ON THE EFFECT OF THE FORM OF THE POSTERIOR APPROXIMATION IN VARIATIONAL LEARNING OF ICA MODELS

Alexander Ilin and Harri Valpola

Laboratory of Computer and Information Science
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Espoo, Finland

ABSTRACT

We show that the choice of posterior approximation of sources affects the solution found in Bayesian variational learning of linear independent component analysis models. Assuming the sources to be independent a posteriori favours a solution which has an orthogonal mixing matrix. A linear dynamic model which uses second-order statistics is considered but the analysis extends to nonlinear mixtures and non-Gaussian source models as well.

1. INTRODUCTION

Recently several methods for variational Bayesian learning of linear ICA models and their extensions have been reported in the literature [1, 2, 3, 4, 5, 6, 7, 8]. The basic idea in these approaches is to approximate the true posterior probability density of the unknown variables by a function which has a restricted form. Typically some type of factorisation is assumed.

In this paper, we study how the choice of the form of posterior approximation affects the solution which is found by variational Bayesian learning of linear ICA models. We investigate two common cases: 1) sources are approximated to be independent a posteriori; and 2) the posterior correlations of the sources are modelled. Note that although ICA models assume sources to be independent a priori, the sources still typically have posterior correlations.

We show that neglecting the posterior correlations of the sources introduces a bias in favour of principal component analysis (PCA) solution. By the PCA solution we mean the solution which has an orthogonal mixing matrix.

The rest of the paper is organised as follows. In Section 2, we briefly introduce variational Bayesian learning. Section 3 discusses the linear dynamic model whose learning we analyse theoretically in Section 4 and experimentally in Section 5. The implications of the analysis are discussed in Section 6.

This research has been funded by the European Commission project BLISS, and the Finnish Center of Excellence Programme (2000–2005) under the project New Information Processing Principles.

2. VARIATIONAL BAYESIAN LEARNING

Variational Bayesian learning techniques are based on approximating the true posterior probability density of the unknown variables of the model by a function with a restricted form. Currently the most common technique is ensemble learning where Kullback-Leibler divergence measures the misfit between the approximation and the true posterior. It has been applied to ICA and its extensions as well as to several other types of models (e.g. [9, 10]).

In ensemble learning, the posterior approximation $q(\boldsymbol{\theta})$ of the unknown variables $\boldsymbol{\theta}$ is required to have a suitably factorial form

$$q(\boldsymbol{\theta}) = \prod_i q(\boldsymbol{\theta}_i)\,, \tag{1}$$

where $\boldsymbol{\theta}_i$ are the subsets of unknown variables. In ICA, at least the sources $\mathbf{S}$ are assumed independent a posteriori of the mixing matrix $\mathbf{A}$ and other parameters:

$$q(\boldsymbol{\theta}) = q(\mathbf{S})q(\mathbf{A})q(\boldsymbol{\theta}_{\rm rest})\,. \tag{2}$$

The misfit between the true posterior $p(\boldsymbol{\theta} \mid \mathbf{X})$ and its approximation $q(\boldsymbol{\theta})$ is measured by Kullback-Leibler divergence which yields a cost function of the form

$$\mathcal{C} = D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{X})) - \log p(\mathbf{X}) \geq -\log p(\mathbf{X})\,.$$

The extra term $-\log p(\mathbf{X})$ is included to the cost function in order to avoid calculation of the model constant $p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. Thus, the minimised expression can be written in the following form:

$$\begin{aligned} \mathcal{C} &= \left\langle \log \frac{q(\mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\rm rest})}{p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\rm rest})} \right\rangle \\ &= \langle \log q(\mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\rm rest}) \rangle - \langle \log p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\rm rest}) \rangle\,, \end{aligned} \tag{3}$$

where $\langle \cdot \rangle$ denotes the expectation over distribution $q(\boldsymbol{\theta})$.

During learning, the factors are typically updated one at a time while keeping others fixed. For each update of the posterior approximation $q(\boldsymbol{\theta}_i)$, the set of variable $\boldsymbol{\theta}_i$ requires the prior distribution $p(\boldsymbol{\theta}_i \mid \text{parents})$ given by its parents and the likelihood

$p(\text{children} \mid \boldsymbol{\theta}_i)$ obtained from its children. The relevant part of the Kullback-Leibler divergence to be minimised is

$$C(q(\boldsymbol{\theta}_i)) = \left\langle \ln \frac{q(\theta)}{p(\theta \mid \text{parents}) p(\text{children} \mid \theta)} \right\rangle . \quad (4)$$

In ensemble learning, conjugate priors are commonly used because they make it very easy to solve the variational minimisation problem of finding the optimal $q(\boldsymbol{\theta}_i)$ which minimises (4).

3. SECOND ORDER ICA MODEL

Linear source models assume the observations to have been generated by sources which are mapped linearly to the observations. The model is

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) , \quad (5)$$

where $\mathbf{n}(t)$ is additive Gaussian noise (sometimes omitted). It is well known that this model has rotational degeneracy if the sources $\mathbf{s}(t)$ have a static Gaussian model (see e.g. [11] for introduction). We can choose any invertible $\mathbf{C}$ and generate a new solution $\mathbf{A}' = \mathbf{A}\mathbf{C}$ and $\mathbf{s}'(t) = \mathbf{C}^{-1}\mathbf{s}(t)$. The sources still remain Gaussian.

In PCA the degeneracy is removed by requiring the mixing matrix $\mathbf{A}$ to be orthogonal. In ICA, the degeneracy can be removed—up to scaling and permutation—by assuming non-Gaussian sources or by introducing a diagonal matrix $\mathbf{B}$ to model the dynamics:

$$\mathbf{s}(t) = \mathbf{B}\mathbf{s}(t-1) + \mathbf{m}(t) , \quad (6)$$

where $\mathbf{m}(t)$ is Gaussian noise. In the latter case, only second-order statistics of the observations are needed [12, 13, 14]. The rotation is identifiable if no two elements of the diagonal of $\mathbf{B}$ are equal. A set of equal elements results in rotational degeneracy among the corresponding set of sources.

In our analysis, we use the linear dynamic model whose learning is based on second-order statistics. The posterior distribution of the sources given a fixed mixing matrix is Gaussian which makes the analysis simple. The overall behaviour will be the same in more complicated cases as well.

4. EFFECT OF POSTERIOR APPROXIMATION: THEORY

In this section we analyse theoretically how the choice of the form of the posterior approximation $q(\mathbf{S})$ of the sources affects the solution which optimises the cost function (4).

First, recall that the idea of the variational approach is to approximate the very complex posterior $p(\boldsymbol{\theta}|\mathbf{X})$ by a simpler and thus tractable parametrised distribution $q(\boldsymbol{\theta})$.

Due to its simplicity, the posterior approximation cannot represent all the different solutions of the model. In order to represent all the degeneracies and permutations, all (nonlinear) correlations of the variables would need to be modelled but this would not be feasible computationally. Instead, the approximation captures a neighbourhood of one particular solution. Each term $q(\boldsymbol{\theta}_i)$ captures the correlations between the variables in the set $\boldsymbol{\theta}_i$ while all posterior correlations with the variables in other sets $\boldsymbol{\theta}_j$ are neglected. In ICA this means that the rotational dependency between the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$ is neglected. Only the neighbourhood of one particular mixing matrix is modelled but not the fact that rotating $\mathbf{A}$ could be compensated by rotating $\mathbf{S}$ correspondingly. Consequently, the uncertainty in the mixing matrix and sources is underestimated. This holds true for all the variational ICA methods cited in this paper.

4.1. Trade-off between posterior mass and posterior misfit

The topic of this paper is the effect which the form of $q(\mathbf{S})$ has on the solution. Ideally the solution should correspond to a model whose neighbourhood contains a large portion of the posterior probability mass. In our case this is fulfilled if 1) the sources and the mixing matrix together explain the observations well and 2) the source dynamics explains the sources well. In other words, the noise covariances of $\mathbf{n}(t)$ and $\mathbf{m}(t)$ should be small. In addition, 3) the solution should be robust. Requirements 1 and 2 imply a high posterior density and 3 guarantees that the solution corresponds to a wide peak in the posterior density. Together these indicate a high probability mass in the neighbourhood of the solution.

Ensemble learning has gained popularity because it is able to find a solution which meets these three requirements. However, the restricted form of the posterior approximation $q(\boldsymbol{\theta})$ results in two additional requirements: 4) the posterior approximation $q(\mathbf{S})$ of the sources and 5) the posterior approximation $q(\mathbf{A})$ of the mixing matrix should match the posterior around the solution. In our case the posterior misfit of the rest of the parameters $\boldsymbol{\theta}_{\text{rest}}$ is not significant in practice but the choice of the functional form of $q(\mathbf{S})$ in particular and $q(\mathbf{A})$ to a lesser extent affects the optimal solution.

In general, there is a trade-off between the amount of posterior mass in the neighbourhood of the solution (requirements 1–3) and the misfit between the approximation and true local probability distribution (requirements 4 and 5). Usually it is desirable that the requirements 4 and 5 affect the solution as little as possible although sometimes it is possible to use them to select an appropriate solution among otherwise degenerate solutions (in [8], source separation is achieved by means of requirement 4 and a proper choice of $q(\mathbf{S})$).

4.2. Factorial $q(\mathbf{S})$ favours orthogonal A

Majority of the applications of ensemble learning to ICA models reported in the literature have assumed a

fully factorised $q(\mathbf{S})$:

$$q(\mathbf{S}) = \prod_{i,t} q(s_i(t)) \,. \tag{7}$$

This results in a computationally efficient learning algorithm but we will now show that it favours an orthogonal $\mathbf{A}$, a characteristic of the PCA solution.

First, we note that with the static ICA model (5) under the restriction (2), the optimal $q(\mathbf{S})$ which minimises (4) can be shown (see, e.g. [7]) to factor into

$$q(\mathbf{S}) = \prod_{t} q(\mathbf{s}(t)) \,. \tag{8}$$

Further, the optimal $q(\mathbf{s}(t))$ can be shown [15] to be Gaussian distributions. Except for the first $q(\mathbf{s}(1))$ and last $q(\mathbf{s}(T))$, each of them has the same covariance

$$\Sigma_{s,\mathrm{opt}} = \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \Sigma_m^{-1} + \mathbf{B}^T \Sigma_m^{-1} \mathbf{B} \right\rangle^{-1} \,, \tag{9}$$

where Σ_n and Σ_m are the noise covariances of $\mathbf{n}(t)$ and $\mathbf{m}(t)$, respectively. Note that the optimal posterior covariance of the sources does not depend directly on the data. This is a characteristic of linear Gaussian models.

The misfit between the factorial approximation (7) and the optimal unrestricted $q(\mathbf{S})$ is minimised when the optimal $q(\mathbf{S})$ agrees with (7). This is the case when the optimal covariance matrix $\Sigma_{s,\mathrm{opt}}$ is diagonal. This, in turn, happens if and only if $\mathbf{A}$ is orthogonal w.r.t. the inverse noise covariance Σ_n^{-1}. Since ensemble learning is trying to minimise the misfit, it favours orthogonal solutions for $\mathbf{A}$.

Figure 1 illustrates the trade-off between the misfit of the posterior approximation of the sources and the accuracy of the model. Let us assume that the data were generated by a process which can be accurately modelled by (5) and (6). Further assume that there are two sources and the mixing matrix $\mathbf{A}$ is not orthogonal. The optimal posterior covariance of the sources could then look like the ones in upper plot of Fig. 1. In the PCA solution, the posterior covariance would be diagonal and the assumption (7) would be valid. The cost of inaccurate assumption would increase towards the ICA solution as shown with dashed line on the second plot of Fig. 1.

According to our assumption, the sources can be accurately modelled in the ICA solution. If the source space is rotated by $\mathbf{S}' = \mathbf{C}\mathbf{S}$ and this is compensated by

$$\mathbf{B}' = \mathbf{C}\mathbf{B}\mathbf{C}^{-1} \,, \tag{10}$$

a model with diagonal $\mathbf{B}$ may no longer be able to capture resulting new dynamics $\mathbf{B}'$. In our two-dimensional case $b_2 = b_1$ yields a diagonal $\mathbf{B}' = \mathbf{B}$ but $b_2 \neq b_1$ will in general result in off-diagonal terms in $\mathbf{B}'$. The further b_2 is away from b_1, the stronger these off-diagonal terms are and the worse the diagonal matrix $\mathbf{B}$ can model the dynamics. This is depicted with solid lines in Fig. 1.

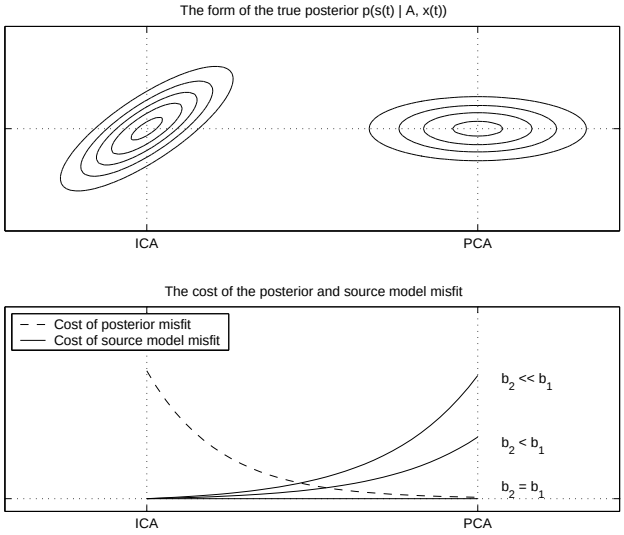

Fig. 1. Schematic illustration of the trade-offs between the ICA and PCA solutions. In the PCA solution, the posterior covariance of the sources is diagonal. This minimises the misfit between the optimal posterior and its approximation. However, the sources are explained better in the ICA solution.

This analysis suggests that the optimal solution is a result of a trade-off between the ICA solution where the explanation of the sources is best and the PCA solution where the posterior approximation of the sources is most accurate. If the mixing matrix is close to orthogonal and the source model is strongly in favour of the ICA solution, the optimal solution can be expected to be close to the ICA solution and vice versa. If the observation noise is not very high, we can expect that the explanation of the observations is not compromised. In other words, linear transformations of $\mathbf{A}$ are appropriately compensated by linear transformations of $\mathbf{S}$.

4.3. Factorial approximation for $q(\mathbf{A})$

The matrices $\mathbf{A}$ and $\mathbf{S}$ appear symmetrically in (5). Consequently, the optimal posterior under the assumption $q(\mathbf{A}) = \prod_i q(\mathbf{A}_{i,:})$ is achieved by Gaussian densities whose covariance resembles (9):

$$\Sigma_{\mathbf{A}_{i,:},\mathrm{opt}} = \left\langle \sum_{t=1}^{N} \mathbf{s}(t)\mathbf{s}^T(t)/\Sigma_{n,i,i} + \Sigma_{\mathbf{A}}^{-1} \right\rangle^{-1} \tag{11}$$

where $\Sigma_{\mathbf{A}}^{-1}$ is the covariance of the Gaussian prior of $\mathbf{A}_{i,:}$.

Often the dimension of the data vectors is much smaller than the number of them. This means that there are far fewer elements in $\mathbf{A}$ than in $\mathbf{S}$ and consequently the posterior approximation $q(\mathbf{A})$ does not play a significant role. However, if the evidence in support of the ICA solution is weak ($b_1 \approx b_2$) and the posterior

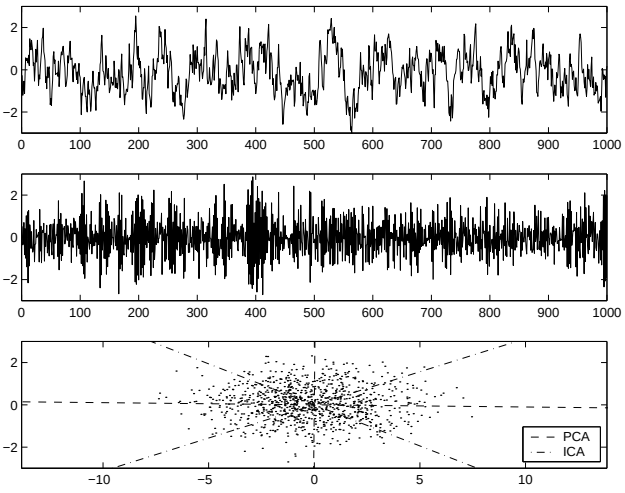

Fig. 2. The two sources with the linear dynamic model ($b_1 = 0.8$ and $b_2 = -0.8$) and their noisy mixture plotted in the subspace spanned by the columns of the mixing matrix. The PCA and ICA directions are also shown on the last plot.

of the sources is allowed to have full covariance, a factorial posterior approximation $q(\mathbf{A}_{i,:}) = \prod_j q(\mathbf{A}_{i,j})$ can change the balance in favour of the PCA solution. This is because (11) has the term $\left\langle \sum_{t=1}^{N} \mathbf{s}(t)\mathbf{s}^T(t) \right\rangle$ which is non-diagonal if the posterior covariance of the sources is non-diagonal. This in turn is the case when the mixing matrix $\mathbf{A}$ is non-orthogonal as discussed earlier.

5. EFFECT OF POSTERIOR APPROXIMATION: EXPERIMENTS

In this section, the trade-off between the ICA and PCA solutions is studied experimentally. We use the linear dynamic model defined by (5) and (6). The model and learning rules are summarised in Appendices A and B, respectively. The data set consists of 10-dimensional observation vectors which were generated by a linear mapping from two sources. The number of samples was 1000.

The element of the diagonal of the matrix $\mathbf{B}$ corresponding to the first source was chosen to be $b_1 = 0.8$ while the other element b_2 was varied in the range $[-0.8,\ 0.8]$. This controls the strength of evidence in favour of the ICA solution present in the data.

Figure 2 shows the original sources and their linear mixture in the subspace defined by the 10×2 mixing matrix $\mathbf{A}$. Note that the ICA directions corresponding to the columns of the mixing matrix are chosen to be non-orthogonal and for clarity they differ very much from the PCA directions plotted in the same figure.

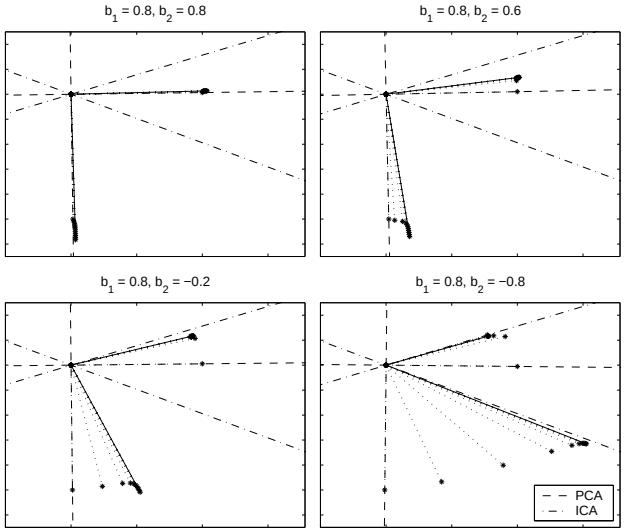

Fig. 3. The results for the diagonal approximation. Four data sets with $b_2 \in \{0.8, 0.6, -0.2, -0.8\}$ are tested. The solution is presented by the estimated columns of $\mathbf{A}$ projected onto the subspace of the true $\mathbf{A}$. The model was initialised with PCA. The dotted lines represent the solution after every 100 iterations. The final solution is plotted with the solid line.

5.1. Factorial approximation $q(\mathbf{s}(t))$

We first use the generated artificial data to test the learning procedure with the maximally factorial posterior approximation $q(\mathbf{S})$ defined by (7).

The model was implemented using the building blocks and learning rules presented in [16]. Then it was learned using 2000 iterations of alternate updates of the parameters of the approximate posterior $q(\boldsymbol{\theta})$.

Figure 3 shows the results of learning for four different data sets with $b_1 = 0.8$ and $b_2 \in \{0.8, 0.6, -0.2, -0.8\}$. The solution is presented by the estimated columns of the mixing matrix projected onto the subspace spanned by the true ICA directions. In the experiments, we tried different initialisations of $\mathbf{A}$ including the PCA and ICA solutions but the simulations converged to the same solutions for all initialisations.

Analysing the results, we see that 1) when the sources have the same dynamics ($b_2 = 0.8$), the PCA solution is found; 2) when the dynamics of the sources differs a lot ($b_2 = -0.8$), the solution is very close to the ICA directions; and 3) when the difference in dynamics is somewhere in between the two extreme cases (e.g., $b_2 = 0.6$ or $b_2 = -0.2$), the found solution lies between PCA and ICA: The more different the source dynamics, the closer the solution is to ICA. The results show that the quality of the solution found with the maximally factorial approximation depends very much on the training data and how well they support the assumed ICA model.

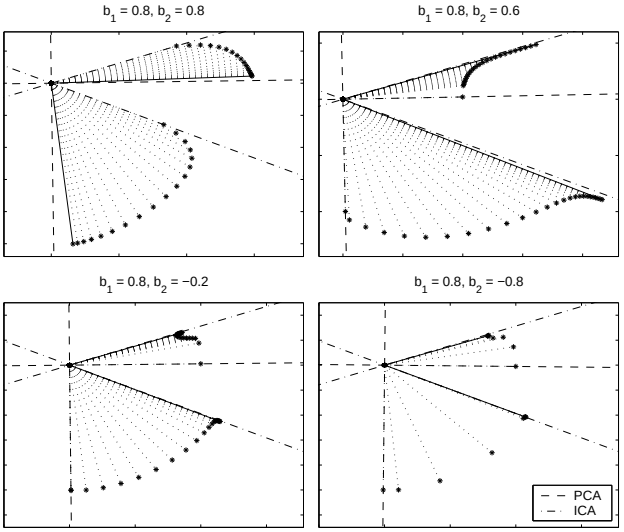

Fig. 4. The results for the unrestricted $q(\mathbf{s}(t))$. The same data sets as in Fig. 3 are tested. The current solution is plotted after every 100 iterations for $b_2 = -0.2, -0.8$, every 1000 iterations for $b_2 = 0.6$ and every 5000 iterations for $b_2 = 0.8$. The rotation of the solution is much slower in the case when the souce dynamics is just slightly different ($b_1 = 0.8$, $b_2 = 0.6$).

5.2. Unrestricted approximation $q(\mathbf{s}(t))$

We then tested the same simulations with unrestricted $q(\mathbf{s}(t))$ which yields Gaussian distributions with full covariance matrix. The rest of the model parameters $\boldsymbol{\theta}$ are modelled with the maximally factorial approximation as previously. The learning rules for the model are presented in Appendix B.

Figure 4 presents the solutions obtained with the full covariance of the source posterior. The results clearly show that the performance of the learning procedure was significantly improved as compared with the case of diagonal approximation: The ICA solution is found except in the case where $b_1 = b_2$ in which case the model converged to the PCA solution despite initialisation to the ICA solution as predicted in Section 4.3.

Note that the similarity of the source dynamics makes the separation problem more difficult. If the autocorrelation coefficients are just slightly different, it is possible to find the ICA directions but the rotation of the solution is much slower.

If the dynamics of the sources are equal, the separation problem becomes ill-posed: Any direction in the observation space has similar dynamic properties and none of them is preferred unless some extra assumptions are made.

6. DISCUSSION

As we have seen, the form of the posterior approximation can strongly affect the result found by ensemble learning. We based the analysis on a linear dynamic Gaussian model for the sake of simplicity. The situation is slightly more complicated with non-Gaussian source models or nonlinear mixtures because then the optimal posterior form is not Gaussian and even if it is restricted to be Gaussian, the posterior covariance of the sources depends on the data and is not the same for all $q(\mathbf{s}(t))$.

However, the overall results of the analysis apply to non-Gaussian and nonlinear cases. With linear models, the expectations $\left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} \right\rangle$ and $\left\langle \mathbf{s}(t) \mathbf{s}^T(t) \right\rangle$ appear just as in our analysis. In nonlinear models, the situation can be approximated by a time-dependent $\mathbf{A}(t)$ if the nonlinear mixture is smooth. Moreover, nonlinear models which are based on multi-layer linear feedforward mappings with elementwise nonlinearities have similar properties as linear models since the first linear mapping from sources to nonlinear nodes can compensate linear transformations of the source space.

To conclude, we do not claim that fully factorised posterior approximations are not useful. After all, we have applied them successfully ourselves. However, one has to be careful. If the mixing matrix cannot be made more orthogonal e.g. by pre-whitening, it is possible to end up close to the PCA solution even though the model should be able to judge the ICA solution to be better. Improving the posterior approximation will help in those situations but the price to pay is increased computational cost.

7. REFERENCES

[1] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[2] H. Lappalainen, "Ensemble learning for independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, (Aussois, France), pp. 7–12, 1999.

[3] J. Miskin and D. MacKay, "Ensemble learning for blind image separation and deconvolution," in *Advances in Independent Component Analysis* (M. Girolami, ed.), pp. 123–141, Springer-Verlag, 2000.

[4] R. Choudrey, W. Penny, and S. Roberts, "An ensemble learning approach to independent component analysis," in *Proc. of the IEEE Workshop on Neural Networks for Signal Processing, Sydney, Australia, December 2000*, IEEE Press, 2000.

[5] H. Valpola, "Nonlinear independent component analysis using ensemble learning: theory," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, (Helsinki, Finland), pp. 251–256, 2000.

[6] K. Chan, T.-W. Lee, and T. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, (San Diego, USA), pp. 492–497, 2001.

[7] K. Chan, T. Lee, and T. J. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," *Journal of Machine Learning Research*, vol. 3, pp. 99–114, August 2002.

[8] H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," *Neural Computation*, vol. 14, no. 11, pp. 2647–2692, 2002.

[9] D. Barber and C. Bishop, "Ensemble learning for multi-layer networks," in *Advances in Neural Information Processing Systems 10* (M. Jordan, M. Kearns, and S. Solla, eds.), pp. 395–401, Cambridge, MA, USA: The MIT Press, 1998.

[10] Z. Ghahramani and G. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 963–996, 2000.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. J. Wiley, 2001.

[12] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. on S.P.*, vol. 45, no. 2, pp. 434–44, 1997.

[13] A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Cuiro, "Artifact reduction in magnetoneurography based on time-delayed second order correlations," Tech. Rep. 31, GMD - Forschungszentrum Informationstechnik GmbH, 1998.

[14] A. Cichocki and R. Thawonmas, "On-line algorithm for blind signal extraction of arbitrarily distributed, but temporally correlated sources using second order statistics," *Neural Processing Letters*, vol. 12, pp. 91–98, 2000.

[15] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," in *Advances in Neural Information Processing Systems 13* (T. Leen, T. Dietterich, and V. Tresp, eds.), (Cambridge, MA, USA), pp. 507–513, The MIT Press, 2001.

[16] H. Valpola, T. Raiko, and J. Karhunen, "Building blocks for hierarchical latent variable models," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, (San Diego, USA), pp. 710–715, 2001.

A. THE DENSITY MODEL

The simple ICA model considered in Section 5:

$$p(\mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{X}|\mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\text{rest}})p(\mathbf{S}|\boldsymbol{\theta}_{\text{rest}})p(\mathbf{A})p(\boldsymbol{\theta}_{\text{rest}})$$

Here, we use the following notation: m is the number of sources; n is the number of observations; N is the number of samples in the data set; α_j, β_j, γ, $\boldsymbol{\sigma}$ are some constants; $\mathbf{D}(\boldsymbol{\sigma})$ denotes a diagonal matrix with the elements of vector $\boldsymbol{\sigma}$ on its main diagonal.

The prior model of the sources and the likelihood:

$$p(\mathbf{S}|\boldsymbol{\theta}_{\text{rest}}) = \mathcal{N}\left(\mathbf{s}(1)|\mathbf{0}, \Sigma_{m_1}\right) \prod_{t=2}^{N} \mathcal{N}\left(\mathbf{s}(t)|\mathbf{B}\mathbf{s}(t-1), \Sigma_m\right)$$

$$p(\mathbf{X}|\mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\text{rest}}) = \prod_{t=1}^{N} \mathcal{N}\left(\mathbf{x}(t)|\mathbf{A}\mathbf{s}(t), \Sigma_n\right)$$

where $\Sigma_{m_1} = \mathbf{D}(\boldsymbol{\sigma})$, $\Sigma_m = \mathbf{D}(e^{-\mathbf{v}_s})$, $\Sigma_n = \mathbf{D}(e^{-\mathbf{v}_x})$.

The prior for the (hyper)parameters:

$$\begin{aligned}
p(\mathbf{A}) &= \prod_{i=1}^{n}\prod_{j=1}^{m} \mathcal{N}\left(a_{ij}|0, \alpha_j^{-1}\right) \\
p(\mathbf{B}) &= \prod_{j=1}^{m} \mathcal{N}\left(b_j|0, \beta_j^{-1}\right) \\
p(\mathbf{v}_x|m_{v_x}, v_{v_x}) &= \prod_{i=1}^{n} \mathcal{N}\left(v_{x,i}|m_{v_x}, e^{-v_{v_x}}\right) \\
p(\mathbf{v}_s|m_{v_s}, v_{v_s}) &= \prod_{j=1}^{m} \mathcal{N}\left(v_{s,j}|m_{v_s}, e^{-v_{v_s}}\right)
\end{aligned}$$

$$m_{v_x}, v_{v_x}, m_{v_s}, v_{v_s} \sim \mathcal{N}(0, \gamma)$$

B. LEARNING RULES

The following recursive learning rules are obtained as a result of using conjugate priors for $\mathbf{s}_t$, $\mathbf{A}$, $\mathbf{B}$. The rest of the parameters ($\mathbf{v}_x$, $\mathbf{v}_s$, m_{v_x}, v_{v_x}, m_{v_s}, v_{v_s}) are updated using the rules presented in [16].

B.1. Update rules for $q(\mathbf{s}_t)$

$$q(\mathbf{s}_t) = \mathcal{N}\left(\mathbf{s}_t|\overline{\mathbf{s}}_t, \Sigma_{\mathbf{s}_t}\right)$$

$$\begin{aligned}
\Sigma_{\mathbf{s}_t} &= \left\langle \mathbf{A}^T\Sigma_n^{-1}\mathbf{A} + \Sigma_m^{-1} + \mathbf{B}^T\Sigma_m^{-1}\mathbf{B} \right\rangle^{-1} \\
\overline{\mathbf{s}}_t &= \Sigma_{\mathbf{s}_t}\left\langle \mathbf{A}^T\Sigma_n^{-1}\mathbf{x}_t + \Sigma_m^{-1}\mathbf{B}\mathbf{s}_{t-1} + \mathbf{B}^T\Sigma_m^{-1}\mathbf{s}_{t+1} \right\rangle
\end{aligned}$$

with the following exceptions: when $t = 1$, the term $+\Sigma_m^{-1}+$ is replaced by $+\Sigma_{m_1}^{-1}+$ and the term with $\mathbf{s}_{t-1}$ is omitted; and when $t = N$, the terms $\mathbf{B}^T \ldots$ are omitted.

B.2. Update rules for $q(\mathbf{A})$

$$q(\mathbf{A}) = \prod_{i=1}^{n}\prod_{j=1}^{m} \mathcal{N}\left(a_{ij}|\overline{a}_{ij}, \widetilde{a}_{ij}\right)$$

$$\widetilde{a}_{ij}^{-1} = \langle \alpha_j \rangle + \langle e^{v_{x,i}} \rangle \sum_{t=1}^{N} \left\langle s_{t,j}^2 \right\rangle$$

$$\overline{a}_{ij} = \widetilde{a}_{ij} \langle e^{v_{x,i}} \rangle \sum_{t=1}^{N} [\mathbf{x}_{t,i} \langle s_{t,j} \rangle - \sum_{k \neq j} \langle a_{ik} \rangle \langle s_{t,k} s_{t,j} \rangle]$$

B.3. Update rules for $q(\mathbf{B})$

$$q(\mathbf{B}) = \prod_{j=1}^{m} \mathcal{N}\left(b_j|\overline{b}_j, \widetilde{b}_j\right)$$

$$\widetilde{b}_j^{-1} = \langle \beta_j \rangle + \langle e^{v_{s,j}} \rangle \sum_{t=2}^{N} \left\langle s_{t-1,j}^2 \right\rangle$$

$$\overline{b}_j = \widetilde{b}_j \langle e^{v_{s,j}} \rangle \sum_{t=2}^{N} \langle s_{t,j} s_{t-1,j} \rangle$$