NEWTON METHOD FOR THE ICA MIXTURE MODEL

J. A. Palmer[1]*, S. Makeig*[1]*, K. Kreutz-Delgado*[2]*, and B. D. Rao*[2]

[1]Swartz Center for Computational Neuroscience
[2]Department of Electrical and Computer Engineering
University of California San Diego, La Jolla, CA 92093

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 yields an algorithm that can accommodate non-stationary environments and arbitrary source densities. We prove asymptotic stability when the source models match the true sources. An example application to EEG segmentation is given.

Index Terms— Independent Component Analysis, Bayesian linear mixture model, Newton method, EEG signal analysis

1. INTRODUCTION

Linear representations are useful in a variety of signal processing applications, including compression, detection, transmission, and others. In non-stationary environments, a single complete basis may be not be sufficient to represent the signal at all times. Overcomplete representations overcome this limitation of complete basis sets, but they are computationally inefficient for large scale sensor arrays such as those used in Electro-encephalography (EEG), requiring iterative nonlinear optimization to estimate the coefficients in the representation, given the observed linear combination.

ICA mixture models [1, 2] offer a useful compromise between the efficiency of (conditional) invertibility of the model, and the need for richer representations in non-stationary environments. However, while feasible to optimize, the standard gradient and natural or relative gradient [3, 4] formulations still require many thousands of 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 [5] 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 [3]. 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 [1, 2, 6] and adaptive mixture sources [7].

2. ICA MIXTURE MODEL

Our starting point is 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 $\mathbf{A}_h$ and "center" $\mathbf{c}_h$, and a random vector of independent sources $\mathbf{s}(t) \sim q_h(\mathbf{s}) = \prod_{i=1}^n q_{hi}(s_i)$, such that,

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

with $h = h_t$. We shall assume that only one of the models 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 (MGSMs), as in [7, 8],

$$q_{hi}\big(s_i(t)\big) = \sum_{j=1}^m \alpha_{hij}\beta_{hij}^{1/2} \, q_{hij}\big(\beta_{hij}^{1/2}\big(s_i(t) - \mu_{hij}\big)\big)$$

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 $j = 1, \ldots, m$.

2.1. Invariances in the model

Invariance, or redundancy, exists in the model in two areas. The first involves 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

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

for all Δ_{hi}. Putting $\Delta_{hi} = \sum_j \alpha_{hij}\mu_{hij}$ makes the sources $\mathbf{s}(t)$ zero mean for each model. The zero mean assumption is used in the calculation of the expected Hessian for the Newton algorithm.

There is also redundancy in the row norms of $\mathbf{W}_h$ and the scale of the source densities. Specifically, $p(\mathbf{X}; \Theta) = p(\mathbf{X}; \Theta')$, where

$\Theta = \{\mathbf{W}_h, \mathbf{c}_h, \mu_{hij}, \beta_{hij}, \ldots\}$, $\Theta' = \{\mathbf{W}'_h, \mathbf{c}'_h, \mu'_{hij}, \beta'_{hij}, \ldots\}$, if for any $\tau_{hi} > 0$,

$$[\mathbf{W}_h]'_{i:} = [\mathbf{W}_h]_{i:}/\tau_{hi}, \;\; [\mathbf{c}_h]'_i = [\mathbf{c}_h]_i/\tau_{hi},$$
$$\mu'_{hij} = \mu_{hij}/\tau_{hi}, \;\; \beta'_{hij} = \beta_{hij}\tau^2_{hi}, \;\; j = 1, \ldots, m$$

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 $\mathbf{c}_h$, since the model centers are redundant given the source means. The reparameterization is carried out after the other parameters have been updated (by EM, Newton, or scaled gradient descent).

3. 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. For the negative log likelihood of the data then (which is to be minimized), we have,

$$L(\mathbf{W}) \;=\; \sum_{t=1}^{N} -\log|\det \mathbf{W}| - \sum_{i=1}^{n} \log q_i(y_{it}) \qquad (1)$$

The gradient of this function is proportional to,

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

where we define,

$$f_i(y_i) \triangleq -\log q_i(y_i)$$

and $f(\mathbf{y}) \triangleq \sum_i f_i(y_i)$.

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

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

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 $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^T)$, we have for arbitrary $\mathbf{V} \in \mathbb{R}^{n\times n}$,

$$\Big\langle \mathbf{V}, \mathbf{V}\mathbf{W}\mathbf{W}^T \Big\rangle = \Big\langle \mathbf{V}\mathbf{W}, \mathbf{V}\mathbf{W} \Big\rangle > 0 \qquad (4)$$

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

3.1. 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,

$$\frac{\partial g_{ij}}{\partial w_{kl}} = [\mathbf{W}^{-1}]_{li}[\mathbf{W}^{-1}]_{jk} + \Big\langle f_i''(\mathbf{w}_k^T\mathbf{x})x_j x_l \delta_{ik} \Big\rangle_N$$

where $\mathbf{w}_k^T$ is the kth row of $\mathbf{W}$, and δ_{ik} is the Kronecker delta symbol. 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,

$$c_{ij} \;=\; \sum_k\sum_l [\mathbf{W}^{-1}]_{li}[\mathbf{W}^{-1}]_{jk} b_{kl} + \Big\langle f_i''(y_i)\, x_j \sum_l b_{il}x_l \Big\rangle_N$$

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

$$\mathbf{C} = \mathcal{H}(\mathbf{B}) = \mathbf{W}^{-T}\mathbf{B}^T\mathbf{W}^{-T} + \frac{1}{N}\sum_{t=1}^{N} \mathrm{diag}(f''(\mathbf{y}_t))\mathbf{B}\mathbf{x}_t\mathbf{x}_t^T \qquad (5)$$

where $\mathrm{diag}(f''(\mathbf{y}_t))$ is the diagonal matrix with diagonal elements $f_i''(y_{it})$. The asymptotic stability of the algorithm is determined by the positivity of the eigenvalues of the expected value of this transformation evaluated at the optimum [5]. Assuming that the model holds, the source estimates at the optimal $\mathbf{W}$ will be independent. We also assume that the (conditional) mean of the data has been removed, so that the sources are (conditionally) zero mean as well.

It will be easier to calculate the expected 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,

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

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,

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

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

$$\tilde{c}_{ii} \;\to\; \tilde{b}_{ii} + E\Big\{f_i''(y_i)\sum_k \tilde{b}_{ik}y_k y_i\Big\} = \tilde{b}_{ii}(1+\eta_i) \qquad (7)$$

where we define $\eta_i \triangleq E\{f''(y_i)y_i^2\}$. The cross terms drop out since the expected value of $\alpha_i y_i y_k$ is zero for $k \neq i$ by the independence and zero mean assumption on the sources. Now we note [4, 5] that the off-diagonal elements of the equation (6) can be paired as follows,

$$\tilde{c}_{ij} \;\to\; \tilde{b}_{ji} + E\Big\{f_i''(y_i)\sum_k \tilde{b}_{ik}y_k y_j\Big\} = \tilde{b}_{ji} + \kappa_i\sigma_j^2\tilde{b}_{ij}$$
$$\tilde{c}_{ji} \;\to\; \tilde{b}_{ij} + E\Big\{f_j''(y_j)\sum_k \tilde{b}_{jk}y_k y_i\Big\} = \tilde{b}_{ij} + \kappa_j\sigma_i^2\tilde{b}_{ji}$$

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,

$$\begin{bmatrix}\tilde{c}_{ij}\\ \tilde{c}_{ji}\end{bmatrix} = \begin{bmatrix}\kappa_i\sigma_j^2 & 1\\ 1 & \kappa_j\sigma_i^2\end{bmatrix}\begin{bmatrix}\tilde{b}_{ij}\\ \tilde{b}_{ji}\end{bmatrix} \qquad (8)$$

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,

$$\mathbf{C} = \mathcal{H}(\mathbf{B}) = \tilde{\mathcal{H}}\Big(\mathbf{B}\mathbf{W}^{-1}\Big)\mathbf{W}^{-T}$$

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 [5] as,

1. $1 + \eta_i > 0, \ \forall\, i$
2. $\kappa_i > 0, \ \forall\, i$, and,
3. $\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1 > 0, \ \ \forall\, i \neq j$

3.2. Asymptotic stability

Using integration by parts, it can be shown that the stability conditions are always satisfied when $f(y) = -\log p(y)$ matches the true source density. The only regularity condition imposed is that $p'(y) = o(1/y^2)$. This must be the case for non-pathological, integrable $p(y)$, since otherwise we would have $p(y) = O(1/y)$ and non-integrable. Specifically, we have the following.

Theorem 1. *If $f_i(y_i) \triangleq -\log q_i(y_i) = -\log p_i(y_i)$, $i = 1, \ldots, n$, i.e. the source density models match the true source densities, and $p_i'(y) = o(1/y^2)$, $i = 1, \ldots, n$, 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

$$-\ y^2 p'(y)\big|_{-\infty}^{\infty} \ - \ \int_{-\infty}^{\infty} f'(y)\big(2y \ - \ y^2 f'(y)\big)\, p(y)\, dy \quad (9)$$

The first term in (9) is zero if $p'(y) = o(1/y^2)$ as $y \to \pm\infty$. 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 $p'(y)$ tends to 0 as $y \to \pm\infty$ by assumption, 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,

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

by the Cauchy Schwartz inequality, with equality only for $f(y) = y$, 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 $p_i(y)$ and $p_j(y)$ is nongaussian. □

3.3. Newton method

The inverse of the Hessian operator will be given by,

$$\mathbf{B} = \mathcal{H}^{-1}(\mathbf{C}) = \tilde{\mathcal{H}}^{-1}\Big(\mathbf{C}\mathbf{W}^T\Big)\mathbf{W} \quad (10)$$

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

$$\tilde{b}_{ii} \ = \ \frac{\tilde{c}_{ii}}{1 + \eta_i}, \ \forall\, i \quad (11)$$

$$\tilde{b}_{ij} \ = \ \frac{\kappa_j \sigma_i^2 \tilde{c}_{ij} - \tilde{c}_{ji}}{\kappa_i \kappa_j \sigma_i^2 \sigma_j^2 - 1}, \ \forall\, i \neq j \quad (12)$$

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

$$\Delta\mathbf{W} = -\tilde{\mathcal{H}}^{-1}\Big(\mathbf{G}\mathbf{W}^T\Big)\mathbf{W} \quad (13)$$

4. EM PARAMETER UPDATES

Define h_t to be the index of the model producing observation $\mathbf{x}(t)$, and define the random variable v_{ht} to equal 1 if $h_t = h$, and 0 otherwise. Define j_{it} to be the source mixture component index chosen (independently of h_t) for the ith source of the h_tth model, and define z_{hijt} to equal 1 if $j_{it} = j$, and 0 otherwise. We make the definitions,

$$y_{hijt}^l \ \triangleq \ \sqrt{\beta_{hij}^l}\,\Big(\big[\mathbf{W}_h^l \mathbf{x}_t - \mathbf{c}_h\big]_i - \mu_{hij}^l\Big) \quad (14)$$

$$Q_{hijt}^l \ \triangleq \ \alpha_{hij}^l \sqrt{\beta_{hij}^l}\, q_{hij}\big(y_{hijt}^l\big) \quad (15)$$

$$L_{ht}^l \ \triangleq \ \gamma_h^l \left|\det \mathbf{W}_h^l\right| \prod_{i=1}^{n} \sum_{j=1}^{m} Q_{hijt}^l \quad (16)$$

The expectations $\hat{z}_{hijt}$ and $\hat{v}_{ht}$ are given by,

$$\hat{z}_{hijt}^l = \frac{Q_{hijt}^l}{\sum_{j'=1}^{m} Q_{hij't}^l}, \quad \hat{v}_{ht}^l = \frac{L_{ht}^l}{\sum_{h'=1}^{M} L_{h't}^l} \quad (17)$$

We define $\hat{r}_{hijt}^l \triangleq \hat{v}_{ht}^l \hat{z}_{hijt}^l$. Optimizing the mixing coefficient parameters in the EM algorithm, we get

$$\gamma_h^{l+1} = \frac{1}{N} \sum_{t=1}^{N} \hat{v}_{ht}^l, \quad \alpha_{hij}^{l+1} = \frac{1}{N \gamma_h^{l+1}} \sum_{t=1}^{N} \hat{r}_{hijt}^l \quad (18)$$

The source density location parameters are updated by,

$$\mu_{hij}^{l+1} = \mu_{hij}^l + \frac{\sum_{t=1}^{N} \hat{r}_{hijt}^l f_{hij}'\big(y_{hijt}^l\big)}{\sqrt{\beta_{hij}^l} \sum_{t=1}^{N} \hat{r}_{hijt}^l \xi_{hijt}^l} \quad (19)$$

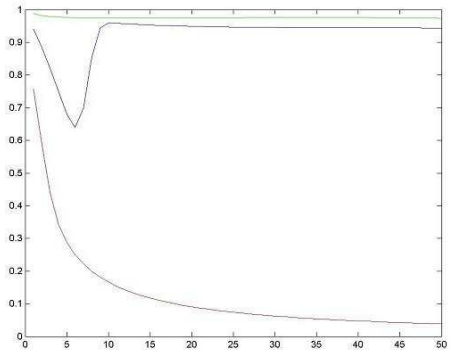


Fig. 1. Newton convergence rate versus gradient and natural gradient in a simulation with a 10×10 mixing matrix with Laplacian sources. Ordinary gradient (top line) and natural gradient (middle line) are linearly convergent with high asymptotic rate, while Newton method (bottom line) is tending toward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

where $\xi^l_{hijt} \triangleq f'_{hij}(y^l_{hijt})/y^l_{hijt}$ (see [7]). The scale parameters are updated by,

$$\beta^{l+1}_{hij} = \frac{\beta^l_{hij} \sum_{t=1}^N \hat{r}^l_{hijt}}{\sum_{t=1}^N \hat{r}^l_{hijt} f'_{hij}(y^l_{hijt}) y^l_{hijt}} \tag{20}$$

The vector $\mathbf{g}^l_{ht} \triangleq \nabla f_{hi}(\mathbf{y}^l_t)$ used in the matrix gradient $\mathbf{G}$ given in (2), is,

$$\left[\mathbf{g}^l_{ht}\right]_i = \sum_{j=1}^m \hat{r}^l_{hijk} \sqrt{\beta^l_{hij}}\, f'_{hij}(y^l_{hijk}) \tag{21}$$

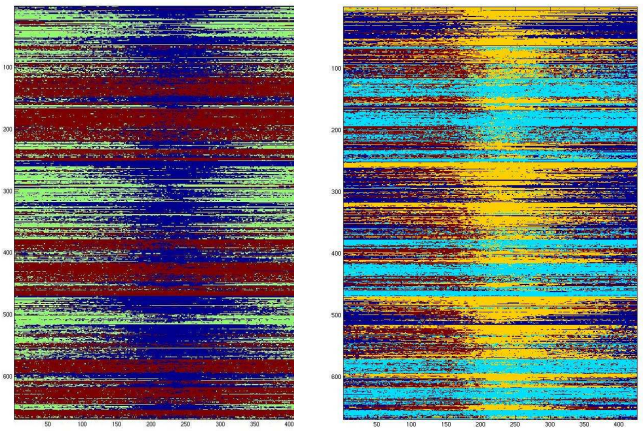
The unmixing matrices are updated according to the Newton method derived in §3.3, with conditional time averages substituted for ensemble averages. Integration by parts is used to write expectations of second derivatives as expectations of squares of first derivatives. The log likelihood of Θ^l given $\mathbf{X}$ is calculated as,

$$L(\Theta^l|\mathbf{X}) \;=\; \sum_{t=1}^N \log\left(\sum_{h=1}^M L^l_{ht}\right) \tag{22}$$

5. 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 [9]. For superlinear algorithms, this ratio tends to zero, and the order of convergence q is the power of the denominator which yields a constant 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 2 shows the result of segmenting an experiment according to the most likely model given the data (MAP). The subject is shown a sequence of letters are indicates whether current letter is the same as letter before last. At $t = 175$ there is feedback as to whether the response was correct or incorrect. Muscle activity (horizontal spanning lines) as well as post-feedback theta activity are segmented. The trials are stacked vertically. Time points are plotted in the color of the model most likely for that point. Three and four models are used. There appears to be consistency in the segmentation, with increased resolution in the four model segmentation.



(a) (b)

Fig. 2. Segmentation of EEG trials: (a) three models (b) four models.

6. REFERENCES

[1] 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.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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*. 2006, MIT Press.

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