Asymptotic Newton Method for the ICA Mixture Model with Adaptive Source Densities

Jason A. Palmer, Ken Kreutz-Delgado, and Scott Makeig

Abstract

We derive an asymptotic Newton algorithm for Quasi Maximum Likelihood estimation of the ICA mixture model, using the ordinary gradient and Hessian. The probabilistic mixture framework can accommodate non-stationary environments and arbitrary source densities. We prove asymptotic stability when the source models match the true sources. An application to EEG segmentation is given.

Index Terms

Independent Component Analysis, Bayesian linear model, mixture model, Newton method, EEG

I. Introduction

The linear model,

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

is widely used in statistics (regression, classification, detection), signal processing (source coding, channel coding, denoising), and machine learning (data mining, etc.). In early statistics research, the linear model was proposed as Factor Analysis. In general the assumption was made for tractability purposes that distributions were Gaussian and conjugate distributions. This assumption was natural given the close relationship between linear dependence and the covariance structure of a Gaussian distribution.

The Gaussian distribution, however, has a singular disadvantage in terms of the linear model, namely that in general it leads to *non-identifiability* of the matrix $\mathbf{A}$. This arises from the fact that a positive definite matrix can be factored in an infinite number of ways, leading to an infinite number of possible $\mathbf{A}$

J. A. Palmer and S. Makeig are with the Swartz Center for Computational Neuroscience, La Jolla, CA, {jason,scott}@sccn.ucsd.edu. K. Kreutz-Delgado is with the ECE Department, Univ. of California San Diego, La Jolla, CA, kreutz@ece.ucsd.edu.

matrices and Gaussian $\mathbf{s}$ yielding the same distribution for $\mathbf{x}$. However, while the Gaussian distribution arises naturally in limits of sums, typical real world $\mathbf{s}$ signals or random variables are generally *not* Gaussian.

It turns out that non-Gaussianity of $\mathbf{s}$ does in fact lead to identifiability of $\mathbf{A}$, and hence to identifiability of $\mathbf{s}$ given a sample of $\mathbf{x}$ [1], [2]. Thus non-Gaussianity, rather than being a disadvantage, is in fact a necessity and a great aid to identification.

Here there is a bridge between *linear representations*, focusing on the matrix $\mathbf{A}$ as a basis with given source density characteristics (such as sparsity), and *source separation*, focusing on the estimation of $\mathbf{W} = \mathbf{A}^{-1}$ and $\mathbf{s} = \mathbf{W}\mathbf{x}$, where the distributions are in fact assumed to be unknown. Thus it is seen that the source separation problem can formulated in the linear model framework if the source densities are modeled *quasi-parametrically*, and the estimation of the source parameters is viewed as *learning the source densities*.

The latter formulation is referred to as *Quasi Maximum Likelihood* estimation [3]. The likelihood cost function can be interpreted in terms of mutual information via the Kullback-Leibler divergence, and the estimation can be seen as minimizing the mutual information or dependence of the estimated sources over the model source densities and the linear model $\mathbf{W} = \mathbf{A}^{-1}$. The interpretation as minimizing mutual information arises from the basic assumption on that the components of the sources or coefficients $\mathbf{s}$ are independent random variables.

The main innovations in the modern formulation over the classical linear model are that the densities of the sources, or coefficients, $\mathbf{s}$, are *non*-Gaussian, and their distributions may themselves be estimated.

As the Gaussian linear model was extended to a mixture of Gaussian linear models, so the Independent Component Analysis (ICA) model was extended to mixtures of bases with independent sources.

Rather than adding basis sets in a mixture model, the matrix $\mathbf{A}$ may be taken to be non-square $m \times n$ with more columns than rows, $n > m$, in which case $\mathbf{A}$ is said to be *overcomplete*. In this case $\mathbf{A}$ is not invertible, and a separate iterative optimization or approximate solution method must be undertaken to produce an estimate of the $\mathbf{s}$ corresponding to a given $\mathbf{x}$ and $\mathbf{A}$. This increase in computational load can be decisive in large scale sensor array applications.

The linear model often assumes the presence of noise,

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\nu}$$

When $\mathbf{s}$ is non-Gaussian, the noise case, like the overcomplete case, again leads to indeterminacy of the sources or coefficients $\mathbf{s}$ given $\mathbf{x}$ and $\mathbf{A}$. Indeed the noise case can be seen as a special case of the

overcomplete noiseless case [4].

When the observations $\mathbf{x}$ are of relatively the small dimension, the computational overhead of iterative MAP estimates of coefficients may be feasible. However, in large dimensional cases, such as that of high-density EEG [5], [6], where the number of channels (hundreds) with a modest sampling rate (tens of Hz) makes the determination of the sources a crucial bottleneck. Even direct multiplication of the observed data $\mathbf{x}$ by the inverse $\mathbf{W}$ takes non-trivial time. Thus the computational savings gained by assuming a complete or undercomplete noiseless mixture seem to greatly outweigh the advantage one may incur from modeling noise. Approximate one-step shrinkage techniques [7] for estimating the sources given noise can also be considered.

ICA mixture models [8], [9] offer a useful compromise between the efficiency of (conditional) invertibility of the model, and the need for richer representations e.g. in non-stationary environments. Mixture models provide the benefit of automatically grouping the components into basis sets that are commonly active. However this is at a cost of efficiency in the estimation in the case of observations being generated from arbitrarily combined basis vectors, such that a large number of overlapping mixture bases would be necessary. In this case the overlapping components would be estimated based on only a fraction of the data that they would be if they were identified, as e.g. in the overcomplete model. In real data, however, the "largely non-overlapping basis sets" assumption is often approximately valid, yielding computationally efficient, improved representation ability in the mixture model. Heuristic approaches to identifying components across models and estimating them accordingly (without loss of efficiency) can also be considered.

However, while feasible to optimize, the standard gradient and natural or relative gradient [10], [11] formulations still require many iterations to converge, as they are ultimately only linearly convergent. For large scale problems, with non-negligible time per iteration, the time required for convergence may be prohibitive.

Amari [12] derived a Newton-based method for optimization of a single ICA model in his stability analysis of the ICA problem. The Newton method differs from the *natural gradient*, also developed by Amari [10]. The natural gradient is still only linearly convergent, while Newton method is quadratically convergent.

In this paper we derive the Newton algorithm for a multiple mixture model [8], [9], [13] and adaptive mixture sources [14].

A. Related Work

The Gaussian liner model approach is described [15]–[17]. Non-Gaussian sources in the form of Gaussian scale mixtures, in particular Student's t distribution, were developed in [18]–[20]. A mixture of Gaussians source model was employed in [9], [21]–[24]. Similar approaches were proposed in [25], [26]. These models generally include noise and involve computationally intensive optimization algorithms. The focus in these models is generally on "variational" methods of automatically determining the number of mixtures in a mixture model during the optimization procedure. There is also overlap between the variational technique used in these methods, and the Gaussian scale mixture approach to representing non-Gaussian densities.

A model similar to that proposed here was presented in [8]. The main distinguishing features of the proposed model are,

1) Mixtures of Gaussian scale mixture sources provide more flexibility than the Gaussian mixture models of [9], [21], or fixed density models used in [8]. Accurate source density modeling is important to take advantage of Newton convergence for the true source model, as well as to distinguish between partially overlapping ICA models by posterior likelihood.
2) Implementation of the Amari Newton method described in [12] greatly improving the convergence, particularly in the multiple model case, in which prewhitening is not possible (in general a different whitening matrix will be required for each unknown model.)
3) The second derivative source density quantities are converted to first derivative quantities using integration by parts related properties of the score function and Fisher Information Matrix. Again accurate modeling of the source densities makes this conversion possible, and makes it robust in the presence of other (interfering) models.

The proposed model is readily extendable to MAP estimation or Variational Bayes or Ensemble Learning approaches, which put conjugate hyperpriors on the parameters. We are interested primarily in the large sample case, so we do not pursue these extensions here.

The probabilistic framework can also be extended to incorporate Markov dependence of state parameters in the ICA and source mixtures.

We have also extended the model to include mixtures of linear processes [13], where blind deconvolution is treated in a manner similar to [27]–[30], as well as complex ICA [31] and dependent sources [31]–[33]. In all of these contexts the adaptive source densities, asymptotic Newton method, and mixture model features can all be maintained.

II. ICA MIXTURE MODEL

In the standard linear model, observations $\mathbf{x}(t) \in \mathbb{R}^m$, $t = 1, \ldots, N$, are modeled as linear combinations of a set of basis vectors $\mathbf{A} \triangleq [\mathbf{a}_1 \cdots \mathbf{a}_n]$ with random and independent coefficients $s_i(t)$, $i = 1, \ldots, n$,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

We assume for simplicity the noiseless case, or that the data has been pre-processed, e.g. by PCA, filtering, etc., to remove noise. The data is assumed however to be non-stationary, so that different linear models may be in effect at different times. Thus for each observation $\mathbf{x}(t)$, there is an index $h_t \in \{1, \ldots, M\}$, with corresponding complete basis set $\mathbf{A}_h$ with "center" $\mathbf{c}_h$, and a random vector of zero mean, independent sources $\mathbf{s}(t) \sim q_h(\mathbf{s})$, where,

$$q_h(\mathbf{s}) = \prod_{i=1}^{n} q_{hi}(s_i)$$

such that,

$$\mathbf{x}(t) = \mathbf{A}_h \mathbf{s}(t) + \mathbf{c}_h$$

with $h = h_t$. We shall assume that only one model is active at each time, and that model h is active with probability γ_h. For simplicity we assume temporal independence of the model indices h_t, $t = 1. \ldots, N$.

Since the model is conditionally linear, the conditional density of the observations is given by,

$$p(\mathbf{x}(t) \,|\, h) = |\det \mathbf{W}_h| \, q_h\big(\mathbf{W}_h(\mathbf{x}(t) - \mathbf{c}_h)\big)$$

where $\mathbf{W}_h \triangleq \mathbf{A}_h^{-1}$.

The sources are taken to be mixtures of (generally *non*gaussian) Gaussian Scale Mixtures (GSMs), as in [14],

$$q_{hi}\big(s_i(t)\big) = \sum_{j=1}^{m} \alpha_{hij} \sqrt{\beta_{hij}} \, q_{hij}\big(\sqrt{\beta_{hij}}(s_i(t) - \mu_{hij}) \,; \rho_{hij}\big)$$

where each q_{hij} is a GSM parameterized by ρ_{hij}.

Thus the density of the observations $\mathbf{X} \triangleq \{\mathbf{x}(t)\}$, $t = 1, \ldots, N$, is given by,

$$p(\mathbf{X}; \Theta) = \prod_{t=1}^{N} \sum_{h=1}^{M} \gamma_h \, p(\mathbf{x}(t) \,|\, h),$$

$\gamma_h \geq 0$, $\sum_{h=1}^{M} \gamma_h = 1$. The parameters to be estimated are,

$$\Theta = \big\{\mathbf{W}_h, \mathbf{c}_h, \gamma_h, \alpha_{hij}, \mu_{hij}, \beta_{hij}, \rho_{hij}\big\},$$

$h = 1, \ldots, M$, $i = 1, \ldots, n$, and $j = 1, \ldots, m$.

A. Invariances in the model

Besides the accepted invariance to permutation of the component indices, invariance or redundancy in the model also exists in two other respects. The first concerns the model centers, $\mathbf{c}_h$, and the source density location parameters μ_{hij}. Specifically, we have $p(\mathbf{X};\Theta) = p(\mathbf{X};\Theta')$, $\Theta = \{\ldots, \mathbf{c}_h, \mu_{hij}, \ldots\}$, $\Theta' = \{\ldots, \mathbf{c}'_h, \mu'_{hij}, \ldots\}$, if

$$\mathbf{c}'_h = \mathbf{c}_h + \Delta\mathbf{c}_h, \quad \mu'_{hij} = \mu_{hij} - [\mathbf{W}_h \Delta\mathbf{c}_h]_i, \;\; j = 1, \ldots, m$$

for any $\Delta\mathbf{c}_h$. Putting $\mathbf{c}'_h = E\{\mathbf{x}(t) \,|\, h\}$, we make the sources $\mathbf{s}(t)$ zero mean given the model. The zero mean assumption is used in the calculation of the expected Hessian for the Newton algorithm.

There is also scale redundancy in the row norms of $\mathbf{W}_h$ and the scale parameters of the source densities. Specifically, $p(\mathbf{X};\Theta) = p(\mathbf{X};\Theta')$, where $\Theta = \{\mathbf{W}_h, \mu_{hij}, \beta_{hij}, \ldots\}$, $\Theta' = \{\mathbf{W}'_h, \mu'_{hij}, \beta'_{hij}, \ldots\}$, if for any $\tau_{hi} > 0$,

$$[\mathbf{W}'_h]_{i:} = [\mathbf{W}_h]_{i:}/\tau_{hi},$$

$$\mu'_{hij} = \mu_{hij}/\tau_{hi}, \;\; \beta'_{hij} = \beta_{hij}\tau_{hi}^2, \;\; j = 1, \ldots, m$$

where $[\mathbf{W}_h]_{i:}$ is the ith row of $\mathbf{W}_h$. We use this redundancy to enforce at each iteration that the rows of $\mathbf{W}_h$ are unit norm by putting $\tau_{hi} = \|[\mathbf{W}_h]_{i:}\|$.

These "reparameterizations" constitute the only updates for the model centers $\mathbf{c}_h$. The centers are redundant parameters given the source means, and are used only to maintain zero posterior source mean given the model.

III. Maximum Likelihood

In this section we assume that the model is given and suppress the subscript h. Given i.i.d. data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we consider the ML estimate of $\mathbf{W} = \mathbf{A}^{-1}$. For the density of $\mathbf{X}$, we have,

$$p(\mathbf{X}) = \prod_{t=1}^{N} |\det \mathbf{W}| \, p_{\mathbf{s}}(\mathbf{W}\mathbf{x}_t)$$

Let $\mathbf{y}_t = \mathbf{W}\mathbf{x}_t$ be the estimate of the sources $\mathbf{s}_t$, and let $q_i(y_i)$ be the density model for the ith source, with $q(\mathbf{y}_t) = \prod_i q_i(y_{it})$. We define,

$$f_i(y_{it}) \triangleq -\log q_i(y_{it})$$

and $f(\mathbf{y}_t) \triangleq \sum_i f_i(y_{it})$. For the negative log likelihood of the data then (which is to be minimized), we have,

$$L(\mathbf{W}) \;=\; \sum_{t=1}^{N} -\log|\det \mathbf{W}| + f(\mathbf{y}_t) \tag{1}$$

The gradient of this function is proportional to,

$$\nabla L(\mathbf{W}) \;\propto\; -\mathbf{W}^{-T} + \frac{1}{N}\sum_{t=1}^{N} \nabla f(\mathbf{y}_t)\mathbf{x}_t^T \tag{2}$$

Note that if we multiply (2) by $\mathbf{W}^T\mathbf{W}$ on the right, we get,

$$\Delta\mathbf{W} = \left(\mathbf{I} - \frac{1}{N}\sum_{t=1}^{N} \mathbf{g}_t\mathbf{y}_t^T\right)\mathbf{W} \tag{3}$$

where $\mathbf{g}_t \triangleq \nabla f(\mathbf{y}_t)$. This transformation is in fact a positive definite linear transformation of the matrix gradient. Specifically, using the standard matrix inner product in $\mathbb{R}^{n\times n}$, $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^T)$, we have for nonzero $\mathbf{V}$,

$$\left\langle \mathbf{V}, \mathbf{V}\mathbf{W}^T\mathbf{W}\right\rangle = \left\langle \mathbf{V}\mathbf{W}^T, \mathbf{V}\mathbf{W}^T\right\rangle > 0 \tag{4}$$

when $\mathbf{W}$ is full rank. The direction (3) is known as the "natural gradient" [10].

A. Hessian

Denote the gradient (2) by $\mathbf{G}$ with elements g_{ij}, each a function of $\mathbf{W}$. Taking the derivative of (2), we find,

$$\frac{\partial g_{ij}}{\partial w_{kl}} = [\mathbf{W}^{-1}]_{li}[\mathbf{W}^{-1}]_{jk} + \left\langle f_i''\big([\mathbf{W}\mathbf{x}_t]_k\big)x_{jt}x_{lt}\delta_{ik}\right\rangle_N$$

where δ_{ik} is the Kronecker delta symbol, and $\left\langle \cdot \right\rangle_N$ denotes the empirical average $\frac{1}{N}\sum \cdot$. To see how this linear Hessian operator transforms an argument $\mathbf{B}$, let $\mathbf{C} = \mathcal{H}(\mathbf{B})$ be the transformed matrix. Then we calculate,

$$c_{ij} = \sum_k\sum_l [\mathbf{W}^{-1}]_{li}[\mathbf{W}^{-1}]_{jk}b_{kl} + \left\langle f_i''(y_{it})\,x_{jt}\sum_l b_{il}x_{lt}\right\rangle_N$$

The first term of c_{ij} can be written,

$$\begin{aligned}\sum_l [\mathbf{W}^{-1}]_{li}[\mathbf{W}^{-1}\mathbf{B}]_{jl} &= \sum_l [\mathbf{W}^{-T}]_{il}[\mathbf{B}^T\mathbf{W}^{-T}]_{lj} \\ &= [\mathbf{W}^{-T}\mathbf{B}^T\mathbf{W}^{-T}]_{ij}\end{aligned}$$

Writing the second term in matrix form as well, we have for the linear transformation $\mathbf{C} = \mathcal{H}(\mathbf{B})$,

$$\mathbf{C} = \mathbf{W}^{-T}\mathbf{B}^T\mathbf{W}^{-T} + \left\langle \mathrm{diag}(f''(\mathbf{y}_t))\mathbf{B}\mathbf{x}_t\mathbf{x}_t^T\right\rangle_N \tag{5}$$

where $\mathrm{diag}(f''(\mathbf{y}_t))$ is the diagonal matrix with diagonal elements $f_i''(y_{it})$. Assuming that the model holds, the source estimates at the optimal $\mathbf{W}$ will be independent. We also assume that the mean has been removed, so that the sources are zero mean, as noted in §II-A.

It will be easier to calculate the asymptotic value of the Hessian if we rewrite the transformation (5) in terms of the source estimates $\mathbf{y}$ since the sources are assumed to be independent and zero mean. At the optimum, we may assume that the source density models $q_i(y_i)$ are equivalent to the true source densities $p_i(s_i)$. We first write,

$$\mathbf{C} = (\mathbf{B}\mathbf{W}^{-1})^T\mathbf{W}^{-T} + \Big\langle \text{diag}\big(f''(\mathbf{y}_t)\big)\mathbf{B}\mathbf{W}^{-1}\mathbf{y}_t\mathbf{y}_t^T\mathbf{W}^{-T}\Big\rangle_N$$

Now if we define $\tilde{\mathbf{C}} \triangleq \mathbf{C}\mathbf{W}^T$ and $\tilde{\mathbf{B}} \triangleq \mathbf{B}\mathbf{W}^{-1}$, then we have,

$$\tilde{\mathbf{C}} = \tilde{\mathbf{B}}^T + \Big\langle \text{diag}\big(f''(\mathbf{y}_t)\big)\tilde{\mathbf{B}}\mathbf{y}_t\mathbf{y}_t^T\Big\rangle_N \tag{6}$$

Writing this equation in component form and letting N go to infinity we find for the diagonal elements,

$$[\tilde{\mathbf{C}}]_{ii} \;=\; [\tilde{\mathbf{B}}]_{ii} + E\big\{f_i''(y_i)\textstyle\sum_k [\tilde{\mathbf{B}}]_{ik}y_k y_i\big\} = [\tilde{\mathbf{B}}]_{ii}(1+\eta_i) \tag{7}$$

where we define $\eta_i \triangleq E\{f''(y_i)y_i^2\}$. The cross terms drop out since the expected value of $f''(y_i)y_iy_k$ is zero for $k \neq i$ by the independence and zero mean assumption on the sources. Now we note [11], [12] that the off-diagonal elements of the equation (6) can be paired as follows,

$$[\tilde{\mathbf{C}}]_{ij} = [\tilde{\mathbf{B}}]_{ji} + E\big\{f_i''(y_i)\textstyle\sum_k [\tilde{\mathbf{B}}]_{ik}y_k y_j\big\} = [\tilde{\mathbf{B}}]_{ji} + \kappa_i\sigma_j^2[\tilde{\mathbf{B}}]_{ij}$$
$$[\tilde{\mathbf{C}}]_{ji} = [\tilde{\mathbf{B}}]_{ij} + E\big\{f_j''(y_j)\textstyle\sum_k [\tilde{\mathbf{B}}]_{jk}y_k y_i\big\} = [\tilde{\mathbf{B}}]_{ij} + \kappa_j\sigma_i^2[\tilde{\mathbf{B}}]_{ji}$$

where we define $\kappa_i \triangleq E\{f_i''(y_i)\}$ and $\sigma_i^2 \triangleq E\{y_i^2\}$. Again the cross terms drop out due to the expectation of independent zero mean random variables. Putting these equations in matrix form, we have,

$$\begin{bmatrix}[\tilde{\mathbf{C}}]_{ij}\\ [\tilde{\mathbf{C}}]_{ji}\end{bmatrix} = \begin{bmatrix}\kappa_i\sigma_j^2 & 1\\ 1 & \kappa_j\sigma_i^2\end{bmatrix}\begin{bmatrix}[\tilde{\mathbf{B}}]_{ij}\\ [\tilde{\mathbf{B}}]_{ji}\end{bmatrix} \tag{8}$$

If we denote the linear transformation defined by equations (7) and (8) by $\tilde{\mathbf{C}} = \tilde{\mathcal{H}}(\tilde{\mathbf{B}})$, then we have,

$$\mathbf{C} \;=\; \mathcal{H}(\mathbf{B}) \;=\; \tilde{\mathcal{H}}\big(\mathbf{B}\mathbf{W}^{-1}\big)\mathbf{W}^{-T} \tag{9}$$

Thus by an argument similar to (4), we see that $\mathcal{H}$ is asymptotically positive definite if and only if $\tilde{\mathcal{H}}$ is asymptotically positive definite and $\mathbf{W}$ is full rank.

The conditions for positive definiteness of $\tilde{\mathcal{H}}$ can be found by inspection of equations (7) and (8). With the definitions,

$$\eta_i \triangleq E\{y_i^2 f_i''(y_i)\}, \quad \kappa_i \triangleq E\{f_i''(y_i)\}, \quad \sigma_i^2 \triangleq E\{y_i^2\}$$

the conditions can be stated [12] as,

1) $1 + \eta_i > 0, \;\; \forall\, i$

2) $\kappa_i > 0, \ \forall\, i, \ $ and,

3) $\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1 > 0, \quad \forall\, i \neq j$

B. Asymptotic stability

Using integration by parts, it can be shown that the stability conditions are always satisfied when $f(y) = -\log p(y)$, i.e. $q(y)$ matches the true source density $p(y)$. Specifically, we have the following.

Theorem 1: If $f_i(y_i) \triangleq -\log q_i(y_i) = -\log p_i(y_i)$, with $\int p_i(y) = 1$, $i = 1, \ldots, n$, i.e. the source density models match the true source densities, and $p_i(y)$ is twice differentiable with $E\{f_i''(y)\}$ and $E\{y_i^2\}$ finite, $i = 1, \ldots, n$, and at most one source is Gaussian, then the stability conditions hold.

Proof: For the first condition, we use integration by parts to evaluate,

$$E\{y^2 f''(y)\} = \int_{-\infty}^{\infty} y^2 f''(y) p(y)\, dy$$

with $u = y^2 p(y)$ and $dv = f''(y)\, dy$. Using the fact that $v = f'(y) = -p'(y)/p(y)$, we get

$$-y^2 p'(y)\big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f'(y)\big(2y - y^2 f'(y)\big)\, p(y)\, dy \tag{10}$$

The first term in (10) is zero if $p'(y) = o(1/y^2)$ as $y \to \pm\infty$. This must be the case for integrable $p(y)$, since otherwise we would have $p'(y) \to C/y^2$, and $p(y) = O(1/y)$ and non-integrable. Then, since $\int p(y) dy = 1$, we have,

$$\begin{aligned} 1 + E\{y^2 f''(y)\} &= \int_{-\infty}^{\infty} \big(y^2 f'(y)^2 - 2y f'(y) + 1\big) p(y) dy \\ &= E\big\{\big(y f'(y) - 1\big)^2\big\} \ \geq \ 0 \end{aligned}$$

where equality holds only if $p(y) = 1/y$, so strict inequality must hold for integrable $p(y)$.

For the second condition,

$$E\{f''(y)\} > 0$$

using integration by parts with $u = p(y)$, $dv = f''(y)\, dy$, and the fact that $p'(y)$ must tend to 0 as $y \to \pm\infty$ for integrable $p(y)$, we get,

$$E\{f''(y)\} = \int_{-\infty}^{\infty} f'(y)^2 p(y)\, dy = E\big\{f'(y)^2\big\} > 0$$

Finally, for the third condition, we have,

$$E\{y^2\} E\{f''(y)\} = E\big\{y^2\big\} E\big\{f'(y)^2\big\} \geq \big(E\{y f'(y)\}\big)^2 = 1$$

by the Cauchy Schwartz inequality, with equality only for $f'(y) \propto y$, i.e. $p(y)$ Gaussian. Thus,

$$E\{y_i^2\} E\{f_i''(y_i)\} E\{y_j^2\} E\{f_j''(y_j)\} > 1$$

whenever at least one of y_i and y_j is nongaussian. ■

C. Newton method

The inverse of the Hessian operator, from (9), will be given by,

$$\mathbf{B} \;=\; \mathcal{H}^{-1}(\mathbf{C}) \;=\; \tilde{\mathcal{H}}^{-1}\big(\mathbf{C}\mathbf{W}^T\big)\mathbf{W} \tag{11}$$

The calculation of $\tilde{\mathbf{B}} = \tilde{\mathcal{H}}^{-1}(\tilde{\mathbf{C}})$ can again be carried out by inspection of (7) and (8),

$$[\tilde{\mathbf{B}}]_{ii} \;=\; \frac{[\tilde{\mathbf{C}}]_{ii}}{1+\eta_i}, \;\; i = 1, \ldots, n \tag{12}$$

$$[\tilde{\mathbf{B}}]_{ij} \;=\; \frac{\kappa_j \sigma_i^2 [\tilde{\mathbf{C}}]_{ij} - [\tilde{\mathbf{C}}]_{ji}}{\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1}, \;\; \forall\, i \neq j \tag{13}$$

The Newton direction is given by taking $\mathbf{C} = -\mathbf{G}$, the gradient (2),

$$\Delta\mathbf{W} = \tilde{\mathcal{H}}^{-1}\big(-\mathbf{G}\mathbf{W}^T\big)\mathbf{W} \tag{14}$$

Let,

$$\mathbf{\Phi} \triangleq \frac{1}{N}\sum_{t=1}^{N} \mathbf{g}_t \mathbf{y}_t^T \tag{15}$$

We have $-\mathbf{G}\mathbf{W}^T = \mathbf{I} - \mathbf{\Phi}$. If we let $\tilde{\mathbf{B}} = \tilde{\mathcal{H}}^{-1}(-\mathbf{G}\mathbf{W}^T)$, then

$$\tilde{b}_{ii} \;=\; \frac{1 - [\mathbf{\Phi}]_{ii}}{1+\eta_i}, \;\; i = 1, \ldots, n \tag{16}$$

$$\tilde{b}_{ij} \;=\; \frac{[\mathbf{\Phi}]_{ji} - \kappa_j \sigma_i^2 [\mathbf{\Phi}]_{ij}}{\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1}, \;\; \forall\, i \neq j \tag{17}$$

Then

$$\Delta\mathbf{W} = \tilde{\mathbf{B}}\mathbf{W} \tag{18}$$

IV. EM Parameter Updates

We define h_t to be the random variable denoting the index of the model chosen at time t, producing the observation $\mathbf{x}(t)$, and define the random variables v_{ht},

$$v_{ht} \triangleq \begin{cases} 1, & h_t = h \\ 0, & \text{otherwise} \end{cases}$$

We define j_{hit} to be the random variable indicating the source density mixture component index that is chosen at time t (independently of h_t) for the ith source of the hth model, and we define the random variables u_{hijt} by,

$$u_{hijt} \triangleq \begin{cases} 1, & j_{hit} = j \\ 0, & \text{otherwise} \end{cases}$$

We employ the EM algorithm by writing the density of $\mathbf{X}$ as a marginal integral over "complete" data, which includes $\mathbf{U}$ and $\mathbf{V}$,

$$p(\mathbf{X};\Theta) = \sum_{\mathbf{U},\mathbf{V}} \prod_{t=1}^{N}\prod_{h=1}^{M} \gamma_h^{v_{ht}} |\det \mathbf{W}_h|^{v_{ht}} \prod_{i=1}^{n}\prod_{j=1}^{m} Q_{hijt}^{u_{hijt}v_{ht}}$$

where we make the definitions,

$$\begin{aligned} \mathbf{b}_{ht} &\triangleq \mathbf{W}_h(\mathbf{x}_t - \mathbf{c}_h) \\ y_{hijt} &\triangleq \sqrt{\beta_{hij}}\,\big([\mathbf{b}_{ht}]_i - \mu_{hij}\big) \\ Q_{hijt} &\triangleq \alpha_{hij}\sqrt{\beta_{hij}}\,q_{hij}\big(y_{hijt}\big) \end{aligned}$$

The function to be minimized in the EM algorithm, sometimes referred to as the variational free energy, is,

$$F^l(\Theta) = \sum_{t=1}^{N}\sum_{h=1}^{M}\Bigg[\hat{v}_{ht}^l\big(-\log\gamma_h - \log|\det\mathbf{W}_h|\big) + \sum_{i=1}^{n}\sum_{j=1}^{m}\hat{z}_{hijt}^l\big(-\log\alpha_{hij} - \tfrac{1}{2}\log\beta_{hij} + f_{hij}(y_{hijt})\big)\Bigg]$$

where $f_{hij} \triangleq -\log q_{hij}$. The $\hat{v}_{ht}^l \triangleq E\{v_{ht}|\mathbf{x}_t;\Theta^l\}$ are given by,

$$\hat{v}_{ht}^l = P\big[v_{ht}=1|\mathbf{x}_t;\Theta^l\big] = \frac{L_{ht}^l}{\sum_{h'=1}^{M} L_{h't}^l} \tag{19}$$

where we define,

$$L_{ht}^l \triangleq \gamma_h^l\,\big|\det\mathbf{W}_h^l\big|\prod_{i=1}^{n}\sum_{j=1}^{m} Q_{hijt}^l \tag{20}$$

For the $\hat{z}_{hijt}^l \triangleq E[u_{hijt}v_{ht}|\mathbf{x}_t;\Theta^l]$, we have,

$$\begin{aligned} \hat{z}_{hijt}^l &= P\big[v_{ht}=1, u_{hijt}=1\,|\,\mathbf{x}_t;\Theta^l\big] \\ &= P\big[u_{hijt}=1\,|\,v_{ht}=1,\mathbf{x}_t;\Theta^l\big]P\big[v_{ht}=1\,|\,\mathbf{x}_t;\Theta^l\big] \\ &= \hat{u}_{hijt}^l\hat{v}_{ht}^l \end{aligned} \tag{21}$$

where $\hat{u}_{hijt}^l \triangleq E\big\{u_{hijt}|v_{ht}=1,\mathbf{x}_k,;\Theta^l\big\}$,

$$\hat{u}_{hijt}^l = P\big[u_{hijt}=1|\mathbf{x}_t,v_{ht}=1;\Theta^l\big] = \frac{Q_{hijt}^l}{\sum_{j'=1}^{m} Q_{hij't}^l} \tag{22}$$

Minimizing F^l over γ_h and α_{hij} subject to $\gamma_h,\alpha_{hij}\geq 0$, $\sum_h\gamma_h = 1$ and $\sum_j\alpha_{hij}=1$, we get,

$$\gamma_h^{l+1} = \frac{1}{N}\sum_{t=1}^{N}\hat{v}_{ht}^l\,, \quad \alpha_{hij}^{l+1} = \frac{\sum_{t=1}^{N}\hat{z}_{hijt}^l}{\sum_{t=1}^{N}\hat{v}_{ht}^l} \tag{23}$$

We make the following definitions, in which the G_t are arbitrary functions of $\mathbf{x}_t$,

$$E_v\{G_t \,|\, h\} \;\triangleq\; \frac{\sum_t \hat{v}_{ht}^l G_t}{\sum_t \hat{v}_{ht}^l}, \quad E_z\{G_t \,|\, h,j\} \;\triangleq\; \frac{\sum_t \hat{z}_{hijt}^l G_t}{\sum_t \hat{z}_{hijt}^l}$$

We can then write the function to be minimized over the remaining parameters as,

$$F^l(\Theta) \quad \propto \quad \sum_{h=1}^{M} \gamma_h^{l+1} \Big[-\log|\det \mathbf{W}_h| \;+\; \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{hij}^{l+1} \Big(-\tfrac{1}{2}\log\beta_{hij} \;+\; E_z\big\{ f_{hij}(y_{hijt}) \,|\, h,j \big\} \Big) \Big]$$

A. ICA mixture model Newton updates

Since F^l is an additive function of the $\mathbf{W}_h$, the Newton updates can be considered separately. The cost function for $\mathbf{W}_h$ is,

$$-\log|\det \mathbf{W}_h| + \sum_{i=1}^{n}\sum_{j=1}^{m} E_v\big\{ \hat{u}_{hijt}^l f_{hij}(y_{hijt}) \,|\, h \big\}$$

The gradient of this function is,

$$-\mathbf{W}_h^{-T} + E_v\big\{ \mathbf{g}_{ht}(\mathbf{x}_t - \mathbf{c}_h)^T \,|\, h \big\} \tag{24}$$

where $\mathbf{g}_{ht}$ is defined by,

$$[\mathbf{g}_{ht}]_i \triangleq \sum_{j=1}^{m} \hat{u}_{hijt}^l \sqrt{\beta_{hij}}\, f'_{hij}\big(y_{hijt}\big) \tag{25}$$

Denote the matrix gradient (24) by $\mathbf{G}_h$. Taking the derivative of $[\mathbf{G}_h]_{i\nu}$ with respect to $[\mathbf{W}_h]_{k\lambda}$, we get,

$$\frac{\partial\,[\mathbf{G}_h]_{i\nu}}{\partial\,[\mathbf{W}_h]_{k\lambda}} \;=\; \big[\mathbf{W}_h^{-1}\big]_{\lambda i}\big[\mathbf{W}_h^{-1}\big]_{\nu k} \;+\; \delta_{ik}\sum_{j=1}^{m} \beta_{hij} E_v\big\{ \hat{u}_{hijt}^l f''_{hij}(y_{hijt})(x_{\nu t} - [\mathbf{c}_h]_\nu)(x_{\lambda t} - [\mathbf{c}_h]_\lambda) \,\big|\, h \big\}$$

For the linear transformation $\mathbf{C} = \mathcal{H}(\mathbf{B})$, we have,

$$\mathbf{C} = \mathbf{W}_h^{-T}\mathbf{B}^T\mathbf{W}_h^{-T} + E_v\big\{ \mathbf{D}_{ht}^l \mathbf{B}(\mathbf{x}_t - \mathbf{c}_h)(\mathbf{x}_t - \mathbf{c}_h)^T \,\big|\, h \big\} \tag{26}$$

where $\mathbf{D}_{ht}$ is the diagonal matrix with diagonal elements

$$\big[\mathbf{D}_{ht}\big]_{ii} = \sum_{j=1}^{m} \hat{u}_{hijt}^l \beta_{hij} f''_{hij}\big(y_{hijt}\big) \tag{27}$$

To simplify the calculation of the asymptotic value of the Hessian, we rewrite the second term on the right-hand side of (26) as,

$$E_v\big\{ \mathbf{D}_{ht}\mathbf{B}\,\mathbf{W}_h^{-1}\mathbf{b}_{ht}\mathbf{b}_{ht}^T\mathbf{W}_h^{-T} \,\big|\, h \big\}$$

If we define $\tilde{\mathbf{C}} \triangleq \mathbf{C}\mathbf{W}_h^T$ and $\tilde{\mathbf{B}} \triangleq \mathbf{B}\mathbf{W}_h^{-1}$, then we have,

$$\tilde{\mathbf{C}} = \tilde{\mathbf{B}}^T + E_v\big\{ \mathbf{D}_{ht}\tilde{\mathbf{B}}\,\mathbf{b}_{ht}\mathbf{b}_{ht}^T \,\big|\, h \big\} \tag{28}$$

Writing the ith row of the second term in (28) as,

$$\sum_{j=1}^{m} \beta_{hij} E_v\big\{ \hat{u}^l_{hijt} f''_{hij}\big(y_{hijt}\big) [\tilde{\mathbf{B}}\, \mathbf{b}_{ht}]_i \, \mathbf{b}^T_{ht} \,\big|\, h \big\} \tag{29}$$

Since $\mathbf{b}_{ht}$ is zero mean conditioned on the model, the Hessian matrix reduces to a 2×2 block diagonal form as in the single model case. In the multiple model case we get,

$$\begin{aligned}
\eta_{hi} &\triangleq \sum_{j=1}^{m} \alpha^{l+1}_{hij} \beta_{hij} E_z\big\{ f''_{hij}(y_{hijt}) [\mathbf{b}_{ht}]^2_i \,\big|\, h, j \big\} \\
\kappa_{hi} &\triangleq \sum_{j=1}^{m} \alpha^{l+1}_{hij} \beta_{hij} E_z\big\{ f''_{hij}(y_{hijt}) \,\big|\, h, j \big\} \\
\sigma^2_{hi} &\triangleq E_v\big\{ [\mathbf{b}_{ht}]^2_i \,|\, h \big\}
\end{aligned}$$

If we define,

$$\begin{aligned}
\eta_{hij} &\triangleq E_z\big\{ f''_{hij}(y_{hijt}) y^2_{hij} \,|\, h, j \big\} \\
\kappa_{hij} &\triangleq E_z\big\{ f''_{hij}(y_{hijt}) \,|\, h, j \big\}
\end{aligned}$$

or, using integration by parts to rewrite the integrals,

$$\begin{aligned}
\lambda_{hij} \triangleq 1 + \eta_{hij} &= E_z\big\{ \big(f'_{hij}(y_{hijt}) y_{hijt} - 1 \big)^2 \,|\, h, j \big\} \\
\kappa_{hij} &= E_z\big\{ f'_{hij}(y_{hijt})^2 \,|\, h, j \big\}
\end{aligned}$$

then the expressions can be simplified to the following,

$$\begin{aligned}
\lambda_{hi} &= \sum_{j=1}^{m} \alpha^{l+1}_{hij} \big(\lambda_{hij} + \beta_{hij} \kappa_{hij} \mu^2_{hij} \big) \\
\kappa_{hi} &= \sum_{j=1}^{m} \alpha^{l+1}_{hij} \beta_{hij} \kappa_{hij} \\
\sigma^2_{hi} &= E_v\big\{ [\mathbf{b}_{ht}]^2_i \,|\, h \big\}
\end{aligned}$$

Define,

$$\mathbf{\Phi}_h \triangleq E_v\big\{ \mathbf{g}_{ht} \mathbf{b}^T_{ht} \,\big|\, h \big\} \tag{30}$$

We have $-\mathbf{G}_h \mathbf{W}^T_h = \mathbf{I} - \mathbf{\Phi}_h$. If we let,

$$\tilde{\mathbf{B}} = \tilde{\mathcal{H}}^{-1}\big(-\mathbf{G}_h \mathbf{W}^T_h \big) = \tilde{\mathcal{H}}^{-1}\big(\mathbf{I} - \mathbf{\Phi}_h \big)$$

then we have,

$$[\tilde{\mathbf{B}}]_{ii} = \frac{1 - [\mathbf{\Phi}_h]_{ii}}{\lambda_{hi}}, \quad i = 1, \ldots, n \tag{31}$$

$$[\tilde{\mathbf{B}}]_{ij} = \frac{[\mathbf{\Phi}_h]_{ji} - \kappa_{hj} \sigma^2_{hi} [\mathbf{\Phi}_h]_{ij}}{\kappa_{hi} \kappa_{hj} \sigma^2_{hi} \sigma^2_{hj} - 1}, \quad \forall\, i \neq j \tag{32}$$

Then

$$\Delta \mathbf{W}_h = \tilde{\mathbf{B}} \mathbf{W}_h \tag{33}$$

B. Density parameter EM updates

The location parameters are updated by (see [14]),

$$\mu_{hij}^{l+1} = \mu_{hij}^{l} + \frac{1}{\sqrt{\beta_{hij}^{l}}} \frac{E_z\{ f'_{hij}(y_{hijt}) \,|\, h, j \}}{E_z\{ f'_{hij}(y_{hijt}) / y_{hijt} \,|\, h, j \}} \tag{34}$$

The scale parameters are updated by,

$$\beta_{hij}^{l+1} = \beta_{hij}^{l} \,/\, E_z\{ f'_{hij}(y_{hijt}) y_{hijt} \,|\, h, j \} \tag{35}$$

The Generalized Gaussian shape parameters are updated by,

$$\Delta \rho_{hij} \;=\; 1 \;-\; \rho_{hij}^{l} \frac{E_z\{ |y_{hijt}|^{\rho_{hij}^{l}} \log |y_{hijt}|^{\rho_{hij}^{l}} \,|\, h, j \}}{\Psi\Big(1 + \frac{1}{\rho_{hij}^{l}}\Big)} \tag{36}$$

The log likelihood of Θ^l given $\mathbf{X}$ is calculated as,

$$L(\Theta^l | \mathbf{X}) \;=\; \sum_{t=1}^{N} \log \left(\sum_{h=1}^{M} L_{ht}^{l} \right) \tag{37}$$

V. EXPERIMENTS

In Figure 1, we plot the ratio $\left\| \mathbf{W}^{l+1} - \mathbf{W}^* \right\| / \left\| \mathbf{W}^{l} - \mathbf{W}^* \right\|$, versus iteration l, where $\mathbf{W}^*$ is the optimum and $\mathbf{W}^l$ is the estimate at iteration l. For linearly convergent algorithms, this ratio tends to a constant [34]. For superlinear algorithms, this ratio tends to zero, and the order of convergence q is the power of the denominator which yields a finite nonzero limit for the ratio $\left\| \mathbf{W}^{l+1} - \mathbf{W}^* \right\| / \left\| \mathbf{W}^{l} - \mathbf{W}^* \right\|^q$. For Newton's method, $q = 2$.

We also present an example of segmentation using the mixture model. Figure 3 shows the result of segmenting an EEG experiment according to the most likely model given the data (MAP). The trials are stacked vertically with time on the horizontal axis, time locked to the feedback at $t = 175$. Time points are plotted in the color of the model most likely for that point. Muscle activity (red and light blue spanning lines) as well as post-feedback theta activity (yellow) are segmented. There appears to be consistency in the 3 and 4 model segmentations, with increased resolution in the 4 model segmentation.

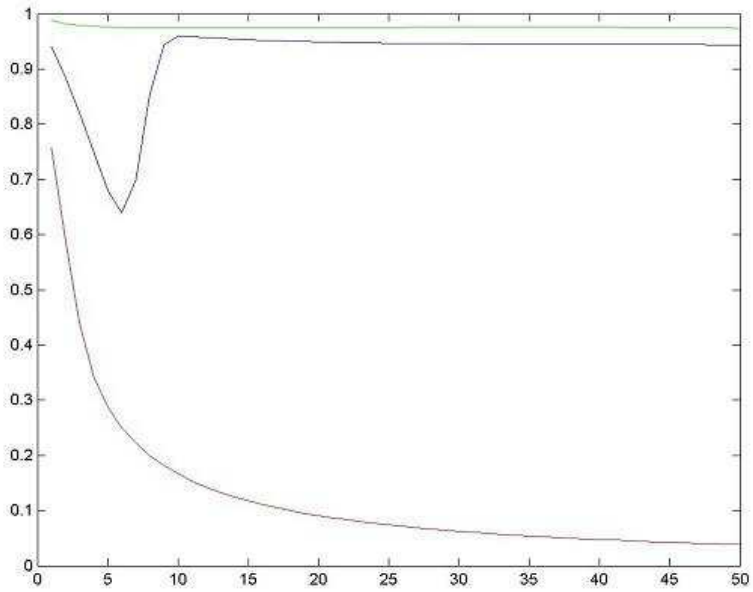


Fig. 1. Newton convergence rate versus gradient and natural gradient in a simulation with a 10×10 mixing matrix with Laplacian sources. Ordinary gradient (top line) and natural gradient (middle line) are linearly convergent with high asymptotic rate, while Newton method (bottom line) is tending toward superlinearity.

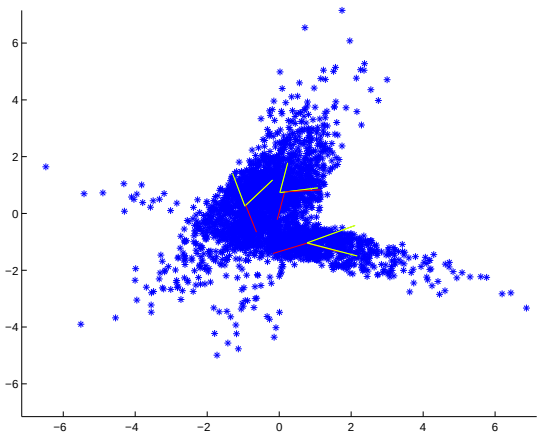
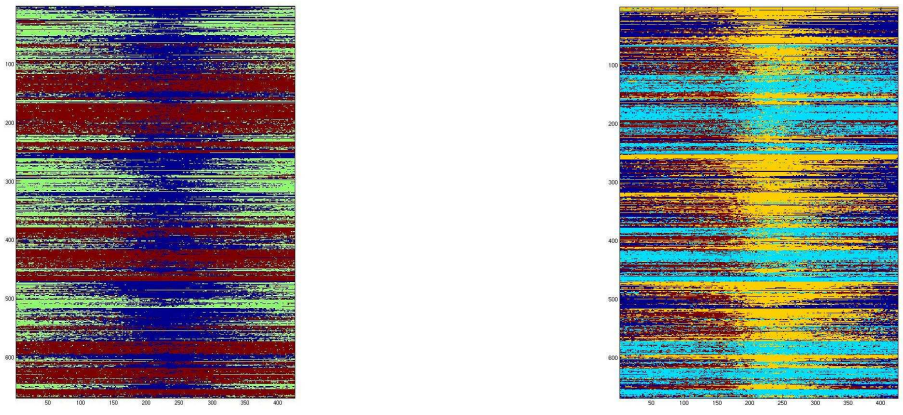


Fig. 2. Toy experiment with three models. True model bases are shown in red, and learned bases are in yellow, centered at the learned centers. Three Generalized Gaussian source mixture densities are used with fixed shape parameters of $\rho = 1.5$ (other choices also work, as does adapting the shape parameter.) The true sources are Laplacian and Generalized Gaussian with shape parameter $\rho = 5$. The adaptive sources are able to separate this combination of sub- and super-gaussian sources. The Newton method greatly speeds convergence, allowing the step size to tend to 1, whereas the natural gradient requires reduction of the step size near the optimum, inducing very slow linear convergence.



(a) (b)

Fig. 3. Segmentation of EEG trials: (a) three models (b) four models. The subject is shown a sequence of letters and responds whether current letter is the same as letter before last. Trials (letter presentation, response, feedback) are time-locked (synchronized): at $t = 175$ there is feedback as to whether the response was correct or incorrect.

REFERENCES

[1] M. Rosenblatt, *Gaussian and Non-Gaussian Linear Time Series and Random Fields*, Springer, 2000.

[2] Q. Cheng, "On the unique representation of non-gaussian linear processes," *The Annals of Statistics*, vol. 20, pp. 1143–1145, 1992.

[3] D. T. Pham and Ph. Garat, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, no. 11, pp. 2768–2779, 1996.

[4] J. A. Palmer and K. Kreutz-Delgado, "A general framework for component estimation," in *Proceedings of the 4th International Symposium on Independent Component Analysis*, 2003.

[5] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. 1996, pp. 145–151, MIT Press, Cambridge, MA.

[6] S. Makeig, T-P. Jung, D. Ghahremani, A. J. Bell, and T. J. Sejnowski, "Blind separation of event-related brain responses into independent components," *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 10979–10984, 1997.

[7] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, pp. 1200–1224, 1995.

[8] T.-W. Lee, M. S. Lewicki, and T. J. Sejnowski, "ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1078–1089, 2000.

[9] R. A. Choudrey and S. J. Roberts, "Variational mixture of Bayesian independent component analysers," *Neural Computation*, vol. 15, no. 1, pp. 213–252, 2002.

[10] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[11] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE. Trans. Sig. Proc.*, vol. 44, no. 12, pp. 3017–3030, 1996.

[12] S.-I. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[13] J. A. Palmer, *Variational and Scale Mixture Representations of Non-Gaussian Densities for Estimation in the Bayesian Linear Model*, Ph.D. thesis, University of California San Diego, 2006, Available at http://sccn.ucsd.edu/∼jason.

[14] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Super-Gaussian mixture source model for ICA," in *Proceedings of the 6th International Conference on Independent Component Analysis*, J. Rosca et al., Ed. 2006, Lecture Notes in Computer Science, Springer-Verlag.

[15] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.

[16] M. E. Tipping and C. M Bishop, "Mixtures of probabilistic principle component analyzers," *Neural Computation*, vol. 11, pp. 443–482, 1999.

[17] Sam Roweis and Zoubin Ghahramani, "A unifying review of linear gaussian models," *Neural Computation*, vol. 11, no. 5, pp. 305–345, 1999.

[18] D. J. C. Mackay, "Comparison of approximate methods for handling hyperparameters," *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[19] M. E. Tipping, "Sparse Bayesian learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[20] M. E. Tipping and N. D. Lawrence, "Variational inference for student's t models: Robust Bayesian interpolation and generalised component analysis," *Neurocomputing*, vol. 69, pp. 123–141, 2005.

[21] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[22] H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems 12*. 2000, MIT Press.

[23] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Advances in Neural Information Processing Systems 12*. 2000, MIT Press.

[24] H. Lappalainen, "Ensemble learning for independent component analysis," in *Proceedings of the First International Workshop on Independent Component Analysis*, 1999.

[25] James W. Miskin, *Ensemble Learning for Independent Component Analysis*, Ph.D. thesis, Dissertation, University of Cambridge, 2000.

[26] K. Chan, T.-W. Lee, and T. J. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," *Journal of Machine Learning Research*, vol. 3, pp. 99–114, 2002.

[27] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution: The dynamic component analysis algorithm," *Neural Computation*, vol. 10, pp. 1373–1424, 1998.

[28] S. C. Douglas, A. Cichocki, and S. Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions," in *Proc. IEEE Workshop on Neural Networks for Signal Processing, Amelia Island Plantation, FL*, 1997, pp. 436–445.

[29] D. T. Pham, "Mutual information approach to blind separation of stationary sources," *IEEE Trans. Information Theory*, vol. 48, no. 7, pp. 1935–1946, 2002.

[30] A. M. Bronstein, M. M. Bronstein, and M. Zibulevsky, "Relative optimization for blind deconvolution," *IEEE Transactions on Signal Processing*, vol. 53, no. 6, pp. 2018–2026, 2005.

[31] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Speech and Audio Processing*, vol. 15, no. 1, 2007.

[32] A. Hyvärinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, 2001.

[33] T. Eltoft, T. Kim, and T.-W. Lee, "Multivariate scale mixture of Gaussians modeling," in *Proceedings of the 6th International Conference on Independent Component Analysis*, J. Rosca et al., Ed. 2006, Lecture Notes in Computer Science, pp. 799–806, Springer-Verlag.

[34] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, 1970.