Journal of Machine Learning Research 0 (0000) 0-0 Submitted 0/00; Published 0/00

An Independent Component Analysis Mixture Model with Adaptive Source Densities

Jason Palmer JAPALMER@UCSD.EDU
Ken Kreutz-Delgado KREUTZ@ECE.UCSD.EDU
Scott Makeig SCOTT@SCCN.UCSD.EDU
Department of Electrical and Computer Engineering
University of California San Diego
La Jolla, CA 90293-0407, USA

Editor: XXXX

Abstract

We propose an extension of the mixture of factor (or independent component) analyzers model to include strongly super-gaussian mixture source densities. This density model greater economy in the representation of densities with (multiple) peaked modes or heavy tails than using several Gaussians to represent these features. Subgaussian features can also be represented using Gaussian mixtures, which are a special case. We exploit an convexity-based inequality to derive an EM algorithm for maximum likelihood estimation of the model, and show that it converges globally to a local optimum of the actual non-gaussian mixture model without requiring any approximations. We also propose adaptive Generalized Gaussian and Student's t mixtures which adapt the shape parameters of the mixture component densities. Experiments verify the validity of the algorithm.

Keywords: Mixture model, ICA, super-gaussian, EM

1. Introduction

In this paper we propose an extension of the mixture of factor (Attias, 1999), or independent component (Choudrey and Roberts, 2002) analyzers model. The extension increases the flexibility of the source density mixture model by employing strongly super-gaussian component densities, while working within a theoretical framework that allows us to maintain the global convergence properties of the Gaussian mixture model. Mixture model source densities allow one to model skewed and multi-modal densities.

The variational Gaussian mixture models, proposed by Ghahramani and Beal (2000), Attias (1999), Choudrey and Roberts (2002), Chan et al. (2002), are ultimately mixtures of Student's t distributions after the random variance is integrated out (Tipping, 2001, Attias, 2000). Lee et al. (2000) propose a mixture generalization of the Infomax algorithm, employing a mixture model over sets of basis vectors but not for the source component density models. The means are updated by gradient descent or by a heuristic approximate EM update. Park and Lee (2004) employ a variance mixture of Laplacians model over the source densities, in which the Laplacian components in each mixture have the same mean, but differing variances. An EM algorithm is derived by exploiting the closed form solution

of the M-step for the variance parameters. Pearlmutter and Parra (1996) estimate a mixture of Logistic source density model by gradient descent.

The proposed model generalizes all of these algorithms, applying to Gaussian, Laplacian, Logistic, as well as Generalized Gaussian, Student's t, and any mixture combination of these densities. The key to the algorithm is the definition of an appropriate class of densities, and showing that the "complete log likelihood" that arises in the EM algorithm can be guaranteed to increase as a result of an appropriate parameter update, which thus guarantees increase in the true likelihood. It is thus a "Generalized EM" (GEM) algorithm (Dempster et al., 1977). For a given number of mixture components, the EM algorithm estimates the location (mode) and scale parameters of each mixture component. We follow Neal and Hinton (1998) and Saul et al. (1996) in deriving the EM algorithm.

The property of strongly super-gaussian densities that we use, namely log-convexity in x^2, has been exploited previously by Jaakkola (1997) and Jaakkola and Jordan (1997) for graphical models, and Girolami (2001) for ICA using the Laplacian density. The model we propose extends the work of Girolami (2001) in applying more generally to the (large) class of strongly super-gaussian densities, as well as mixtures of these densities. We also take the approach of Attias (2000) in allowing the scale of the sources to vary (actually a necessity in the mixture case,) and fixing the scale of the un-mixing filters at unity by an appropriate transformation at each iteration, in order to avoid the scale ambiguity inherent in factor analysis models.

Using the natural gradient (Amari, 1998) to update the un-mixing matrices (the inverses of the basis or mixing matrices), we can further guarantee (in principle) increase of the likelihood. Furthermore, it is possible, for densities that are parameterized besides the location and scale parameters such that all densities in a range of the additional parameter are strongly super-gaussian, e.g. Generalized Gaussian shape parameters less than 2, to update these parameters according to the gradient of the free energy, remaining in the GEM framework and guaranteeing increase in the data likelihood under the model. The de-mixing matrices and any other shape parameters will require a step size to be specified in advance, but the mixture component locations and scales will be updated in closed form. In the Gaussian case, the algorithm reduces to the classical EM algorithm for Gaussian mixtures.

The practical situation in which we shall be interested is the analysis of EEG/MEG, the characteristics of which are large number of channels and data points, and mildly skewed occasionally multi-modal source densities. The large number of channels constrains the algorithm to be scalable. This along with the large number of data points suggests the natural gradient maximum likelihood approach, which is scalable and asymptotically efficient. The large amount of data also dictates that we limit computational and storage overhead to only what is necessary or actually beneficial, rather than doing Bayesian MAP estimation of all parameters as in the variational Bayes algorithms (Attias, 2000, Choudrey and Roberts, 2002). Also for computational reasons we consider only noiseless mixtures of complete bases so that inverses exist.

In §2 we define strongly super-gaussian densities and mixtures of these densities. In §3-5 we derive the EM algorithm for density estimation. In §6 we introduce adaptive Generalized Gaussian and Student's t algorithms. §7 contains experimental verification of the theory.

2. Strongly Super-Gaussian Mixtures

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

An equivalent definition is given by Benveniste et al. (1980), which defines $p(x) = \exp(-f(x))$ to be super-gaussian (sub-gaussian) if $f'(x)/x$ is decreasing (increasing) on $(0,\infty)$. This condition is equivalent to $f(x) = g(x^2)$ with g concave, i.e. g' decreasing, where $g'(x^2) = f'(x)/x$.

Palmer et al. (2005) discuss these densities in some detail, and derive relationships between them and the hyperprior representation used in the evidence framework (Mackay, 1999) and the Variational Bayes framework (Attias, 1999). 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, $\forall\, s,t$,

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

Examples of densities satisfying this criterion are: (i) Generalized Gaussian $\propto \exp(-|s|^\beta)$, $0 < \beta \le 2$, (ii) Logistic $\propto 1/\cosh^2(s/2)$, (iii) Student's $t \propto (1 + s^2/\nu)^{-(\nu+1)/2}$, $\nu > 0$, (iv) symmetric α-stable densities (having characteristic function $\exp(-|\omega|^\alpha)$, $0 < \alpha \le 2$), (v) all Gaussian scale mixtures (Keilson and Steutel, 1974, Palmer et al., 2005). The property of being strongly sub- or super-gaussian is independent of scale.

Density Name	Density Form	$\xi = f'(y)/y$
Generalized Gaussian, $0 < \rho \le 2$	$\exp(-\lvert y\rvert^\rho)$	$\lvert y\rvert^{\rho-2}$
Student's t, $\nu > 0$	$(1 + y^2/\nu)^{-(\nu+1)/2}$	$(\nu+1)/(\nu + y^2)$
Jeffrey's prior	$1/y$	$1/y^2$
Logistic	$1/\cosh^2(y/2)$	$\tanh(y/2)/y$
Symmetric α-stable	no closed form	no closed form

Table 1: Variational weight parameter for common strongly super-gaussian densities.

The theory is developed for the class of densities that are strongly super-gaussian mixtures. Mixture densities have the form,

$$p(s) = \sum_{j=1}^{m} \alpha_j \sqrt{\beta_j}\, p_j\Big(\sqrt{\beta_j}\big(s - \mu_j\big)\Big)\,, \quad \sum_j \alpha_j = 1,\, \alpha_j \ge 0,\, \beta_j > 0$$

We first consider a single square mixing matrix $\mathbf{A}$ with super-gaussian mixture sources, so that the $p_j(s)$ are assumed to be strongly super-gaussian. Note that $p(s)$ is not necessarily super-gaussian, only the mixture components densities $p_j(s)$. Later we extend the model to mixtures over mixing or basis matrices. Initially, the jth source mixture component density of the ith source will be denoted $p_{ij}(s_{ij})$ with mode (location) μ_{ij} and inverse square scale β_{ij}. In the Gaussian case $p_{ij}(s) = \mathcal{N}(s\,;\mu_{ij},\beta_{ij}^{-1})$, μ_{ij} is the mean and β_{ij} is the inverse variance. For general strongly super-gaussian densities, μ_{ij} is the mean only if the mean exists, and β_{ij} is the inverse variance divided by $\int s^2 p_{ij}(s) ds$ only when the latter exists.

3. The EM Algorithm

We follow the framework of Neal and Hinton (1998) and Saul et al. (1996) in deriving the EM algorithm, which was first derived rigorously by Dempster et al. (1977). The log likelihood of the data decomposes as follows,

$$\begin{aligned} \log p(\mathbf{x};\theta) &= \int q(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{z},\mathbf{x};\theta)}{q(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} \;+\; D\big(q(\mathbf{z}|\mathbf{x}) \big\| \, p(\mathbf{z}|\mathbf{x};\theta)\big) \\ &\equiv -F(q;\theta) \;+\; D(q\,\|\,p_\theta) \end{aligned} \tag{2}$$

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* (Saul et al., 1996, Neal and Hinton, 1998). 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 (2), 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,

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

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 $F(q^l;\theta^{l+1}) \le F(q^l;\theta^l)$, i.e. that F decreases as a result of updating θ. This leads to the Generalized EM (GEM) algorithm (Dempster et al., 1977), which we employ in this paper.

To guarantee a decrease in $F(q;\theta)$ with respect to θ, we use the inequality (1) 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',

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

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

4. ICA with Strongly Super-Gaussian Mixture Sources

Let the data $\mathbf{x}_k$, $k = 1,\ldots,N$ be given, and consider the instantaneous model,

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

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is non-singular, and the sources s_i, $i = 1,\ldots,n$, are independent with strongly super-gaussian mixture densities. We allow the number of source mixture components m_i to differ for different sources.

We wish to estimate the parameter $\mathbf{W} = \mathbf{A}^{-1}$ and the parameters of the source mixtures,

$$\theta = \{\mathbf{w}_i, \alpha_{ij}, \mu_{ij}, \beta_{ij}\}, \quad i = 1, \ldots, n\,,\ j = 1, \ldots, m_i$$

where the vector $\mathbf{w}_i$ is the ith column of $\mathbf{W}^T$. We define $\mathbf{X} \equiv [\mathbf{x}_1 \cdots \mathbf{x}_N]$.

The source mixture model is equivalent to a scenario in which for each source s_i, a mixture component j_i is drawn from the discrete probability distribution $P[j_i = j] = \alpha_{ij}$, $1 \le j \le m_i$, then s_i is drawn from the mixture component density p_{ij_i}. We define j_{ik} to be the index chosen for the ith source in the kth sample.

To use the EM algorithm, we define the random variables z_{ijk} as follows,

$$z_{ijk} \equiv \begin{cases} 1, & j_{ik} = j \\ 0, & \text{otherwise} \end{cases}$$

Let $\mathbf{Z} = \{z_{ijk}\}$. Then we have,

$$p(\mathbf{X}; \theta) = \sum_{\mathbf{Z}} \prod_{k=1}^{N} |\det \mathbf{W}| \prod_{i=1}^{n} \prod_{j=1}^{m_i} \left[\alpha_{ij} \sqrt{\beta_{ij}}\, p_{ij}\Big(\sqrt{\beta_{ij}}\big(\mathbf{w}_i^T \mathbf{x}_k - \mu_{ij}\big)\Big) \right]^{z_{ijk}}$$

For the variational free energy, we have $F(q^l; \theta) = F^l(\theta) + H(\mathbf{Z}; \theta^l)$, where $H(\mathbf{Z}; \theta^l)$ is the entropy of the $\mathbf{Z}$ evaluated for $\theta = \theta^l$, and $F^l(\theta)$ is given by,

$$F^l(\theta) = -N \log |\det \mathbf{W}| \ + \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \hat{z}_{ijk}^l \Big[-\log \alpha_{ij} - \tfrac{1}{2} \log \beta_{ij} + f_{ij}\Big(\sqrt{\beta_{ij}}\big(\mathbf{w}_i^T \mathbf{x}_k - \mu_{ij}\big)\Big)\Big]$$

where we define $f_{ij} \equiv -\log p_{ij}$ and $\hat{z}_{ijk}^l \equiv E[z_{ijk} | \mathbf{x}_k; \theta^l]$. We also define $y_{ijk} \equiv \sqrt{\beta_{ij}}\big(\mathbf{w}_i^T \mathbf{x}_k - \mu_{ij}\big)$, and,

$$y_{ijk}^l \equiv \sqrt{\beta_{ij}^l}\left(\mathbf{w}_i^{lT} \mathbf{x}_k - \mu_{ij}^l \right) \tag{3}$$

The $\hat{z}_{ijk}^l = P[z_{ijk} = 1 | \mathbf{x}_k; \theta^l]$ are determined as in the usual Gaussian EM algorithm,

$$\hat{z}_{ijk}^l = \frac{p(\mathbf{x}_k | z_{ijk} = 1; \theta^l) P[z_{ijk} = 1; \theta^l]}{\sum_{j'=1}^{m_i} p(\mathbf{x}_k | z_{ij'k} = 1; \theta^l) P[z_{ij'k} = 1; \theta^l]} = \frac{\alpha_{ij}^l \sqrt{\beta_{ij}^l}\, p_{ij}(y_{ijk}^l)}{\sum_{j'=1}^{m_i} \alpha_{ij'}^l \sqrt{\beta_{ij'}^l}\, p_{ij'}(y_{ij'k}^l)} \tag{4}$$

The new α_{ij} are found by maximizing $F^l(\theta)$ such that $\sum_{j=1}^{m_i} \alpha_{ij} = 1$, $\alpha_{ij} > 0$, yielding,

$$\alpha_{ij}^{l+1} \ = \ \frac{\sum_{k=1}^{N} \hat{z}_{ijk}^l}{\sum_{j'=1}^{m_i} \sum_{k=1}^{N} \hat{z}_{ij'k}^l} \ = \ \frac{1}{N} \sum_{k=1}^{N} \hat{z}_{ijk}^l \tag{5}$$

which is equivalent to the update in the ordinary Gaussian mixture model EM algorithm.

To update the source mixture component parameters, we define,

$$\xi_{ijk}^l \ \equiv \ \frac{f_{ij}'(y_{ijk}^l)}{y_{ijk}^l} \tag{6}$$

and use the inequality (1) to replace $f_{ij}(y_{ijk})$ in $F^l(\theta)$ by $\frac{1}{2}\xi^l_{ijk}y^2_{ijk}$ to get,

$$\tilde{F}^l(\theta) = -N\log|\det \mathbf{W}| \; + \sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{z}^l_{ijk}\left[-\log\alpha_{ij} - \tfrac{1}{2}\log\beta_{ij} + \tfrac{1}{2}\xi^l_{ijk}\beta_{ij}\left(\mathbf{w}_i^T\mathbf{x}_k - \mu_{ij}\right)^2\right]$$

Minimizing $\tilde{F}^l$ with respect to μ_{ij} and β_{ij} guarantees, using the inequality (1), 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_{ij} does not depend on β_{ij}. The updates, using the definitions (3), (4) and (6), are found to be,

$$\mu^{l+1}_{ij} \;=\; \frac{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}\mathbf{w}_i^{lT}\mathbf{x}_k}{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}} \;=\; \mu^l_{ij} + \frac{\sum_{k=1}^N \hat{z}^l_{ijk}f'_{ij}(y^l_{ijk})}{\sqrt{\beta^l_{ij}}\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}} \tag{7}$$

and,

$$\beta^{l+1}_{ij} \;=\; \frac{\sum_{k=1}^N \hat{z}^l_{ijk}}{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}\left(\mathbf{w}_i^{lT}\mathbf{x}_k - \mu^l_{ij}\right)^2} \;=\; \frac{\beta^l_{ij}\sum_{k=1}^N \hat{z}^l_{ijk}}{\sum_{k=1}^N \hat{z}^l_{ijk}f'_{ij}(y^l_{ijk})y^l_{ijk}} \tag{8}$$

We adapt $\mathbf{W}$ according to the natural gradient of F. Defining the vector $\mathbf{u}^l_k$ such that,

$$\left[\mathbf{u}^l_k\right]_i \equiv \sum_{j=1}^{m_i}\hat{z}^l_{ijk}\sqrt{\beta^l_{ij}}f'_{ij}(y^l_{ijk}) \tag{9}$$

we have,

$$\Delta\mathbf{W} = \left(\mathbf{I} - \frac{1}{N}\sum_{k=1}^{N}\mathbf{u}^l_k\mathbf{x}_k^T\mathbf{W}^{l\,T}\right)\mathbf{W}^l \tag{10}$$

5. ICA Mixture Model with Strongly Super-Gaussian Mixture Sources

We now consider the case in which the data is generated by a mixture model over a set of mixing matrices, $\mathbf{A}_h = \mathbf{W}_h^{-1}$, $h = 1,\ldots,M$,

$$p(\mathbf{x}_k;\theta) = \sum_{h=1}^{M}\gamma_h p_h(\mathbf{x}_k;\theta)\,,\;\; \gamma_h \ge 0,\; \sum_{h=1}^{M}\gamma_h = 1$$

The parameters to be estimated are,

$$\theta = \left\{\gamma_h, \mathbf{W}_h, \alpha_{hij}, \mu_{hij}, \beta_{hij}\right\},\; h = 1,\ldots,M,\; i = 1,\ldots,n_h,\; j = 1,\ldots,m_{hi}$$

The EM algorithm for the full mixture model is derived similarly to the case of source mixtures. In this model, each (independent) sample $\mathbf{x}_k$ is generated by drawing a mixture component h' from the discrete probability distribution $P[h' = h] = \gamma_h$, $1 \le h \le M$, then drawing $\mathbf{x}$ from $p_{h'}(\mathbf{x};\theta)$.

We define h_k to be the index chosen for the kth sample, and we define the random variable,

$$v_{hk} \equiv \begin{cases} 1, & h_k = h \\ 0, & \text{otherwise} \end{cases}$$

Let $\mathbf{V} \equiv \{v_{hk}\}$. We define j_{hik} to be the source mixture component index chosen (independently of h_k) for the ith source of the hth model in the kth sample, and we define the random variables z_{hijk} by,

$$z_{hijk} \equiv \begin{cases} 1, & j_{hik} = j \\ 0, & \text{otherwise} \end{cases}$$

with $\mathbf{Z} \equiv \{z_{hijk}\}$. Now, for the likelihood of θ, we can write,

$$p(\mathbf{X};\theta) = \sum_{\mathbf{V},\mathbf{Z}} \prod_{k=1}^{N} \prod_{h=1}^{M} \gamma_h^{v_{hk}} |\det \mathbf{W}_h|^{v_{hk}} \prod_{i=1}^{n_h} \prod_{j=1}^{m_{hi}} \left[\alpha_{hij} \sqrt{\beta_{hij}}\, p_{hij}\Big(\sqrt{\beta_{hij}}\big(\mathbf{w}_{hi}^T \mathbf{x}_k - \mu_{hij}\big)\Big)\right]^{v_{hk} z_{hijk}}$$

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. We now have,

$$F^l(\theta) = \sum_{k=1}^{N} \sum_{h=1}^{M} \left[\sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} E\big[v_{hk} z_{hijk} | \mathbf{x}_k;\theta^l\big] \Big(-\log \alpha_{hij} - \tfrac{1}{2} \log \beta_{hij} + f_{hij}\Big(\sqrt{\beta_{hij}}\big(\mathbf{w}_{hi}^T \mathbf{x}_k - \mu_{hij}\big)\Big)\Big)\right]$$
$$+\; E\big[v_{hk} | \mathbf{x}_k;\theta^l\big] \Big(-\log \gamma_h - \log |\det \mathbf{W}_h| \Big)$$

where we define $f_{hij} \equiv -\log p_{hij}$. We define y_{hijk}^l and ξ_{hijk}^l as in (3) and (6), and we define $\hat{z}_{hijk}^l$ to be the conditional expectation of z_{hijk},

$$\hat{z}_{hijk}^l \;\equiv\; E\big[z_{hijk} \,\big|\, v_{hk}=1, \mathbf{x}_k, ;\theta^l\big] \;=\; \frac{\alpha_{hij}^l \sqrt{\beta_{hij}^l}\, p_{hij}\big(y_{hijk}^l\big)}{\sum_{j'=1}^{m_{hi}} \alpha_{hij'}^l \sqrt{\beta_{hij'}^l}\, p_{hij'}\big(y_{hij'k}^l\big)} \tag{11}$$

The $\hat{v}_{hk}^l \equiv E[v_{hk}|\mathbf{x}_k;\theta^l]$ are given by,

$$\begin{aligned} \hat{v}_{hk}^l &= \frac{p(\mathbf{x}_k|v_{hk}=1;\theta^l) P[v_{hk}=1;\theta^l]}{\sum_{h'=1}^{M} p(\mathbf{x}_k|v_{h'k}=1;\theta^l) P[v_{h'k}=1;\theta^l]} \\ &= \frac{\gamma_h^l\, |\det \mathbf{W}_h^l| \prod_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \alpha_{hij}^l \sqrt{\beta_{hij}^l}\, p_{hij}\big(y_{hijk}^l\big)}{\sum_{h'=1}^{M} \gamma_{h'}^l\, |\det \mathbf{W}_{h'}^l| \prod_{i=1}^{n_{h'}} \sum_{j=1}^{m_{h'i}} \alpha_{h'ij}^l \sqrt{\beta_{h'ij}^l}\, p_{h'ij}\big(y_{h'ijk}^l\big)} \end{aligned}$$

Defining $\hat{r}_{hijk}^l \equiv E[v_{hk} z_{hijk} | \mathbf{x}_k;\theta^l]$, we have,

$$\begin{aligned} \hat{r}_{hijk}^l &= P\big[v_{hk}=1, z_{hijk}=1 \,|\, \mathbf{x}_k;\theta^l\big] \\ &= P\big[z_{hijk}=1 \,|\, v_{hk}=1, \mathbf{x}_k;\theta^l\big] P\big[v_{hk}=1 \,|\, \mathbf{x}_k;\theta^l\big] \\ &= \hat{z}_{hijk}^l \hat{v}_{hk}^l \end{aligned} \tag{12}$$

Minimizing F over γ_h and α_{hij}, we get,

$$\gamma_h^{l+1} = \frac{1}{N}\sum_{k=1}^{N} \hat{v}_{hk}^l\,, \quad \alpha_{hij}^{l+1} = \frac{1}{N\gamma_h^{l+1}}\sum_{k=1}^{N} \hat{r}_{hijk}^l \tag{13}$$

The remaining parameters are updated as before,

$$\mu_{hij}^{l+1} \;=\; \frac{\sum_{k=1}^N \hat{r}_{hijk}^l \xi_{hijk}^l \mathbf{w}_{hi}^{l\,T}\mathbf{x}_k}{\sum_{k=1}^N \hat{r}_{hijk}^l \xi_{hijk}^l} \;=\; \mu_{hij}^l + \frac{\sum_{k=1}^N \hat{r}_{hijk}^l f'_{hij}\big(y_{hijk}^l\big)}{\sqrt{\beta_{hij}^l}\sum_{k=1}^N \hat{r}_{hijk}^l \xi_{ijk}^l} \tag{14}$$

and,

$$\beta_{hij}^{l+1} \;=\; \frac{\sum_{k=1}^N \hat{r}_{hijk}^l}{\sum_{k=1}^N \hat{r}_{hijk}^l \xi_{hijk}^l \big(\mathbf{w}_{hi}^{l\,T}\mathbf{x}_k - \mu_{hij}^l\big)^2} \;=\; \frac{\beta_{hij}^l \sum_{k=1}^N \hat{r}_{hijk}^l}{\sum_{k=1}^N \hat{r}_{hijk}^l f'_{hij}\big(y_{hijk}^l\big) y_{hijk}^l} \tag{15}$$

Defining the vector $\mathbf{u}_{hk}^l$ such that,

$$\big[\mathbf{u}_{hk}^l\big]_i \equiv \sum_{j=1}^{m_{hi}} \hat{r}_{hijk}^l \sqrt{\beta_{hij}^l}\, f'_{hij}(y_{hijk}^l) \tag{16}$$

we have,

$$\Delta\mathbf{W}_h = \left(\gamma_h^{l+1}\mathbf{I} - \frac{1}{N}\sum_{k=1}^{N} \mathbf{u}_{hk}^l \mathbf{x}_k^T \mathbf{W}_h^{l\,T}\right)\mathbf{W}_h^l \tag{17}$$

If we make the definitions,

$$C_{hijk}^l \;\equiv\; \alpha_{hij}^l \sqrt{\beta_{hij}^l}\, p_{hij}\big(y_{hijk}^l\big), \quad L_{hk}^l \;\equiv\; \gamma_h^l \left|\det \mathbf{W}_h^l\right| \prod_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} C_{hijk}^l \tag{18}$$

then the $\hat{z}_{hijk}$ and $\hat{v}_{hk}$ updates become,

$$\hat{z}_{hijk}^l = \frac{C_{hijk}^l}{\sum_{j'=1}^{m_{hi}} C_{hij'k}^l}\,, \quad \hat{v}_{hk}^l = \frac{L_{hk}^l}{\sum_{h'=1}^{M} L_{h'k}^l} \tag{19}$$

The log likelihood of θ^l given $\mathbf{X}$, which we denote by $\bar{L}^l$, is calculated as,

$$\bar{L}^l \;=\; \sum_{k=1}^{N} \log\left(\sum_{h=1}^{M} L_{hk}^l\right) \tag{20}$$

$\bar{L}^l$ increases monotonically with iteration l.

6. Adaptive Strong Super-Gaussians

We can obtain further flexibility in the source model by adapting the mixture component densities within a parameterized family of strongly super-gaussian densities.

6.1 Generalized Gaussians with adaptive shape parameter, ρ

In this section we consider the case of Generalized Gaussian mixtures, with source mixture component densities,

$$p(s_{hij};\mu_{hij},\beta_{hij},\rho_{hij}) = \frac{\sqrt{\beta_{hij}}}{2\,\Gamma\Big(1+\frac{1}{\rho_{hij}}\Big)} \exp\Big(-\Big|\sqrt{\beta_{hij}}\big(s_{hij}-\mu_{hij}\big)\Big|^{\rho_{hij}}\Big)$$

The parameters ρ_{hij} are adapted by scaled gradient descent. The gradient of F with respect to ρ_{hij} is,

$$\frac{\partial F}{\partial \rho_{hij}} \;=\; \sum_{k=1}^{N} \hat{r}_{hijk}\Bigg[|y_{hijk}|^{\rho_{hij}}\log|y_{hijk}| - \frac{1}{\rho_{hij}^2}\Psi\Big(1+\frac{1}{\rho_{hij}}\Big)\Bigg]$$

We have found that scaling this by $\rho_{hij}^2/\Big(\Psi\Big(1+\frac{1}{\rho_{hij}}\Big)\sum_{k=1}^{N}\hat{r}_{hijk}\Big)$, which is positive for $0<\rho_{hij}\leq 2$, leads to faster convergence. The update then becomes,

$$\Delta\rho_{hij} \;=\; 1 \;-\; \frac{{\rho_{hij}^l}^2 \sum_{k=1}^{N}\hat{r}_{hijk}^l\,|y_{hijk}^l|^{\rho_{hij}^l}\log|y_{hijk}^l|}{\Psi\Big(1+\frac{1}{\rho_{hij}^l}\Big)\sum_{k=1}^{N}\hat{r}_{hijk}^l} \tag{21}$$

6.2 Student's t densities with adaptive degrees of freedom parameter, ν

$$p(s_{hij};\mu_{hij},\beta_{hij},\nu_{hij}) = \frac{\sqrt{\beta_{hij}}\;\Gamma\Big(\frac{\nu_{hij}+1}{2}\Big)}{\sqrt{\pi\nu_{hij}}\;\Gamma\Big(\frac{\nu_{hij}}{2}\Big)}\Bigg(1+\frac{\beta_{hij}}{\nu_{hij}}s_{hij}^2\Bigg)^{-\frac{\nu_{hij}+1}{2}}$$

The parameters ν_{hij} are adapted by scaled gradient descent. The gradient of F with respect to ν_{hij} is,

$$\frac{\partial F}{\partial \nu_{hij}} \;=\; \frac{1}{2}\sum_{k=1}^{N}\hat{r}_{hijk}\Bigg[\Psi\Big(\frac{\nu_{hij}}{2}\Big) - \Psi\Big(\frac{\nu_{hij}+1}{2}\Big) + \frac{\nu_{hij}+1}{\nu_{hij}+y_{hijk}^2} + \log\Big(1+\frac{y_{hijk}^2}{\nu_{hij}}\Big) - 1\Bigg]$$

Dividing this by $\frac{1}{2}\Big(1+\Psi\Big(\frac{\nu_{hij}^l+1}{2}\Big)-\Psi\Big(\frac{\nu_{hij}^l}{2}\Big)\Big)\sum_{k=1}^{N}\hat{r}_{hijk}$, which is positive for $\nu_{hij}>0$, the update becomes,

$$\Delta\nu_{hij} \;=\; 1 \;-\; \frac{\sum_{k=1}^{N}\hat{r}_{hijk}^l\Big[\frac{\nu_{hij}^l+1}{\nu_{hij}^l+{y_{hijk}^l}^2} + \log\Big(1+\frac{{y_{hijk}^l}^2}{\nu_{hij}^l}\Big)\Big]}{\Big(1+\Psi\Big(\frac{\nu_{hij}^l+1}{2}\Big)-\Psi\Big(\frac{\nu_{hij}^l}{2}\Big)\Big)\sum_{k=1}^{N}\hat{r}_{hijk}^l} \tag{22}$$

7. Experiments

7.1 Artificial data

We verified the convergence of the algorithm with synthetic data generated from Generalized Gaussian mixtures with randomly generated parameters. as well as on real EEG data. We show an example of a super-gaussian mixture that was learned by the adaptive Generalized

Gaussian mixture algorithm, including the shape parameter update. The shape parameters ρ_{hij} were initialized to 1.5, and the location and scale parameters were randomly initialized.

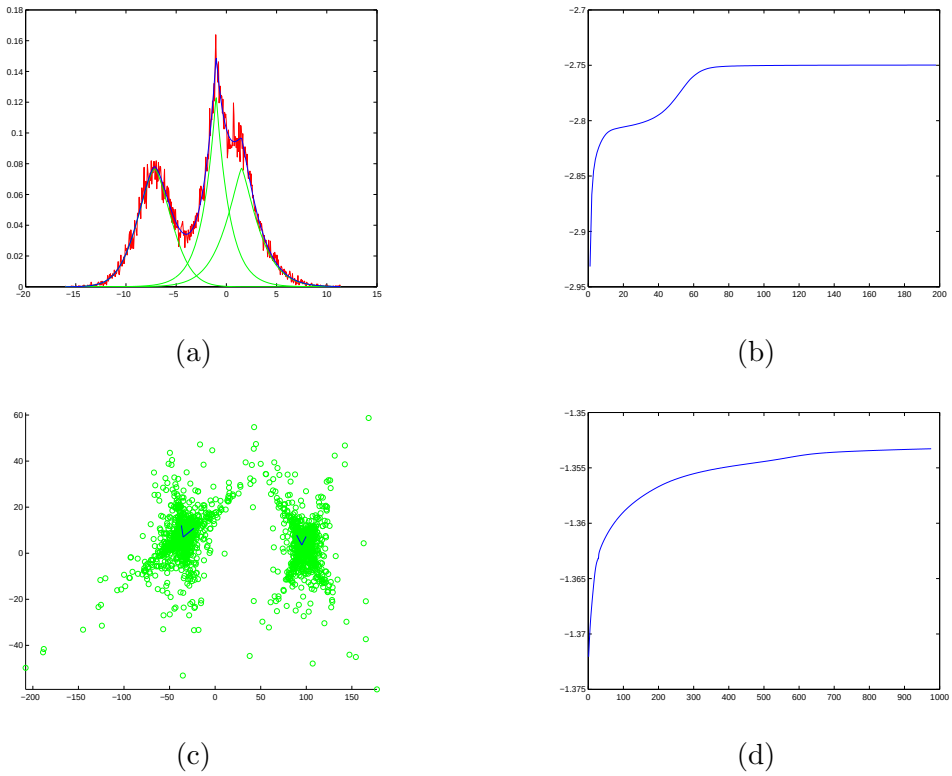


Figure 1: (a) Example of adaptive convergence of a artificial one-dimensional super-gaussian mixture. (b) Log likelihood for this run is seen to be monotonically increasing. (c) Example of full mixture model with two models, each having single component super-gaussian source densities. (d) Log likelihood for the full adaptive Generalized Gaussian model.

7.2 EEG data

We give an example of a super-gaussian mixture that was learned by the adaptive Generalized Gaussian mixture algorithm, including the shape parameter update, on a real EEG separation problem. The data consisted of 719 epochs (segments of length 750) of recordings from a 71 channel scalp electrode cap. Five mixture components per were used per source model. The Generalized Gaussian methods was employed, with shape parameters were initialized to 1.5, the location and scale parameters were randomly initialized. The data was sphered prior to running the algorithm, and the unmixing matrices were initialized to identity plus very low Gaussian noise.

References

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.

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

A. Benveniste, M. Goursat, and G. Ruget. Robust identification of a nonminimum phase system. *IEEE Transactions on Automatic Control*, 25(3):385–399, 1980.

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

R. A. Choudrey and S. J. Roberts. Variational mixture of Bayesian independent component analysers. *Neural Computation*, 15(1):213–252, 2002.

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.

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

M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13:2517–2532, 2001.

T. S. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, Massachusetts Institute of Technology, 1997.

T. S. Jaakkola and M. I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. In *Proceedings of the 1997 Conference on Artificial Intelligence and Statistics*, 1997.

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.

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*, 22(10):1078–1089, 2000.

D. J. C. Mackay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

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.

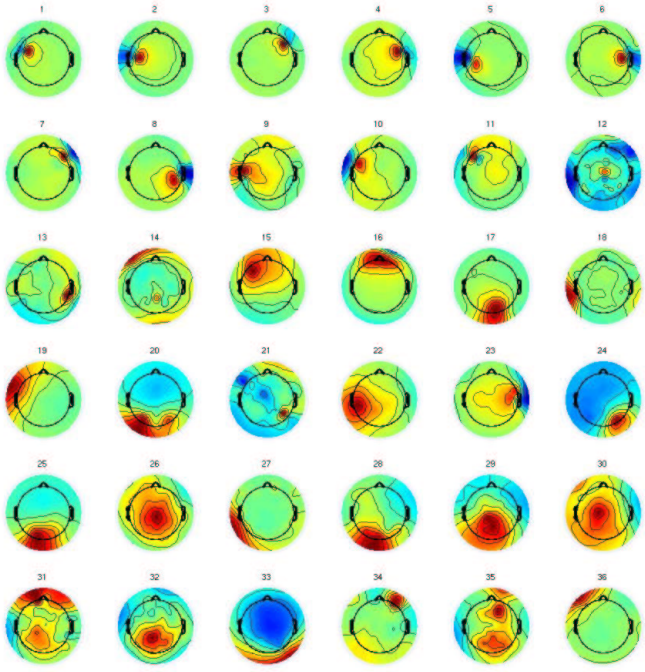
J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*. MIT Press, 2005. Available at http://dsp.ucsd.edu/∼japalmer/.

H.-J. Park and T.-W. Lee. Modeling nonlinear dependencies in natural images using mixture of Laplacian distribution. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2004. MIT Press.

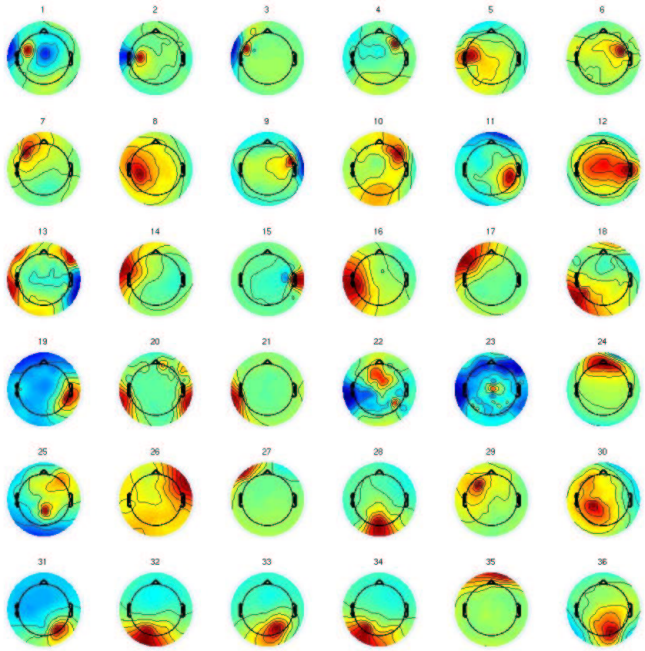
B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In M. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*. MIT Press, 1996.

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.

M. E. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.



(a)



(b)

Figure 2: Estimated dipolar scalp maps for the two models learned from EEG data. The components are arranged in order of maximum mutual information between the component activation y_{hik} and the probability signal v_{hk}. For two models, $h = 1, 2$, we have $v_{1k} = 1 - v_{2k}$, so there is really only one model probability signal. The models seem to be distinct, with model (a) favoring what are apparently muscle components.

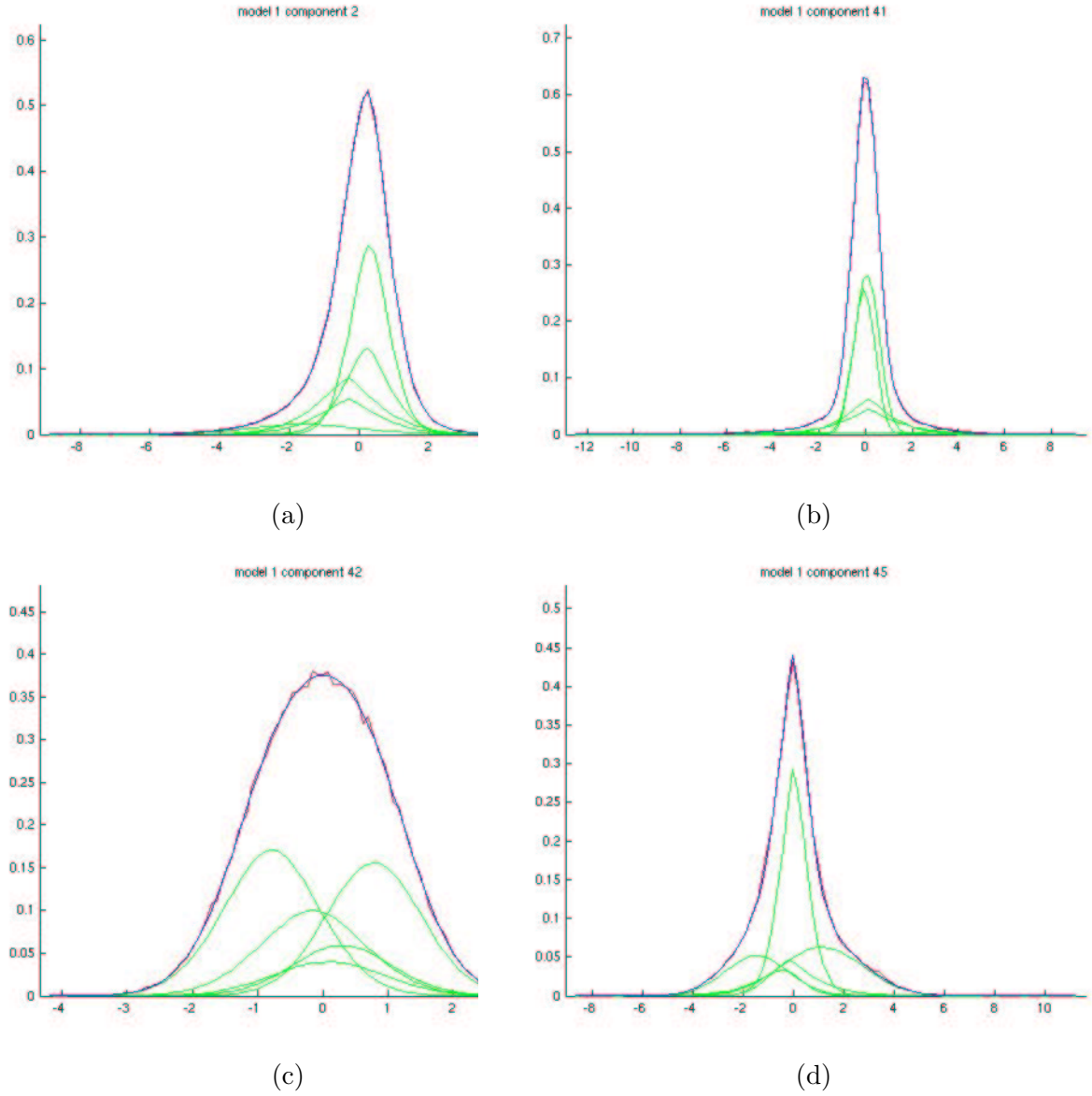


Figure 3: Converged source distributions for some components, showing (a) skewed, (b) heavy-tailed (c) sub-gaussian, and (d) sharply peaked densities. The model density is plotted over the empirical histogram showing exact agreement, i.e. zero Kullback-Leibler divergence.

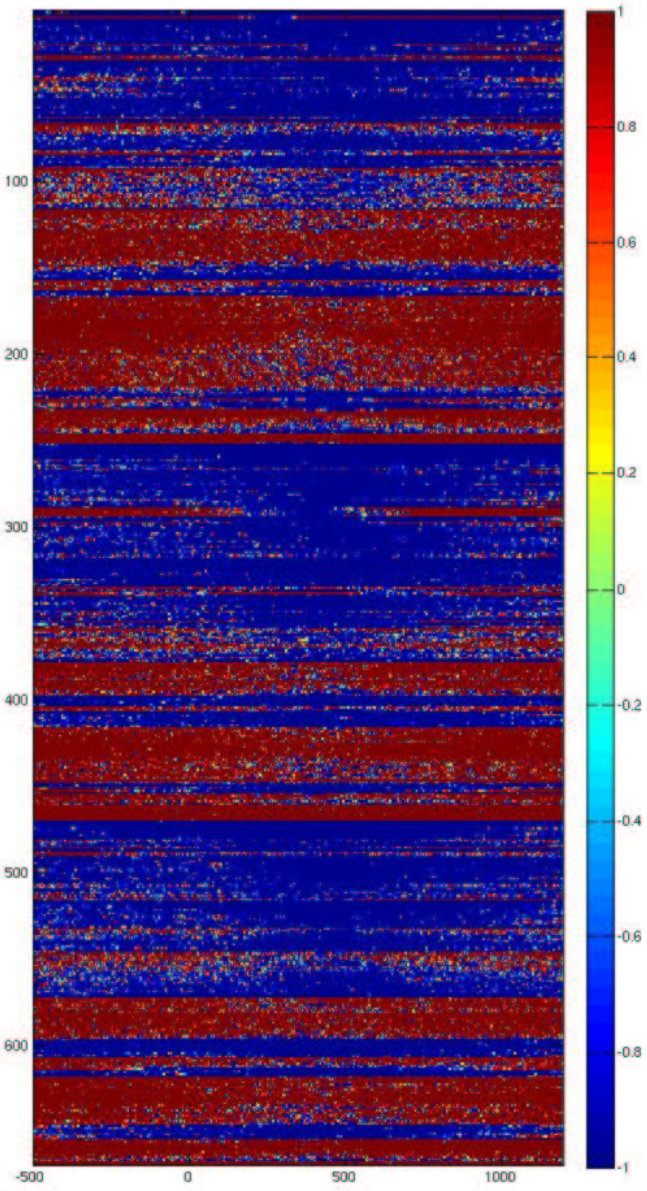


Figure 4: Difference between model probabilities. Horizontal axis is time, and vertical axis is trial number. Stimulus onset occurs at time 0. The trials seem to be divided into distinct groups, with probability very close to 1 for one or the other model. An event related potential is detectable after the stimulus onset, but the overall preference of the trials for one model or the other seems to be independent of the stimulus. The trials are arranged in temporal order, and preference for one model or the other also seems to persist over many trials.