Variational Bayesian Independent Component Analysis

Neil D. Lawrence
7 Kingsfold Close,
Billingshurst, West Sussex
RH14 9HG, U.K.
neil.lawrence@cl.cam.ac.uk

Christopher M. Bishop
Microsoft Research
St. George House, 1 Guildhall Street
Cambridge, CB2 3NH, U.K.
cmbishop@microsoft.com

Abstract

Blind separation of signals through the info-max algorithm may be viewed as maximum likelihood learning in a latent variable model. In this paper we present an alternative approach to maximum likelihood learning in these models, namely Bayesian inference. It has already been shown how Bayesian inference can be applied to determine latent dimensionality in principal component analysis models (Bishop, 1999a). In this paper we derive a similar approach for removing unecessary source dimensions in an independent component analysis model. We present results on a toy data-set and on some artificially mixed images.

1 Introduction

The objective of independent component analysis is to find a representation for a data-set which is based on probabilistically independent components. One way to achieve such a representation is to fit the data to a latent variable model where the latent variables are constrained to be independent.

We consider a model, $\mathcal{M}$, in which there are I latent dimensions, P observed dimensions and our data set contains N samples. In the ICA literature the latent dimensions are often refered to as 'sources'. Since we are seeking representations for our data in which the latent variables, $\mathbf{x}$, are generated independently, we take

$$p(\mathbf{x}_n) = \prod_{i=1}^{I} p(x_{in}) \tag{1}$$

for any particular data point n. Assuming Gaussian noise, the probability of each instantiation of the observed variables, $\mathbf{t}_n$, may then be taken to be

$$p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \beta, \boldsymbol{\mu}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(\frac{\beta}{2}\|\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}\|^2\right), \tag{2}$$

where $\mathbf{W}$ is a $P \times I$ matrix of parameters, β represents an inverse noise variance and $\boldsymbol{\mu}$ a vector of means.

1.1 The Source Distribution

As is well known in independent component analysis, the choice of the latent distribution is important. In particularit must be non-Gaussian. Non-Gaussian source distributions may be split into two classes, those with positive kurtosis or 'heavy tails' and those with negative kurtosis or 'light tails'. The former are known as super-Gaussian distributions and the latter as sub-Gaussian. If our true source distributions belong in either of these two classes we may attempt to separate them. For our ICA model, we follow Attias (1998) who chooses a flexible source distribution which may take a super or sub-Gaussian form. The resulting model will then be applicable for both eventualities. Attias selected a factorised distribution in which each factor is a mixture of M Gaussians,

$$p(\mathbf{x}_n) = \prod_{i=1}^{I} \left[\sum_{m=1}^{M} \pi_m \mathcal{N}(x_{ni}|m_m, \sigma_m^2) \right], \tag{3}$$

where $\{\pi_m\}$ are the mixing coefficients and each component is governed by a mean m_m and a variance σ_m^2. Attias refered to the model as independent factor analysis. We may now write down a likelihood which is a function of the parameters $\mathbf{W}$, β and $\boldsymbol{\mu}$

$$p(\mathbf{t}|\mathbf{W}, \beta, \boldsymbol{\mu}) = \prod_{n=1}^{N} \int p(\mathbf{t}_n|\mathbf{x}_n, \mathbf{W}, \beta, \boldsymbol{\mu}) p(\mathbf{x}_n) d\mathbf{x}_n. \tag{4}$$

This function could now be maximised with respect to the parameters in order to determine the independent components. Traditionally this optimisation is performed in the limit as β tends to zero. This approach to blind source separation was introduced by Bell and Sejnowski (1995) as an information maximisation algorithm. The relationship with maximum likelihood was pointed out by various authors including Cardoso (1997) and MacKay (1996).

2 A Bayesian Formalism of ICA

In this paper we propose, rather than learning the parameters through maximum likelihood, to follow the Bayesian approach of inferring the parameterisation of the model. This requires us to place priors over the model parameters.

We aim to show how through a particular selection of our prior distribution $p(\mathbf{W})$ we may automatically determine the number of sources which have produced our data. We are inspired in our approach by the Bayesian PCA of Bishop (1999a) which aims to determine the dimensionality of the principal sub-space automatically.

We choose to treat the noise precision, β, with a gamma prior

$$p(\beta) = \mathrm{gam}(\beta|a_\beta, b_\beta) \tag{5}$$

where we define the gamma-distribution as

$$\mathrm{gam}(\tau|a, b) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) \tag{6}$$

For the mixing matrix, $\mathbf{W}$, we consider a Gaussian prior. In particular the relevance of each input may be determined through the use of the automatic relevance determination (ARD) prior (Neal, 1996; MacKay, 1995b)

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{I} \prod_{p=1}^{P} \mathcal{N}(w_{ip}|0, \alpha_i^{-1}) \tag{7}$$

where the prior is governed by a vector of hyper-parameters, $\boldsymbol{\alpha}$, of length I. The parameters associated with each input of the network are governed by an element of the vector which determines the its 'relevance'.

The hyper-parameters $\boldsymbol{\alpha}$ may be inferred through a hierarchical Bayesian framework. We therefore place gamma distributed hyper-prioes across these parameters,

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{I} \mathrm{gam}(\alpha_i | a_{\alpha_i}\, b_{\alpha_i}). \tag{8}$$

Finally we place a Gaussian prior over the means $\boldsymbol{\mu}$

$$p(\boldsymbol{\mu}) = \prod_{p=1}^{P} \mathcal{N}(\boldsymbol{\mu}_p | 0, \tau) \tag{9}$$

where τ represents the inverse variance of the prior.

We may now define our model likelihood

$$\begin{aligned} p(\mathbf{t}|\mathcal{M}) \quad = \quad & \int p(\mathbf{t}|\mathbf{W}, \beta, \mathbf{x}) p(\mathbf{x}) p(\mathbf{W}|\boldsymbol{\alpha}) \\ & \times p(\boldsymbol{\alpha}) p(\beta) p(\boldsymbol{\mu}) d\mathbf{x} d\mathbf{W} d\boldsymbol{\alpha} d\beta d\boldsymbol{\mu}. \end{aligned} \tag{10}$$

3 The Variational Approach

In Bayesian inference we aim to infere posterior distributions for our parameters. Integrals of the type shown in Eqn 10 are important in this process. Unfortunately, for Bayesian ICA as we have described it, this integral is intractable and we must look to approximations to make progress. We choose to take a variational approach (Jordan et al., 1998; Lawrence, 2000). The variational approach involves developing an approximation, $q(\mathcal{H})$, to the true posterior distribution, $p(\mathcal{H}|\mathcal{V})$, of the latent variables, $\mathcal{H}$, given the observed variables, $\mathcal{V}$. Variational inference can provide a rigourous lower bound on marginalised log-likelihoods of the form,

$$\ln \int p(\mathcal{H}, \mathcal{V}) d\mathcal{H} \geq \int q(\mathcal{H}) \ln \frac{p(\mathcal{H}, \mathcal{V})}{q(\mathcal{H})} d\mathcal{H} \tag{11}$$

The difference between this bound and the true marginalised likelihood can be shown to be the Kullback-Leibler (KL) divergence between the true posterior distribution and the approximation.

$$\mathrm{KL}(q||p) = \int q(\mathcal{H}) \ln \frac{p(\mathcal{H}|\mathcal{V})}{q(\mathcal{H})} d\mathcal{H} \tag{12}$$

If we were to utilise an unrestricted approximation, $q(\mathcal{H})$, and perform a free-form maximisation of bound 11 with respect to $q(\mathcal{H})$ we would recover $q(\mathcal{H}) = p(\mathcal{H}|\mathcal{V})$ and the bound would become exact. This approach is the expectation step of an expectation-maximisation algorithm. However in our model, if such a choice were made, the intractabilities would not be resolved. Instead, by placing restrictions on the form of the approximating distribution, we hope to make minimisation of the KL-divergence achievable.

The choice of the variational q-distribution is important. We seek a choice which is simple enough to make our computations tractable, but one which gives enough flexibility to make the bound 11 tight. There are various approaches to determining a useful approximation to the posterior. Lappalainen (1999), for example, imposed specific parameterised functional forms on his variational distributions and then minimised the KL-divergence by gradient based optimisation of their parameters. In this paper we prefer to consider free-form optimisations of our approximating distributions.

As we have already mentioned if we were to allow completely unconstrained free-form optimisation of the posterior approximation, we would merely recover the true posterior distribution and the intractabilities would not have been circumvented. We therefore must impose constraints on the form of the approximation. Consider a model where the latent variables, $\mathcal{H}$, are split into exclusive sub-sets $\mathcal{H}_i$. If we impose separability constraints across these sub-sets on our approximation to the posterior,

$$q(\mathcal{H}) = \prod_i q(\mathcal{H}_i), \tag{13}$$

it is straightforward to show (MacKay, 1995a; Lawrence, 2000) that the optimum form for each component of posterior distribution is

$$q(\mathcal{H}_j) \propto \exp \left\langle \ln p(\mathcal{H}) \right\rangle_{\prod_{i \neq j} q(\mathcal{H}_i)} . \tag{14}$$

Here we have used the notation $\left\langle \cdot \right\rangle_q$ to denote an expectation under the distribution q.

By utilising an approximation that factorises across the parameters of the model,

$$q(\mathbf{x}, \mathbf{W}, \boldsymbol{\alpha}, \beta, \boldsymbol{\mu}) = q_x(\mathbf{x}) q_w(\mathbf{W}) q_\alpha(\boldsymbol{\alpha}) q_\beta(\beta) q_\mu(\boldsymbol{\mu}), \tag{15}$$

we may obtain a lower bound on the model log-likelihood of the form

$$\begin{aligned} \ln p(\mathbf{t}|\mathcal{M}) \quad \geq \quad & \left\langle \ln p(\mathbf{t}|\mathbf{W}, \beta, \mathbf{x}, \boldsymbol{\mu}) \right\rangle_{q_x q_w q_\beta q_\mu} \\ & + \left\langle \ln p(\mathbf{x}) \right\rangle_{q_x} + \left\langle \ln p(\mathbf{W}|\boldsymbol{\alpha}) \right\rangle_{q_w q_\alpha} \\ & + \left\langle \ln p(\boldsymbol{\alpha}) \right\rangle_{q_\alpha} + \left\langle \ln p(\beta) \right\rangle_{q_\beta} + \left\langle \ln p(\boldsymbol{\mu}) \right\rangle_{q_\mu} \\ & + S(q_x) + S(q_w) + S(q_\alpha) + S(q_\beta) \\ & + S(q_\mu) \end{aligned} \tag{16}$$

where $S(p(\ \cdot\ ))$ is the entropy of distribution $p(\ \cdot\ )$The difference between this bound and $\ln p(\mathbf{t}|\mathcal{M})$ is the KL-divergence between the true and approximating posterior. For the Bayesian ICA model, as we have described it, all the necessary expectations in Eqn 14 may be performed analytically giving the following results

$$\begin{aligned} q_x &= \prod_{n=1}^{N} \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \frac{\tilde{\pi}_{\mathbf{m}}^{(n)}}{Z_n} \mathcal{N}(x_n | \mathbf{f}_{\mathbf{m}}^{(n)}, \mathbf{D}_{\mathbf{m}}), & (17) \\ q_\mu &= \mathcal{N}(\boldsymbol{\mu} | \mathbf{m}_\mu, \boldsymbol{\Sigma}_\mu) & (18) \\ q_w &= \prod_{p=1}^{P} \mathcal{N}(\mathbf{w}_p | \mathbf{m}_w^{(p)}, \boldsymbol{\Sigma}_w) & (19) \\ q_\alpha &= \prod_{i=1}^{I} \mathrm{gam}(\alpha_i | \tilde{a}_\alpha, \tilde{b}_\alpha^{(i)}) & (20) \\ q_\beta &= \mathrm{gam}(\beta | \tilde{a}_\beta, \tilde{b}_\beta), & (21) \end{aligned}$$

where the parameters of the latent distribution q_x: $\mathbf{D}_{\mathbf{m}}, \mathbf{f}_{\mathbf{m}}^{(n)}, \tilde{\pi}_{\mathbf{m}}^{(n)}$ and Z_n are given in Section A. The other parameters are straightforward to derive and may be found in Bishop (1999b).

The solution for the optimal factors of the q-distribution is an implicit one. Each distribution depends on moments of the other solutions. A solution may be determined numerically by starting with a suitable initial guess and cycling through through each distribution using the update equations given in Section A and Bishop (1999b).

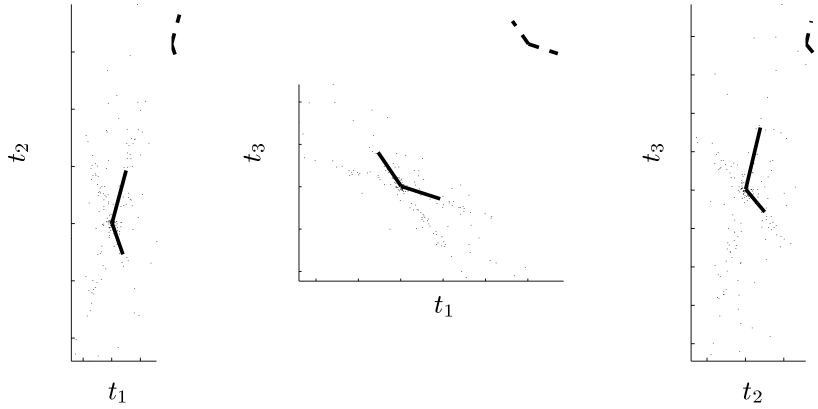

Figure 1: Scatter plots of samples from the model. The discovered embedded dimensions are shown on the centre of the plots as solid lines while the true embedded dimensions are shown to the upper right of each plot as dashed lines.

4 Results

4.1 Toy Problem

For an evaluation of the ability of our model to determine the latent dimensionality, we sampled a toy data-set $\{\mathbf{t}_n\}$ from a randomly parameterised model. The two source distributions were chosen to contain two components, both with mean zero. The variances of each component, σ_m^2, were taken to be 1 and 50. The number of observed data dimensions was then taken to be three, and the inverse variance of the noise was set at 1×10^{-4}. The values of the mixing matrix $\mathbf{W}$ were randomly sampled from a Gaussian with unit variance. We obtained 200 samples from the model and then sought to determine the number of sources and the noise variance by inference in a model with three source dimensions. The sources were of identical type to those in the generating model. The values of $\mathbf{W}$ were initialised with a PCA solution. Optimisation of the distributions then proceeded by first updating the moments of $\boldsymbol{\mu}$ then the moments of $\mathbf{x}$ and finally the moments of $\mathbf{W}$. These moments were updated until convergence or for a maximum of 100 times. The moments of $\boldsymbol{\alpha}$ and β were then updated. This whole process was repeated fifty times.

The selected independent components are shown in Figure 1. Note that the true number of source dimensions has been correctly determined. Shown in Figure 2 is the evolution of each $\log_{10} \alpha_i$ during learning. Note that very quickly one of the hyper-parameters becomes three orders of magnitude larger than the others effectively switching off one of the source dimensions.

4.2 Real Data

As a further test of our model we analysed two images. The images are from The Corel Gallery 1,000,000 Collection and were averaged across their red, green and blue channels to obtain grey-scale images. The kurtosis of both images was negative, showing that both images are sub-Gaussian. We sub-sampled 1000 data-points from the images and then mixed them with the same matrix as for the toy problem. Gaussian noise was then added with a variance which was 7.8% of the source signal size. We then implemented a model containing three source dimensions. The factors of the source distributions contained two

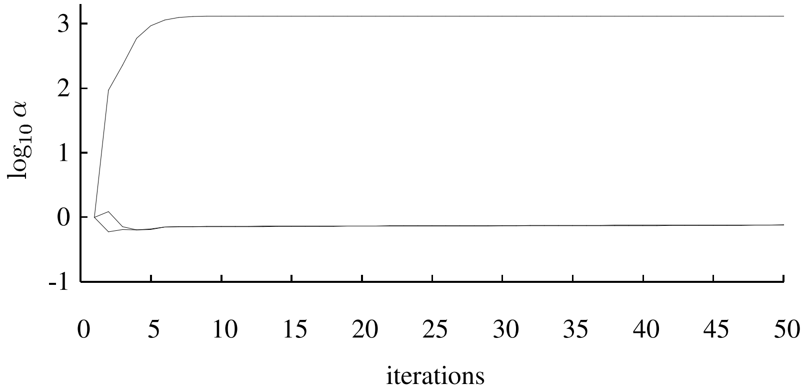

Figure 2: Evolution of the base 10 logarithm of the hyper-parameters during learning.

components, both with unit variance and with means of 1 and -1. The updates of the variational distributions were carried out in a similar manner to the model trained on the toy data.

Figure 3 shows the results for the model. This model did not remove the third source automatically. However we have shown the image with the third output dimension (the one associated with the highest α_i) manually removed.

Figure 4 shows scatter plots of the sub-sampled data projected along the three possible pairings of the output dimensions. Also shown, in the upper right corners of each plot, are the true directions of the mixing matrix and in the centre of the plot the expected value of the mixing matrix determined by the algorithm.

5 Discussion

The suggested model performed well when trained on toy data which had been sampled from similar source distributions. However in experiments on images the independent component analysis model was unable to determine the true number of sources for the real data. This problem perhaps arises because the assumed forms of the latent distributions are not matched to the true source distributions[1]. As a result the model may always better explain the data by adding 'phantom' source distributions. This is because the embedded observation space may obtain more and more complex forms through the addition of more latent dimensions. This is in contrast to the situation for Bayesian PCA. In Bayesian PCA the latent distributions are taken to be Gaussian and any sub-space embedded within a Gaussian distribution will also be Gaussian. As a result it does not exhibit these problems.

We might take the approach of removing the source associated with the highest hyper-parameter and checking whether the lower bound on the model likelihood increases. Unfortunately, although we are maximising a lower bound on the model-likelihood for Bayesian ICA, we are unable to determine the value of this bound. This is a consequence of the posterior approximations for the latent variables being a mixture of Gaussians. To calculate the bound 16 we are required to evaluate the entropy of the mixture distribution. There is some hope for progress because while exact evaluation of this entropy is intractable, it may be lower bounded (Lawrence & Azzouzi, 1999).

[1] We are not referring to the issue of sub-Gaussian versus super-Gaussian distributions, we have assumed that the sign of the kurtosis of the sources was known in these experiments and have selected our latent distributions accordingly.

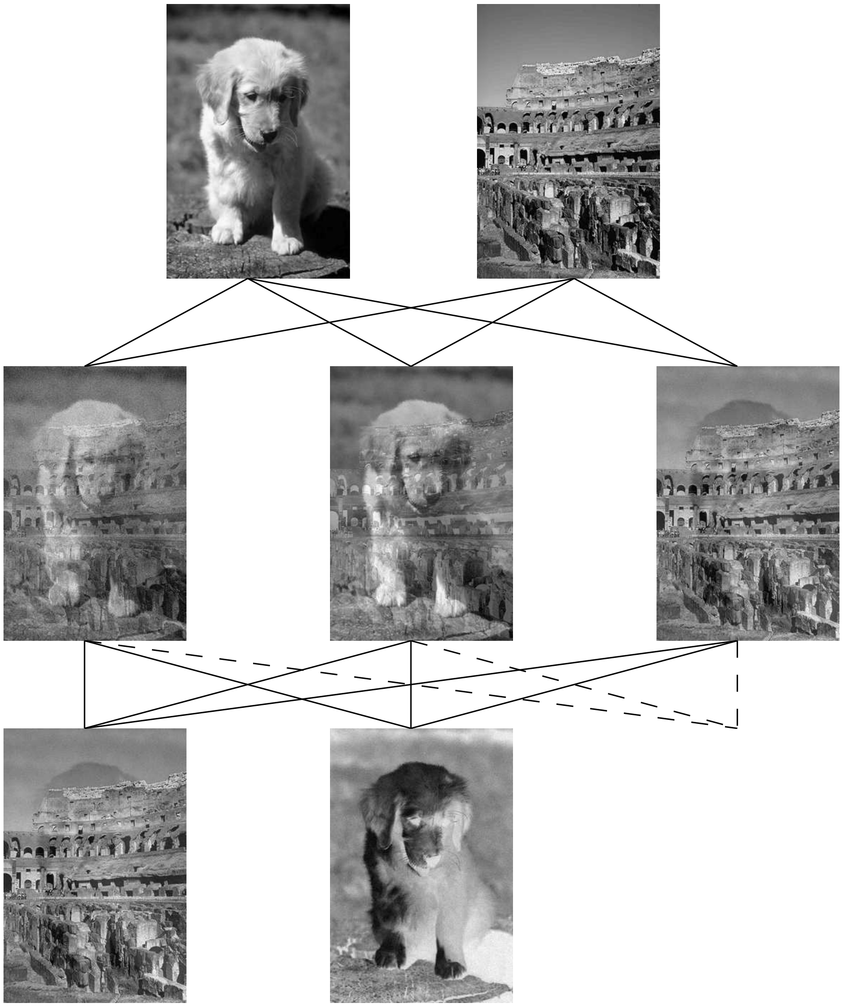

Figure 3: Results when the model is left to determine the latent dimensionality. The recovered sources with the phantom latent dimension manually removed. Note that there is a stronger ghost of the puppy on the image of the Colosseum.

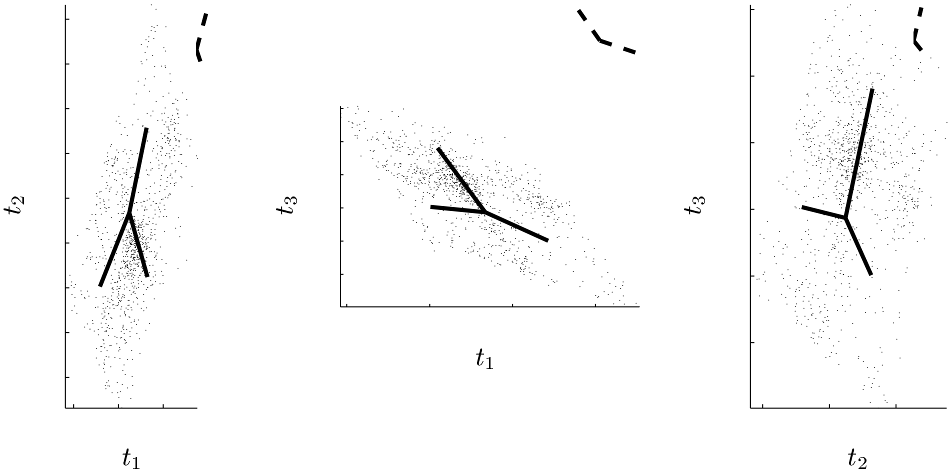

Figure 4: Scatter plots of sub-samples of the observed images for the experiment attempting to determine the true number of sources. The discovered embedded dimensions are shown on the centre of the plots as solid lines while the true embedded dimensions are shown to the upper right of each plot as dashed lines.

Another solution, and the one we would propose, would be to adaptively modify the latent distributions as part of the optimisation process. This could be achieved through modification of the latent variable distributions as shown by, for example, Attias (1998).

A final alternative is to use a simpler model to determine the latent dimensionality. We could apply Bayesian PCA to determine the principal sub-space and the associated latent dimensionality. Then a simple maximum likelihood ICA model could be applied within that embedded space. We prefer our suggested solution though as it would be able to handle problems where there are more sources than observed dimensions and is therefore more general. Another and very different approach to determining the sources has been suggested by Hyvärinen and Oja (1997). This approach is not based on a likelihood model and is related to projection pursuit.

We note that the model we have described should also be able to handle more source dimensions than there are data dimensions. This has been considered a difficult problem in the independent component analysis literature.

An additional problem with our approach is the exponential growth in the number of components as a function of the latent space dimensionality. This makes the approach impractical for models with large numbers of sources.

A Source q-distribution for Bayesian ICA

First of all we note that the joint distribution factorises across the data-points and we may therefore take

$$q_x(\mathbf{X}) = \prod_{n=1}^{N} q_x^{(n)}(\mathbf{x}_n). \tag{22}$$

In Eqn 14 we need only take expectations of the components of $\ln p(D, \boldsymbol{\theta})$ which contain $\mathbf{x}_n$, as we will look to take care of the other terms, which are constant in $\mathbf{x}_n$, through the normalising constant.

$$\begin{aligned}\ln q_x^{(n)}(\mathbf{x}) &= \sum_{i=1}^{I} \ln p(\mathbf{x}_n) - \left\langle \frac{\beta}{2} \left\|\mathbf{t}_n - \mathbf{w}\mathbf{x}_n - \mu\right\|^2 \right\rangle_{\boldsymbol{\theta} \neq \mathbf{x}} \\ &\quad + \text{const} \end{aligned} \tag{23}$$

where $p(\mathbf{x})$ is the latent variable model which we took to be a factorised distribution in which each factor is a mixture of M Gaussians,

$$p(\mathbf{x}_n) = \prod_{i=1}^{I} \sum_{m=1}^{M} \pi_m \mathcal{N}_{im}, \tag{24}$$

where the notation $\mathcal{N}_{im}$ indicates the Gaussian distribution $\mathcal{N}(x_{ni}|m_m, \sigma_m^2)$. The second term of Eqn 23 may be expanded and rewritten in terms of the expectations of each parameter

$$-\frac{\langle\beta\rangle}{2} \mathbf{x}_n^{\mathrm{T}} \left\langle \mathbf{w}^{\mathrm{T}}\mathbf{w} \right\rangle \mathbf{x}_n + \langle\beta\rangle \mathbf{x}_n^{\mathrm{T}} \langle\mathbf{w}\rangle \left(\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle\right)^{\mathrm{T}} \equiv A(\mathbf{x}_n). \tag{25}$$

Exponentiating both sides we obtain

$$q_x^{(n)}(\mathbf{x}_n) \propto \exp[A(\mathbf{x}_n)] \prod_{i=1}^{I} \sum_{m=1}^{M} \pi_m \mathcal{N}_{im}. \tag{26}$$

The product of the sums of Gaussians may be rewritten in terms of sums of products of Gaussians

$$\prod_{i=1}^{I} \sum_{m=1}^{M} \pi_m \mathcal{N}_{im} = \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \prod_{i=1}^{I} \pi_{m_i} \mathcal{N}_{im_i}. \tag{27}$$

This means $q_x^{(n)}(\mathbf{x}_n)$ is a mixture of Gaussians with M^I terms

$$\begin{aligned} q_x^{(n)}(\mathbf{x}_n) &\propto \exp\left\{A(\mathbf{x}_n)\right\} \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \prod_{i=1}^{I} \pi_{m_i} \mathcal{N}_{im_i} \\ &\propto \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \pi_{\mathbf{m}} \frac{\prod_{i=1}^{I} \beta_{m_i}^{\frac{1}{2}}}{(2\pi^{\frac{1}{2}})} \exp\left\{ A(\mathbf{x}_n) \right. \\ &\qquad \left. - \sum_{i=1}^{I} \frac{1}{2\sigma_{m_i}^2} (x_{ni} - m_{m_i})^2 \right\} \end{aligned} \tag{28}$$

where $\pi_{\mathbf{m}} = \prod_{i=1}^{I} \pi_{m_i}$. The argument of the exponential may be rewritten as

$$-E_{q_x} \equiv A(\mathbf{x}_n) - \frac{1}{2}(\mathbf{x}_n - \mathbf{m}_{\mathbf{m}})^{\mathrm{T}} \mathbf{B}_{\mathbf{m}}^{-1} (\mathbf{x}_n - \mathbf{m}_{\mathbf{m}}) \tag{29}$$

where $\mathbf{B}_{\mathbf{m}} = \mathrm{diag}(\sigma_{m_i}^2)$ and $\mathbf{m}_{\mathbf{m}} = \{m_{m_i}\}^{\mathrm{T}}$. We rewrite E_{q_x}

$$\begin{aligned} E_{q_x} &= \frac{1}{2}(\mathbf{x}_n - \mathbf{f}_{\mathbf{m}}^{(n)})^{\mathrm{T}} \mathbf{D}_{\mathbf{m}}^{-1} (\mathbf{x}_n - \mathbf{f}_{\mathbf{m}}^{(n)}) \\ &\quad + \frac{1}{2} \mathbf{m}_{\mathbf{m}}^{\mathrm{T}} \mathbf{B}_{\mathbf{m}}^{-1} \mathbf{m}_{\mathbf{m}} - \frac{1}{2} \mathbf{f}_{\mathbf{m}}^{(n)^{\mathrm{T}}} \mathbf{D}_{\mathbf{m}}^{-1} \mathbf{f}_{\mathbf{m}}^{(n)} \end{aligned} \tag{30}$$

where $\mathbf{D}_{\mathbf{m}}$ and $\mathbf{f}_{\mathbf{m}}$ are defined as

$$\mathbf{D}_{\mathbf{m}} = \left(\langle\beta\rangle \left\langle \mathbf{w}^{\mathrm{T}}\mathbf{w} \right\rangle + \mathbf{B}_{\mathbf{m}}^{-1}\right)^{-1} \tag{31}$$

$$\mathbf{f}_{\mathbf{m}}^{(n)} = \mathbf{D}_{\mathbf{m}} \left[\langle\beta\rangle \left\langle \mathbf{w}^{\mathrm{T}} \right\rangle (\mathbf{t}_n - \langle\boldsymbol{\mu}\rangle) + \mathbf{B}_{\mathbf{m}}^{-1} \mathbf{m}_{\mathbf{m}} \right]. \tag{32}$$

Thus each factor of the variational distribution is:

$$q_x^{(n)}(\mathbf{x}_n) = \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \frac{\tilde{\pi}_{\mathbf{m}}^{(n)}}{Z_n} \mathcal{N}(x_n | \mathbf{f}_{\mathbf{m}}^{(n)}, \mathbf{D}_{\mathbf{m}}), \tag{33}$$

where

$$\tilde{\pi}_{\mathbf{m}}^{(n)} = \frac{\pi_{\mathbf{m}} e^{-\frac{1}{2}\mathbf{m}_{\mathbf{m}}^{\mathrm{T}} \mathbf{B}_{\mathbf{m}}^{-1} \mathbf{m}_{\mathbf{m}}} e^{\frac{1}{2}\mathbf{f}_{\mathbf{m}}^{(n)^{\mathrm{T}}} \mathbf{D}_{\mathbf{m}}^{-1} \mathbf{f}_{\mathbf{m}}^{(n)}}}{|\mathbf{B}_{\mathbf{m}}|^{\frac{1}{2}} |\mathbf{D}_{\mathbf{m}}|^{-\frac{1}{2}}} \tag{34}$$

$$Z_n = \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \tilde{\pi}_{\mathbf{m}}^{(n)}. \tag{35}$$

The required moments are therefore evaluated to give

$$\langle \mathbf{x}_n \rangle = \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \frac{\tilde{\pi}_{\mathbf{m}}^{(n)}}{Z_n} \mathbf{f}_{\mathbf{m}}^{(n)} \tag{36}$$

$$\left\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \right\rangle = \sum_{m_1=1}^{M} \cdots \sum_{m_I=1}^{M} \frac{\tilde{\pi}_{\mathbf{m}}^{(n)}}{Z_n} \left(\mathbf{D}_{\mathbf{m}} + \mathbf{f}_{\mathbf{m}}^{(n)} \mathbf{f}_{\mathbf{m}}^{(n)\mathrm{T}} \right). \tag{37}$$

References

Attias, H. (1998). Independent factor analysis. *Neural Computation*, *11*, 803–851.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Bishop, C. M. (1999a). Bayesian PCA. *Advances in Neural Information Processing Systems* (pp. 482–388). Cambridge, MA: MIT Press.

Bishop, C. M. (1999b). Variational principal components. *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99* (pp. 509–514).

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Letters on Signal Processing*, *4*, 112–114.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483–1492.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1998). An introduction to variational methods for graphical models. *Learning in Graphical Models* (pp. 105–162). Dordrecht, The Netherlands: Kluwer.

Lappalainen, H. (1999). Ensemble learning for independent component analysis. *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation* (pp. 7–12).

Lawrence, N. D. (2000). *Variational inference in probabilistic models*. Doctoral dissertation, Computer Laboratory, University of Cambridge, New Museums Site, Pembroke Street, Cambridge, CB2 3QG, U.K.

Lawrence, N. D., & Azzouzi, M. (1999). A variational Bayesian committee of neural networks. Submitted to *Neural Networks*.

MacKay, D. J. C. (1995a). Developments in probabilistic modelling with neural networks—ensemble learning. *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks, Nijmegen, Netherlands, 14-15 September 1995* (pp. 191–198). Berlin: Springer.

MacKay, D. J. C. (1995b). Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.

MacKay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis. Unpublished manuscript, available from `http://wol.ra.phy.cam.ac.uk/mackay/homepage.html`.

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer. Lecture Notes in Statistics 118.