theta^{l+1} = \arg\min_\theta F\big(q^l;\theta\big)$$

This algorithm is guaranteed to increase the likelihood since,

$$-\log p(\mathbf{x};\theta^{l+1}) \;=\; F(q^{l+1};\theta^{l+1}) \;\leq\; F(q^l;\theta^{l+1}) \;\leq\; F(q^l;\theta^l) \;=\; -\log p(\mathbf{x};\theta^l)$$

Note that it is not necessary to find the actual minimum of F with respect to θ in order to guarantee that the likelihood increases. It is enough to guarantee that $F(q^l;\theta^{l+1}) \leq F(q^l;\theta^l)$, i.e. that F decreases as a result of updating θ. This leads to the Generalized EM (GEM) algorithm [4], which we employ in this paper.

To guarantee a decrease in $F(q;\theta)$ with respect to θ, we use the inequality (6) to define a function $\tilde{F}(q;\theta)$ which it is possible to minimize with respect to θ, and which satisfies, for all θ, θ',

$$F(q;\theta') - F(q;\theta) \;\leq\; \tilde{F}(q;\theta') - \tilde{F}(q;\theta)$$

Setting θ^{l+1} to minimize $\tilde{F}(q^l;\theta)$ over θ then guarantees, using the inequality (6), that,

$$F(q^l;\theta^{l+1}) - F(q^l;\theta^l) \;\leq\; \tilde{F}(q^l;\theta^{l+1}) - \tilde{F}(q^l;\theta^l) \;\leq\; 0$$

and thus that $F(q^l;\theta)$ is decreased as required by the GEM algorithm.

6 Convolutive Mixture model

We extend the instantaneous model described in [12] to include convolutive mixing. Assuming independent segments $\mathbf{X}_t$, $t = 1,\ldots,T$, we have for the likelihood,

$$p\big(\{\mathbf{X}_t\}\big) = \prod_{t=1}^{T}\sum_{h=1}^{M}\gamma_h p\big(\mathbf{X}_t \,|\, h\big)$$

The parameters to be estimated are $\theta = \big\{\gamma_h, \mathbf{W}_{hk}, \alpha_{hij}, \mu_{hij}, \beta_{hij}\big\}$, $k = -L,\ldots,L$, $h = 1,\ldots,M$, $i = 1,\ldots,n_h$, $j = 1,\ldots,m_{hi}$. In this model, each segment $\mathbf{X}_t$ is generated (independently) by drawing a mixture component h' from the discrete probability distribution $P[h' = h] = \gamma_h$, $1 \leq h \leq M$, then drawing $\mathbf{X}_t$ from $p_{h'}(\mathbf{X};\theta)$.

We define h_t to be the index chosen for the tth segment, and we define the random variable $v_{ht} = 1$ if $h_t = h$ and 0 otherwise. Let $\mathbf{V} \equiv \{v_{hk}\}$. Now, for the complete log likelihood of $\{\mathbf{X}_t\}_{t=1}^T$ and $\mathbf{V}$, we can write,

$$p\big(\{\mathbf{X}_t\},\mathbf{V};\theta\big) = \prod_{t=1}^{T}\prod_{h=1}^{M}\gamma_h^{v_{ht}} p(\mathbf{X}_t|h;\theta)^{v_{ht}}$$

We define $j_{hit\tau}$ to be the source mixture component index chosen (independently of h_t) for the ith source of the hth model in τth index of the tth segment, and we define the random variables $z_{hijt\tau}$ by, $z_{hijt\tau} = 1$ if $j_{hit\tau} = j$ and 0 otherwise, with $\mathbf{Z} \equiv \{z_{hijt\tau}\}$. We define,

$$y_{hijt\tau} \equiv \sqrt{\beta_{hij}}\left(\sum_{k=-L}^{L}\mathbf{w}_{hik}^T\mathbf{x}_{t+\tau-k} - \mu_{hij}\right)$$

Then we have

$$p(\mathbf{X}_t,\mathbf{Z}|h;\theta) = \exp\left(\frac{N_1}{2\,\pi}\int_{-\pi}^{\pi}\log|\det\mathbf{W}_h(\omega)|\,d\omega\right)\;\times \prod_{\tau=-N+L}^{N-L}\prod_{i=1}^{n_h}\prod_{j=1}^{m_{hi}}\Big[\alpha_{hij}\sqrt{\beta_{hij}}\,q_{hij}(y_{hijt\tau})\Big]^{z_{hijt\tau}}$$

where $N_1 = 2(N-L)+1$ and $\mathbf{W}_h(\omega) \equiv \sum_k \mathbf{W}_k\, e^{-i\omega k}$. For the joint distribution, or "complete likelihood," we have then,

$$p\big(\{\mathbf{X}_t\},\mathbf{V},\mathbf{Z};\theta\big) = \prod_{t=1}^{T}\prod_{h=1}^{M} \gamma_h^{v_{ht}}\, p(\mathbf{X}_t,\mathbf{Z}|h;\theta)^{v_{ht}}$$

For the variational free energy we have $F(q^l;\theta) = F^l(\theta) + H(\mathbf{V};\theta^l) + H(\mathbf{Z};\theta^l)$, where $H(\mathbf{V};\theta^l)$ and $H(\mathbf{Z};\theta^l)$ are the entropies of $\mathbf{V}$ and $\mathbf{Z}$ with the parameters set to θ^l, and,

$$F^l(\theta) \equiv \sum_{t=1}^{T}\sum_{h=1}^{M}\Bigg[\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\sum_{\tau=-N+L}^{N-L} E\big[v_{ht}z_{hijt\tau}|\mathbf{X}_t;\theta^l\big]\ \times \Big(-\log\alpha_{hij} - \tfrac{1}{2}\log\beta_{hij} + f_{hij}(y_{hijt\tau})\Big)\Bigg] + E\big[v_{ht}|\mathbf{X}_t;\theta^l\big]\Big(-\log\gamma_h - \frac{N_1}{2\pi}\int_{-\pi}^{\pi}\log\big|\det\mathbf{W}_h(\omega)\big|\,d\omega\Big)$$

where we define $f_{hij} \equiv -\log q_{hij}$. We define $\hat{z}^l_{hijk}$ to be the conditional expectation,

$$\hat{z}^l_{hijt\tau} \;\equiv\; E\big[z_{hijt\tau}\big|\, v_{ht}{=}1, \mathbf{X}_t,;\theta^l\big] \;=\; \frac{\alpha^l_{hij}\sqrt{\beta^l_{hij}}\; q_{hij}\big(y^l_{hijt\tau}\big)}{\sum_{j'=1}^{m_{hi}}\alpha^l_{hij'}\sqrt{\beta^l_{hij'}}\; q_{hij'}\big(y^l_{hij't\tau}\big)}$$

where we use Bayes' rule to evaluate the (discrete) posterior distribution. Simlarly, the $\hat{v}^l_{ht} \equiv E[v_{ht}|\mathbf{X}_t;\theta^l]$ are given by,

$$\hat{v}^l_{ht} \;=\; \frac{p\,(\mathbf{X}_t|v_{ht}=1;\theta^l)P[v_{ht}=1;\theta^l]}{\sum_{h'=1}^{M}p\,(\mathbf{X}_t|v_{h't}=1;\theta^l)P[v_{h't}=1;\theta^l]} \;=\; \frac{\gamma^l_h p(\mathbf{X}_t|h;\theta^l)}{\sum_{h'=1}^{M}\gamma^l_{h'}p(\mathbf{X}_t|h';\theta^l)}$$

and we have,

$$\begin{aligned} E\big[v_{ht}z_{hijt\tau}|\mathbf{X}_t;\theta^l\big] &= P\big[v_{ht}=1, z_{hijt\tau}{=}1\,|\,\mathbf{X}_t;\theta^l\big] \\ &= P\big[v_{ht}{=}1\,|\,\mathbf{X}_t;\theta^l\big]\,P\big[z_{hijt\tau}{=}1\,|\,v_{ht}=1,\mathbf{X}_t;\theta^l\big] \\ &= \hat{v}^l_{ht}\hat{z}^l_{hijt\tau} \end{aligned}$$

Minimizing F over γ_h and α_{hij} subject to the positivity and normalization constraints, we get,

$$\gamma^{l+1}_h = \frac{1}{T}\sum_{t=1}^{T}\hat{v}^l_{ht}\,, \quad \alpha^{l+1}_{hij} = \frac{1}{TN_1\gamma^{l+1}_h}\sum_{t=1}^{T}\hat{v}^l_{ht}\sum_{\tau=-N+L}^{N-L}\hat{z}^l_{hijt\tau}$$

where $N_1 = 2(N-L)+1$.

Now, to determine the updates for μ_{hij} and β_{hij}, we use the inequality (6) to replace $f_{hij}(y_{hijt\tau})$ in $F^l(\theta)$ by $\frac{1}{2}\xi^l_{hijt\tau}y^2_{hijt\tau}$, where,

$$\xi^l_{hijt\tau} \;\equiv\; \frac{f'_{hij}(y^l_{hijt\tau})}{y^l_{hijt\tau}} \tag{8}$$

Our "surrogate" free energy is then,

$$\tilde{F}^l(\theta) \;=\; \sum_{t=1}^{T}\sum_{h=1}^{M} \hat{v}_{ht}^l \Bigg[\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} \sum_{\tau=-N+L}^{N-L} \hat{z}_{hijt\tau}^l \times \Big(-\log\alpha_{hij} \;-\; \tfrac{1}{2}\log\beta_{hij} \;+\; \tfrac{1}{2}\xi_{hijt\tau}^l y_{hijt\tau}^2 \Big) \Bigg] + \hat{v}_{ht}^l \Big(-\log\gamma_h - \frac{N_1}{2\pi}\int_{-\pi}^{\pi} \log\big|\det \mathbf{W}_h(\omega)\big|\, d\omega \Big)$$

Minimizing $\tilde{F}^l$ with respect to μ_{hij} and β_{hij} guarantees, using the inequality (6), that,

$$F(q^l;\theta^{l+1}) - F(q^l;\theta^l) \;\le\; \tilde{F}(q^l;\theta^{l+1}) - \tilde{F}(q^l;\theta^l) \;\le\; 0$$

and thus that $F(q^l;\theta)$ is decreased as required by the EM algorithm. As in the Gaussian mixture case, the optimal value of μ_{hij} does not depend on β_{hij}. The updates are found to be,

$$\mu_{hij}^{l+1} = \mu_{hij}^l + \frac{\sum_{t=1}^{T} \hat{v}_{ht}^l \sum_{\tau=-N+L}^{N-L} \hat{z}_{hijt\tau}^l f'_{hij}(y_{hijt\tau}^l)}{\sqrt{\beta_{hij}^l}\sum_{t=1}^{T} \hat{v}_{ht}^l \sum_{\tau=-N+L}^{N-L} \hat{z}_{hijt\tau}^l \xi_{hijt\tau}^l}$$

$$\beta_{hij}^{l+1} = \frac{\beta_{hij}^l \sum_{t=1}^{T} \hat{v}_{ht}^l \sum_{\tau=-N+L}^{N-L} \hat{z}_{hijt\tau}^l}{\sum_{t=1}^{T} \hat{v}_{ht}^l \sum_{\tau=-N+L}^{N-L} \hat{z}_{hijt\tau}^l f'_{hij}(y_{hijt\tau}^l) y_{hijt\tau}^l}$$

Since $\xi = f'(y)/y$ may go to infinity at $y = 0$ for strongly super-gaussian densities, we have eliminated it from the updates except in the denominator of the μ update, where ξ becoming infinite (with $f'(y)$ remaining bounded) has the effect of keeping μ constant.

Now we make the definitions,

$$\mathbf{b}_{ht\tau}^l \equiv \sum_{k=-L}^{L} \mathbf{W}_{hk}^l \mathbf{x}_{t+\tau-k}$$

for $\tau = -N+L, \ldots, N-L$, and

$$\mathbf{u}_{ht\tau}^l \equiv \sum_{k=-L}^{L} {\mathbf{W}_{hk}^l}^T \mathbf{b}_{ht(\tau+k)}^l$$

for $\tau = -N+2L, \ldots, N-2L$. We define

$$y_{hijt\tau}^l \equiv \sqrt{\beta_{hij}^l}\Big(b_{hit\tau}^l - \mu_{hij} \Big)$$

for $\tau = -N+L, \ldots, N-L$ and we define the vector $\mathbf{g}_{ht\tau}^l$ such that,

$$g_{hit\tau}^l \equiv \hat{v}_{ht}^l \sum_{j=1}^{m_{hi}} \hat{z}_{hijt\tau}^l \sqrt{\beta_{hij}^l}\, f'_{hij}(y_{hijt\tau}^l)$$

Then the natural gradient direction for $\mathbf{W}_{hk}$ is given by,

$$\Delta\mathbf{W}_{hk} = \gamma_h^{l+1}\mathbf{W}_{hk}^l - \frac{1}{TN_2}\sum_{t=1}^{T}\sum_{\tau=-N+2L}^{N-2L}\mathbf{g}_{ht(\tau+k)}^l\mathbf{u}_{ht\tau}^{l\,T}$$

where we have replaced N_1 by N_2 to reduce the bias since we are discarding the edges. To express this in a form more efficient for computation, we define the subset of $\mathbf{X}_t$, $\tilde{\mathbf{X}}_{t,k} \equiv [\mathbf{x}_{t-k-N+L}\cdots\mathbf{x}_{t-k+N-L}]$. Then, for $\mathbf{B}_t \equiv [\mathbf{b}_{-N+L}\cdots\mathbf{b}_{N-L}]$, we have,

$$\mathbf{B}_t = \sum_{k=-L}^{L}\mathbf{W}_k\tilde{\mathbf{X}}_{t,k}$$

Now if we define $\tilde{\mathbf{B}}_{t,k} \equiv [\mathbf{b}_{k-N+2L}\cdots\mathbf{b}_{k+N-2L}]$ and put $\mathbf{U}_t \equiv [\mathbf{u}_{-N+2L}\cdots\mathbf{u}_{N-2L}]$, then we have,

$$\mathbf{U}_t = \sum_{k=-L}^{L}\mathbf{W}_k^T\tilde{\mathbf{B}}_{t,k}$$

Finally, if we put $\tilde{\mathbf{G}}_{t,k} \equiv [\mathbf{g}_{k-N+2L}\cdots\mathbf{g}_{k+N-2L}]$, then we have for the natural gradient with respect to $\mathbf{W}_k$,

$$\Delta\mathbf{W}_k = \gamma_h^{l+1}\mathbf{W}_k - \frac{1}{TN_2}\sum_{t=1}^{T}\tilde{\mathbf{G}}_{t,k}\mathbf{U}_t^T$$

where $N_2 = 2(N-2L)+1$.

We approximate the integral in (2) by a Riemann sum using the DFT, which samples the DTFT defined by the integral. If we make the definitions,

$$\begin{aligned}
Q_{hijt\tau}^l &\equiv \alpha_{hij}^l\sqrt{\beta_{hij}^l}\,q_{hij}\big(y_{hijt\tau}^l\big)\\
D_{hn}^l &\equiv \frac{2\pi}{N_F}\log\left|\det\sum_{k=-L}^{L}\mathbf{W}_{hk}^l\,e^{-i2\pi nk/N_F}\right|,\;\; n=1,\ldots,N_F\\
P_{ht}^l &\equiv \gamma_h^l\exp\Big(N_1\sum_{n=1}^{N_F}D_{hn}^l\Big)\prod_{\tau=-N+L}^{N-L}\prod_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}Q_{hijt\tau}^l
\end{aligned}$$

where N_F is the DFT length, then the $\hat{v}_{ht}$ and $\hat{z}_{hijt\tau}$ updates can be written,

$$\hat{v}_{ht}^l = \frac{P_{ht}^l}{\sum_{h'=1}^{M}P_{h't}^l}\,,\quad \hat{z}_{hijt\tau}^l = \frac{Q_{hijt\tau}^l}{\sum_{j'=1}^{m_{hi}}Q_{hij't\tau}^l} \tag{9}$$

Then,

$$\gamma_h^{l+1} = \frac{1}{T}\sum_{t=1}^{T}\hat{v}_{ht}^l\,,\quad \alpha_{hij}^{l+1} = \frac{1}{TN_1\gamma_h^{l+1}}\sum_{t=1}^{T}\hat{v}_{ht}^l\sum_{\tau=-N+L}^{N-L}\hat{z}_{hijt\tau}^l$$

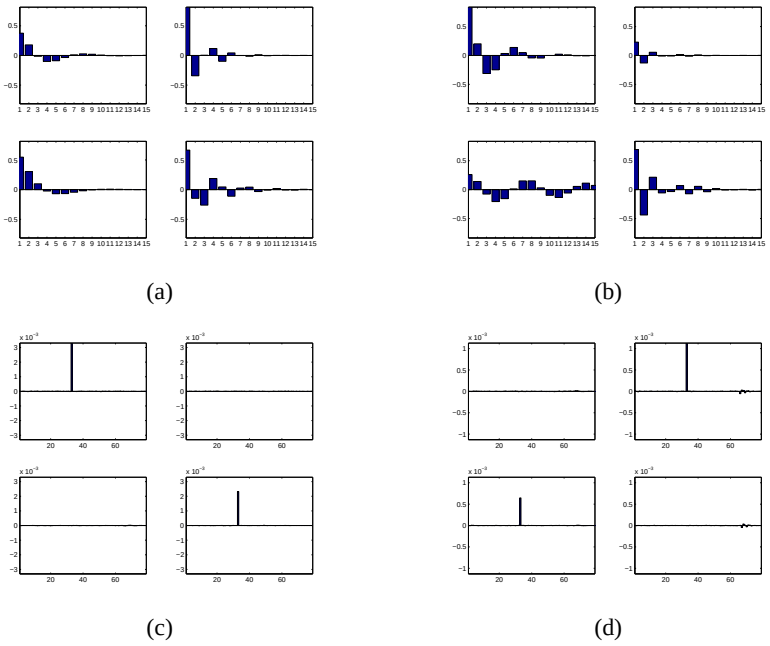

Fig. 1. Toy experiment with two mixing multichannel filters (a) and (b). In (c) and (d) are plotted the multiple convolution with the learned deconvolving filters.

7 Experiments

See Figures 1 and 2.

References

1. S.-I. Amari and A. Cichocki. Adaptive blind signal processing—neural network approaches. *Proceedings of the IEEE*, 86(10):2026–2047, 1998.
2. H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.
3. A. Benveniste, M. Goursat, and G. Ruget. Robust identification of a nonminimum phase system. *IEEE Transactions on Automatic Control*, 25(3):385–399, 1980.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
5. D. L. Donoho. On miminum entropy deconvolution. In D. F. Findlay, editor, *Applied Time Series II*, New York, 1981. Academic Press.
6. H. Gazzah, P. A. Regalia, and J.-P. Delmas. Asymptotic eigenvalue distribution of block Toeplitz matrices and application to blind SIMO channel identification. *IEEE Trans. Information Theory*, 47(3):1243–1251, 2001.
7. E. J. Hannan. *Mutltiple Time Series*. John Wiley & Sons, Inc., New York, 1970.
8. J. Keilson and F. W. Steutel. Mixtures of distributions, moment inequalities, and measures of exponentiality and Normality. *The Annals of Probability*, 2:112–130, 1974.

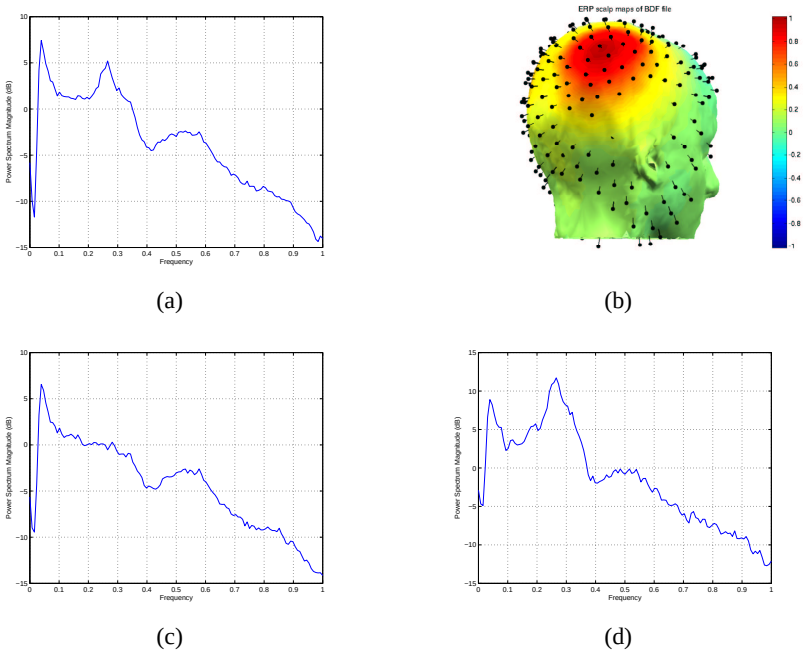

Fig. 2. Experiment with EEG data. (a) shows the original spectral density of the brain component with dipole estimate shown in (b). A mixture of deconvolving filters is applied to this source. The resulting psd's in (c) and (d) clearly show a division into alpha and non-alpha segments.

9. D. J. C. Mackay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.
10. M. Miranda and P. Tilli. Asymptotic spectra of Hermitian block Toeplitz matrices and preconditioning results. *SIAM Journal of Matrix Analysis and Applications*, 21(3):867–881, 2000.
11. R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer, 1998.
12. J. A. Palmer, K. Kreutz-Delgado, and S. Makeig. Super-Gaussian mixture source model for ICA. In *Proceedings of the 6th International Symposium on Independent Component Analysis and Blind Source Separation*, Lecture Notes in Computer Science. Springer, 2006.
13. J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*. MIT Press, 2005. Available at http://dsp.ucsd.edu/∼japalmer/.
14. D. T. Pham. Mutual information approach to blind separation of stationary sources. *IEEE Trans. Information Theory*, 48(7):1935–1946, 2002.
15. M. Rosenblatt. *Gaussian and Non-Gaussian Linear Time Series and Random Fields*. Springer, 2000.
16. L. K. Saul, T. S. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.