AN EM ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATION OF BARNDORFF-NIELSEN'S GENERALIZED HYPERBOLIC DISTRIBUTION

Jason A. Palmer[1], *Ken Kreutz-Delgado*[2], *and Scott Makeig*[1]

[1]Swartz Center for Computational Neuroscience
Institute for Neural Computation
[2]Dept. of Electrical and Computer Engineering
University of California San Diego

ABSTRACT

We present an EM algorithm for Maximum Likelihood (ML) estimation of the location, structure matrix, skew or drift, and shape parameters of Barndorff-Nielsen's Generalized Hyperbolic distribution, which is the Gaussian Location Scale mixture (or Normal Variance Mean Mixture) with Generalized Inverse Gaussian (GIG) scale mixing distribution. We use the GLSM representation along with the closed form posterior expectations possible with the GIG distribution to derive an EM algorithm for computing ML parameter estimates.

Index Terms— Generalized Inverse Gaussian, Gaussian Location-Scale Mixtures, multivariate, non-gaussian, non-ellipsoidal, quasi-parametric density estimation

1. INTRODUCTION

The Gaussian mixture model and its EM parameter estimation are well-known. Less well known, is the use of the EM algorithm for parameter estimation in Gaussian Scale Mixture (GSM) and Gaussian Location Scale Mixture (GLSM) [1], though the former was indeed described by the proposers of the EM algorithm [2]. The GLSM generative model takes the form,

$$\mathbf{s} = \xi^{1/2}\mathbf{z} + \xi\boldsymbol{\theta} + \boldsymbol{\mu} \tag{1}$$

where ξ is a non-negative scalar random variable, $\mathbf{z}$ is Gaussian, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\theta}, \boldsymbol{\mu} \in \mathbb{R}^n$. If $\boldsymbol{\theta} = \mathbf{0}$, then the density of $\mathbf{s}$ is called simply a Gaussian Scale Mixture (GSM). A particularly flexible yet tractable scale mixing distribution model is the Generalized Inverse Gaussian (GIG) density [3, §9.3],[1], which has the form,

$$p(\xi; \lambda, \delta, \kappa) = \frac{(\kappa/\delta)^{\lambda}}{2K_{\lambda}(\delta\kappa)}\xi^{\lambda-1}\exp(-\tfrac{1}{2}\delta^2\xi^{-1} - \tfrac{1}{2}\kappa^2\xi) \tag{2}$$

where K_λ is the Bessel K function (modified Bessel function of second kind) [4], and $\delta, \kappa > 0$. Limiting cases are: the ordinary Gamma distribution for $\lambda > 0, \delta \to 0$, and the Inverse Gamma distribution for $\lambda < 0, \kappa \to 0$. The GIG density has the very convenient features that all of its moments can be computed, and its conjugate posterior density is also GIG, with computable moments. Most of the derivations amount to application of the fact that,

$$\int_0^\infty \xi^{\lambda-1}\exp(-\tfrac{1}{2}\delta^2\xi^{-1} - \tfrac{1}{2}\kappa^2\xi)\,d\xi = 2\,(\delta/\kappa)^{\lambda}K_{\lambda}(\delta\kappa) \tag{3}$$

As an example we have the rth moment computation, for any r, $\xi \sim \mathrm{GIG}(\lambda, \delta^2, \kappa^2)$,

$$E\{\xi^r\} = \left(\frac{\delta}{\kappa}\right)^r \frac{K_{\lambda+r}(\delta\kappa)}{K_{\lambda}(\delta\kappa)} \tag{4}$$

Now, if we consider the density of the random vector defined in (1), we have, $p(\mathbf{s}) = \int \mathcal{N}(\boldsymbol{\mu} + \xi\boldsymbol{\theta}, \xi\boldsymbol{\Sigma})p(\xi; \lambda, \delta, \kappa)d\xi$,

$$\begin{aligned} p(\mathbf{s}) \;&=\; (2\pi)^{-n/2}|\det\boldsymbol{\Sigma}|^{-1/2}\exp\big(\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{s}-\boldsymbol{\mu})\big) \times \\ &\int_0^\infty \xi^{-n/2}\exp\big(-\tfrac{1}{2}\xi^{-1}\|\mathbf{s}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} - \tfrac{1}{2}\xi\,\|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}\big)p(\xi)\,d\xi \end{aligned}$$

Using (3), we see that the general form of Barndorff-Nielsen's Generalized Hyperbolic distribution $p(\mathbf{s}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}, \lambda, \delta, \kappa)$ is given by,

$$\begin{aligned} p(\mathbf{s}) \;&=\; (2\pi)^{-n/2}|\det\boldsymbol{\Sigma}|^{-1/2}\exp\big(\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{s}-\boldsymbol{\mu})\big) \\ &\times\;\; \frac{(\kappa/\delta)^{\lambda}}{K_{\lambda}(\delta\kappa)}\left(\frac{\sqrt{\delta^2 + \|\mathbf{s}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}}}{\sqrt{\kappa^2 + \|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}}}\right)^{\lambda-n/2} \\ &\times\;\; K_{\lambda-n/2}\Big(\sqrt{\delta^2 + \|\mathbf{s}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}}\sqrt{\kappa^2 + \|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}}\Big) \end{aligned} \tag{5}$$

To eliminate a scale indeterminacy, as in [1], we define the density with the constraint $|\det\boldsymbol{\Sigma}| = 1$.

The GLSM (or NVMM) associated with the GIG scale mixing density was described by Barndorff-Nielsen [1]. Barndorff-Nielsen et al [1] do not seem to have proposed the EM algorithm or any other particular algorithm for parameter estimation. Given the widely applicable flexibility of the model, with controlled and limited degrees of freedom, and the very simple and convenient EM updates provided by Gaussian location scale mixture model representation, the EM algorithm for the Generalized Hyperbolic model is likely to be of considerable general interest, the despite the initially daunting appearance of the form of the multivariate density itself.

In §3 we derive the complete log likelihood associated with the EM algorithm, along with posterior mixing variable expectations. In §4 we derive the parameter updates, first considering the straightforward location and scale updates μ and σ, and then deriving a variational bound related to the θ objective leading to closed form (constrained) maximum. In §5 we derive the updates arising when this model is embedded in a finite mixture model EM framework, with the further constraint the overall signal mean be zero, which is important for simplifying second order ICA algorithms.

This research was partially funded by the Swartz Foundation.

2. COMPLETE LOG LIKELIHOOD AND POSTERIOR MOMENTS

The complete log likelihood $\log\prod_t p(\mathbf{s}_t|\xi_t)p(\xi_t)$ of N i.i.d. samples, in terms of the location, scale, and drift parameters is, neglecting a constant term,

$$\begin{aligned}\sum_{t=1}^{N} \big(\zeta - \tfrac{1}{2}\big)\log|\boldsymbol{\Sigma}| + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{s}_t-\boldsymbol{\mu}) - \tfrac{1}{2}\,\xi_t\big(\kappa^2+\|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}\big)\\ - \tfrac{1}{2}\,\xi_t^{-1}\big(\delta^2+\|\mathbf{s}_t-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\big) + \lambda\log(\xi_t\kappa/\delta) - \log K_\lambda(\delta\kappa)\end{aligned} \tag{6}$$

where ζ is the Lagrange multiplier for the constraint $\log|\boldsymbol{\Sigma}| = 0$. Note that we need to compute the posterior expectations of ξ_t^{-1}, ξ_t, and for the λ update, $\log\xi_t$. The posterior distribution of ξ_t given $\mathbf{s}_t$ is GIG($\lambda - n/2$, $\delta^2 + \|\mathbf{s}_t - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}$, $\kappa^2 + \|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}$). Thus using (4), and defining $\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \sqrt{\delta^2 + \|\mathbf{s}_t-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}}$, and $\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma}) \triangleq \sqrt{\kappa^2+\|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}^{-1}}}$,

$$E\{\xi_t|\mathbf{s}_t\} = \frac{\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})}\,\frac{K_{\lambda-n/2+1}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}{K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}$$

$$E\{\xi_t^{-1}|\mathbf{s}_t\} = \frac{\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})}{\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})}\,\frac{K_{\lambda-n/2-1}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}{K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}$$

And using the limiting relation $\log\xi = \lim_{a\to 0}(\xi^a - 1)/a$, we find,

$$\begin{aligned}E\{\log\xi_t|\mathbf{s}_t\} &= \frac{\frac{d}{d\lambda}K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}{K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}\\ &\approx \epsilon^{-1}\left(\left(\frac{\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})}\right)^{\epsilon}\frac{K_{\lambda-n/2+\epsilon}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)}{K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)} - 1\right)\end{aligned}$$

In practice we will use the latter finite difference formula with $\epsilon = 10^{-6}$ since the derivative of the Bessel function with respect to order is not generally available for practical computation.

Define $\nu_t \triangleq E\{\xi_t^{-1}|\mathbf{s}_t\}$, $\gamma_t \triangleq E\{\xi_t|\mathbf{s}_t\}$, and $\eta_t \triangleq E\{\log\xi_t|\mathbf{s}_t\}$, and their sample averages $\hat{\nu} \triangleq N^{-1}\sum_{t=1}^{N}\nu_t$, $\hat{\gamma} \triangleq N^{-1}\sum_{t=1}^{N}\gamma_t$, and $\hat{\eta} \triangleq N^{-1}\sum_{t=1}^{N}\eta_t$. Let $\mathbf{m} \triangleq N^{-1}\sum_{t=1}^{N}\mathbf{s}_t$. Then the complete log likelihood (6) scaled by N^{-1} can be written,

$$\begin{aligned}\big(\zeta-\tfrac{1}{2}\big)\log|\boldsymbol{\Sigma}| + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) - \tfrac{1}{2}\hat{\gamma}\,\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - \tfrac{1}{2}\hat{\gamma}\kappa^2 - \tfrac{1}{2}\hat{\nu}\delta^2\\ + \hat{\eta}\lambda + \lambda\log(\kappa/\delta) - \log K_\lambda(\delta\kappa) - \frac{1}{N}\sum_{t=1}^{N}\tfrac{1}{2}\,\nu_t\,\|\mathbf{s}_t-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\end{aligned} \tag{7}$$

Alternating the evaluation of the expectations $\hat{\gamma}$, $\hat{\nu}$, etc., with the maximization of this complete log likelihood with respect to the parameters, allows us ultimately to maximize true log likelihood, $N^{-1}\log\prod_{t=1}^{N}p(\mathbf{s}_t)$,

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N}\log p(\mathbf{s}_t) = -\tfrac{1}{2}n\log(2\pi) + \lambda\log(\kappa/\delta) - \log K_\lambda(\delta\kappa)\\ + \boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) + \frac{1}{N}\sum_{t=1}^{N}\left((\lambda-n/2)\log\left(\frac{\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})}{\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})}\right)\right.\\ \left. + \;\log K_{\lambda-n/2}\big(\delta(\mathbf{s}_t,\boldsymbol{\mu},\boldsymbol{\Sigma})\kappa(\boldsymbol{\theta},\boldsymbol{\Sigma})\big)\right)\end{aligned} \tag{8}$$

3. CLOSED FORM PARAMETER UPDATES

We derive closed form formulae for all parameter updates yielding parameters with monotonically increasing likelihood with no step sizes needed to be set.

3.1. Updates for location and drift: $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$

We define the weighted mean,

$$\mathbf{m}_\nu \triangleq \frac{1}{\hat{\nu}N}\sum_{t=1}^{N}\nu_t\mathbf{s}_t \tag{9}$$

The optimal $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ satisfy,

$$\begin{array}{ccccc}\hat{\nu}\,\boldsymbol{\mu} & + & \boldsymbol{\theta} & = & \hat{\nu}\,\mathbf{m}_\nu\\ \boldsymbol{\mu} & + & \hat{\gamma}\,\boldsymbol{\theta} & = & \mathbf{m}\end{array}$$

leading to the updates,

$$\boldsymbol{\mu} = (1-\hat{\nu}\,\hat{\gamma})^{-1}(\mathbf{m} - \hat{\gamma}\,\hat{\nu}\,\mathbf{m}_\nu) \tag{10}$$

$$\boldsymbol{\theta} = (1-\hat{\nu}\,\hat{\gamma})^{-1}\hat{\nu}\,(\mathbf{m}_\nu - \mathbf{m}) \tag{11}$$

3.2. Update for structure matrix: $\boldsymbol{\Sigma}$

We also define the weighted covariance,

$$\boldsymbol{\Sigma}_\nu \triangleq \frac{1}{\hat{\nu}N}\sum_{t=1}^{N}\nu_t(\mathbf{s}_t-\boldsymbol{\mu})(\mathbf{s}_t-\boldsymbol{\mu})^T \tag{12}$$

Taking the gradient of the complete log likelihood with respect to the symmetric matrix $\boldsymbol{\Sigma}$, we get $\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{B}\boldsymbol{\Sigma}^{-1}$, where,

$$\mathbf{B} = (2\zeta-1)\boldsymbol{\Sigma} - (\mathbf{m}-\boldsymbol{\mu})\boldsymbol{\theta}^T - \boldsymbol{\theta}(\mathbf{m}-\boldsymbol{\mu})^T + \hat{\gamma}\,\boldsymbol{\theta}\boldsymbol{\theta}^T + \hat{\nu}\,\boldsymbol{\Sigma}_\nu$$

At a stationary point then, $\boldsymbol{\Sigma}$ satisfies,

$$(1-2\zeta)\boldsymbol{\Sigma} = \hat{\nu}\,\boldsymbol{\Sigma}_\nu - (\mathbf{m}-\boldsymbol{\mu})\boldsymbol{\theta}^T - \boldsymbol{\theta}(\mathbf{m}-\boldsymbol{\mu})^T + \hat{\gamma}\,\boldsymbol{\theta}\boldsymbol{\theta}^T \tag{13}$$

with $|\det\boldsymbol{\Sigma}| = 1$. Satisfaction of the constraint thus amounts to a simple determinant normalization of the update.

3.3. Shape parameter updates: δ and κ

We use the following variational representation of the log Bessel function,

$$-\log K_\lambda(t) = \sup_v\; v\log t - h(v)$$

where h is a relative conjugate function whose explicit form is not needed. For the optimal v, using the fact that $\frac{d}{dx}\log K_\lambda(x) = \lambda/x - K_{\lambda+1}(x)/K_\lambda(x)$, we have,

$$\tilde{v} = \delta\kappa\frac{K_{\lambda+1}(\delta\kappa)}{K_\lambda(\delta\kappa)} - \lambda \tag{14}$$

and the following surrogate cost function for δ and κ,

$$-\tfrac{1}{2}\hat{\gamma}\kappa^2 - \tfrac{1}{2}\hat{\nu}\delta^2 \;-\; \lambda\log(\delta/\kappa) + \tilde{v}\log(\delta\kappa)$$

which yields for the optima,

$$\delta^2 = \frac{\tilde{v}-\lambda}{\hat{\nu}}, \quad \kappa^2 = \frac{\tilde{v}+\lambda}{\hat{\gamma}} \tag{15}$$

Note that κ^2 is positive for all $\lambda\in\mathbb{R}$, and since

$$xK_{\lambda+1}(x)/K_\lambda(x) > 2\lambda$$

for all $\lambda\in\mathbb{R}$ and $x > 0$, we have δ^2 also always positive.

3.4. Shape parameter update: λ

To bound the log likelihood as a function of λ, we use the following,

$$\log K_\lambda(t) \;=\; \inf_u \;\tfrac{1}{2}u\lambda^2 - g(u)$$

for a relative conjugate function g where again only the optimal u is needed for the algorithm,

$$\tilde{u} = \frac{1}{\lambda\,\epsilon}\log\frac{K_{\lambda+\epsilon}(\delta\kappa)}{K_\lambda(\delta\kappa)} \tag{16}$$

where the derivative of the Bessel function with respect to order is approximated by a small finite difference. This yields the surrogate likelihood cost function for λ,

$$\hat{\eta}\lambda + \lambda\log(\kappa/\delta) - \tfrac{1}{2}\tilde{u}\lambda^2$$

which has optimum,

$$\lambda = \frac{\log(\kappa/\delta) + \hat{\eta}}{\tilde{u}} \tag{17}$$

4. FINITE MIXTURE MODEL

We can extend the algorithm to estimate a finite mixture model incorporating the standard mixture model EM algorithm to estimate the mixture parameters using the explicit formula for the likelihood. Let us define the parameters of the jth model,

$$\boldsymbol{\Theta}_j \triangleq \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{\theta}_j, \lambda_j, \delta_j, \kappa_j\}$$

The mixture model has the form,

$$p\big(\mathbf{s}\,;\mathbf{b},\{\alpha_j,\boldsymbol{\Theta}_j\}_{j=1}^M\big) = \sum_{j=1}^M \alpha_j\, p\big(\mathbf{s}-\mathbf{b}\,;\boldsymbol{\Theta}_j\big)$$

subject to the constraints $\alpha_j, \delta_j, \kappa_j > 0$, $\lambda_j \in \mathbb{R}$, $|\det\boldsymbol{\Sigma}_j| = 1$, $j = 1,\ldots,M$, $\sum_{j=1}^M \alpha_j = 1$, $\sum_{j=1}^M \alpha_j(\boldsymbol{\mu}_j + E\{\xi_j\}\boldsymbol{\theta}_j) = \mathbf{0}$. Define the hidden model index for time t, $j_t \in \{1,\ldots,M\}$, the hidden model indicator variables,

$$z_{jt} = \begin{cases} 1, & j_t = j \\ 0, & j_t \neq j \end{cases}$$

and the posterior expectations $\hat{z}_{jt} = E\{z_{jt}|\,\mathbf{s}_t\}$. Then we have the standard mixture model updates,

$$\hat{z}_{jt} = \frac{\alpha_j^\ell p(\,\mathbf{s}_t - \mathbf{b}\,;\boldsymbol{\Theta}_j)}{\sum_{j'=1}^M \alpha_{j'}^\ell p(\,\mathbf{s}_t-\mathbf{b}\,;\boldsymbol{\Theta}_{j'})} \tag{18}$$

and $\alpha_j^{\ell+1} = \frac{1}{N}\sum_{t=1}^N \hat{z}_{jt}$, where we use ℓ to indicate the iteration number.

Now defining $\nu_{jt} \triangleq E\{\xi_{jt}^{-1}|\mathbf{s}_t, j_t = j\}$, etc., leaving out the zero mixture mean constraint for the moment, the complete log likelihood is,

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^N\sum_{j=1}^M \hat{z}_{jt}\bigg(& -\tfrac{1}{2}\log|\boldsymbol{\Sigma}_j| \;+\; \boldsymbol{\theta}_j^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{s}_t - \mathbf{b} - \boldsymbol{\mu}_j) \\ & - \tfrac{1}{2}\gamma_{jt}\,\boldsymbol{\theta}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\theta}_j \;-\; \tfrac{1}{2}\,\nu_{jt}\,\|\mathbf{s}_t - \mathbf{b} - \boldsymbol{\mu}_j\|^2_{\boldsymbol{\Sigma}_j^{-1}} - \tfrac{1}{2}\nu_{jt}\delta_j^2 \;+\; \eta_{jt}\lambda_j \\ & - \tfrac{1}{2}\gamma_{jt}\kappa_j^2 + \lambda_j\log(\kappa_j/\delta_j) \;-\; \log K_{\lambda_j}(\delta_j\kappa_j)\bigg) + \zeta_j\log|\boldsymbol{\Sigma}_j|\end{aligned}$$

where ζ_j are the Lagrange multipliers for the unit determinant constraints. Defining $\mathbf{m}_j \triangleq (N\alpha_j^{\ell+1})^{-1}\sum_{t=1}^N \hat{z}_{jt}\mathbf{s}_t$, and $\hat{\nu}_j \triangleq (N\alpha_j^{\ell+1})^{-1}\sum_{t=1}^N \hat{z}_{jt}\nu_{jt}$, etc., the complete log likelihood can be written,

$$\begin{aligned}\sum_{j=1}^M \alpha_j^{\ell+1}\bigg(& -\tfrac{1}{2}\log|\boldsymbol{\Sigma}_j| + \boldsymbol{\theta}_j^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{m}_j - \mathbf{b} - \boldsymbol{\mu}_j) - \tfrac{1}{2}\hat{\gamma}_j\kappa_j^2 \\ & - \tfrac{1}{2}\hat{\gamma}_j\,\boldsymbol{\theta}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\theta}_j \;-\; \hat{\eta}_j\lambda_j - \lambda_j\log(\delta_j/\kappa_j) - \log K_{\lambda_j}(\delta_j\kappa_j) \\ & -\tfrac{1}{2}\hat{\nu}_j\delta_j^2 - \frac{1}{N\alpha_j^{\ell+1}}\sum_{t=1}^N \tfrac{1}{2}\hat{z}_{jt}\nu_{jt}\,\|\mathbf{s}_t - \mathbf{b} - \boldsymbol{\mu}_j\|^2_{\boldsymbol{\Sigma}_j^{-1}}\bigg) + \zeta_j\log|\boldsymbol{\Sigma}_j|\end{aligned}$$

We first derive the update for the overall bias by maximizing with respect to $\mathbf{b}$. Define the weighted model means,

$$\mathbf{m}_{j\nu} \triangleq (N\alpha_j^{\ell+1}\hat{\nu}_j)^{-1}\sum_{t=1}^N \hat{z}_{jt}\nu_{jt}\,\mathbf{s}_t \tag{19}$$

Then the optimal $\mathbf{b}$ is found to be,

$$\mathbf{b} = \bigg(\sum_{j=1}^M \alpha_j^{\ell+1}\hat{\nu}_j\,\boldsymbol{\Sigma}_j^{-1}\bigg)^{-1}\sum_{j=1}^M \alpha_j^{\ell+1}\boldsymbol{\Sigma}_j^{-1}\big(\hat{\nu}_j(\mathbf{m}_{j\nu} - \boldsymbol{\mu}_j) - \boldsymbol{\theta}_j\big) \tag{20}$$

Now, the zero mean mixture model constraint has the form $\sum_{j=1}^M \alpha_j\big(\boldsymbol{\mu}_j + \hat{\xi}_j\boldsymbol{\theta}_j\big) = \mathbf{0}$, where we define the model scale means $\hat{\xi}_j \triangleq E\{\xi_j\}$, i.e.,

$$\hat{\xi}_j = \frac{\delta_j}{\kappa_j}\frac{K_{\lambda_j+1}(\delta_j\kappa_j)}{K_{\lambda_j}(\delta_j\kappa_j)} \tag{21}$$

If we let $\boldsymbol{\beta}$ be the Lagrangian multiplier vector corresponding to this constraint, then the Lagrangian equations to be solved for $\boldsymbol{\mu}_j, \boldsymbol{\theta}_j$, $j = 1,\ldots,M$, along with the constraint, can be written,

$$\begin{array}{ccccccc} \hat{\nu}_j\,\boldsymbol{\mu}_j & + & \boldsymbol{\theta}_j & + & \boldsymbol{\Sigma}_j\boldsymbol{\beta} & = & \hat{\nu}_j(\mathbf{m}_{j\nu} - \mathbf{b}) \\ \boldsymbol{\mu}_j & + & \hat{\gamma}_j\boldsymbol{\theta}_j & + & \hat{\xi}_j\boldsymbol{\Sigma}_j\boldsymbol{\beta} & = & \mathbf{m}_j \;-\; \mathbf{b} \end{array}$$

Defining $\mathbf{c}_j \triangleq \hat{\nu}_j(\mathbf{m}_{j\nu} - \mathbf{b})$ and $\mathbf{d}_j \triangleq \mathbf{m}_j - \mathbf{b}$, the equations to be solved can be put in the following matrix form,

$$\begin{bmatrix} \hat{\nu}_1\mathbf{I} & \mathbf{I} & & & & \boldsymbol{\Sigma}_1 \\ \mathbf{I} & \hat{\gamma}_1\mathbf{I} & & & & \hat{\xi}_1\boldsymbol{\Sigma}_1 \\ & & \ddots & & & \vdots \\ & & & \hat{\nu}_M\mathbf{I} & \mathbf{I} & \boldsymbol{\Sigma}_M \\ & & & \mathbf{I} & \hat{\gamma}_M\mathbf{I} & \hat{\xi}_M\boldsymbol{\Sigma}_M \\ \hat{\alpha}_1\mathbf{I} & \hat{\alpha}_1\hat{\xi}_1\mathbf{I} & \ldots & \hat{\alpha}_M\mathbf{I} & \hat{\alpha}_M\hat{\xi}_M\mathbf{I} & \end{bmatrix}\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\mu}_M \\ \boldsymbol{\theta}_M \\ \boldsymbol{\beta}\end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{d}_1 \\ \vdots \\ \mathbf{c}_M \\ \mathbf{d}_M \\ \mathbf{0}\end{bmatrix}$$

Define the Schur complement of the null corner block,

$$\mathbf{S} \triangleq \sum_{j=1}^M \alpha_j\frac{\hat{\nu}_j\hat{\xi}_j^2 - 2\hat{\xi}_j + \hat{\gamma}_j}{1 - \hat{\nu}_j\hat{\gamma}_j}\boldsymbol{\Sigma}_j \tag{22}$$

Solving for $\boldsymbol{\beta}$, we get,

$$\boldsymbol{\beta} = \mathbf{S}^{-1}\sum_{j=1}^M \hat{\alpha}_j\left(\frac{\hat{\gamma}_j - \hat{\xi}_j}{1 - \hat{\nu}_j\hat{\gamma}_j}\,\mathbf{c}_j + \frac{\hat{\nu}_j\hat{\xi}_j - 1}{1 - \hat{\nu}_j\hat{\gamma}_j}\,\mathbf{d}_j\right) \tag{23}$$

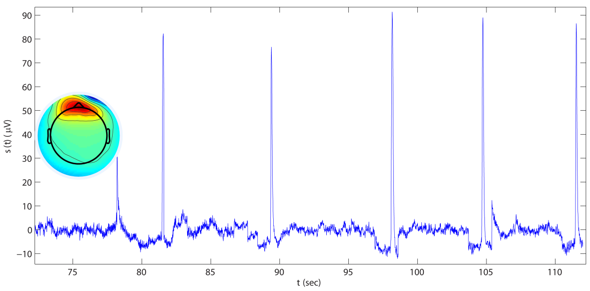

Fig. 1. Topographic map and plot of eyeblink electric potential. Spikes correspond to blinks.

Now if we let $\tilde{\mathbf{c}}_j \triangleq \mathbf{c}_j - \boldsymbol{\Sigma}_j\boldsymbol{\beta}$ and $\tilde{\mathbf{d}}_j \triangleq \mathbf{d}_j - \hat{\xi}_j\boldsymbol{\Sigma}_j\boldsymbol{\beta}$, then,

$$\boldsymbol{\mu}_j = \frac{1}{\hat{\nu}_j\hat{\gamma}_j - 1}\left(\hat{\gamma}_j\tilde{\mathbf{c}}_j - \tilde{\mathbf{d}}_j\right) \tag{24}$$

$$\boldsymbol{\theta}_j = \frac{1}{\hat{\nu}_j\hat{\gamma}_j - 1}\left(\hat{\nu}_j\tilde{\mathbf{d}}_j - \tilde{\mathbf{c}}_j\right) \tag{25}$$

For the structure matrix update, if we define,

$$\boldsymbol{\Sigma}_{j\nu} \triangleq \left(N\alpha_j^{\ell+1}\hat{\nu}_j\right)^{-1}\sum_{t=1}^{N}\hat{z}_{jt}\nu_{jt}(\mathbf{s}_t - \mathbf{b} - \boldsymbol{\mu}_j)(\mathbf{s}_t - \mathbf{b} - \boldsymbol{\mu}_j)^T \tag{26}$$

then the update for $\boldsymbol{\Sigma}_j$ is,

$$\boldsymbol{\Sigma}_j = \hat{\nu}_j\boldsymbol{\Sigma}_{j\nu} - (\mathbf{m}_j - \mathbf{b} - \boldsymbol{\mu}_j)\boldsymbol{\theta}_j^T - \boldsymbol{\theta}_j(\mathbf{m}_j - \mathbf{b} - \boldsymbol{\mu}_j)^T + \hat{\gamma}_j\boldsymbol{\theta}_j\boldsymbol{\theta}_j^T \tag{27}$$

followed by unit determinant normalization.

For the shape parameters δ_j and κ_j, using the definition of $\hat{\xi}_j$, we have,

$$\delta_j^2 = \frac{\hat{\xi}_j\kappa_j^2 - 2\lambda_j}{\hat{\nu}_j}, \quad \kappa_j^2 = \frac{\hat{\xi}_j}{\hat{\gamma}_j}\kappa_j^2 \tag{28}$$

And finally for the λ_j shape parameters,

$$\tilde{u}_j = \frac{1}{\lambda_j\,\epsilon}\log\frac{K_{\lambda_j+\epsilon}(\delta_j\kappa_j)}{K_{\lambda_j}(\delta_j\kappa_j)}, \quad \lambda_j = \frac{\log(\kappa_j/\delta_j) + \hat{\eta}_j}{\tilde{u}_j} \tag{29}$$

5. EXPERIMENTS

Monotonic convergence of the algorithm without the need to set or modify step sizes has been verified. We consider a case study application to the eyeblink muscle signal component typically seen in EEG recordings (see Figure 5). In figure 5 we show the determination of model order using the Generalized Likelihood Ratio Test approach, where twice the change in log likelihood is distributed $\chi^2(k)$ where k is the difference in the number of degrees of freedom. The red lines plot the likelihood increases to reject. The log likelihood is seen to exhibit the expected non-significant linear increase after an appropriate model order is reached.

6. CONCLUSION

We have derived an EM algorithm to compute Maximum Likelihood parameter estimates for Barndorff-Nielsen's flexible, multivariate Generalized Hyperbolic distribution which includes many distributions as limiting cases. We used two novel variational bounds to derive shape parameter updates. Finally we derived finite mixture model parameter updates.

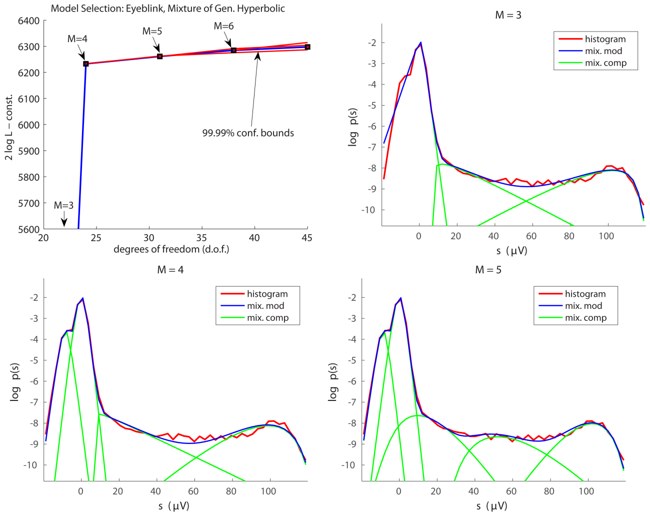

Fig. 2. Generalized Hyperbolic mixture model fits of eyeblink source distribution with $M = 2, 3, 5, 7$.

REFERENCES

[1] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review*, vol. 50, pp. 145–159, 1982.

[2] 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*, vol. 39, pp. 1–38, 1977.

[3] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions, Volume 1*, John Wiley & Sons, Inc.: New York, 1994.

[4] M. Abramaowitz and I. A. Stegun, Eds., *Handbook of mathematical functions*, National Bureau of Mathematics, 1964.