LETTER **Communicated by Erkki Oja**

Variational Bayesian Learning of ICA with Missing Data

Kwokleung Chan
kwchan@salk.edu
Computational Neurobiology Laboratory, Salk Institute, La Jolla, CA 92037, U.S.A.

Te-Won Lee
tewon@salk.edu
Institute for Neural Computation, University of California at San Diego, La Jolla, CA 92093, U.S.A.

Terrence J. Sejnowski
terry@salk.edu
Computational Neurobiology Laboratory, Salk Institute, La Jolla, CA 92037, U.S.A., and Department of Biology, University of California at San Diego, La Jolla, CA 92093, U.S.A.

Missing data are common in real-world data sets and are a problem for many estimation techniques. We have developed a variational Bayesian method to perform independent component analysis (ICA) on high-dimensional data containing missing entries. Missing data are handled naturally in the Bayesian framework by integrating the generative density model. Modeling the distributions of the independent sources with mixture of gaussians allows sources to be estimated with different kurtosis and skewness. Unlike the maximum likelihood approach, the variational Bayesian method automatically determines the dimensionality of the data and yields an accurate density model for the observed data without overfitting problems. The technique is also extended to the clusters of ICA and supervised classification framework.

1 Introduction

Data density estimation is an important step in many machine learning problems. Often we are faced with data containing incomplete entries. The data may be missing due to measurement or recording failure. Another frequent cause is difficulty in collecting complete data. For example, it could be expensive and time-consuming to perform some biomedical tests. Data scarcity is not uncommon, and it would be very undesirable to discard those data points with missing entries when we already have a small data set. Traditionally, missing data are filled in by mean imputation or regression imputation during preprocessing. This could introduce biases into the data

 © 2003 Massachusetts Institute of Technology

cloud density and adversely affect subsequent analysis. A more principled way would be to use probability density estimates of the missing entries instead of point estimates. A well-known example of this approach is the use of the expectation-maximization (EM) algorithm in fitting incomplete data with a single gaussian density (Little & Rubin, 1987).

Independent component analysis (ICA; Hyvarinen, Karhunen, & Oja, 2001) assumes the observed data $\mathbf{x}$ are generated from a linear combination of independent sources $\mathbf{s}$:

$$\mathbf{x} = \mathbf{A}\,\mathbf{s} + \boldsymbol{\nu}, \tag{1.1}$$

where $\mathbf{A}$ is the mixing matrix, which can be nonsquare. The sources $\mathbf{s}$ have nongaussian density such as $p(s_l) \propto \exp(-|s_l|^q)$. The noise term $\boldsymbol{\nu}$ can have nonzero mean. ICA tries to locate independent axes within the data cloud and was developed for blind source separation. It has been applied to speech separation and analyzing fMRI and EEG data (Jung et al., 2001). ICA is also used to model data density, describing data as linear mixtures of independent features and finding projections that may uncover interesting structure in the data. Maximum likelihood learning of ICA with incomplete data has been studied by Welling and Weber (1999) in the limited case of a square mixing matrix and predefined source densities.

Many real-world data sets have intrinsic dimensionality smaller than that of the observed data. With missing data, principal component analysis cannot be used to perform dimension reduction as preprocessing for ICA. Instead, the variational Bayesian method applied to ICA can handle small data sets with high observed dimension (Chan, Lee, & Sejnowski, 2002; Choudrey & Roberts, 2001; Miskin, 2000). The Bayesian method prevents overfitting and performs automatic dimension reduction. In this article, we extend the variational Bayesian ICA method to problems with missing data. More important, the probability density estimate of the missing entries can be used to fill in the missing values. This allows the density model to be refined and made more accurate.

2 Model and Theory

2.1 ICA Generative Model with Missing Data. Consider a data set of T data points in an N-dimensional space: $\mathbf{X} = \{\mathbf{x}_t \in \mathcal{R}^N\}$, t in $\{1, \ldots, T\}$. Assume a noisy ICA generative model for the data,

$$P(\mathbf{x}_t \mid \theta) = \int \mathcal{N}(\mathbf{x}_t \mid \mathbf{A}\mathbf{s}_t + \boldsymbol{\nu}, \boldsymbol{\Psi}) P(\mathbf{s}_t \mid \theta_s)\, d\mathbf{s}_t, \tag{2.1}$$

where $\mathbf{A}$ is the mixing matrix and $\boldsymbol{\nu}$ and $[\boldsymbol{\Psi}]^{-1}$ are the observation mean and diagonal noise variance, respectively. The hidden source $\mathbf{s}_t$ is assumed

to have L dimensions. Similar to the independent factor analysis of Attias (1999), each component of $\mathbf{s}_t$ will be modeled by a mixture of K gaussians to allow for source densities of various kurtosis and skewness,

$$P(\mathbf{s}_t \mid \theta_s) = \prod_l^L \left(\sum_{k_l}^K \pi_{lk_l} \mathcal{N}\left(\mathbf{s}_t(l) \mid \phi_{lk_l}, \beta_{lk_l}\right) \right). \tag{2.2}$$

Split each data point into a missing part and an observed part: $\mathbf{x}_t^\top = (\mathbf{x}_t^{o\top}, \mathbf{x}_t^{m\top})$. In this article, we consider only the random missing case (Ghahramani & Jordan, 1994), that is, the probability for the missing entries $\mathbf{x}_t^m$ is independent of the value of $\mathbf{x}_t^m$, but could depend on the value of $\mathbf{x}_t^o$. The likelihood of the data set is then defined to be

$$\mathcal{L}(\theta; \mathbf{X}) = \prod_t P(\mathbf{x}_t^o \mid \theta), \tag{2.3}$$

where

$$\begin{aligned} P(\mathbf{x}_t^o \mid \theta) &= \int P(\mathbf{x}_t \mid \theta)\, d\mathbf{x}_t^m \\ &= \int \left[\int \mathcal{N}(\mathbf{x}_t \mid \mathbf{A}\mathbf{s}_t + \boldsymbol{\nu}, \boldsymbol{\Psi})\, d\mathbf{x}_t^m \right] P(\mathbf{s}_t \mid \theta_s)\, d\mathbf{s}_t \\ &= \int \mathcal{N}(\mathbf{x}_t^o \mid [\mathbf{A}\mathbf{s}_t + \boldsymbol{\nu}]_t^o, [\boldsymbol{\Psi}]_t^o) P(\mathbf{s}_t \mid \theta_s)\, d\mathbf{s}_t. \end{aligned} \tag{2.4}$$

Here we have introduced the notation $[\cdot]_t^o$, which means taking only the observed dimensions (corresponding to the tth data point) of whatever is inside the square brackets. Since equation 2.4 is similar to equation 2.1, the variational Bayesian ICA (Chan et al., 2002; Choudrey & Roberts, 2001; Miskin, 2000) can be extended naturally to handle missing data, but only if care is taken in discounting missing entries in the learning rules.

2.2 Variational Bayesian Method. In a full Bayesian treatment, the posterior distribution of the parameters θ is obtained by

$$P(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta) P(\theta)}{P(\mathbf{X})} = \frac{\prod_t P(\mathbf{x}_t^o \mid \theta) P(\theta)}{P(\mathbf{X})}, \tag{2.5}$$

where $P(\mathbf{X})$ is the marginal likelihood and given as

$$P(\mathbf{X}) = \int \prod_t P(\mathbf{x}_t^o \mid \theta) P(\theta)\, d\theta. \tag{2.6}$$

The ICA model for $P(\mathbf{X})$ is defined with the following priors on the parameters $P(\theta)$,

$$P(A_{nl}) = \mathcal{N}(A_{nl} \mid 0, \alpha_l) \qquad P(\boldsymbol{\pi}_l) = \mathcal{D}(\boldsymbol{\pi}_l \mid \mathbf{d}_o(\boldsymbol{\pi}_l))$$
$$P(\alpha_l) = \mathcal{G}(\alpha_l \mid a_o(\alpha_l), b_o(\alpha_l)) \qquad P(\phi_{lk_l}) = \mathcal{N}(\phi_{lk_l} \mid \mu_o(\phi_{lk_l}), \Lambda_o(\phi_{lk_l})) \quad (2.7)$$
$$P(\beta_{lk_l}) = \mathcal{G}(\beta_{lk_l} \mid a_o(\beta_{lk_l}), b_o(\beta_{lk_l}))$$
$$P(\nu_n) = \mathcal{N}(\nu_n \mid \mu_o(\nu_n), \Lambda_o(\nu_n)) \quad P(\Psi_n) = \mathcal{G}(\Psi_n \mid a_o(\Psi_n), b_o(\Psi_n)), \quad (2.8)$$

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

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sqrt{\frac{|\boldsymbol{\Lambda}|}{(2\pi)^N}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})}; \quad (2.9)$$

$$\mathcal{G}(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}; \quad (2.10)$$

$$\mathcal{D}(\boldsymbol{\pi} \mid \mathbf{d}) = \frac{\Gamma(\sum d_k)}{\prod \Gamma(d_k)} \pi_1^{d_1-1} \times \cdots \times \pi_K^{d_K-1}. \quad (2.11)$$

Here $a_o(\cdot)$, $b_o(\cdot)$, $\mathbf{d}_o(\cdot)$, $\mu_o(\cdot)$, and $\Lambda_o(\cdot)$ are prechosen hyperparameters for the priors. Notice that $\boldsymbol{\Lambda}$ in the normal distribution is an inverse covariance parameter.

Under the variational Bayesian treatment, instead of performing the integration in equation 2.6 to solve for $P(\theta \mid \mathbf{X})$ directly, we approximate it by $Q(\theta)$ and opt to minimize the Kullback-Leibler distance between them (Mackay, 1995; Jordan, Ghahramani, Jaakkola, & Saul, 1999):

$$\begin{aligned} -KL(Q(\theta) \mid P(\theta \mid \mathbf{X})) &= \int Q(\theta) \log \frac{P(\theta \mid \mathbf{X})}{Q(\theta)} \, d\theta \\ &= \int Q(\theta) \left[\sum_t \log P(\mathbf{x}_t^o \mid \theta) + \log \frac{P(\theta)}{Q(\theta)} \right] d\theta \\ &\quad - \log P(\mathbf{X}). \end{aligned} \quad (2.12)$$

Since $-KL(Q(\theta) \mid P(\theta \mid \mathbf{X})) \leq 0$, we get a lower bound for the log marginal likelihood,

$$\log P(\mathbf{X}) \geq \int Q(\theta) \sum_t \log P(\mathbf{x}_t^o \mid \theta) \, d\theta + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} \, d\theta, \quad (2.13)$$

which can also be obtained by applying Jensen's inequality to equation 2.6. $Q(\theta)$ is then solved by functional maximization of the lower bound. A sep-

arable approximate posterior $Q(\theta)$ will be assumed:

$$Q(\theta) = Q(\boldsymbol{\nu})Q(\boldsymbol{\Psi}) \times Q(\mathbf{A})Q(\boldsymbol{\alpha}) \times \prod_l \left[Q(\boldsymbol{\pi}_l) \prod_{k_l} Q(\phi_{lk_l})Q(\beta_{lk_l}) \right]. \tag{2.14}$$

The second term in equation 2.13, which is the negative Kullback-Leibler divergence between approximate posterior $Q(\theta)$ and prior $P(\theta)$, is then expanded as

$$\begin{aligned}
&\int Q(\theta) \log \frac{P(\theta)}{Q(\theta)}\, d\theta \\
&= \sum_l \int Q(\boldsymbol{\pi}_l) \log \frac{P(\boldsymbol{\pi}_l)}{Q(\boldsymbol{\pi}_l)}\, d\boldsymbol{\pi}_l \\
&\quad + \sum_{l\,k_l} \int Q(\phi_{lk_l}) \log \frac{P(\phi_{lk_l})}{Q(\phi_{lk_l})}\, d\phi_{lk_l} + \sum_{l\,k_l} \int Q(\beta_{lk_l}) \log \frac{P(\beta_{lk_l})}{Q(\beta_{lk_l})}\, d\beta_{lk_l} \\
&\quad + \int\int Q(\mathbf{A})Q(\boldsymbol{\alpha}) \log \frac{P(\mathbf{A} \mid \boldsymbol{\alpha})}{Q(\mathbf{A})}\, d\mathbf{A}\, d\boldsymbol{\alpha} + \int Q(\boldsymbol{\alpha}) \log \frac{P(\boldsymbol{\alpha})}{Q(\boldsymbol{\alpha})}\, d\boldsymbol{\alpha} \\
&\quad + \int Q(\boldsymbol{\nu}) \log \frac{P(\boldsymbol{\nu})}{Q(\boldsymbol{\nu})}\, d\boldsymbol{\nu} + \int Q(\boldsymbol{\Psi}) \log \frac{P(\boldsymbol{\Psi})}{Q(\boldsymbol{\Psi})}\, d\boldsymbol{\Psi}.
\end{aligned} \tag{2.15}$$

2.3 Special Treatment for Missing Data. Thus far, the analysis follows almost exactly that of the variational Bayesian ICA on complete data, except that $P(\mathbf{x}_t \mid \theta)$ is replaced by $P(\mathbf{x}_t^o \mid \theta)$ in equation 2.6, and consequently the missing entries are discounted in the learning rules. However, it would be useful to obtain $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$, that is, the approximate distribution on the missing entries, which is given by

$$Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o) = \int Q(\theta) \int \mathcal{N}(\mathbf{x}_t^m \mid [\mathbf{A}\mathbf{s}_t + \boldsymbol{\nu}]_t^m, [\boldsymbol{\Psi}]_t^m) Q(\mathbf{s}_t)\, d\mathbf{s}_t\, d\theta\,. \tag{2.16}$$

As noted by Welling and Weber (1999), elements of $\mathbf{s}_t$ given $\mathbf{x}_t^o$ are dependent. More important, under the ICA model, $Q(\mathbf{s}_t)$ is unlikely to be a single gaussian. This is evident from Figure 1, which shows the probability density functions of the data $\mathbf{x}$ and hidden variable $\mathbf{s}$. The inserts show the sample data in the two spaces. Here the hidden sources assume density of $P(s_l) \propto \exp(-|s_l|^{0.7})$. They are mixed noiselessly to give $P(\mathbf{x})$ in the upper graph. The cut in the upper graph represents $P(x_1 \mid x_2 = -0.5)$, which transforms into a highly correlated and nongaussian $P(\mathbf{s} \mid x_2 = -0.5)$.

Unless we are interested in only the first- and second-order statistics of $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$, we should try to capture as much structure as possible of

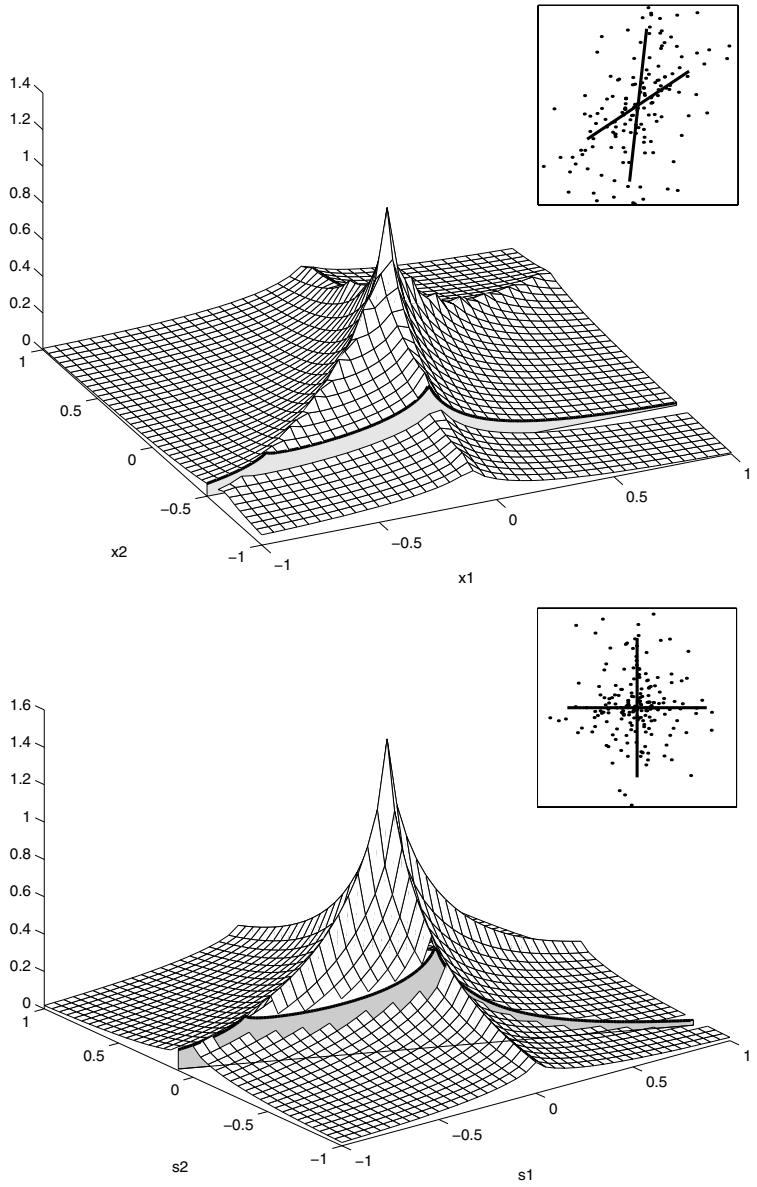

Figure 1: Probability density functions for the data $\mathbf{x}$ (top) and hidden sources $\mathbf{s}$ (bottom). Inserts show the sample data in the two spaces. The "cuts" show $P(x_1 \mid x_2 = -0.5)$ and $P(\mathbf{s} \mid x_2 = -0.5)$.

$P(\mathbf{s}_t \mid \mathbf{x}_t^o)$ in $Q(\mathbf{s}_t)$. In this article, we take a slightly different route from Chan et al. (2002) or Choudrey and Roberts (2001) when performing variational Bayesian learning. First, we break down $P(\mathbf{s}_t)$ into a mixture of K^L gaussians in the L-dimensional $\mathbf{s}$ space:

$$\begin{aligned} P(\mathbf{s}_t) &= \prod_l^L \left(\sum_{k_l} \pi_{lk_l} \mathcal{N}(\mathbf{s}_t(l) \mid \phi_{lk_l} \beta_{lk_l}) \right) \\ &= \sum_{k_1} \cdots \sum_{k_L} [\pi_{1k_1} \times \cdots \times \pi_{Lk_L} \\ &\qquad \times \mathcal{N}(\mathbf{s}_t(1) \mid \phi_{1k_1} \beta_{1k_1}) \times \cdots \times \mathcal{N}(\mathbf{s}_t(L) \mid \phi_{Lk_L} \beta_{Lk_L})] \\ &= \sum_{\mathbf{k}} \pi_{\mathbf{k}} \mathcal{N}(\mathbf{s}_t \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}}). \end{aligned} \tag{2.17}$$

Here we have defined $\mathbf{k}$ to be a vector index. The "$\mathbf{k}$th" gaussian is centered at $\boldsymbol{\phi}_{\mathbf{k}}$, of inverse covariance $\boldsymbol{\beta}_{\mathbf{k}}$, in the source $\mathbf{s}$ space,

$$\begin{aligned} \mathbf{k} &= (k_1, \ldots, k_l, \ldots, k_L)^\top, \quad k_l = 1, \ldots, K \\ \boldsymbol{\phi}_{\mathbf{k}} &= (\phi_{1k_1}, \ldots, \phi_{lk_l}, \ldots, \phi_{Lk_L})^\top \\ \boldsymbol{\beta}_{\mathbf{k}} &= \begin{pmatrix} \beta_{1k_1} & & \\ & \ddots & \\ & & \beta_{Lk_L} \end{pmatrix} \\ \pi_{\mathbf{k}} &= \pi_{1k_1} \times \cdots \times \pi_{Lk_L}. \end{aligned} \tag{2.18}$$

Log likelihood for $\mathbf{x}_t^o$ is then expanded using Jensen's inequality,

$$\begin{aligned} \log P(\mathbf{x}_t^o \mid \theta) &= \log \int P(\mathbf{x}_t^o \mid \mathbf{s}_t, \theta) \sum_{\mathbf{k}} \pi_{\mathbf{k}} \mathcal{N}(\mathbf{s}_t \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}}) \, d\mathbf{s}_t \\ &= \log \sum_{\mathbf{k}} \pi_{\mathbf{k}} \int P(\mathbf{x}_t^o \mid \mathbf{s}_t, \theta) \, \mathcal{N}(\mathbf{s}_t \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}}) \, d\mathbf{s}_t \\ &\geq \sum_{\mathbf{k}} Q(\mathbf{k}_t) \log \int P(\mathbf{x}_t^o \mid \mathbf{s}_t, \theta) \mathcal{N}(\mathbf{s}_t \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}}) \, d\mathbf{s}_t \\ &\quad + \sum_{\mathbf{k}} Q(\mathbf{k}_t) \log \frac{\pi_{\mathbf{k}}}{Q(\mathbf{k}_t)}. \end{aligned} \tag{2.19}$$

Here, $Q(\mathbf{k}_t)$ is a short form for $Q(\mathbf{k}_t = \mathbf{k})$. $\mathbf{k}_t$ is a discrete hidden variable, and $Q(\mathbf{k}_t = \mathbf{k})$ is the probability that the tth data point belongs to the $\mathbf{k}$th gaussian. Recognizing that $\mathbf{s}_t$ is just a dummy variable, we introduce $Q(\mathbf{s}_{\mathbf{k}t})$,

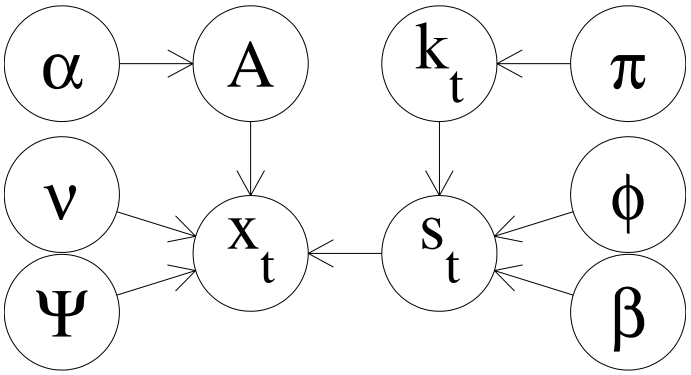

Figure 2: A simplified directed graph for the generative model of variational ICA. $\mathbf{x}_t$ is the observed variable, $\mathbf{k}_t$ and $\mathbf{s}_t$ are hidden variables, and the rest are model parameters. The $\mathbf{k}_t$ indicates which of the K^L expanded gaussians generated $\mathbf{s}_t$.

apply Jensen's inequality again, and get

$$\log P(\mathbf{x}_t^o \mid \theta) \geq \sum_{\mathbf{k}} Q(\mathbf{k}_t) \left[\int Q(\mathbf{s}_{\mathbf{k}t}) \log P(\mathbf{x}_t^o \mid \mathbf{s}_{\mathbf{k}t}, \theta)\, d\mathbf{s}_{\mathbf{k}t} + \int Q(\mathbf{s}_{\mathbf{k}t}) \log \frac{\mathcal{N}(\mathbf{s}_{\mathbf{k}t} \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}})}{Q(\mathbf{s}_{\mathbf{k}t})}\, d\mathbf{s}_{\mathbf{k}t} \right] + \sum_{\mathbf{k}} Q(\mathbf{k}_t) \log \frac{\pi_{\mathbf{k}}}{Q(\mathbf{k}_t)}. \tag{2.20}$$

Substituting $\log P(\mathbf{x}_t^o \mid \theta)$ back into equation 2.13, the variational Bayesian method can be continued as usual. We have drawn in Figure 2 a simplified graphical representation for the generative model of variational ICA. $\mathbf{x}_t$ is the observed variable, $\mathbf{k}_t$ and $\mathbf{s}_t$ are hidden variables, and the rest are model parameters, where $\mathbf{k}_t$ indicates which of the K^L expanded gaussians generated $\mathbf{s}_t$.

3 Learning Rules

Combining equations 2.13, 2.15, and 2.20, we perform functional maximization on the lower bound of the log marginal likelihood, $\log P(\mathbf{X})$, with regard to $Q(\theta)$ (see equation 2.14), $Q(\mathbf{k}_t)$ and $Q(\mathbf{s}_{\mathbf{k}t})$ (see equation 2.20)—for example,

$$\log Q(\boldsymbol{\nu}) = \log P(\boldsymbol{\nu}) + \int Q(\theta^{\setminus \nu}) \sum_t \log P(\mathbf{x}_t^o \mid \theta)\, d\theta^{\setminus \nu} + \text{const.}, \tag{3.1}$$

where $\theta^{\backslash\nu}$ is the set of parameters excluding $\boldsymbol{\nu}$. This gives

$$
\begin{aligned}
Q(\boldsymbol{\nu}) &= \prod_n \mathcal{N}(\nu_n \mid \mu(\nu_n), \Lambda(\nu_n)) \\
\Lambda(\nu_n) &= \Lambda_o(\nu_n) + \langle \Psi_n \rangle \sum_t o_{nt} \\
\mu(\nu_n) &= \frac{\Lambda_o(\nu_n)\mu_o(\nu_n) + \langle \Psi_n \rangle \sum_t o_{nt} \sum_{\mathbf{k}} Q(\mathbf{k}_t) \langle (x_{nt} - \mathbf{A}_{n\cdot}\mathbf{s}_{\mathbf{k}t}) \rangle}{\Lambda(\nu_n)}.
\end{aligned} \tag{3.2}
$$

Similarly,

$$
\begin{aligned}
Q(\boldsymbol{\Psi}) &= \prod_n \mathcal{G}(\Psi_n \mid a(\Psi_n), b(\Psi_n)) \\
a(\Psi_n) &= a_o(\Psi_n) + \frac{1}{2} \sum_t o_{nt} \\
b(\Psi_n) &= b_o(\Psi_n) + \frac{1}{2} \sum_t o_{nt} \sum_{\mathbf{k}} Q(\mathbf{k}_t) \langle (x_{nt} - \mathbf{A}_{n\cdot}\mathbf{s}_{\mathbf{k}t} - \nu_n)^2 \rangle.
\end{aligned} \tag{3.3}
$$

$$
\begin{aligned}
Q(\mathbf{A}) &= \prod_n \mathcal{N}(\mathbf{A}_{n\cdot} \mid \boldsymbol{\mu}(\mathbf{A}_{n\cdot}), \boldsymbol{\Lambda}(\mathbf{A}_{n\cdot})) \\
\boldsymbol{\Lambda}(\mathbf{A}_{n\cdot}) &= \begin{pmatrix} \langle \alpha_1 \rangle & & \\ & \ddots & \\ & & \langle \alpha_L \rangle \end{pmatrix} + \langle \Psi_n \rangle \sum_t o_{nt} \sum_{\mathbf{k}} Q(\mathbf{k}_t) \langle \mathbf{s}_{\mathbf{k}t} \mathbf{s}_{\mathbf{k}t}^\top \rangle \\
\boldsymbol{\mu}(\mathbf{A}_{n\cdot}) &= \left(\langle \Psi_n \rangle \sum_t o_{nt} (x_{nt} - \langle \nu_n \rangle) \sum_{\mathbf{k}} Q(\mathbf{k}_t) \langle \mathbf{s}_{\mathbf{k}t}^\top \rangle \right) \boldsymbol{\Lambda}(\mathbf{A}_{n\cdot})^{-1}.
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
Q(\boldsymbol{\alpha}) &= \prod_l \mathcal{G}(\alpha_l \mid a(\alpha_l), b(\alpha_l)) \\
a(\alpha_l) &= a_o(\alpha_l) + \frac{N}{2} \\
b(\alpha_l) &= b_o(\alpha_l) + \frac{1}{2} \sum_n \langle A_{nl}^2 \rangle.
\end{aligned} \tag{3.5}
$$

$$
\begin{aligned}
Q(\boldsymbol{\pi}_l) &= \mathcal{D}(\boldsymbol{\pi} \mid \mathbf{d}(\boldsymbol{\pi}_l)) \\
d(\pi_{lk}) &= d_o(\pi_{lk}) + \sum_t \sum_{\mathbf{k}_l = k} Q(\mathbf{k}_t).
\end{aligned} \tag{3.6}
$$

$$
\begin{aligned}
Q(\phi_{lk_l}) &= \mathcal{N}(\phi_{lk_l} \mid \mu(\phi_{lk_l}), \Lambda(\phi_{lk_l})) \\
\Lambda(\phi_{lk_l}) &= \Lambda_o(\phi_{lk_l}) + \langle \beta_{lk_l} \rangle \sum_t \sum_{\mathbf{k}_l = k} Q(\mathbf{k}_t) \\
\mu(\phi_{lk_l}) &= \frac{\Lambda_o(\phi_{lk_l})\mu_o(\phi_{lk_l}) + \langle \beta_{lk_l} \rangle \sum_t \sum_{\mathbf{k}_l = k} Q(\mathbf{k}_t) \langle s_{\mathbf{k}t}(l) \rangle}{\Lambda(\phi_{lk_l})}.
\end{aligned} \tag{3.7}
$$

$$
\begin{aligned}
Q(\beta_{lk_l}) &= \mathcal{G}(\beta_{lk_l} \mid a(\beta_{lk_l}), b(\beta_{lk_l})) \\
a(\beta_{lk_l}) &= a_o(\beta_{lk_l}) + \frac{1}{2} \sum_t \sum_{\mathbf{k}_l = k} Q(\mathbf{k}_t) \\
b(\beta_{lk_l}) &= b_o(\beta_{lk_l}) + \frac{1}{2} \sum_t \sum_{\mathbf{k}_l = k} Q(\mathbf{k}_t) \langle (\mathbf{s}_{\mathbf{k}t}(l) - \phi_{lk_l})^2 \rangle.
\end{aligned} \tag{3.8}
$$

$$
\begin{aligned}
Q(\mathbf{s}_{\mathbf{k}t}) &= \mathcal{N}(\mathbf{s}_{\mathbf{k}t} \mid \boldsymbol{\mu}(\mathbf{s}_{\mathbf{k}t}), \boldsymbol{\Lambda}(\mathbf{s}_{\mathbf{k}t})) \\
\boldsymbol{\Lambda}(\mathbf{s}_{\mathbf{k}t}) &= \begin{pmatrix} \langle \beta_{1\mathbf{k}_1} \rangle & & \\ & \ddots & \\ & & \langle \beta_{L\mathbf{k}_L} \rangle \end{pmatrix} \\
&\quad + \left\langle \mathbf{A}^\top \begin{pmatrix} o_{1t}\Psi_1 & & \\ & \ddots & \\ & & o_{Nt}\Psi_N \end{pmatrix} \mathbf{A} \right\rangle \\
\boldsymbol{\Lambda}(\mathbf{s}_{\mathbf{k}t})\boldsymbol{\mu}(\mathbf{s}_{\mathbf{k}t}) &= \begin{pmatrix} \langle \beta_{1\mathbf{k}_1} \phi_{1\mathbf{k}_1} \rangle \\ \vdots \\ \langle \beta_{L\mathbf{k}_L} \phi_{L\mathbf{k}_L} \rangle \end{pmatrix} \\
&\quad + \left\langle \mathbf{A}^\top \begin{pmatrix} o_{1t}\Psi_1 & & \\ & \ddots & \\ & & o_{Nt}\Psi_N \end{pmatrix} (\mathbf{x}_t - \boldsymbol{\nu}) \right\rangle.
\end{aligned} \tag{3.9}
$$

In the above equations, $\langle \cdot \rangle$ denotes the expectation over the posterior distributions $Q(\cdot)$. $\mathbf{A}_{n\cdot}$ is the nth row of the mixing matrix $\mathbf{A}$, $\sum_{\mathbf{k}_l = k}$ means picking out those gaussians such that the lth element of their indices $\mathbf{k}$ has the value of k, and $\mathbf{o}_t$ is an indicator variable for observed entries in $\mathbf{x}_t$:

$$
o_{nt} = \begin{cases} 1, & \text{if } x_{nt} \text{ is observed} \\ 0, & \text{if } x_{nt} \text{ is missing} \end{cases}. \tag{3.10}
$$

For a model of equal noise variance among all the observation dimensions, the summation in the learning rules for $Q(\boldsymbol{\Psi})$ would be over both t and

n. Note that there exists scale and translational degeneracy in the model, as given by equation 2.1 and 2.2. After each update of $Q(\boldsymbol{\pi}_l)$, $Q(\phi_{lk_l})$, and $Q(\beta_{lk_l})$, it is better to rescale $P(\mathbf{s}_t(l))$ to have zero mean and unit variance. $Q(\mathbf{s}_{\mathbf{k}t})$, $Q(\mathbf{A})$, $Q(\boldsymbol{\alpha})$, $Q(\boldsymbol{\nu})$, and $Q(\boldsymbol{\Psi})$ have to be adjusted correspondingly. Finally, $Q(\mathbf{k}_t)$ is given by

$$\begin{aligned}\log Q(\mathbf{k}_t) = {} & \langle \log P(\mathbf{x}_t^o \mid \mathbf{s}_{\mathbf{k}t}, \theta) \rangle + \langle \log \mathcal{N}(\mathbf{s}_{\mathbf{k}t} \mid \boldsymbol{\phi}_{\mathbf{k}}, \boldsymbol{\beta}_{\mathbf{k}}) \rangle \\ & - \langle \log Q(\mathbf{s}_{\mathbf{k}t}) \rangle + \langle \log \boldsymbol{\pi}_{\mathbf{k}} \rangle - \log z_t, \end{aligned} \tag{3.11}$$

where z_t is a normalization constant. The lower bound $\mathcal{E}(\mathbf{X}, Q(\theta))$ for the log marginal likelihood, computed using equations 2.13, 2.15, and 2.20, can be monitored during learning and used for comparison of different solutions or models. After some manipulation, $\mathcal{E}(\mathbf{X}, Q(\theta))$ can be expressed as

$$\mathcal{E}(\mathbf{X}, Q(\theta)) = \sum_t \log z_t + \int Q(\theta) \log \frac{P(\theta)}{Q(\theta)} \, d\theta. \tag{3.12}$$

4 Missing Data

4.1 Filling in Missing Entries. Recovering missing values while performing demixing is possible if we have $N > L$. More specifically, if the number of observed dimensions in $\mathbf{x}_t$ is greater than L, the equation

$$\mathbf{x}_t^o = [\mathbf{A}]_t^o \cdot \mathbf{s}_t \tag{4.1}$$

would be overdetermined in $\mathbf{s}_t$ unless $[\mathbf{A}]_t^o$ has a rank smaller than L. In this case, $Q(\mathbf{s}_t)$ is likely to be unimodal and peaked, point estimates of $\mathbf{s}_t$ would be sufficient and reliable, and the learning rules of Chan et al. (2002), with small modification to account for missing entries, would give a reasonable approximation. When $Q(\mathbf{s}_t)$ is a single gaussian, the exponential growth in complexity is avoided. However, if the number of observed dimensions in $\mathbf{x}_t$ is less than L, equation 4.1 is now underdetermined in $\mathbf{s}_t$, and $Q(\mathbf{s}_t)$ would have a broad, multimodal structure. This corresponds to overcomplete ICA where single gaussian approximation of $Q(\mathbf{s}_t)$ is undesirable and the formalism discussed in this article is needed to capture the higher-order statistics of $Q(\mathbf{s}_t)$ and produce a more faithful $Q(\mathbf{x}_t^m \mid x_t^o)$. The approximate distribution $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ can be obtained by

$$Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o) = \sum_{\mathbf{k}} Q(\mathbf{k}_t) \int \delta(\mathbf{x}_t^m - \mathbf{x}_{\mathbf{k}t}^m) Q(\mathbf{x}_{\mathbf{k}t}^m \mid \mathbf{x}_t^o, \mathbf{k}) \, d\mathbf{x}_{\mathbf{k}t}^m, \tag{4.2}$$

where $\delta(\cdot)$ is the delta function, and

$$Q(\mathbf{x}_{\mathbf{k}t}^m \mid \mathbf{x}_t^o, \mathbf{k}) = \int Q(\theta) \int \mathcal{N}(\mathbf{x}_{\mathbf{k}t}^m \mid [\mathbf{A}\mathbf{s}_{\mathbf{k}t} + \boldsymbol{\nu}]_t^m, [\boldsymbol{\Psi}]_t^m) Q(\mathbf{s}_{\mathbf{k}t})\, d\mathbf{s}_{\mathbf{k}t}\, d\theta$$

$$= \int\int Q(\mathbf{A}) Q(\boldsymbol{\Psi}) \mathcal{N}(\mathbf{x}_{\mathbf{k}t}^m \mid \boldsymbol{\mu}(\mathbf{x}_{\mathbf{k}t}^m), \boldsymbol{\Lambda}(\mathbf{x}_{\mathbf{k}t}^m))\, d\mathbf{A}\, d\boldsymbol{\Psi} \tag{4.3}$$

$$\boldsymbol{\mu}(\mathbf{x}_{\mathbf{k}t}^m) = [\mathbf{A}\boldsymbol{\mu}(\mathbf{s}_{\mathbf{k}t}) + \boldsymbol{\mu}(\boldsymbol{\nu})]_t^m \tag{4.4}$$

$$\boldsymbol{\Lambda}(\mathbf{x}_{\mathbf{k}t}^m)^{-1} = [\mathbf{A}\boldsymbol{\Lambda}(\mathbf{s}_{\mathbf{k}t})^{-1}\mathbf{A}^\top + \boldsymbol{\Lambda}(\boldsymbol{\nu})^{-1} + \mathrm{diag}(\boldsymbol{\Psi})^{-1}]_t^m. \tag{4.5}$$

Unfortunately, the integration over $Q(\mathbf{A})$ and $Q(\boldsymbol{\Psi})$ cannot be carried out analytically, but we can substitute $\langle \mathbf{A} \rangle$ and $\langle \boldsymbol{\Psi} \rangle$ as an approximation. Estimation of $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ using the above equations is demonstrated in Figure 3. The shaded area is the exact posterior $P(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ for the noiseless mixing in Figure 1 with observed $x_2 = -2$ and the solid line is the approximation by equations 4.2 through 4.5. We have modified the variational ICA of Chan et al. (2002) by discounting missing entries. This is done by replacing $\sum_t$

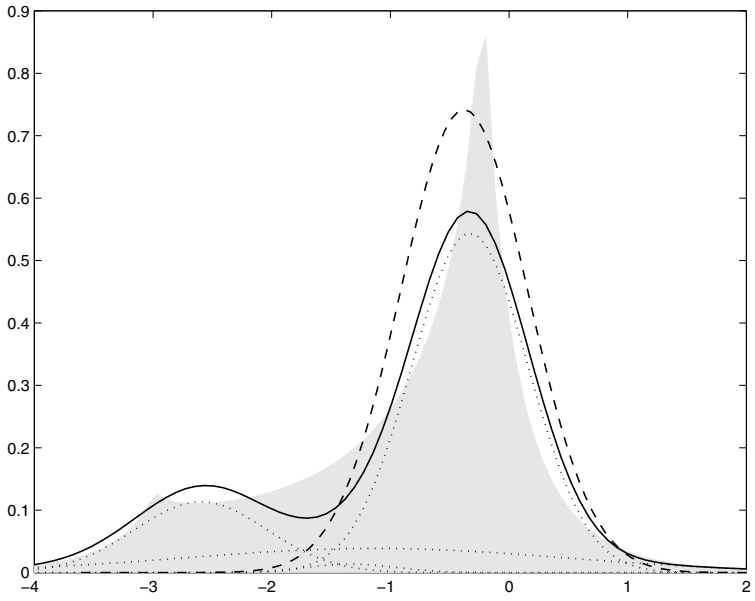

Figure 3: The approximation of $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ from the full missing ICA (solid line) and the polynomial missing ICA (dashed line). The shaded area is the exact posterior $P(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ corresponding to the noiseless mixture in Figure 1 with observed $x_2 = -2$. Dotted lines are the contribution from the individual $Q(\mathbf{x}_{\mathbf{k}t}^m \mid \mathbf{x}_t^o, \mathbf{k})$.

with $\sum_t o_{nt}$ and Ψ_n with $o_{nt}\Psi_n$ in their learning rules. The dashed line is the approximation $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ from this modified method, which we refer to as *polynomial missing ICA*. The treatment of fully expanding the K^L hidden source gaussians discussed in section 2.3 is named *full missing ICA*. The full missing ICA gives a more accurate fit for $P(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ and a better estimate for $\langle \mathbf{x}_t^m \mid \mathbf{x}_t^o \rangle$. From equation 2.16,

$$Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o) = \int Q(\theta) \int \mathcal{N}(\mathbf{x}_t^m \mid [\mathbf{A}\mathbf{s}_t + \boldsymbol{\nu}]_t^m, [\boldsymbol{\Psi}]_t^m) Q(\mathbf{s}_t)\, d\mathbf{s}_t\, d\theta, \tag{4.6}$$

and the above formalism, $Q(\mathbf{s}_t)$, becomes

$$Q(\mathbf{s}_t) = \sum_{\mathbf{k}} Q(\mathbf{k}_t) \int \delta(\mathbf{s}_t - \mathbf{s}_{\mathbf{k}t}) Q(\mathbf{s}_{\mathbf{k}t})\, d\mathbf{s}_{\mathbf{k}t}, \tag{4.7}$$

which is a mixture of K^L gaussians. The missing values can then be filled in by

$$\langle \mathbf{s}_t \mid \mathbf{x}_t^o \rangle = \int \mathbf{s}_t Q(\mathbf{s}_t)\, d\mathbf{s}_t = \sum_{\mathbf{k}} Q(\mathbf{k}_t) \boldsymbol{\mu}(\mathbf{s}_{\mathbf{k}t}) \tag{4.8}$$

$$\begin{aligned} \langle \mathbf{x}_t^m \mid \mathbf{x}_t^o \rangle &= \int \mathbf{x}_t^m Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)\, d\mathbf{x}_t^m \\ &= \sum_{\mathbf{k}} Q(\mathbf{k}_t) \boldsymbol{\mu}(\mathbf{x}_{\mathbf{k}t}^m) = [\mathbf{A}]_t^m \langle \mathbf{s}_t \mid \mathbf{x}_t^o \rangle + [\boldsymbol{\mu}(\boldsymbol{\nu})]_t^m, \end{aligned} \tag{4.9}$$

where $\boldsymbol{\mu}(\mathbf{s}_{\mathbf{k}t})$ and $\boldsymbol{\mu}(\mathbf{x}_{\mathbf{k}t}^m)$ are given in equations 3.9 and 4.4. Alternatively, a maximum a posterior (MAP) estimate on $Q(\mathbf{s}_t)$ and $Q(\mathbf{x}_t^m \mid \mathbf{x}_t^o)$ may be obtained, but then numerical methods are needed.

4.2 The "Full" and "Polynomial" Missing ICA. The complexity of the full variational Bayesian ICA method is proportional to $T \times K^L$, where T is the number of data points, L is the number of hidden sources assumed, and K is the number of gaussians used to model the density of each source. If we set $K = 2$, the five parameters in the source density model $P(\mathbf{s}_t(l))$ are already enough to model the mean, variance, skewness, and kurtosis of the source distribution. The full missing ICA should always be preferred if memory and computational time permit. The "polynomial missing ICA" converges more slowly per epoch of learning rules and suffers from many more local maxima. It has an inferior marginal likelihood lower bound. The problems are more serious at high missing data rates, and a local maximum solution is usually found instead. In the full missing ICA, $Q(\mathbf{s}_t)$ is a mixture of gaussians. In the extreme case, when all entries of a data point are missing, that is, empty $\mathbf{x}_t^o$, $Q(\mathbf{s}_t)$ is the same as $P(\mathbf{s}_t \mid \theta)$ and would not interfere with the learning of $P(\mathbf{s}_t \mid \theta)$ from other data point. On the other hand, the single gaussian $Q(\mathbf{s}_t)$ in the polynomial missing ICA would drive $P(\mathbf{s}_t \mid \theta)$ to become gaussian too. This is very undesirable when learning ICA structure.

5 Clusters of ICA

The variational Bayesian ICA for missing data described above can be easily extended to model data density with C clusters of ICA. First, all parameters θ and hidden variables $\mathbf{k}_t$, $s_{\mathbf{k}t}$ for each cluster are given a superscript index c. Parameter $\boldsymbol{\rho} = \{\rho^1, \ldots, \rho^C\}$ is introduced to represent the weights on the clusters. $\boldsymbol{\rho}$ has a Dirichlet prior (see equation 2.11). $\Theta = \{\boldsymbol{\rho}, \theta^1, \ldots, \theta^C\}$ is now the collection of all parameters. Our density model in equation 2.1 becomes

$$\begin{aligned} P(\mathbf{x}_t \mid \Theta) &= \sum_c P(c_t = c \mid \boldsymbol{\rho}) P(\mathbf{x}_t \mid \theta^c) \\ &= \sum_c P(c_t = c \mid \boldsymbol{\rho}) \int \mathcal{N}(\mathbf{x}_t \mid \mathbf{A}^c \mathbf{s}_t^c + \boldsymbol{\nu}^c, \boldsymbol{\Psi}^c) P(\mathbf{s}_t^c \mid \theta_s^c)\, d\mathbf{s}_t^c. \end{aligned} \tag{5.1}$$

The objective function in equation 2.13 remains the same, but with θ replaced by Θ. The separable posterior $Q(\Theta)$ is given by

$$Q(\Theta) = Q(\boldsymbol{\rho}) \prod_c Q(\theta^c) \tag{5.2}$$

and similar to equation 2.15,

$$\begin{aligned} \int Q(\Theta) \log \frac{P(\Theta)}{Q(\Theta)}\, d\Theta = &\int Q(\boldsymbol{\rho}) \log \frac{P(\boldsymbol{\rho})}{Q(\boldsymbol{\rho})}\, d\boldsymbol{\rho} \\ &+ \sum_c \int Q(\theta^c) \log \frac{P(\theta^c)}{Q(\theta^c)}\, d\theta^c. \end{aligned} \tag{5.3}$$

Equation 2.20 now becomes,

$$\begin{aligned} \log P(\mathbf{x}_t^o \mid \Theta) \geq &\sum_c Q(c_t) \log \frac{P(c_t)}{Q(c_t)} + \sum_{c,\mathbf{k}} Q(c_t) Q(\mathbf{k}_t^c) \\ &\times \left[\int Q(\mathbf{s}_{\mathbf{k}t}^c) \log P(\mathbf{x}_t^o \mid \mathbf{s}_{\mathbf{k}t}^c, \theta^c)\, d\mathbf{s}_{\mathbf{k}t}^c \right. \\ &\left. + \int Q(\mathbf{s}_{\mathbf{k}t}^c) \log \frac{\mathcal{N}(\mathbf{s}_{\mathbf{k}t}^c \mid \boldsymbol{\phi}_{\mathbf{k}}^c, \boldsymbol{\beta}_{\mathbf{k}}^c)}{Q(\mathbf{s}_{\mathbf{k}t}^c)}\, d\mathbf{s}_{\mathbf{k}t}^c \right] \\ &+ \sum_{c,\mathbf{k}} Q(c_t) Q(\mathbf{k}_t^c) \log \frac{\pi_{\mathbf{k}}^c}{Q(\mathbf{k}_t^c)}. \end{aligned} \tag{5.4}$$

We have introduced one more hidden variable c_t, and $Q(c_t)$ is to be interpreted in the same fashion as $Q(\mathbf{k}_t^c)$. All learning rules in section 3 remain

the same, only with $\sum_t$ replaced by $\sum_t Q(c_t)$. Finally, we need two more learning rules,

$$d(\rho^c) = d_o(\rho^c) + \sum_t Q(c_t) \tag{5.5}$$

$$\log Q(c_t) = \langle \log \rho^c \rangle + \log z_t^c - \log Z_t, \tag{5.6}$$

where z_t^c is the normalization constant for $Q(\mathbf{k}_t^c)$ (see equation 3.11) and Z_t is for normalizing $Q(c_t)$.

6 Supervised Classification

It is generally difficult for discriminative classifiers such as multilayer perceptron (Bishop, 1995) or support vector machine (Vapnik, 1998) to handle missing data. In this section, we extend the variational Bayesian technique to supervised classification.

Consider a data set $(\mathbf{X}_T, \mathbf{Y}_T) = \{(\mathbf{x}_t, y_t), t \text{ in } (1, \ldots, T)\}$. Here, $\mathbf{x}_t$ contains the input attributes and may have missing entries. $y_t \in \{1, \ldots, y, \ldots, Y\}$ indicates which of the Y classes $\mathbf{x}_t$ is associated with. When given a new data point $\mathbf{x}_{T+1}$, we would like to compute $P(y_{T+1} \mid \mathbf{x}_{T+1}, \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})$,

$$\begin{aligned} &P(y_{T+1} \mid \mathbf{x}_{T+1}, \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M}) \\ &\quad = \frac{P(\mathbf{x}_{T+1} \mid y_{T+1}, \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M}) P(y_{T+1} \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})}{P(\mathbf{x}_{T+1} \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})}. \end{aligned} \tag{6.1}$$

Here $\mathcal{M}$ denotes our generative model for observation $\{\mathbf{x}_t, y_t\}$:

$$P(\mathbf{x}_t, y_t \mid \mathcal{M}) = P(\mathbf{x}_t \mid y_t, \mathcal{M}) P(y_t \mid \mathcal{M}). \tag{6.2}$$

$P(\mathbf{x}_t \mid y_t, \mathcal{M})$ could be a mixture model as given by equation 5.1.

6.1 Learning of Model Parameters. Let $P(\mathbf{x}_t \mid y_t, \mathcal{M})$ be parameterized by Θ_y and $P(y_t \mid \mathcal{M})$ be parameterized by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_Y)$,

$$P(\mathbf{x}_t \mid y_t = y, \mathcal{M}) = P(\mathbf{x}_t \mid \Theta_y) \tag{6.3}$$

$$P(y_t \mid \mathcal{M}) = P(y_t = y \mid \boldsymbol{\omega}) = \omega_y. \tag{6.4}$$

If $\boldsymbol{\omega}$ is given a Dirichlet prior, $P(\boldsymbol{\omega} \mid \mathcal{M}) = \mathcal{D}(\boldsymbol{\omega} \mid \mathbf{d}_o(\boldsymbol{\omega}))$, its posterior has also a Dirichlet distribution:

$$P(\boldsymbol{\omega} \mid \mathbf{Y}_T, \mathcal{M}) = \mathcal{D}(\boldsymbol{\omega} \mid \mathbf{d}(\boldsymbol{\omega})) \tag{6.5}$$

$$d(\omega_y) = d_o(\omega_y) + \sum_t I(y_t = y). \tag{6.6}$$

$I(\cdot)$ is an indicator function that equals 1 if its argument is true and 0 otherwise.

Under the generative model of equation 6.2, it can be shown that

$$P(\Theta_y \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M}) = P(\Theta_y \mid \mathbf{X}_y), \tag{6.7}$$

where $\mathbf{X}_y$ is a subset of $\mathbf{X}_T$ but contains only those $\mathbf{x}_t$ whose training labels y_t have value y. Hence, $P(\Theta_y \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})$ can be approximated with $Q(\Theta_y)$ by applying the learning rules in sections 3 and 5 on subset $\mathbf{X}_y$.

6.2 Classification. First, $P(y_{T+1} \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})$ in equation 6.1 can be computed by

$$\begin{aligned} P(y_{T+1} = y \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M}) &= \int P(y_{T+1} = y \mid \omega_y) P(\omega_y \mid \mathbf{X}_T, \mathbf{Y}_T)\, d\omega_y \\ &= \frac{d(\omega_y)}{\sum_y d(\omega_y)}. \end{aligned} \tag{6.8}$$

The other term $P(\mathbf{x}_{T+1} \mid y_{T+1}, \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})$ can be computed as

$$\begin{aligned} &\log P(\mathbf{x}_{T+1} \mid y_{T+1} = y, \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M}) \\ &\quad = \log P(\mathbf{x}_{T+1} \mid \mathbf{X}_y, \mathcal{M}) \\ &\quad = \log P(\mathbf{x}_{T+1}, \mathbf{X}_y \mid \mathcal{M}) - \log P(\mathbf{X}_y \mid \mathcal{M}) \qquad (6.9) \\ &\quad \approx \mathcal{E}(\{\mathbf{x}_{T+1}, \mathbf{X}_y\}, Q'(\Theta_y)) - \mathcal{E}(\mathbf{X}_y, Q(\Theta_y)). \qquad (6.10) \end{aligned}$$

The above requires adding $\mathbf{x}_{T+1}$ to $\mathbf{X}_y$ and iterating the learning rules to obtain $Q'(\Theta_y)$ and $\mathcal{E}(\{\mathbf{x}_{T+1}, \mathbf{X}_y\}, Q'(\Theta_y))$. The error in the approximation is the difference $KL(Q'(\Theta_y), P(\Theta_y \mid \{\mathbf{x}_{T+1}, \mathbf{X}_y\})) - KL(Q(\Theta_y), P(\Theta_y \mid \mathbf{X}_y))$. If we assume further that $Q'(\Theta_y) \approx Q(\Theta_y)$,

$$\begin{aligned} \log P(\mathbf{x}_{T+1} \mid \mathbf{X}_y, \mathcal{M}) &\approx \int Q(\Theta_y) \log P(\mathbf{x}_{T+1} \mid \Theta_y)\, d\Theta_y \\ &= \log Z_{T+1}, \end{aligned} \tag{6.11}$$

where Z_{T+1} is the normalization constant in equation 5.6.

7 Experiment

7.1 Synthetic Data. In the first experiment, 200 data points were generated by mixing four sources randomly in a seven-dimensional space. The generalized gaussian, gamma, and beta distributions were used to represent source densities of various skewness and kurtosis (see Figure 5). Noise

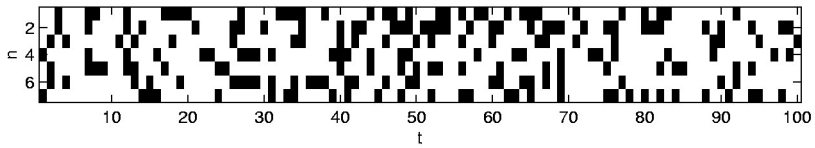

Figure 4: In the first experiment, 30% of the entries in the seven-dimensional data set are missing as indicated by the black entries. (The first 100 data points are shown.)

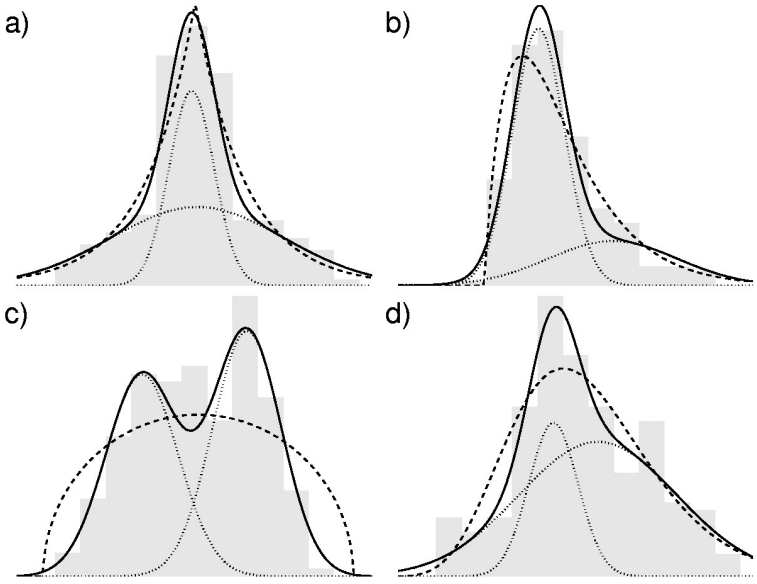

Figure 5: Source density modeling by variational missing ICA of the synthetic data. Histograms: recovered sources distribution; dashed lines: original probability densities; solid line: mixture of gaussians modeled probability densities; dotted lines: individual gaussian contribution.

at -26 dB level was added to the data, and missing entries were created with a probability of 0.3. The data matrix for the first 100 data points is plotted in Figure 4. Dark pixels represent missing entries. Notice that some data points have fewer than four observed dimensions. In Figure 5, we plotted the histograms of the recovered sources and the probability density functions (pdf) of the four sources. The dashed line is the exact pdf used to generate the data, and the solid line is the modeled pdf by mixture of two one-dimensional gaussians (see equation 2.2). This shows that the two gaussians gave adequate fit to the source histograms and densities.

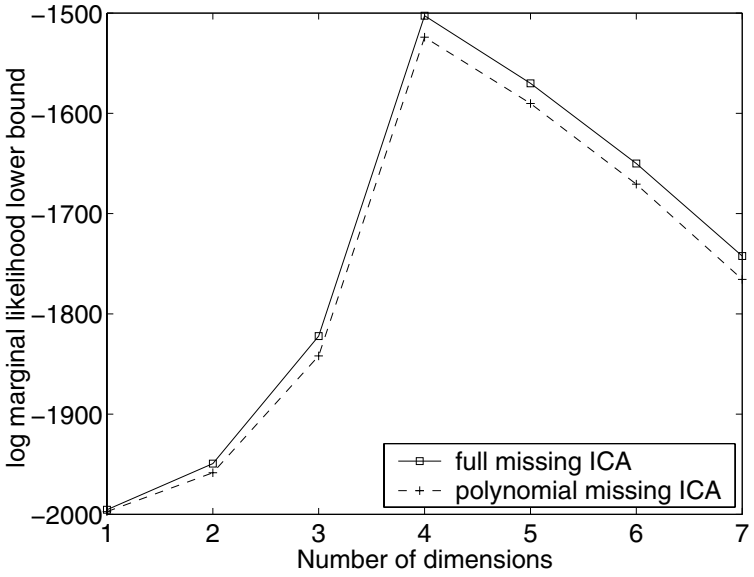

Figure 6: $\mathcal{E}(\mathbf{X}, Q(\theta))$ as a function of hidden source dimensions. *Full missing ICA* refers to the full expansions of gaussians discussed in section 2.3, and *polynomial missing ICA* refers to the Chan et al. (2002) method with minor modification.

Figure 6 plots the lower bound of log marginal likelihood (see equation 3.12) for models assuming different numbers of intrinsic dimensions. As expected, the Bayesian treatment allows us to the infer the intrinsic dimension of the data cloud. In the figure, we also plot the $\mathcal{E}(\mathbf{X}, Q(\theta))$ from the polynomial missing ICA. Since a less negative lower bound represents a smaller Kullback-Leibler divergence between $Q(\theta)$ and $P(\mathbf{X} \mid \theta)$, it is clear from the figure that the full missing ICA gave a better fit to the data density.

7.2 Mixing Images. This experiment demonstrates the ability of the proposed method to fill in missing values while performing demixing. This is made possible if we have more mixtures than hidden sources, or $N > L$. The top row in Figure 7 shows the two original 380×380 pixel images. They were linearly mixed into three images, and -20 dB noise was added. Missing entries were introduced randomly with probability 0.2. The denoised mixtures are shown in the third row of Figure 7, and the recovered sources are in the bottom row. Only 0.8% of the pixels were missing from all three mixed images and could not be recovered; 38.4% of the pixels were missing from only one mixed image, and their values could be filled in with low

Figure 7: A demonstration of recovering missing values when $N > L$. The original images are in the top row. Twenty percent of the pixels in the mixed images (second row) are missing at random. Only 0.8% are missing from the denoised mixed images (third row) and separated images (bottom).

uncertainty; and 9.6% of the pixels were missing from any two of the mixed images. Estimation of their values is possible but would have high uncertainty. From Figure 7, we can see that the source images were well separated and the mixed images were nicely denoised. The signal-to-noise ratio (SNR) in the separated images was 14 dB. We have also tried filling in the missing pixels by EM with a gaussian model. Variational Bayesian ICA was then applied on the "completed" data. The SNR achieved in the unmixed images was 5 dB. This supports that it is crucial to have the correct density model when filling in missing values and important to learn the density model and missing values concurrently. The denoised mixed images in this example were meant only to illustrate the method visually. However, if (x_1, x_2, x_3) represent cholesterol, blood sugar, and uric acid level, for example, it would be possible to fill in the third when only two are available.

7.3 Survival Prediction. We demonstrate the supervised classification discussed in section 6 with an echocardiogram data set downloaded from the UCI Machine Learning Repository (Blake & Merz, 1998). Input variables are *age-at-heart-attack*, *fractional-shortening*, *epss*, *lvdd*, and *wall-motion-index*. The goal is to predict survival of the patient one year after heart attack. There are 24 positive and 50 negative examples. The data matrix has a missing rate of 5.4%. We performed leave-one-out cross-validation to evaluate our classifier. Thresholding the output $P(y_{T+1} \mid \mathbf{X}_T, \mathbf{Y}_T, \mathcal{M})$, computed using equation 6.10, at 0.5, we got a true positive rate of 16/24 and a true negative rate of 42/50.

8 Conclusion

In this article, we derived the learning rules for variational Bayesian ICA with missing data. The complexity of the method is proportional to $T \times K^L$, where T is the number of data points, L is the number of hidden sources assumed, and K is the number of 1D gaussians used to model the density of each source. However, this exponential growth in complexity is manageable and worthwhile for small data sets containing missing entries in a high-dimensional space. The proposed method shows promise in analyzing and identifying projections of data sets that have a very limited number of expensive data points yet contain missing entries due to data scarcity. The extension to model data density with clusters of ICA was discussed. The application of the technique in a supervised classification setting was also covered. We have applied the variational Bayesian missing ICA to a primates' brain volumetric data set containing 44 examples in 57 dimensions. Very encouraging results were obtained and will be reported in another article.

References

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

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press.

Blake, C., & Merz, C. (1998). *UCI repository of machine learning databases*. Irvine, CA: University of California.

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

Choudrey, R. A., & Roberts, S. J. (2001). Flexible Bayesian independent component analysis for blind source separation. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation* (pp. 90–95). San Diego, CA: Institute for Neural Computation.

Ghahramani, Z., & Jordan, M. (1994). *Learning from incomplete data* (Tech. Rep. CBCL Paper No. 108). Cambridge, MA: Center for Biological and Computational Learning, MIT.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*(2), 183–233.

Jung, T.-P., Makeig, S., McKeown, M. J., Bell, A., Lee, T.-W., & Sejnowski, T. J. (2001). Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE, 89*(7), 1107–1122.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Mackay, D. J. (1995). *Ensemble learning and evidence maximization* (Tech. Rep.). Cambridge: Cavendish Laboratory, University of Cambridge.

Miskin, J. (2000). *Ensemble learning for independent component analysis*. Unpublished doctoral dissertation, University of Cambridge.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Welling, M., & Weber, M. (1999). Independent component analysis of incomplete data. In *1999 6th Joint Symposium on Neural Compuatation Proceedings* (Vol. 9, pp. 162–168). San Diego, CA: Institute for Neural Computation.

Received July 18, 2002; accepted January 30, 2003.