AMICA: An Adaptive Mixture of Independent Component Analyzers with Shared Components

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

A. Related Work

The Gaussian liner model approach is described [1]–[3]. Non-Gaussian sources in the form of Gaussian scale mixtures, in particular Student's t distribution, were developed in [4]–[6]. A mixture of Gaussians source model was employed in [7]–[11]. Similar approaches were proposed in [12], [13]. 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 [14]. The main distinguishing features of the proposed model are,

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.

1) Mixtures of Gaussian scale mixture sources provide more flexibility than the Gaussian mixture models of [7], [11], or fixed density models used in [14]. 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 [15] 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 [16], where blind deconvolution is treated in a manner similar to [17]–[20], as well as complex ICA [21] and dependent sources [21]–[23]. 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 [24],

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

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 $\mathbf{V}$,

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

when $\mathbf{W}$ is full rank. The direction (3) is known as the "natural gradient" [25].

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,

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

where δ_{ik} is the Kronecker delta symbol, and $\big\langle \cdot \big\rangle_N$ 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,

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

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 $\mathbf{C} = \mathcal{H}(\mathbf{B})$,

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

where $\mathrm{diag}(f''(\mathbf{y}_t))$ is the diagonal matrix with diagonal elements $f_i''(y_{it})$.

This equation can be simplified as follows. First, let us rewrite the transformation (5) in terms of the source estimates $\mathbf{y}$. We first write,

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

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,

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

At the optimum, we may assume that the source density models $q_i(y_i)$ are equivalent to the true source densities $p_i(s_i)$. Writing (6) in component form and letting N go to infinity we find for the diagonal elements,

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

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, as in [15], [26], [27], 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 [15] 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 $p(y) \propto 1/y$, 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 $\tilde{\mathbf{B}} = \tilde{\mathcal{H}}^{-1}(\tilde{\mathbf{C}})$ is easily carried out by inverting the transformation (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. Cramer-Rao Lower Bound

By the theorem on asymptotic efficiency of Maximum Likelihood estimation, we have that the asymptotic and minimum error covariance of the estimation of $\mathbf{W}$ is given by the inverse Fisher Information matrix. Also, since the asymptotic distribution is Gaussian, we can determine the asymptotic distribution linear transformations of $\hat{\mathbf{W}}$. In particular, we see that the asymptotic distribution of $\mathbf{C} \triangleq \hat{\mathbf{W}}\mathbf{A}$, where $\mathbf{A} = \mathbf{W}^{-1}$ is the true (unknown) mixing matrix. Remarkably, we find that this asymptotic distribution does not depend on $\mathbf{A}$, but only on the source density statistics.

Specifically, the inverse Fisher information matrix is $\tilde{\mathcal{H}}^{-1}$, and the asymptotic minimum error covariance in the estimate is $N^{-1}\tilde{\mathcal{H}}^{-1}$. The minimum error covariance matrix for the pair of off-diagonals c_{ij} and c_{ji} with a sample size N is given by,

$$\frac{1}{N}\,\frac{1}{\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1}\begin{bmatrix} \kappa_j \sigma_i^2 & -1 \\ -1 & \kappa_i \sigma_j^2 \end{bmatrix}$$

In particular, the asymptotic distribution of $\mathbf{C}$ is Normal with mean $\mathbf{I}$, and variance,

$$E\{\hat{c}_{ij}^2\} \geq \frac{1}{N}\frac{\kappa_j\sigma_i^2}{\kappa_i\kappa_j\sigma_i^2\sigma_j^2 - 1}, \quad E\{(\hat{c}_{ii}-1)^2\} \geq \frac{1}{N}\frac{1}{1+\eta_i}$$

For the Generalized Gaussian density, we have,

$$\kappa_i = \frac{\rho_i^2\,\Gamma(2-1/\rho_i)}{\Gamma(1/\rho_i)}, \quad \sigma_i^2 = \frac{\Gamma(3/\rho_i)}{\Gamma(1/\rho_i)}, \quad \eta_i = \rho_i - 1$$

V. 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},

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

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,

$$u_{thij} \triangleq \begin{cases} 1, & j_{thi} = j \text{ and } h_t = h \\ 0, & \text{otherwise} \end{cases}$$

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

$$\begin{aligned} p(\mathbf{X};\Theta) &= \sum_{\mathbf{V}}\prod_{t=1}^{N}\prod_{h=1}^{M} P_{th}^{v_{th}} = \sum_{\mathbf{V}}\prod_{t=1}^{N}\exp\Big(\sum_{h=1}^{M} v_{th}L_{th}\Big) \\ &= \sum_{\mathbf{U},\mathbf{V}}\exp\Big(\sum_{t=1}^{N}\sum_{h=1}^{M} v_{th}\big(\log\gamma_h + \log|\det\mathbf{W}_h|\big) + \sum_{i=1}^{d}\sum_{j=1}^{m} u_{thij}Q_{thij}\Big) \end{aligned}$$

where we make the definitions,

$$\begin{aligned} \mathbf{b}_{th} &\triangleq \mathbf{W}_h(\mathbf{x}_t - \mathbf{c}_h) \\ y_{thij} &\triangleq \sqrt{\beta_{k_{hi}j}}\,\big([\mathbf{b}_{th}]_i - \mu_{k_{hi}j}\big) \\ \exp(Q_{thij}) &\triangleq \alpha_{k_{hi}j}\sqrt{\beta_{k_{hi}j}}\,q_{k_{hi}j}\big(y_{thij}\big) \\ P_{th} &\triangleq \gamma_h\big|\det\mathbf{W}_h\big|\prod_{i=1}^{d}\sum_{j=1}^{m}\exp(Q_{thij}) \triangleq \exp(L_{th}) \end{aligned}$$

The posterior expectation, $\hat{v}_{ht}^l$ is given by,

$$\hat{v}_{th}^l = E\{v_{th}|\mathbf{x}_t;\Theta^l\} = P\big[v_{th}=1|\mathbf{x}_t;\Theta^l\big] = \frac{P_{th}^l}{\sum_{h'=1}^{M}P_{th'}^l} = \frac{\exp(L_{th}^l)}{\sum_{h'=1}^{M}\exp(L_{th'}^l)} \tag{19}$$

For the $\hat{u}^l_{hijt}$, we have,

$$\begin{aligned}\hat{u}^l_{thij} &= P\big[u_{thij}=1 \,|\, \mathbf{x}_t;\Theta^l\big] \\ &= P\big[u_{thij}=1 \,|\, v_{ht}=1,\mathbf{x}_t;\Theta^l\big]P\big[v_{ht}=1 \,|\, \mathbf{x}_t;\Theta^l\big] \\ &= \hat{z}^l_{thij}\hat{v}^l_{th} \end{aligned} \tag{20}$$

where $\hat{z}^l_{thij} \triangleq E\big\{u_{thij}|\, v_{th}=1,\mathbf{x}_t,;\Theta^l\big\}$:

$$\hat{z}^l_{thij} = \frac{\exp(Q^l_{thij})}{\sum_{j'=1}^m \exp(Q^l_{thij'})} \tag{21}$$

The function to be maximized in the EM algorithm is then,

$$\sum_{t=1}^N\sum_{h=1}^M \Bigg[\hat{v}^l_{th}\big(\log\gamma_h + \log|\det \mathbf{W}_h|\big) \;+\; \sum_{i=1}^d\sum_{j=1}^m \hat{u}^l_{thij}\big(\log\alpha_{k_{hi}j} + \tfrac{1}{2}\log\beta_{k_{hi}j} - f_{hij}(y_{thij})\big)\Bigg]$$

where $f_{k_{hi}j} \triangleq -\log q_{k_{hi}j}$. Maximizing with respect to γ_h subject to $\gamma_h \geq 0$, $\sum_h \gamma_h = 1$, we get,

$$\gamma^{l+1}_h = \frac{1}{N}\sum_{t=1}^N \hat{v}^l_{th} \tag{22}$$

We can then rearrange the likelihood as,

$$\sum_{h=1}^M \gamma^{l+1}_h \log|\det \mathbf{W}_h| \;+\; \sum_{k=1}^n\sum_{j=1}^m\sum_{h:i_{kh}>0} \hat{u}^l_{thi_{kh}j}\big(\log\alpha_{kj} + \tfrac{1}{2}\log\beta_{kj} - f_{kj}(y_{thi_{kh}j})\big)\Big)$$

Maximizing with respect to α_{kj} subject to $\alpha_{kj} \geq 0$, $\sum_j \alpha_{kj} = 1$, we get,

$$\alpha^{l+1}_{kj} = \frac{\sum_{h:i_{kh}>0}\sum_{t=1}^N \hat{v}^l_{th}\hat{z}^l_{thi_{kh}j}}{\sum_{h:i_{kh}>0}\sum_{t=1}^N \hat{v}^l_{th}} \tag{23}$$

We define the following expectations conditioned on the model h, in which G_t are arbitrary functions of $\mathbf{x}_t$

$$E_v\{G_t \,|\, h\} \;\triangleq\; \frac{\sum_t \hat{v}^l_{th}G_t}{\sum_t \hat{v}^l_{th}}, \quad E_u\{G_t \,|\, h,j\} \;\triangleq\; \frac{\sum_t \hat{u}^l_{thij}G_t}{\sum_t \hat{u}^l_{thij}}$$

We also define the following, conditioned on the component k, in which G_{th} are arbitrary functions of $\mathbf{x}_t$ and parameters of model an arbitrary model h,

$$E_v\{G_{th} \,|\, k\} \;\triangleq\; \frac{\sum_{h:i_{kh}>0}\sum_t \hat{v}^l_{th}G_{th}}{\sum_{h:i_{kh}>0}\sum_t \hat{v}^l_{th}}, \quad E_u\{G_{th} \,|\, k,j\} \;\triangleq\; \frac{\sum_{h:i_{kh}>0}\sum_t \hat{u}^l_{thi_{kh}j}G_{th}}{\sum_{h:i_{kh}>0}\sum_t \hat{u}^l_{thi_{kh}j}}$$

Now we can rearrange the likelihood as,

$$\sum_{h=1}^M \gamma^{l+1}_h \log|\det \mathbf{W}_h| \;+\; \sum_{k=1}^n \zeta^{l+1}_k \sum_{j=1}^m \alpha^{l+1}_{kj}\Big(\tfrac{1}{2}\,\log\beta_{kj} - E_u\big\{f_{kj}\big(\sqrt{\beta_{kj}}\,([\mathbf{b}_{th}]_{i_{kh}} - \mu_{kj})\big)|\, k,j\big\}\Big)$$

where we define $\zeta_k = \sum_{h:i_{kh}>0} \gamma_h$ to be the total probability of a context containing component k,

$$\zeta_k^{l+1} \triangleq \sum_{h:i_{kh}>0} \gamma_h^{l+1} \tag{24}$$

If the source mixture component densities are strongly super-Gaussian, then we can maximize the surrogate likelihood,

$$\sum_{h=1}^{M} \gamma_h^{l+1} \log|\det \mathbf{W}_h| \;+\; \sum_{k=1}^{n} \zeta_k^{l+1} \sum_{j=1}^{m} \alpha_{kj}^{l+1} \Big(\tfrac{1}{2} \log \beta_{kj} - \tfrac{1}{2}\, \beta_{kj}\, E_u\big\{ \xi_{thi_{kh}j}^{l} \big([\mathbf{b}_{th}]_{i_{kh}} - \mu_{kj} \big)^2 \big\} \Big)$$

where,

$$\xi_{thi_{kh}j}^{l} \triangleq \frac{f'_{kj}(y_{thi_{kh}j}^{l})}{y_{thi_{kh}j}^{l}}$$

The location and scale parameter updates are then given by,

$$\mu_{kj}^{l+1} \;=\; \frac{E_u\big\{ \xi_{thi_{kh}j}^{l} [\mathbf{b}_{th}]_{i_{kh}} \,|\, k, j \big\}}{E_u\big\{ \xi_{thi_{kh}j}^{l} \,|\, k, j \big\}} \;=\; \mu_{kj}^{l} + \frac{1}{\sqrt{\beta_{kj}^{l}}} \frac{E_u\big\{ f'_{kj}(y_{thi_{kh}j}^{l}) \,|\, k, j \big\}}{E_u\big\{ f'_{kj}(y_{thi_{kh}j}^{l}) / y_{thi_{kh}j}^{l} \,|\, k, j \big\}} \tag{25}$$

and,

$$\beta_{kj}^{l+1} \;=\; \frac{1}{E_u\big\{ \xi_{thi_{kh}j}^{l} \big([\mathbf{b}_{th}]_{i_{kh}} - \mu_{kj}^{l} \big)^2 \big| \, k, j \big\}} \;=\; \frac{\beta_{kj}^{l}}{E_u\big\{ f'_{kj}(y_{thi_{kh}j}^{l})\, y_{thi_{kh}j}^{l} \,|\, k, j \big\}} \tag{26}$$

The Generalized Gaussian shape parameters are updated by,

$$\Delta\rho_{hij} \;=\; 1 \;-\; \big(\rho_{kj}^{l} / \Psi\big(1 + 1/\rho_{kj}^{l}\big) \big)\, E_u\big\{ |y_{hi_{kh}jt}|^{\rho_{kj}^{l}} \log |y_{hi_{kh}jt}|^{\rho_{kj}^{l}} \,\big|\, k, j \big\} \tag{27}$$

or if $\rho_{kj} > 2$,

$$\Delta\rho_{kj} \;=\; \Psi\big(1 + 1/\rho_{hij}^{l}\big)/\rho_{kj}^{l} \;-\; E_u\big\{ |y_{hijt}|^{\rho_{kj}^{l}} \log |y_{hijt}|^{\rho_{kj}^{l}} \,\big|\, k, j \big\} \tag{28}$$

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,

$$\log|\det \mathbf{W}_h| - E_v\Big\{ \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{z}_{thij}^{l}\, f_{k_{hi}j}(y_{thij}) \,\Big|\, h \Big\}$$

The gradient of this function is,

$$-\mathbf{W}_h^{-T} + E_v\big\{ \mathbf{g}_{th} (\mathbf{x}_t - \mathbf{c}_h)^T \,|\, h \big\} \tag{29}$$

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

$$[\mathbf{g}_{th}]_i \triangleq \sum_{j=1}^{m} \hat{z}_{thij}^{l} \sqrt{\beta_{k_{hi}j}}\, f'_{k_{hi}j}\big(y_{thij}\big) \tag{30}$$

Denote the matrix gradient (29) by $\mathbf{G}_h$. Taking the derivative of $[\mathbf{G}_h]_{i\nu}$ with respect to $[\mathbf{W}_h]_{k\lambda}$, we get,

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

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

$$\mathbf{C}=\mathbf{W}_h^{-T}\mathbf{B}^T\mathbf{W}_h^{-T}+E_v\big\{\mathbf{D}_{th}\mathbf{B}\,(\mathbf{x}_t-\mathbf{c}_h)(\mathbf{x}_t-\mathbf{c}_h)^T\,\big|\,h\big\} \tag{31}$$

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

$$\big[\mathbf{D}_{th}\big]_{ii}=\sum_{j=1}^{m}\hat{z}_{thij}^{l}\,\beta_{k_{hi}j}\,f''_{hij}\big(y_{thij}\big) \tag{32}$$

To simplify the calculation of the asymptotic value of the Hessian, we rewrite the second term on the right hand side of (31) as,

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

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,

$$\tilde{\mathbf{C}}=\tilde{\mathbf{B}}^T+E_v\big\{\mathbf{D}_{th}\tilde{\mathbf{B}}\,\mathbf{b}_{th}\mathbf{b}_{th}^T\,\big|\,h\big\} \tag{33}$$

Now we write the ith row of the second term in (33) as,

$$\sum_{j=1}^{m}\beta_{k_{hi}j}E_v\big\{\hat{z}_{thij}^{l}f''_{k_{hi}j}\big(y_{thij}\big)\,[\tilde{\mathbf{B}}\,\mathbf{b}_{th}]_i\,\mathbf{b}_{th}^T\,\big|\,h\big\} \tag{34}$$

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

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

where $\bar{\alpha}_{hij}^{l+1}=E_v\{\hat{z}_{thij}^{l}|\,h\}$ (sum weighted only by model h likelihood). If we define,

$$\begin{aligned}
\eta_{hij} &\triangleq E_u\big\{f''_{k_{hi}j}(y_{thij})\,y_{thij}^2\,|\,h,j\big\}\\
\kappa_{hij} &\triangleq E_u\big\{f''_{k_{hi}j}(y_{thij})\,|\,h,j\big\}
\end{aligned}$$

or, using integration by parts to rewrite the integrals,

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

then the expressions can be simplified to the following,

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

Define,

$$\mathbf{\Phi}_h \triangleq E_v\big\{\mathbf{g}_{th}\mathbf{b}_{th}^T \,\big|\, h\big\} \tag{35}$$

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

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

then we have,

$$[\tilde{\mathbf{B}}]_{ii} = \frac{1-[\mathbf{\Phi}_h]_{ii}}{\lambda_{hi}},\quad i=1,\ldots,n \tag{36}$$

$$[\tilde{\mathbf{B}}]_{ij} = \frac{[\mathbf{\Phi}_h]_{ji}-\kappa_{hj}\sigma_{hi}^2[\mathbf{\Phi}_h]_{ij}}{\kappa_{hi}\kappa_{hj}\sigma_{hi}^2\sigma_{hj}^2-1},\quad \forall\, i\neq j \tag{37}$$

Then

$$\Delta\mathbf{W}_h = \tilde{\mathbf{B}}\mathbf{W}_h \tag{38}$$

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

$$L\big(\Theta^l|\mathbf{X}\big) = \sum_{t=1}^{N}\log\Big(\sum_{h=1}^{M}\exp(L_{th}^l)\Big) \tag{39}$$

VI. Experiments

VII. Conclusion

References

[1] 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.

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

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

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

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

[6] 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.

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

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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/$\sim$jason.

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

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

[27] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *IEEE Trans. Signal Processing*, vol. 45, no. 7, pp. 1712–1725, 1997.