Super-Gaussian Mixture Source Model for ICA

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

[1] Department of Electrical and Computer Engineering
University of California San Diego, La Jolla, CA 92093
{japalmer, kreutz}@ece.ucsd.edu
[2] Swartz Center for Computational Neuroscience
University of California San Diego, La Jolla, CA 92093
scott@sccn.ucsd.edu

Abstract. We propose an extension of the mixture of factor (or independent component) analyzers model to include strongly super-gaussian mixture source densities. This allows greater economy in representation of densities with (multiple) peaked modes or heavy tails than using several Gaussians to represent these features. We derive an EM algorithm to find the maximum likelihood estimate of the model, and show that it converges globally to a local optimum of the actual non-gaussian mixture model without needing any approximations. This extends considerably the class of source densities that can be used in exact estimation, and shows that in a sense super-gaussian densities are as natural as Gaussian densities. We also derive an adaptive Generalized Gaussian algorithm that learns the shape parameters of Generalized Gaussian mixture components. Experiments verify the validity of the algorithm.

1 Introduction

We propose an extension of the mixture of factor [2], or independent component [6] analyzers model that enlarges the flexibility of the source density mixture model while maintaining mixtures of strongly super-gaussian densities. Mixture model source densities allow one to model skewed and multi-modal densities, and optimization of these models is subject to convergence to local optima, the mixture model is a generalization of the unimodal model and may be built up by starting with uni- or bi-modal source models, then adding components and monitoring the change in likelihood [8, 3, 6].

Variational Gaussian mixture models, proposed in [8, 2, 6, 5], are ultimately mixtures of Student's t distributions after the random variance is integrated out [19, 3]. In [12] a mixture generalization of the Infomax algorithm is proposed in which a mixture model is employed over sets basis vectors but not for the source component density models. The means are updated by gradient descent or by a heuristic approximate EM update. In [16] a variance mixture of Laplacians model is employed 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. In [17] a mixture of Logistic source density model is estimated by gradient descent.

The property of strongly super-gaussian densities that we use, namely $\log$-convexity in x^2, has been exploited previously by Jaakkola [10, 11] in graphical models, and Girolami [9] for ICA using the Laplacian density. The model we propose extends the work in [9] 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 [3] in allowing the scale of the sources vary (actually a necessity in the mixture case) and fixing the scale of the de-mixing filters to unity by an appropriate transformation at each iteration in order to avoid the scale ambiguity inherent in factor analysis models.

The proposed model generalizes all of these algorithms, including 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 [7]. For a given number of mixture components, the EM algorithm estimates the location (mode) and scale parameters of the mixture component.

Using the natural gradient [1] to update the un-mixing matrices (the inverses of the basis 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 complete log likelihood, remaining within the GEM framework and guaranteeing increase in the data likelihood under the model. The un-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 a 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 [3, 6]. 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 an adaptive generalized Gaussian algorithm. §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 in [4], where the authors define $p(x) = \exp(-f(x))$ to be super-gaussian (sub-gaussian) if $f'(x)/x$ is increasing (decreasing) on $(0, \infty)$. This condition is equivalent to $f(x) = g(x^2)$ with g concave, i.e. g' decreasing, where $g'(x^2) = f'(x)/x$.

In [15] we have discussed these densities in some detail, and derived relationships between them and the hyperprior representation used in the evidence framework [13] and the Variational Bayes framework [2]. 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, $\forall t$,

$$f(t) - f(s) \;=\; g(t^2) - g(s^2) \;\leq\; g'(s^2)(t^2 - s^2) \;=\; \frac{1}{2}\frac{f'(s)}{s}\big(t^2 - s^2\big) \qquad (1)$$

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

Mixture densities have the form,

$$p(s) = \sum_{j=1}^{m} \alpha_j \, p_j\!\left(\frac{s - \mu_j}{\sigma_j}\right), \quad \sum_j \alpha_j = 1\,,\ \sigma_j > 0$$

The probability density of the j_ith mixture component of the ith source is denoted $p_{ij_i}(s_{ij_i})$, with mode μ_{ij_i}, and scale σ_{ij_i}.

3 The EM Algorithm

We follow the framework of [18, 14] in deriving the EM algorithm, which was originally derived rigorously in [7]. The log likelihood of the data decomposes as follows,

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

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* [18, 14]. This representation is useful if $F(q;\theta)$ can easily be minimized with respect to θ. Since the KL divergence is non-negative, we have,

$$-\log p(\mathbf{x};\theta) = \min_q F(q;\theta)$$

where equality is obtained if and only if $q(\mathbf{z}|\mathbf{x};\theta') = p(\mathbf{z}|\mathbf{x};\theta)$. The EM algorithm at the lth iteration, given q^l and θ^l, performs coordinate descent in q and θ,

$$\theta^{l+1} = \min_\theta F\big(q^l;\theta\big)\,, \quad q^{l+1} = p\big(\mathbf{z}|\mathbf{x};\theta^{l+1}\big)$$

This algorithm is guaranteed to increase the likelihood since,

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

Note however, that it is not necessary to actually minimize F to guarantee that the likelihood increases. It is enough simply to guarantee that $F(q^l;\theta^l) \geq F(q^l;\theta^{l+1})$, i.e. to guarantee that F decreases as a result of updating θ. This leads to the Generalized EM (GEM) algorithm [7], and is the approach we follow here. We maintain the global convergence (to a local optimum) property of the EM algorithm however by guaranteeing a decrease in F by an efficient closed form update for the source density parameters.

4 ICA with Strongly Super-Gaussian Mixture Sources

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

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

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is non-singular, and the sources are independent mixtures of independent strongly super-gaussian random variables s_{ij_i}, $j_i = 1, \ldots, m_i$, where we allow the number of source mixture components m_i to differ for different sources.

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 $P[j_i = j] = \alpha_{ij}$, $1 \leq j \leq m_i$, 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.

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

$$\theta = \{\mathbf{w}_i, \alpha_{ij_i}, \mu_{ij_i}, \sigma_{ij_i}\}, \quad i = 1, \ldots, n\,,\ j_i = 1, \ldots, m_i$$

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

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

$$z_{ij_ik} = \begin{cases} 1, & j_{ik} = j_i \\ 0, & \text{otherwise} \end{cases}$$

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

$$p(\mathbf{X};\theta) = \sum_{\mathbf{Z}} \prod_{k=1}^{N} |\det \mathbf{W}| \prod_{i=1}^{n} \prod_{j_i=1}^{m_i} \alpha_{ij_i}^{z_{ij_ik}} \left[\frac{1}{\sigma_{ij_i}} p_{ij_i}\left(\frac{\mathbf{w}_i^T\mathbf{x}_k - \mu_{ij_i}}{\sigma_{ij_i}}\right)\right]^{z_{ij_ik}}$$

For the variational free energy, F, we have,

$$F(q;\theta) = \sum_{k=1}^{N} \sum_{i=1}^{n} \sum_{j_i=1}^{m_i} \hat{z}_{ij_ik} \left[-\log \alpha_{ij_i} - \log \sigma_{ij_i} + f_{ij_i}\left(\frac{\mathbf{w}_i^T\mathbf{x}_k - \mu_{ij_i}}{\sigma_{ij_i}}\right)\right] - N \log |\det \mathbf{W}| \quad (2)$$

where q is the discrete distribution defining the expectation $\hat{z}_{ij_ik} = E[z_{ij_ik}|\mathbf{x}_k]$, and where we define $f_{ij_i} = -\log p_{ij_i}$.

Let us define,

$$y^l_{ij_ik} \equiv \frac{\mathbf{w}_i^{l\,T}\mathbf{x}_k - \mu^l_{ij_i}}{\sigma^l_{ij_i}} \tag{3}$$

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

$$\hat{z}^l_{ij_ik} = \frac{p(\mathbf{x}_k|z_{ij_ik} = 1; \theta^l)P[z_{ij_ik} = 1; \theta^l]}{\sum_{j'_i=1}^{m_i} p(\mathbf{x}_k|z_{ij'_ik} = 1; \theta^l)P[z_{ij'_ik} = 1; \theta^l]} = \frac{p_{ij_i}(y^l_{ij_ik})\,\alpha^l_{ij_i}/\sigma^l_{ij_i}}{\sum_{j'_i=1}^{m_i} p_{ij'_i}(y^l_{ij'_ik})\,\alpha^l_{ij'_i}/\sigma^l_{ij'_i}} \tag{4}$$

as are the optimal α_{ij_i},

$$\alpha^{l+1}_{ij_i} \;=\; \frac{\sum_{k=1}^N \hat{z}^l_{ij_ik}}{\sum_{j'_i=1}^{m_i}\sum_{k=1}^N \hat{z}^l_{ij'_ik}} \;=\; \frac{1}{N}\sum_{k=1}^N \hat{z}^l_{ij_ik}$$

Now, since the p_{ij_i} are strongly super-gaussian, we can use the inequality (1) to replace $f_{ij_i}(y_{ij_ik})$ in (2) by $\big(f'_{ij_i}(y^l_{ij_ik})/2y^l_{ij_ik}\big)\big(y^2_{ij_ik} - y^{l\,2}_{ij_ik}\big)$. Defining,

$$\xi^l_{ij_ik} \;\equiv\; \frac{f'_{ij_i}(y^l_{ij_ik})}{y^l_{ij_ik}} \tag{5}$$

we replace F by,

$$\tilde{F}(q;\theta) = \sum_{k=1}^N\sum_{i=1}^n\sum_{j_i=1}^{m_i} \hat{z}_{ij_ik}\left[-\log\alpha_{ij_i} - \log\sigma_{ij_i} + \frac{\xi^l_{ij_ik}}{2}\left(\frac{\mathbf{w}_i^T\mathbf{x}_k - \mu_{ij_i}}{\sigma_{ij_i}}\right)^2\right] - N\log|\det\mathbf{W}|$$

Minimizing $\tilde{F}$ with respect to μ_{ij_i} and σ_{ij_i} guarantees, using the inequality (1), that,

$$F(q;\theta^{l+1}) - F(q;\theta^l) \;\le\; \tilde{F}(q;\theta^{l+1}) - \tilde{F}(q;\theta^l) \;\le\; 0$$

and thus that $F(q;\theta)$ is decreased as required by the EM algorithm.

As in the Gaussian case, the optimal value of μ_{ij_i} does not depend on σ_{ij_i}, and we can optimize with respect to μ_{ij_i}, then optimize with respect to σ_{ij_i} given μ_{ij_i}, and guarantee an overall increase in the likelihood. The updates, using the definitions (3), (4) and (5), are found to be,

$$\mu^{l+1}_{ij} = \frac{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}\mathbf{w}_i^{l\,T}\mathbf{x}_k}{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}}\,, \quad \sigma^{l+1}_{ij} = \left(\frac{\sum_{k=1}^N \hat{z}^l_{ijk}\xi^l_{ijk}(\mathbf{w}_i^{l\,T}\mathbf{x}_k - \mu^{l+1}_{ij})^2}{\sum_{k=1}^N \hat{z}^l_{ijk}}\right)^{1/2} \tag{6}$$

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

$$\big[\mathbf{u}^l_k\big]_i \equiv \sum_{j_i=1}^{m_i} \hat{z}^l_{ij_ik} f'_{ij_i}(y^l_{ij_ik})/\sigma^l_{ij_i} \tag{7}$$

we have,

$$\Delta \mathbf{W} = \left(\frac{1}{N} \sum_{k=1}^{N} \mathbf{u}_k^l \mathbf{x}_k^T \mathbf{W}^{l\,T} - \mathbf{I} \right) \mathbf{W}^l \tag{8}$$

5 Full ICA Mixture Model with Super-Gaussian Mixture Sources

We now consider the case where the data are generated by a mixture of mixing matrices,

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

where now we have,

$$\theta = \{\gamma_h, \mathbf{w}_{hi}, \alpha_{hij}, \mu_{hij}, \sigma_{hij}\} \,, \; h = 1, \ldots, M, \, i = 1, \ldots, n, \, j = 1, \ldots, m_{hi}$$

The EM algorithm for the full mixture model is derived similarly to the case of source mixtures. Due to space constraints the details are omitted.

6 Adaptive Generalized Gaussian Mixture Model

We can obtain further flexibility in the source model by adapting mixtures of a parameterized family of strongly super-gaussian densities. In this section we consider the case of Generalized Gaussian mixtures,

$$p(s_{ij_i}; \mu_{ij_i}, \sigma_{ij_i}, \beta_{ij_i}) = \frac{1}{2\,\sigma_{ij_i} \Gamma\left(1 + \frac{1}{\beta_{ij_i}}\right)} \exp\left(- \left| \frac{s_{ij_i} - \mu_{ij_i}}{\sigma_{ij_i}} \right|^{\beta_{ij_i}} \right)$$

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

$$\frac{dF}{d\beta_{ij_i}} \;=\; \sum_{k=1}^{N} \hat{z}_{ij_ik} \left[|y_{ij_ik}|^{\beta_{ij_i}} \log |y_{ij_ik}| - \frac{1}{\beta_{ij_i}^2} \Psi\!\left(1 + \frac{1}{\beta_{ij_i}}\right) \right]$$

We have found that scaling this by $\beta_{ij_i}^2 \Big/ \left(\Psi(1 + \frac{1}{\beta_{ij_i}}) \sum_{k=1}^{N} \hat{z}_{ij_ik} \right)$, which is positive, leads to faster convergence. The update is then,

$$\Delta \beta_{ij_i} = \frac{\beta_{ij_i}^2 \sum_{k=1}^{N} \hat{z}_{ij_ik} \, |y_{ij_ik}|^{\beta_{ij_i}} \log |y_{ij_ik}|}{\Psi\!\left(1 + \frac{1}{\beta_{ij_i}}\right) \sum_{k=1}^{N} \hat{z}_{ij_ik}} - 1$$

7 Experiments

We verified the convergence of the algorithm with toy data generated from Generalized Gaussian mixtures with randomly generated parameters. The algorithm is subject to local optima, and thus it is advisable to begin with simpler models and add complexity. The "see-saw" method [2, 6] in which $\mathbf{W}$ is adapted with fixed source parameters, then the source model adapted with fixed $\mathbf{W}$ seems better able to handle local optima than the "chase" method in which the updates are interleaved. Below 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 were initialized to 2, the location and scale parameters were randomly initialized. The log likelihood is monotonically increasing as expected.

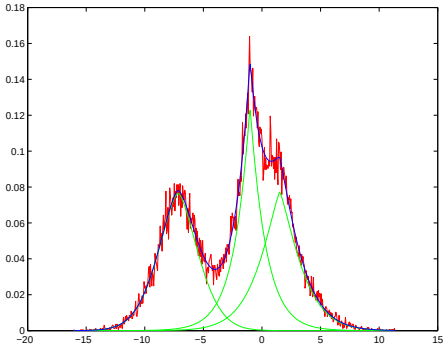

Fig. 1. Example of adaptive convergence of super-gaussian mixture model.

References

1. S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
2. H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.
3. H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.
4. A. Benveniste, M. Goursat, and G. Ruget. Robust identification of a nonminimum phase system. *IEEE Transactions on Automatic Control*, 25(3):385–399, 1980.
5. 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.
6. R. A. Choudrey and S. J. Roberts. Variational mixture of Bayesian independent component analysers. *Neural Computation*, 15(1):213–252, 2002.

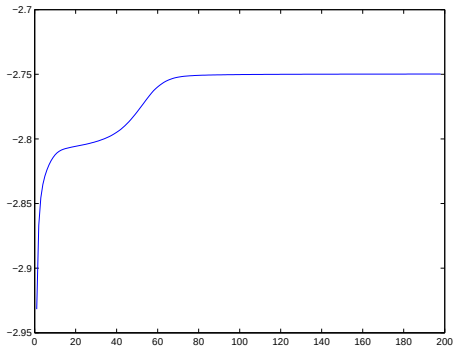

Fig. 2. Log likelihood is monotonically increasing.

7. 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.
8. 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.
9. M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13:2517–2532, 2001.
10. T. S. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models*. PhD thesis, Massachusetts Institute of Technology, 1997.
11. 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.
12. 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.
13. D. J. C. Mackay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.
14. 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.
15. 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/.
16. 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.
17. 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.
18. 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.
19. M. E. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.