Journal of Machine Learning Research 3 (2002) 99-114 Submitted 5/02; Published 8/02

# Variational Learning of Clusters of Undercomplete Nonsymmetric Independent Components

**Kwokleung Chan** KWCHAN@SALK.EDU
*Computational Neurobiology Laboratory*
*The Salk Institute*
*10010 North Torrey Pines Road*
*La Jolla, CA 92037, USA*

**Te-Won Lee** TEWON@SALK.EDU
*Institute for Neural Computation*
*University of California at San Diego*
*La Jolla, CA 92093, USA*

**Terrence J. Sejnowski** TERRY@SALK.EDU
*Computational Neurobiology Laboratory*
*The Salk Institute*
*10010 North Torrey Pines Road*
*La Jolla, CA 92037, USA*
*and*
*Department of Biology*
*University of California at San Diego*
*La Jolla, CA 92093, USA*

**Editor:** Michael I. Jordan

## Abstract

We apply a variational method to automatically determine the number of mixtures of independent components in high-dimensional datasets, in which the sources may be nonsymmetrically distributed. The data are modeled by clusters where each cluster is described as a linear mixture of independent factors. The variational Bayesian method yields an accurate density model for the observed data without overfitting problems. This allows the dimensionality of the data to be identified for each cluster. The new method was successfully applied to a difficult real-world medical dataset for diagnosing glaucoma.

**Keywords:** Density Estimations, Mixture Models, Bayesian Learning, ICA

## 1. Introduction

The performance of a method for pattern classification is often determined by how well it can model the underlying statistical distribution of the data. Independent component analysis (ICA) models non-Gaussian structure, e.g., platykurtic or leptokurtic probability density functions. In ICA (Hyvarinen et al., 2001), the observed data $\mathbf{x}$ are assumed to be generated from a linear combination of independent sources $\mathbf{s}$:

$$\mathbf{x} = \mathbf{A}\,\mathbf{s} + \boldsymbol{\nu}\,,$$

where $\mathbf{A}$ is the mixing matrix, which can be non-square. The sources $\mathbf{s}$ have non-Gaussian density such as $p(s_m) \propto \exp(-|s_m|^q)$. The noise term $\boldsymbol{\nu}$ can have non-zero mean. ICA is a generalization of principal component analysis (PCA), in which columns of $\mathbf{A}$ are constrained to be orthogonal and $p(s_m) \propto \exp(-s_m^2)$. ICA has been applied to speech separation and analyzing fMRI and EEG data (Jung et al., 2001). When modeling the data density, ICA tries to locate independent axes within the data cloud and finds projections that may uncover interesting structure in the data.

Real data sets often have clusters; unsupervised clustering could help uncover the structure of the data density. Clusters of data can be described by an ICA mixture model (Lee et al., 2000). In addition to modeling the data density, ICA can also discover interesting hidden features in data (Olshausen and Field, 1996). This valuable property could be used to study the features of each cluster.

In unsupervised classification, one is interested in obtaining a close fit to the observed data distribution without overfitting. However, maximum likelihood (ML) may overfit the data, especially in high dimensional spaces. Dimensionality of the data may be reduced by first performing PCA, but the intrinsic dimensionality within each cluster could be different. More importantly, the subspace of the clusters may orient differently with respect to each other. It would be desirable to determine the dimensionality of each cluster, instead of performing a single dimensionality reduction on the whole dataset.

The Bayesian approach helps find number of clusters and number of sources in each cluster. In a full Bayesian treatment, the parameters $\theta$ in a model $\mathcal{H}$ are given a prior distribution $P(\theta|\mathcal{H})$. Instead of the likelihood $P(\mathbf{X}|\theta, \mathcal{H})$ on a dataset $\mathbf{X}$, the marginal likelihood, $P(\mathbf{X}|\mathcal{H}) = \int P(\mathbf{X}|\theta, \mathcal{H})P(\theta|\mathcal{H})\ d\theta$, for different models are compared. An overly-complex or flexible model $\mathcal{H}$ will have a low marginal likelihood and can be ruled out. The solution for the chosen model is given as a posterior distribution over the parameter $\theta$, $P(\theta|\mathbf{X}, \mathcal{H})$, rather than a point estimate $\hat{\theta}$ as in the maximum likelihood approach. However, full Bayesian treatments are rarely tractable. The variational method (Jordan et al., 1999, Ghahramani and Beal, 2000) provides an approximate analytical solution by a mean field approach to $P(\theta|\mathbf{X}, \mathcal{H})$ and $P(\mathbf{X}|\mathcal{H})$. Besides performing model selection and avoiding overfitting, the variational Bayesian treatment provides uncertainty estimates for the model parameters, which is not directly available with the ML or MAP approach.

In many treatments of ICA, the form of the source distribution $p(s_m)$ (or equivalently the "non-linearity") is fixed and assumed symmetric. Real data sets often contain both super and sub-Gaussian sources. These sources may also be skewed and therefore non-symmetric. Recently, Karvanen et al. (2000) used non-symmetric source density models to perform ICA under the maximum likelihood approach. We use mixtures of Gaussians (Moulines et al., 1997, Attias, 1999a, Welling and Weber, 2001) for source densities instead of assuming fixed distributions. The use of mixtures of Gaussians allows closed form solution in the variational Bayesian learning while maintaining a flexible model for non-symmetric sources with mixed kurtosis.

The variational approach to ICA has been studied by a number of researchers (Lappalainen, 1999, Miskin, 2000, Højen-Sørensen et al., 2002). In this paper, we extend the mixture model of Ghahramani and Beal (2000) and ICA model of Miskin (2000), and propose a mixture of undercomplete non-symmetric ICA solution to describe the underlying distribution of small but high dimensional dataset. A preliminary version of these results were described by Chan et al. (2001). A similar variational method for ICA clusters was independently proposed by Choudrey and Roberts (2001).

## 2. Theory and Method

In this section, we describe the density model for ICA clusters and study the application of variational Bayesian learning to the density model.

### 2.1 Density model

Observations $\mathbf{X} = \{\mathbf{x}_t \in \mathcal{R}^N\}, t = [1, \cdots, T]$ are assumed to be generated from one of $C$ clusters centered at $\boldsymbol{\nu}^c$ and with diagonal Gaussian noise variance $[\boldsymbol{\Psi}^c]^{-1}$:

$$\begin{aligned} P(\mathbf{x}_t|\boldsymbol{\rho}, \mathbf{A}^c, \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) &= \sum_c^C P(c_t{=}c|\rho_c) P(\mathbf{x}_t|\mathbf{A}^c, \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) \\ P(c_t{=}c|\rho_c) &= \rho_c \,, \end{aligned} \tag{1}$$

$\boldsymbol{\rho} = (\rho_1, \ldots, \rho_C)$ are the weights on the $C$ clusters. Inside each cluster, observation $\mathbf{x}_t$ is a linear combination of $M$ independent sources $\mathbf{s}_t^c = (s_{1t}^c, \ldots, s_{Mt}^c)^\top$:

$$P(\mathbf{x}_t|\mathbf{A}^c, \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) = \int \mathcal{N}(\mathbf{x}_t|\mathbf{A}^c\mathbf{s}_t^c + \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) P(\mathbf{s}_t^c)\, d\mathbf{s}_t^c \,. \tag{2}$$

To allow for non-symmetric sources, the density of each $s_{mt}^c$ is modeled by a mixture of $K$ Gaussians

$$P(s_{mt}^c|\pi_{mk}^c, \phi_{mk}^c, \beta_{mk}^c) \;\; = \;\; \textstyle\sum_k^K \pi_{mk}^c \mathcal{N}(s_{mt}^c|\phi_{mk}^c, \beta_{mk}^c) \,, \tag{3}$$

where $\beta$ is an inverse variance parameter. Following Bishop (1999), we assume an *automatic relevance determination* (ARD) Gaussian density (Neal, 1996) for $\mathbf{A}^c$,

$$P(A_{nm}^c|\alpha_m^c) = \mathcal{N}(A_{nm}^c|0, \alpha_m^c) \,. \tag{4}$$

The generative model described by the above equations can be summarized by the simplified directed graph in Figure 1. The observed variable is $\mathbf{x}_t$ and the hidden variables are $k_{mt}$, $\mathbf{s}_t$ and $c_t$. The rest are model parameters. Nodes inside the dashed box should be repeated for each of the $C$ ICA clusters. The condition dependencies between the variables is evident from the graph. For example, once given the value of hidden variable $c_t$, the generation of $\mathbf{x}_t$ is independent of the value of $\boldsymbol{\rho}$. We use $\theta$ to denote the collection of the parameters $\boldsymbol{\rho}$, $\mathbf{A}^c$, $\boldsymbol{\nu}^c$, $\boldsymbol{\Psi}^c$, $\boldsymbol{\alpha}^c$, $\boldsymbol{\pi}_m^c$, $\phi_{mk}^c$ and $\beta_{mk}^c$.

### 2.2 Variational Bayesian learning

Instead of maximizing the likelihood of the data $\mathcal{L}(\theta; \mathbf{X}) = P(\mathbf{X}|\theta)$ and finding the maximum likelihood estimate of $\theta$, we use a Bayesian approach to obtain the posterior distribution $P(\theta|\mathbf{X})$ over $\theta$:

$$P(\theta|\mathbf{X})P(\mathbf{X}) = P(\mathbf{X}|\theta)P(\theta) \,, \tag{5}$$

where $P(\theta)$ is the prior distribution for $\theta$, defined in Appendix A. Using the variational learning approach of Miskin (2000) and Ghahramani and Beal (2000), we introduce posterior probability $Q(\theta)$ and employ Jensen's inequality (see Appendix B). The log of the marginal likelihood is then lower bounded by

$$\log P(\mathbf{X}) \geq \int Q(\theta) \sum_t \log P(\mathbf{x}_t|\theta)\, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \,. \tag{6}$$

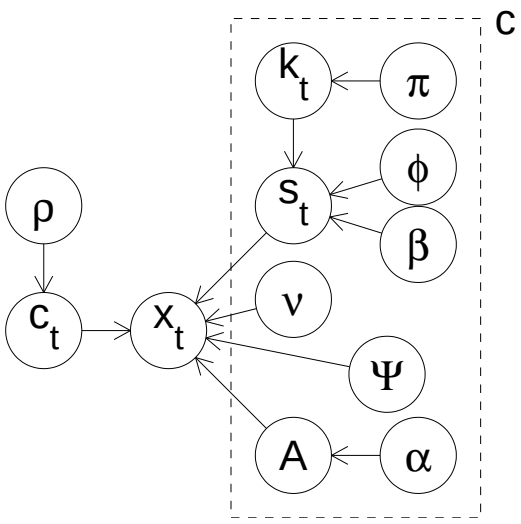


Figure 1: A simplified graphical representation for the generative model of the mixture of variational ICA. $\mathbf{x}_t$ is the observed variable, $k_{mt}$, $\mathbf{s}_t$ and $c_t$ are hidden variables and the rest are model parameters. The nodes inside the dashed box should be repeated for each of the $C$ ICA clusters.

Repeatly introducing $Q(c_t)$, $Q(\mathbf{s}_t^c)$, we arrive at

$$\log P(\mathbf{X}) \geq \int Q(\theta) \sum_t \sum_c Q(c_t) \int Q(\mathbf{s}_t^c) \left[ \log P(\mathbf{x}_t|\mathbf{s}_t^c, \theta) + \log \frac{P(\mathbf{s}_t^c|\theta)}{Q(\mathbf{s}_t^c)} \right] d\mathbf{s}_t^c \, d\theta + \int Q(\theta) \sum_t \sum_c Q(c_t) \log \frac{P(c_t{=}c|\rho_c)}{Q(c_t)} \, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} \, d\theta \,. \quad (7)$$

And finally $\log P(\mathbf{s}_t^c|\theta)$ is replaced as

$$\log P(\mathbf{s}_t^c|\theta) = \sum_m \log P(s_{mt}^c|\theta) \geq \sum_{m\,k} Q(k_{mt}^c) \left[ \log \mathcal{N}(s_{mt}^c|\phi_{mk}^c, \beta_{mk}^c) + \log \frac{\pi_{mk}^c}{Q(k_{mt}^c)} \right] \quad (8)$$

to complete the expansion. Notice that $Q(k_{mt}^c)$ is short for $Q(k_{mt}^c = k)$ and similarly $Q(c_t)$ is short for $Q(c_t = c)$. Learning is accomplished by functional maximization of the lower bound of $\log P(\mathbf{X})$ over $Q(\theta)$, $Q(\mathbf{s}_t^c)$, $Q(c_t)$ and $Q(k_{mt}^c)$. We assume a separable posterior $Q(\theta)$:

$$Q(\theta) = Q(\boldsymbol{\rho}) \prod_c \left[ Q(\boldsymbol{\nu}^c) Q(\boldsymbol{\Psi}^c) Q(\mathbf{A}^c) Q(\boldsymbol{\alpha}^c) \prod_m Q(\boldsymbol{\pi}_m^c) \prod_{mk} Q(\phi_{mk}^c) Q(\beta_{mk}^c) \right] , \quad (9)$$

in order to obtain analytical solutions. Learning rules for $Q(\theta)$ are given in Appendix C.

To compare different solutions resulting from different initial conditions, we computed their corresponding lower bounds $\mathcal{E}(\mathbf{X}, Q(\theta))$ on the log marginal likelihood $P(\mathbf{X})$ using (7) and (8). After some manipulations, $\mathcal{E}(\mathbf{X}, Q(\theta))$ can be expressed as

$$\mathcal{E}(\mathbf{X}, Q(\theta)) = \sum_t \log Z_t + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} \, d\theta \,,$$

where $Z_t$ is defined in (15). The solution with highest $\mathcal{E}(\mathbf{X}, Q(\theta))$ is preferred. The appearance of the normalization term $Z_t$ has no surprise since the marginal likelihood $P(\mathbf{X})$ is itself the normalization term for modifying $P(\mathbf{X}|\theta)P(\theta)$ to $P(\theta|\mathbf{X})$.

### 2.3 Some heuristics

Translational and scale degeneracies exist in the model as described by (1) to (4). After each update of $Q(\boldsymbol{\pi}_m^c)$, $Q(\phi_{mk}^c)$ and $Q(\beta_{mk}^c)$, we rescale $P(\mathbf{s}_{mt}^c)$ to be zero mean and unit variances. The distributions $Q(\mathbf{A}^c)$, $Q(\boldsymbol{\alpha}^c)$ and $Q(\boldsymbol{\nu}^c)$ etc. are adjusted accordingly. This removes the degeneracies and speeds up convergence.

Local maxima of $\mathcal{E}(\mathbf{X}, Q(\theta))$ exist since each cluster is itself a mixture of (correlated) Gaussians. For example, during learning, two clusters may be regarded as one containing a bimodal sub-Gaussian source. This adversely affects the effectiveness of identifying other sources. We employ two heuristics to detect the presence of heavily bimodal source and split the cluster along that axis. The first heuristic is similar to the Fisher's discriminant

$$J = |\phi_k - \phi_{k'}| \times \sqrt{\pi_k \beta_k + \pi_{k'} \beta_{k'}}\,.$$

This index compares the distance $|\phi_k - \phi_{k'}|$ between adjacent Gaussians in $P(s_{mt}^c)$ to their averaged spread: $1/\sqrt{\pi_k \beta_k + \pi_{k'} \beta_{k'}}$. For each source $s_m^c$, $J$ is computed for each pair of adjacent Gaussians $kk'$ in $P(s_{mt}^c)$. The cluster $c$ is then split along $\mathbf{A}_m^c$ whenever $J > J_o$ ($J_o = 5$ was effective in our experiments). The second heuristic numerically computes the height of each "valley" in $P(s_{mt}^c)$ and compares it to the height of the shorter peak next to it. A ratio of 1:5 serves as the splitting criteria. These two heuristics worked well together in our experiments.

By the Central Limit Theorem, linearly mixing arbitrary sources of finite variances leads to a near-Gaussian density. Early in learning with $\mathbf{A}^c$ randomly initialized, $P(s_{mt}^c)$ was sometimes driven to a single Gaussian, especially when the sources to be learned contain some near Gaussian components. To counteract this, the sources were reinitialized when all but one of the Gaussians in $P(s_{mt}^c)$ died, keeping the $\mathbf{A}^c$ unchanged.

### 2.4 Choosing priors

In performing cluster analysis, the number of clusters required and the intrinsic dimension of each cluster can be obtained by comparing the marginal likelihood $P(\mathbf{X}|\mathcal{H})$ for different possible candidate models $\mathcal{H}$. For models of up to $C$ clusters and $M$ dimensions, there are $f(M, C)$ models to consider[1]. The Bayesian method provides another way to perform model selection. Expanding on the R.H.S. of (14) in Appendix B, we get

$$\begin{aligned}
\log P(\mathbf{X}) &\geq \int Q(\theta) \log P(\mathbf{X}|\theta)\, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \\
&= \langle \log P(\mathbf{X}|\theta) \rangle_Q - KL(Q(\theta), P(\theta)) \\
&= \langle \log \mathcal{L}(\theta; \mathbf{X}) \rangle_Q - \text{penality}\,.
\end{aligned}$$

Here $\mathcal{L}(\theta; \mathbf{X})$ is the data likelihood function. In contrast to the maximum likelihood, the Bayesian method automatically incorporates a penalty term rooted in a measurement of "distance" between

1. $f(M, C) = 1 + f(M-1, C) + f(M, C-1)$ and $f(M, 1) = M$, $f(1, C) = C$

the approximate posterior $Q(\theta)$ and the prior distribution $P(\theta)$. $Q(\theta)$ maximizes the averaged log likelihood $\langle \log P(\mathbf{X}|\theta) \rangle_Q$, while minimizing the penality. The set of parameters $\theta'$ not getting support from the data $\mathbf{X}$ (or equivalently contributing little towards $\langle \log \mathcal{L}(\theta; \mathbf{X}) \rangle_Q$) will result in $Q(\theta')$ close to $P(\theta')$. Hence, proper choice of $P(\theta)$ helps prevent overfitting and facilitates model selection.

We used non-informative priors, also know as maximum entropy priors (Jaynes, 1983), for the model parameters. The non-informative prior for a location parameter $\mu$, is uniform in $\mu$, i.e. $p(\mu) \propto 1$. This was approximated by a Gaussian with very large variance. The non-informative prior for a scale parameter $\sigma$ is uniform in $\log \sigma$, i.e. $p(\sigma) \propto 1/\sigma$. This transforms into $p(\Lambda) \propto 1/\Lambda$ for $\Lambda = 1/\sigma^2$. Finally, for proportion parameter $\boldsymbol{\rho} = \{\rho_1, \cdots, \rho_c, \cdots, \rho_C\}$, the non-informative prior is $p(\boldsymbol{\rho}) \propto 1/\Pi\rho_c$. It is interesting that all these priors are improper, i.e., they are not normalizable. They are approximated by limiting the Gaussian, gamma and Dirichlet distributions respectively (Appendix A).

### 2.5 Relationship to MAP

The *maximum a posterior* (MAP) solution arrives when $Q(\theta)$ is constrained to be the delta function $\delta(\theta - \theta')$ in the functional maximization of the log marginal likelihood lower bound (equation 6):

$$\begin{aligned} \max_{Q(\theta)=\delta(\theta-\theta')} \left[ \int Q(\theta) \log P(\mathbf{X}|\theta)\, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \right] & \\ \triangleq \max_{\theta'} \int \delta(\theta - \theta') \log \left[ P(\mathbf{X}|\theta) P(\theta) \right] d\theta & \\ = \max_{\theta'} \log \left[ P(\mathbf{X}|\theta') P(\theta') \right] . & \end{aligned}$$

Here $\triangleq$ denotes 'is equivalent to' since the entropy term $\int Q(\theta) \log Q(\theta)\, d\theta$ is constant and can be dropped when $Q(\theta) = \delta(\theta - \theta')$. Thus MAP further simplifies the approximation of $P(\theta|X)$ by $Q(\theta)$ to a point estimate. One advantage of variational Bayesian method over MAP is that the $Q(\theta)$ also carries information about the uncertainty in $\theta$, although the estimate could be rough due to the assumption that $Q(\theta)$ is separable. Notice that $\mathbf{\Lambda}$ is the precision parameter for the normal distribution, while $a$ and $\mathbf{d}$ are the precision parameters for the gamma and Dirichlet distributions respectively (equations 10–12). From the learning rules in Appendix C, these hyperparameters are roughly proportional to the number of data points supporting their corresponding parameters. Parameters receiving few votes from the data points will then have a low precision and high uncertainty. Besides, model comparison and averaging is not available in MAP since computation of $P(\mathbf{X})$ is deliberately avoided. If we start out with an oversized model, MAP would return the best parameter values under the Bayesian framework, but give no information on how well a selected model is over its alternatives. On the other hand, the lower bound to $P(\mathbf{X})$ in variational approach can be used to compare and combine models of different structures.

## 3. Experiments

In this section, we demonstrate the ability of the proposed Bayesian ICA clustering algorithm to model arbitrary source densities, reduce dimensionality and perform unsupervised clustering.

### 3.1 Sources of various non-symmetric densities

In this experiment, we mixed sources of various skewness and kurtosis: Laplacian, uniform, gamma, beta, generalized Gaussian ($\propto \exp(-|x|^q)$), and rectified generalized Gaussian, to form five clusters in a two dimensional space. The number of points in each cluster ranged from 200 to 400 and –50 dB noise was added to the data. The model was initialized with 8 to 10 clusters. On most trials a five cluster solution was obtained. Figure 2 shows the densities of the 10 sources recovered. For most of the sources densities, three Gaussians fitted the source histograms well. Discrepancies from the true distribution arose from randomness in samples generation and misestimation of variances in some difficult distributions. In particular, since data points at tail were assigned to other clusters, the estimated mean of the rectified super-exponential distribution in Figure 2b) was shifted away from its tail and hence the variance was under-estimated. Figure 3 shows the initial and final configurations. In Figure 4, the evolution of the lower bound of log marginal likelihood is plotted. Dips correspond to splitting of clusters, and large jumps correspond to vanishing of some clusters. The average signal to noise ratio (SNR) for the mixed sources was 9 dB and the SNR for the recovered sources was on average 38 dB.

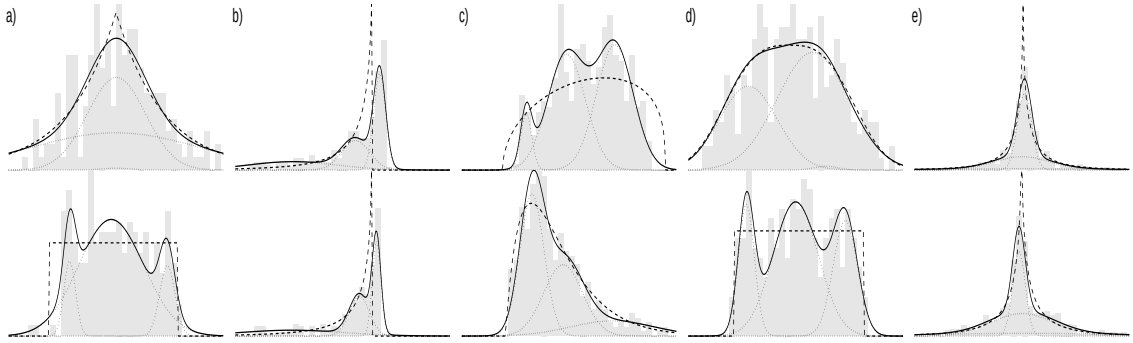


Figure 2: Source densities for the five clusters in the synthetic data experiment. Histograms: recovered sources distribution; dashed lines: original probability densities; solid line: mixture of Gaussians modeled probability densities; dotted lines: contribution of component Gaussians.

### 3.2 Dimensionality reduction

In this experiment, three clusters containing two, three and four sources respectively were embedded in a four dimensional space. The Laplacian, uniform, gamma, beta and generalized Gaussian distributions were again used as source densities. Each cluster had 250 data points and –33 dB noise. The correct number of clusters and their intrinsic dimension were obtained in all trials. The original and learned mixing matrix $\mathbf{A}$'s from one run are shown in Table 1. Both the columns of $\mathbf{A}$ and the corresponding rows of $s_{mt}^c$ were negligible values for 'killed' components. The signal to noise ratios (SNRs) for the mixed and recovered sources of each cluster are listed in Table 2. The average SNR for mixed and recovered sources were 5 dB and 22 dB respectively.

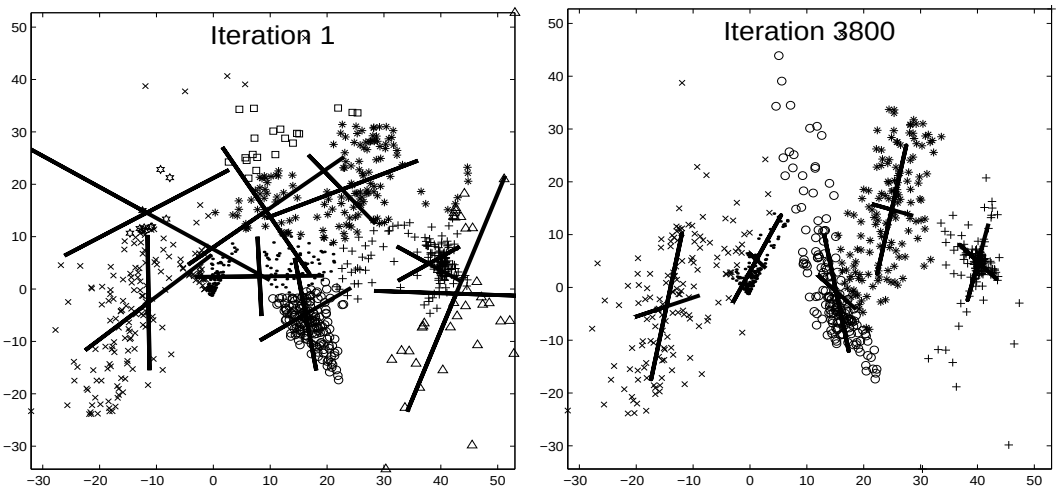


Figure 3: Initial and final configurations of one typical run in the synthetic data experiment.

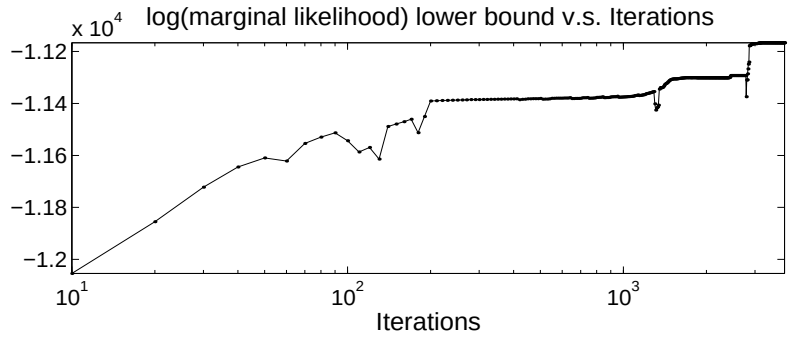


Figure 4: Evolution of $\mathcal{E}(\mathbf{X}, Q(\theta))$ as function of number of iterations for the sample run shown in Figure 3.

### 3.3 Medical data set analysis: Glaucoma

To evaluate the unsupervised classification ability of the Bayes learning algorithm on high dimensional data, we applied it to a glaucoma data set. Glaucoma is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field (Hitchings and Spaeth, 1977). Standard automated perimetry (SAP) is currently the visual function test most relied upon to measure visual function in glaucoma (Johnson, 1997). Automated threshold perimetry gives detailed quantitative data. A commonly used procedure worldwide is the full threshold SAP test of the Humphrey Visual Field Analyzer (HFA, Humphrey-Zeiss, Dublin, CA). Figure 5 shows part of a sample printout from the HFA. In the middle is the absolute visual field sensitivities (in dB) over the retina. On the right is a smoothed gray scale plot.

The data vector is composed of the 52 visual sensitivities over the visual field and the patient's age. The dataset included of 189 normal fields and 156 glaucomatous fields, as defined by the presence of glaucomatous optic neuropathy (GON). Supervised classification on the data by various

Table 1: Original and learned mixing matrix $\mathbf{A}$'s for the three clusters in experiment 2

| CLUSTER | ORIGINAL A | LEARNED A |
|---|---|---|
| 1 | $\begin{pmatrix} -3.0 & 2.0 & 0.1 & 0.0 \\ 2.0 & 2.0 & -3.0 & 3.0 \\ 0.0 & 3.0 & 1.0 & 2.0 \\ 1.0 & 1.0 & 0.5 & 0.0 \end{pmatrix}$ | $\begin{pmatrix} 3.04 & 0.11 & 0.23 & -1.99 \\ -1.95 & 2.56 & 3.28 & -1.80 \\ -0.15 & -1.26 & 2.36 & -2.63 \\ -0.96 & -0.72 & 0.14 & -0.91 \end{pmatrix}$ |
| 2 | $\begin{pmatrix} 2.0 & 2.0 & 3.0 \\ 1.0 & 3.0 & -1.0 \\ -3.0 & 0.0 & 2.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix}$ | $\begin{pmatrix} 2.51 & -0.00 & 2.72 & -1.93 \\ 2.88 & 0.00 & -1.30 & -0.76 \\ 0.20 & 0.00 & 1.96 & 2.85 \\ 1.18 & -0.00 & 0.88 & -0.95 \end{pmatrix}$ |
| 3 | $\begin{pmatrix} -3.0 & 2.0 \\ 2.0 & -3.0 \\ 1.0 & 3.0 \\ 2.0 & -4.0 \end{pmatrix}$ | $\begin{pmatrix} 3.15 & 1.80 & 0.00 & 0.00 \\ -2.25 & -2.85 & -0.00 & -0.00 \\ -0.71 & 3.04 & 0.00 & 0.00 \\ 1.62 & -4.13 & -0.00 & -0.00 \end{pmatrix}$ |

Table 2: Signal to noise ratio (SNR) of mixed and recovered sources in experiment 2

| CLUSTER | MIXTURE (dB) | RECOVERED (dB) |
|---|---|---|
| 1 | 6 5 1 2 | 21 18 24 19 |
| 2 | 5 8 3 | 31 21 19 |
| 3 | 5 11 | 25 24 |

machine learning classifiers was studied previously (Chan et al., 2002). Here we explore hidden structure in the data by fitting a compact density model. With the variational ICA clusters algorithm, we started with one cluster and look for 20 or less sources. The most stable solution consisted of two clusters after some splitting and deletions. Figure 6 shows the strength of the sources in the two clusters and the density distribution of the leading sources. It suggests 12 dimensions in cluster 1 and 6 dimensions in cluster 2. When matching the two clusters to the unseen label GON (cluster 1=glaucoma, 2=normal), we get a true positive rate (sensitivity) of 105/156=67% and a true negative rate (specificity) of 185/189=98%. The log marginal likelihood lower bound from this model was $-1.451 \times 10^4$.

We also performed unsupervised classification using a variational mixture of factor analysis (MFA) (Ghahramani and Beal, 2000, Attias, 1999b). This was done by setting $K$ to be one in the density model. Starting from different random initial conditions, solutions containing more than three clusters were obtained. In these solutions, the normal group was modeled by one cluster while the glaucoma group was divided into few clusters. The best classification had a sensitivity of 69%, a specificity of 97% with four clusters. However, a two cluster solution was preferred by the maximum $\mathcal{E}(\mathbf{X}, Q(\theta)|\mathcal{H})$ criteria. The best two cluster solution had a log marginal likelihood lower bound of $-1.459 \times 10^4$, sensitivity of 63% and a specificity of 98%.

Besides providing the raw measurement data, the Humphrey Field Analyzer (HFA) comes with a statistical analysis package that performs specialized statistical analyses related to diagnosing glaucoma. The purpose of these analyses is to aid the clinicians in interpretation of the visual field. One of the traditionally used index is glaucoma hemifield test (GHT). GHT yielded a sensitivity of

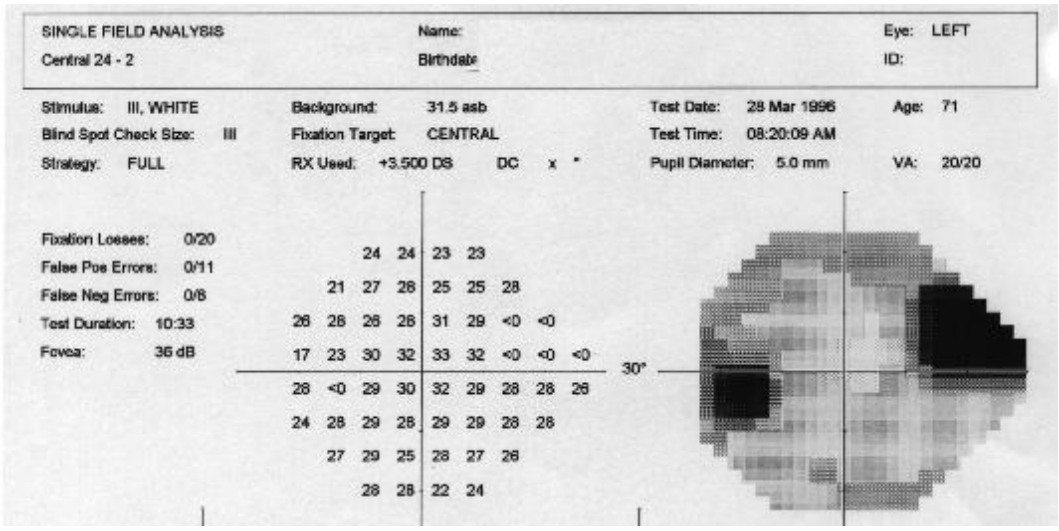


Figure 5: A sample of partial printout from the HFA showing absolute visual sensitivities and gray scale plot over the 54 locations on the retina. Measurement at two locations corresponding to the blind spot were excluded in the analysis.

67% and a specificity of 100% on our dataset. Specificity of $> 95\%$ is desired in the glaucoma community. The large difference between sensitivity and specificity occurs because the glaucoma class contains large number of 'normal looking' examples, while the normal class data is relatively pure.

### 3.4 Comparing to variational MFA

In the experiment on glaucoma data, the difference in log marginal likelihood lower bound, sensitivity and specificity between the variational mixture of ICA and MFA was not significant. They performed equally well in our glaucoma dataset in terms of density modeling and unsupervised classification. However, since MFA broke down the glaucoma class into smaller Gaussians, information about the intrinsic dimension of that cluster was lost. Instead, the variational mixture ICA located independent axes inside the glaucoma class. These axes were represented by the column vectors of mixing matrix $\mathbf{A}$. On the right of Figure 6 are the gray scale plots of values of column $\mathbf{A}_1$ mapped onto the retina for the two clusters. It is interesting to see that the first principal source for the glaucoma cluster indicates a depression in visual sensitivity of the lower fields, while the normal group shows a relatively uniform visual sensitivity. Contrast in visual sensitivities of the upper and lower hemifields of the retina is a common phenomenon found in glaucoma patients and is exploited by GHT to detect glaucoma. In Figure 7, we show the linear decomposition of a sample visual field into defects with different patterns by variational ICA. The identification of the independent patterns aids glaucoma experts in decomposing and generalizing visual fields losses seen in glaucoma patients. This would not be achievable with the MFA clustering. Although the true generative mechanism for dataset may not be linear as described by (1), (2) and (3), locating

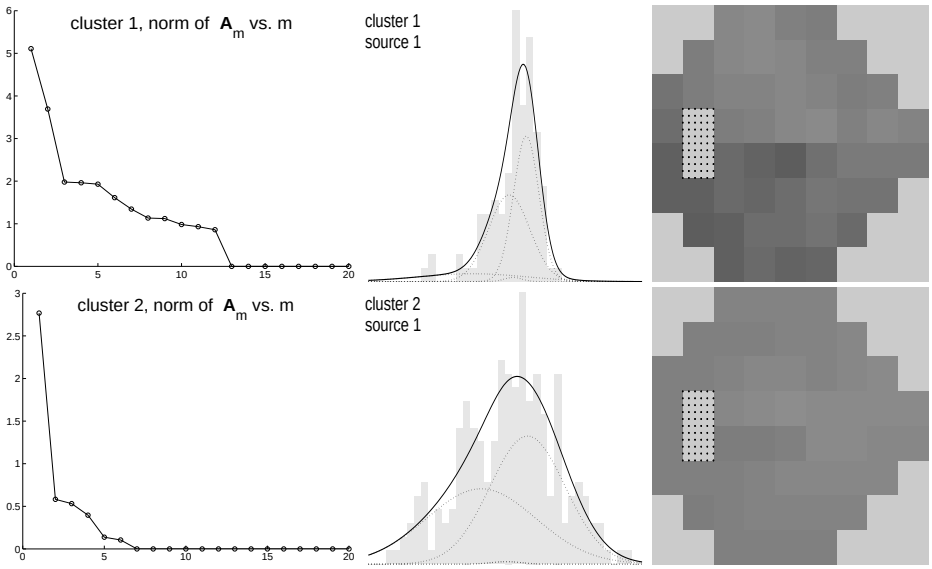


Figure 6: A 2 cluster solution for the glaucoma dataset. Cluster 1 corresponds to the GON group and cluster 2 corresponds to the normal group. Left: standard deviation ($\propto |\mathbf{A}_m|$) of the sources. It shows an intrinsic dimensions of 12 for cluster 1 and 6 for cluster 2. Center column: marginal density distributions for the first source of each cluster. Right: grey scale visual fields map of basic function $\mathbf{A}_1$.

non-trivial projections by ICA would enhance understanding and reveal interesting features in the data set.

## 4. Conclusion

In this paper, we derived the learning rules for the variational learning of mixtures of undercomplete non-symmetric ICA solution. Modeling independent source densities by a mixture of Gaussians is common. Here we extend the algorithm to the multi-clusters case and study its use in unsupervised classification. This is a generalization of Ghahramani and Beal (2000)'s variational learning of mixture of factor analysers, Miskin (2000) and Lappalainen (1999)'s ensemble learning of ICA and Attias (1999a)'s independent factor analysis. The proposed model was successfully applied to a glaucoma data set to identify hidden sources and perform unsupervised classification. The features of the visual fields discovered for the glaucoma data are supported by physiological evidence since they are commonly used by physicians to determine the disease.

Correctly identifying the number of sources in signal mixtures has always been an important and challenging issue. In particular, different numbers of components may be identified for different clusters. A common way to obtain undercomplete ICA solution is to perform complete ICA on PCA reduced data. Although some efficient methods (e.g. Amari, 1999) have been proposed for performing undercomplete ICA skipping PCA, there are no general guidelines on how many sources to expect. This paper exploits the automatic dimensionality reduction of the Bayesian method to

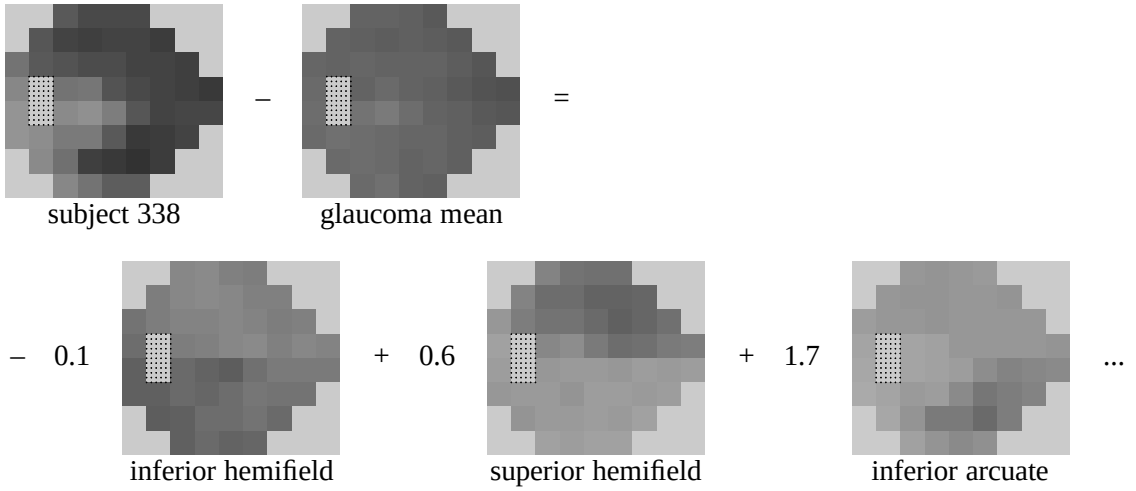


Figure 7: The mixture of ICA separates the visual field into linear combinations of independent patterns. Shown above is the decomposition of subject 338's visual field into defects of different areas.

identify the number of sources in undercomplete noisy ICA. The use of arbitrary source densities allows a flexible linear model for data densities fitting.

## Appendix A. The Density Model

Besides the mixture density (1) and (2), sources $\mathbf{s}_t^c$ (3) and the mixing matrix $\mathbf{A}^c$ (4), we employ the following priors on the parameters and hyper-parameters.

$$\begin{aligned}
P(\boldsymbol{\pi}_m^c) &= \mathcal{D}\left(\pi_{m1}^c, \cdots, \pi_{mK}^c | d_o(\pi_{m1}^c), \cdots, d_o(\pi_{mK}^c)\right) \\
P(\phi_{mk}^c) &= \mathcal{N}(\phi_{mk}^c | \mu_o(\phi_{mk}^c), \Lambda_o(\phi_{mk}^c)) \\
P(\beta_{mk}^c) &= \mathcal{G}(\beta_{mk}^c | a_o(\beta_{mk}^c), b_o(\beta_{mk}^c)) \\
P(\alpha_m^c) &= \mathcal{G}(\alpha_m^c | a_o(\alpha_m^c), b_o(\alpha_m^c))
\end{aligned}$$

$$\begin{aligned}
P(\nu_n^c) &= \mathcal{N}(\nu_n^c | \mu_o(\nu_n^c), \Lambda_o(\nu_n^c)) \\
P(\Psi_n^c) &= \mathcal{G}(\Psi_n^c | a_o(\Psi_n^c), b_o(\Psi_n^c)) \\
P(\boldsymbol{\rho}) &= \mathcal{D}(\rho_1, \cdots, \rho_C | d_o(\rho_1), \cdots, d_o(\rho_C)) \,.
\end{aligned}$$

Here $\mathcal{N}(\cdot)$, $\mathcal{G}(\cdot)$ and $\mathcal{D}(\cdot)$ are the normal, gamma and Dirichlet distributions respectively,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sqrt{\frac{|\boldsymbol{\Lambda}|}{(2\pi)^N}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})} \tag{10}$$

$$\mathcal{G}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{11}$$

$$\mathcal{D}(\boldsymbol{\pi}|\mathbf{d}) = \frac{\Gamma(\sum d_k)}{\prod \Gamma(d_k)} \pi_1^{d_1-1} \times \cdots \times \pi_K^{d_K-1} \,. \tag{12}$$

And we use the following values for the hyper-parameter in the priors: $\mu_o(\nu_n^c) = 0, \Lambda_o(\nu_n^c) = 0.001$, $d_o(\rho_c) = d_o(\pi_{mk}^c) = 0.001$, $\mu_o(\phi_{mk}^c) = 0, \Lambda_o(\phi_{mk}^c) = 1$, $a_o(\beta_{mk}^c) = 1.2, b_o(\beta_{mk}^c) = 0.1$, $a_o(\alpha_m^c) = b_o(\alpha_m^c) = 0.001$ and $a_o(\Psi_n^c) = b_o(\Psi_n^c) = 0.001$.

## Appendix B. Variational Bayesian Method

Taking log on both sides of (5) and introducing normalized density $Q(\theta)$, we have

$$\begin{aligned} \log \frac{P(\theta|\mathbf{X})}{Q(\theta)} + \log P(\mathbf{X}) &= \log \frac{P(\mathbf{X}|\theta)P(\theta)}{Q(\theta)} \\ \int Q(\theta) \log \frac{P(\theta|\mathbf{X})}{Q(\theta)}\, d\theta + \log P(\mathbf{X}) &= \int Q(\theta) \log \frac{P(\mathbf{X}|\theta)P(\theta)}{Q(\theta)}\, d\theta\,. \end{aligned} \tag{13}$$

The first term on the left is identified as the negative Kullback-Leibler Divergence between $Q(\theta)$ and true posterior $P(\theta|\mathbf{X})$,

$$KL(Q(\theta), P(\theta|\mathbf{X})) = \int Q(\theta) \log \frac{Q(\theta)}{P(\theta|\mathbf{X})}\, d\theta\,.$$

Functional minimization of $KL(Q(\theta), P(\theta|\mathbf{X}))$ provides an alternative way of analyzing the Bayes rule (5). Since $KL(Q(\theta), P(\theta|\mathbf{X})) \geq 0$, the Jensen's inequality:

$$\begin{aligned} \log P(\mathbf{X}) &= \log \int Q(\theta) \frac{P(\mathbf{X}|\theta)P(\theta)}{Q(\theta)}\, d\theta \\ &\geq \int Q(\theta) \log \frac{P(\mathbf{X}|\theta)P(\theta)}{Q(\theta)}\, d\theta \end{aligned} \tag{14}$$

is recovered by dropping it from (13). The lower bound of $\log P(\mathbf{X})$ from the Jensen's inequality can then be used as the objective function for solving $Q(\theta)$, which is an approximation to $P(\theta|\mathbf{X})$.

## Appendix C. Learning Rules

Using the separable posterior $Q(\theta)$ (equation 9) together with the posterior on the hidden variables $Q(c_t)$, $Q(\mathbf{s}_t^c)$ and $Q(k_{mt}^c)$, we perform functional maximization on the lower bound of the marginal likelihood (equations 7 and 8) to obtain the following recursive learning rules. Because of the choice of conjugate priors, free-form optimization results in the same form of $Q(\cdot)$ as $P(\cdot)$, but of different hyper-parameters. The only exception is $Q(\mathbf{s}_t^c)$:

$$\begin{aligned} Q(\mathbf{s}_t^c) &= \mathcal{N}(\mathbf{s}_t^c | \boldsymbol{\mu}(\mathbf{s}_t^c), \boldsymbol{\Lambda}(\mathbf{s}_t^c)) \\ \boldsymbol{\Lambda}(\mathbf{s}_t^c) &= \langle \mathbf{A}^{c\top} \boldsymbol{\Psi}^c \mathbf{A}^c \rangle + \mathrm{diag}(\textstyle\sum_k Q(k_{mt}^c) \langle \beta_{mk}^c \rangle) \\ \boldsymbol{\Lambda}(\mathbf{s}_t^c)\boldsymbol{\mu}(\mathbf{s}_t^c) &= \langle \mathbf{A}^{c\top} \boldsymbol{\Psi}^c (\mathbf{x}_t - \boldsymbol{\nu}^c) \rangle + \sum_k \begin{pmatrix} Q(k_{1t}^c) \langle \boldsymbol{\beta}_{1k}^c \phi_{1k}^c \rangle \\ \vdots \\ Q(k_{Mt}^c) \langle \boldsymbol{\beta}_{Mk}^c \phi_{Mk}^c \rangle \end{pmatrix} \end{aligned}$$

$$\begin{aligned} Q(\phi_{mk}^c) &= \mathcal{N}(\phi_{mk}^c | \mu(\phi_{mk}^c), \Lambda(\phi_{mk}^c)) \\ \Lambda(\phi_{mk}^c) &= \Lambda_o(\phi_{mk}^c) + \textstyle\sum_t Q(c_t) Q(k_{mt}^c) \langle \beta_{mk}^c \rangle \\ \mu(\phi_{mk}^c) &= [\Lambda_o(\phi_{mk}^c)\mu_o(\phi_{mk}^c) + \textstyle\sum_t Q(c_t) Q(k_{mt}^c) \langle \beta_{mk}^c s_{mt}^c \rangle] / \Lambda(\phi_{mk}^c) \end{aligned}$$

$$
\begin{aligned}
Q(\beta_{mk}^c) &= \mathcal{G}(\beta_{mk}^c | a(\beta_{mk}^c), b(\beta_{mk}^c)) \\
a(\beta_{mk}^c) &= a_o(\beta_{mk}^c) + \frac{1}{2}\sum_t Q(c_t) Q(k_{mt}^c) \\
b(\beta_{mk}^c) &= b_o(\beta_{mk}^c) + \frac{1}{2}\sum_t Q(c_t) Q(k_{mt}^c) \langle (s_{mt}^c - \phi_{mk}^c)^2 \rangle
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\pi}_m^c) &= \mathcal{D}(\boldsymbol{\pi}_m^c | \mathbf{d}(\boldsymbol{\pi}_m^c)) \\
d(\pi_{mk}^c) &= d_o(\pi_{mk}^c) + \textstyle\sum_t Q(c_t) Q(k_{mt}^c)
\end{aligned}
$$

$$
\begin{aligned}
Q(\mathbf{A}^c) &= \prod_n \mathcal{N}(\mathbf{A}_n^c | \boldsymbol{\mu}(\mathbf{A}_n^c), \boldsymbol{\Lambda}(\mathbf{A}_n^c)) \\
\boldsymbol{\Lambda}(\mathbf{A}_n^c) &= \mathrm{diag}(\langle \alpha_1^c \rangle, \cdots, \langle \alpha_m^c \rangle) + \textstyle\sum_t Q(c_t) \langle \Psi_n^c \rangle \langle \mathbf{s}_t^c \mathbf{s}_t^{c\top} \rangle \\
\boldsymbol{\mu}(\mathbf{A}_n^c) &= [\langle \Psi_n^c \rangle \textstyle\sum_t Q(c_t) \langle (x_{nt} - \nu_n) \mathbf{s}_t^{c\top} \rangle] (\boldsymbol{\Lambda}(\mathbf{A}_n^c))^{-1}
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\alpha}^c) &= \prod_m \mathcal{G}(\alpha_m^c | a(\alpha_m^c), b(\alpha_m^c)) \\
a(\alpha_m^c) &= a_o(\alpha_m^c) + \frac{N}{2} \\
b(\alpha_m^c) &= b_o(\alpha_m^c) + \frac{1}{2}\sum_n \langle A_{nm}^2 \rangle
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\nu}^c) &= \prod_n \mathcal{N}(\nu_n^c | \mu(\nu_n^c, \Lambda(\nu_n^c)) \\
\Lambda(\nu_n^c) &= \Lambda_o(\nu_n^c) + \textstyle\sum_t Q(c_t) \langle \Psi_n^c \rangle \\
\mu(\nu_n^c) &= [\Lambda_o(\nu_n^c) \mu_o(\nu_n^c) + \textstyle\sum_t Q(c_t) \langle (x_{nt} - \mathbf{A}_n^c \mathbf{s}_t^c) \Psi_n \rangle] / \Lambda(\nu_n^c)
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\Psi}^c) &= \prod_n \mathcal{G}(\Psi_n^c | a(\Psi_n^c), b(\Psi_n^c)) \\
a(\Psi_n^c) &= a_o(\Psi_n^c) + \frac{1}{2}\sum_t Q(c_t) \\
b(\Psi_n^c) &= b_o(\Psi_n^c) + \frac{1}{2}\sum_t Q(c_t) \langle (x_{nt} - \mathbf{A}_n^c \mathbf{s}_t^c - \nu_n^c)^2 \rangle
\end{aligned}
$$

$$
\begin{aligned}
Q(\boldsymbol{\rho}) &= \mathcal{D}(\boldsymbol{\rho} | \mathbf{d}(\boldsymbol{\rho})) \\
d(\rho_c) &= d_o(\rho_c) + \textstyle\sum_t Q(c_t) \,.
\end{aligned}
$$

$\langle \cdot \rangle$ denote the expectation of over the posterior distributions $Q(\cdot)$. Hidden variables distributions $Q(c_t)$ and $Q(k_{mt}^c)$ are given by

$$
\log Q(k_{mt}^c) = \langle \log \pi_{mk}^c \rangle + \langle \log \sqrt{\frac{\beta_{mk}^c}{2\pi}} \rangle - \frac{1}{2} \langle \beta_{mk}^c (s_{mt}^c - \phi_{mk}^c)^2 \rangle - \log z_{mt}^c \,;
$$

$$
\log Q(c_t) = \langle \log \rho_c \rangle + \langle \log P(\mathbf{x}_t | \mathbf{s}_t^c, \mathbf{A}^c, \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) \rangle - \langle \log Q(\mathbf{s}_t^c) \rangle + \textstyle\sum_m \log z_{mt}^c - \log Z_t \,. \quad (15)
$$

where $z_{mt}^c$ and $Z_t$ are the normalization constants.

## References

S.-I. Amari. Natural gradient learning for over- and undercomplete bases in ICA. *Neural Computation*, 11(8):1875–1883, Nov 1999.

Hagai Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999a.

Hagai Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30, 1999b.

Christopher M. Bishop. Variational PCA. In *Proc. Ninth Int. Conf. on Artificial Neural Networks*, pages 509–514. ICANN, 1999.

Kwokleung Chan, Te-Won Lee, Pamela A. Sample, Michael Goldbaum, Robert N. Weinreb, and Terrence J. Sejnowski. Comparison of machine learning and traditional classifier in glaucoma diagnosis. *IEEE Transactions on Biomedical Engineering*, 2002. (accepted).

Kwokleung Chan, Te-Won Lee, and Terrence J Sejnowski. Variational learning of clusters of undercomplete nonsymmetric independent components. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 492–497, San Diego, Dec. 09-12 2001.

Rizwan A. Choudrey and Stephen J. Roberts. Variational mixture of Bayesian independent component analysers. Technical Report PARG-01-04, Department of Engineering Science, University of Oxford, 2001.

Zoubin Ghahramani and Matthew J. Beal. Variational inference for Bayesian mixtures of factor analysers. In S. Solla, Todd K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, 2000.

R. A. Hitchings and G. L. Spaeth. The optic disc in glaucoma, II: correlation of appearance of the optic disc with the visual field. *Br J Ophthalmol.*, 61:107–113, 1977.

Pedro A.d.F.R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14(4):889–918, 2002.

Aapo Hyvarinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. J. Wiley, New York, 2001.

Edwin T. Jaynes. *E.T. Jaynes : papers on probability, statistics, and statistical physics*. Kluwer, Boston, 1983.

C. A. Johnson. Perimetry and visual field testing. In K. Zadnik, editor, *The Ocular Examination: Measurements and Findings*. W.B. Saunders, Philadelphia, 1997.

Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Tzyy-Ping Jung, Scott Makeig, Martin J. McKeown, Anthony Bell, Te-Won Lee, and Terrence J. Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–22, 2001.

J. Karvanen, J. Eriksson, and V Koivunen. Maximum likelihood estimation of ICA model for wide class of source distributions. In B. Widrow, L. Guan, K. Paliwa, T. Adali, J. Larsen, E. Wilson, and S Douglas, editors, *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, volume 1, pages 445–454, Piscataway, NJ, USA, Dec. 11- 13, 2000. IEEE.

H. Lappalainen. Ensemble learning for independent component analysis. In *International Workshop on ICA and Blind Signal Separation*, pages 7–12, Aussois, Jan. 11-15 1999.

T-W. Lee, M. S. Lewicki, and T. J. Sejnowski. ICA mixture models for unsupervised classification with non-Gaussian sources and automatic context switching in blind signal separation. *IEEE Transactions on Pattern Recognition and Machine Learning*, 22(10):1–12, Oct 2000.

James Miskin. Ensemble learning for independent component analysis. Ph.D. thesis, Department of Physics, University of Cambridge, UK, Jun. 2000.

Èric Moulines, Jean-Franşois Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proceedings of the ICASSP'97*, volume 5, pages 3617–3620, Munich, Germany, 1997.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, New York, 1996.

B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Max Welling and Markus Weber. A constrained EM algorithm for independent component analysis. *Neural Computation*, 13(3):677–689, 2001.