On the Effect of the Form of the Posterior Approximation in Variational Learning of ICA Models

Alexander Ilin (alexander.ilin@hut.fi)
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, 02015 HUT, Finland

Harri Valpola (harri.valpola@hut.fi)
Laboratory of Computational Engineering
Helsinki University of Technology
P.O. Box 9203, 02015 HUT, Finland

Abstract. We show that the choice of posterior approximation 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 orthogonal mixing vectors. Linear mixing models with either temporally correlated sources or non-Gaussian source models are considered but the analysis extends to nonlinear mixtures as well.

Keywords: variational Bayesian learning, independent component analysis

Abbreviations: ICA – Independent component analysis; MoG – Mixture of Gaussians; PCA – Principal component analysis

1. Introduction

Recently several methods for variational Bayesian learning of linear ICA models and their extensions have been reported in the literature (Attias, 1999; Lappalainen, 1999; Miskin and MacKay, 2000; Choudrey et al., 2000; Valpola, 2000; Chan et al., 2002; Chan et al., 2003; Valpola and Karhunen, 2002). 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 in detail 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 where the mixing vectors, columns of mixing matrix $\mathbf{A}$, are orthogonal with respect to the inverse of the estimated noise covariance Σ_n, that is $\mathbf{A}^T\Sigma_n^{-1}\mathbf{A}$ is a diagonal matrix. The preliminary results of this study were reported in (Ilin and Valpola, 2003).

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. Section 6 extends the analysis to non-Gaussian source models and the implications of the analysis are discussed in Section 7.

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. (Barber and Bishop, 1998; Ghahramani and Hinton, 2000)).

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 $\mathbf{S} = \{\mathbf{s}(t)|t\}$ are assumed independent a posteriori of the mixing matrix $\mathbf{A}$ and other parameters:

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

Here, $\boldsymbol{\theta}_{\text{rest}}$ are, for instance, variance parameters of the observation noise and various hyperparameters. Given the observed data $\mathbf{X} = \{\mathbf{x}(t)|t\}$, 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:

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

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

The overall probability $p(\mathbf{X}, \mathbf{S}, \mathbf{A}, \boldsymbol{\theta}_{\text{rest}})$ usually has a simple factorial form, for example

$$p(\mathbf{X} \mid \boldsymbol{\theta}_1) p(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) \dots p(\boldsymbol{\theta}_{N-1} \mid \boldsymbol{\theta}_N) p(\boldsymbol{\theta}_N) \,, \tag{4}$$

and therefore the cost function (4) splits into a sum of simple terms

$$\begin{aligned} \mathcal{C} \ = \ & \sum_{i=1}^{N} \langle \log q(\boldsymbol{\theta}_i) \rangle - \\ & \langle \log p(\mathbf{X} \mid \boldsymbol{\theta}_1) \rangle - \sum_{i=1}^{N-1} \langle \log p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{i+1}) \rangle - \langle \log p(\boldsymbol{\theta}_N) \rangle \ . \end{aligned} \tag{5}$$

During learning, the factors $q(\boldsymbol{\theta}_i)$ are typically updated one at a time while keeping others fixed. For each update of the posterior approximation $q(\boldsymbol{\theta}_i)$, only the terms with the prior distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{i+1})$ and the likelihood $p(\boldsymbol{\theta}_{i-1} \mid \boldsymbol{\theta}_i)$ are relevant. The part of the Kullback-Leibler divergence to be minimised is then

$$\mathcal{C}(q(\boldsymbol{\theta}_i)) = \left\langle \log \frac{q(\boldsymbol{\theta}_i)}{p(\boldsymbol{\theta}_{i-1} \mid \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{i+1})} \right\rangle \ . \tag{6}$$

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 (6).

3. ICA model with temporally correlated sources

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

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \,, \tag{7}$$

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., (Hyvärinen et al., 2001) 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 vectors (columns of $\mathbf{A}$) to be orthogonal. In ICA, the degeneracy can be removed—up to scaling and permutation—by assuming non-Gaussian sources or, for example, by introducing a diagonal matrix $\mathbf{B}$ to model the dynamics:

$$\mathbf{s}(t) = \mathbf{B}\mathbf{s}(t-1) + \mathbf{m}(t)\,, \tag{8}$$

where $\mathbf{m}(t)$ is Gaussian noise. In the latter case, only second-order statistics of the observations are needed (Belouchrani et al., 1997; Ziehe et al., 1998; Tong et al., 1990). 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. In Section 6, the analysis is extended to non-Gaussian distributions. 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 posterior approximation form for the sources and the mixing matrix 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 $q(\boldsymbol{\theta})$ 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 (Valpola and Karhunen, 2002), 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 mixing vectors

Majority of the applications of ensemble learning to ICA models reported in the literature have assumed a fully factorised $q(\mathbf{S})$:

$$q(\mathbf{S}) = \prod_{i,t} q(s_i(t)) \,. \tag{9}$$

This results in a computationally efficient learning algorithm but we will show that it favours orthogonal mixing vectors, a characteristic of the PCA solution.

First, we note that with the static ICA model (7) under the restriction (2), the optimal $q(\mathbf{S})$ which minimises (4) can be shown (see, e.g., (Chan et al., 2002)) to factor into

$$q(\mathbf{S}) = \prod_{t=1}^{N} q(\mathbf{s}(t)) \,. \tag{10}$$

Further, the optimal $q(\mathbf{s}(t))$ can be shown (Ghahramani and Beal, 2001) to be Gaussian distributions. Except for the first $q(\mathbf{s}(1))$ and last $q(\mathbf{s}(N))$, each of them has the same covariance

$$\Sigma_{\mathbf{s},\mathrm{opt}} = \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \Sigma_m^{-1} + \mathbf{B}^T \Sigma_m^{-1} \mathbf{B} \right\rangle^{-1} , \tag{11}$$

where Σ_n and Σ_m are the noise covariances of $\mathbf{n}(t)$ and $\mathbf{m}(t)$, respectively.[1] 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 (9) and the optimal unrestricted $q(\mathbf{S})$ is minimised when the optimal $q(\mathbf{S})$ agrees with (9). This is the case when the optimal covariance matrix $\Sigma_{\mathbf{s},\mathrm{opt}}$ is diagonal. This, in turn, happens if and only if the columns of $\mathbf{A}$ are 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 (7) and (8). Further assume that there are two sources and the mixing vectors, columns of $\mathbf{A}$, are not orthogonal. The optimal posterior covariance of the sources could then look like the ones in the upper plot of Figure 1. In the PCA solution, the posterior covariance would be diagonal and the assumption (9) would be valid. The cost of inaccurate assumption would increase towards the ICA solution as shown with dashed line on the second plot of Figure 1.

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

$$\mathbf{B}' = \mathbf{CBC}^{-1} \,, \tag{12}$$

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 $\mathbf{B}'$.

[1] The full form of $q(\mathbf{s}(t))$ for all t is given in Appendix A.2.

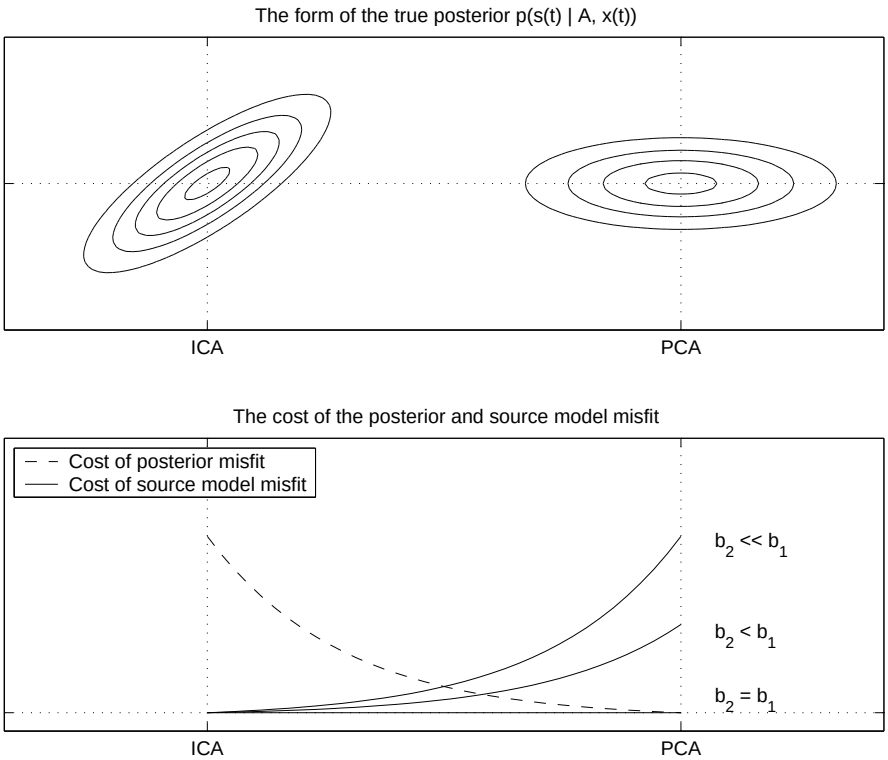


Figure 1. Schematic illustration of the trade-offs between the ICA and PCA solutions. In the PCA solution, the posterior covariance of the sources is diagonal. This minimises the misfit between the optimal posterior and its approximation. However, the sources are explained better in the ICA solution.

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 Figure 1.

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 vectors are 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 $q(\mathbf{A})$ favours orthogonal sources

Rewriting (7) in the matrix form

$$\mathbf{X} = \mathbf{AS} + \text{noise} \tag{13}$$

shows that the matrices $\mathbf{A}$ and $\mathbf{S}$ appear symmetrically in the model. Consequently, the optimal posterior under the assumption $q(\mathbf{A}) = \prod_i q(\mathbf{A}_{i,:})$ (where $\mathbf{A}_{i,:}$ are the rows of the mixing matrix) is achieved by Gaussian densities whose covariance resembles (11):

$$\Sigma_{\mathbf{A}_{i,:},\text{opt}} = \left\langle \sum_{t=1}^{N} \mathbf{s}(t)\mathbf{s}^T(t)/\Sigma_{n,i,i} + \Sigma_{\mathbf{A}}^{-1} \right\rangle^{-1} \tag{14}$$

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 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 (14) 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 columns of the mixing matrix $\mathbf{A}$ are 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 (7) and (8). The model and learning rules are summarised in Appendix A. 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

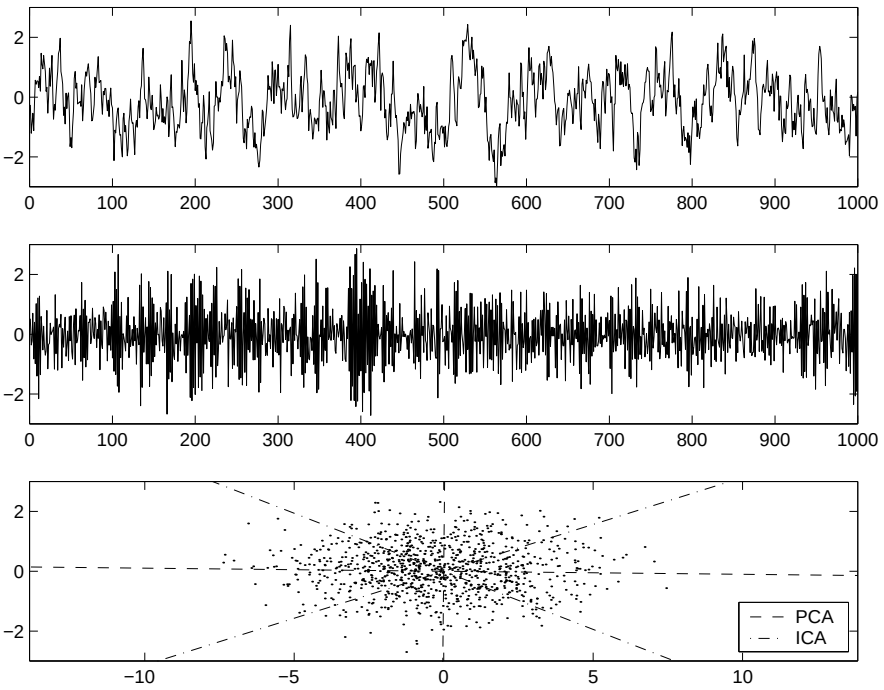


Figure 2. The two sources with the linear dynamic model ($b_1 = 0.8$ and $b_2 = -0.8$) and their noisy mixture plotted in the subspace spanned by the columns of the mixing matrix. The PCA and ICA directions are also shown on the last plot.

to be non-orthogonal and for clarity they differ very much from the PCA directions plotted in the same figure.

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 (9).

The model was implemented using the building blocks software (Valpola et al., 2003) based on the learning rules presented in (Valpola et al., 2001). 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.

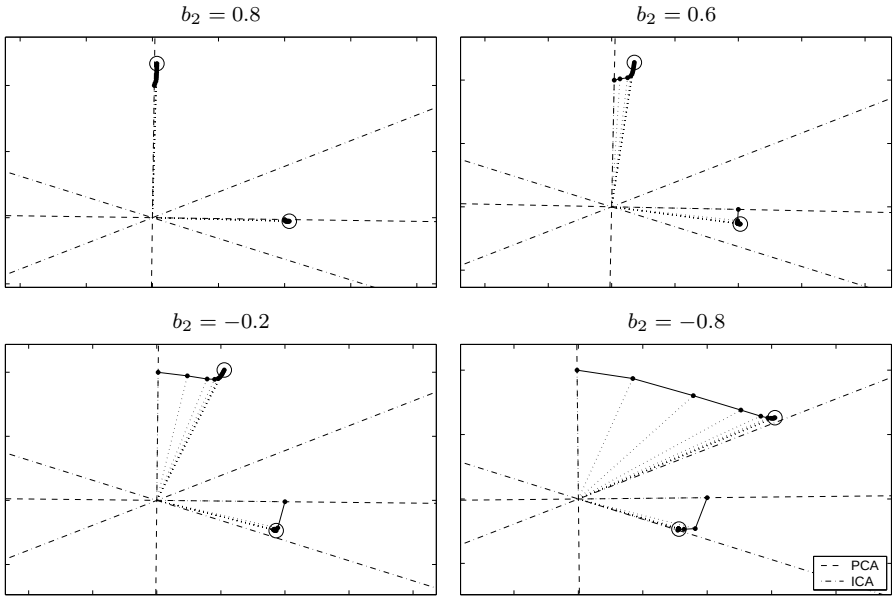


Figure 3. ICA model with temporally correlated sources. The results achieved with the factorial $q(\mathbf{s}(t))$ for four data sets with $b_2 \in \{0.8, 0.6, -0.2, -0.8\}$. The solution is presented by the estimated columns of $\mathbf{A}$ projected onto the subspace of the true $\mathbf{A}$. The model was initialised with PCA. The dotted lines represent the solution after every 100 iterations. The final solution is circled. The intervals between the ticks on all axes are equal, the scale is arbitrary due to the scaling indeterminacy.

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.

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

We performed the same simulations with the unrestricted $q(\mathbf{s}(t))$ which yields Gaussian distributions with a 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 A.2 and the sofware implementation is available on-line (Ilin, 2004).

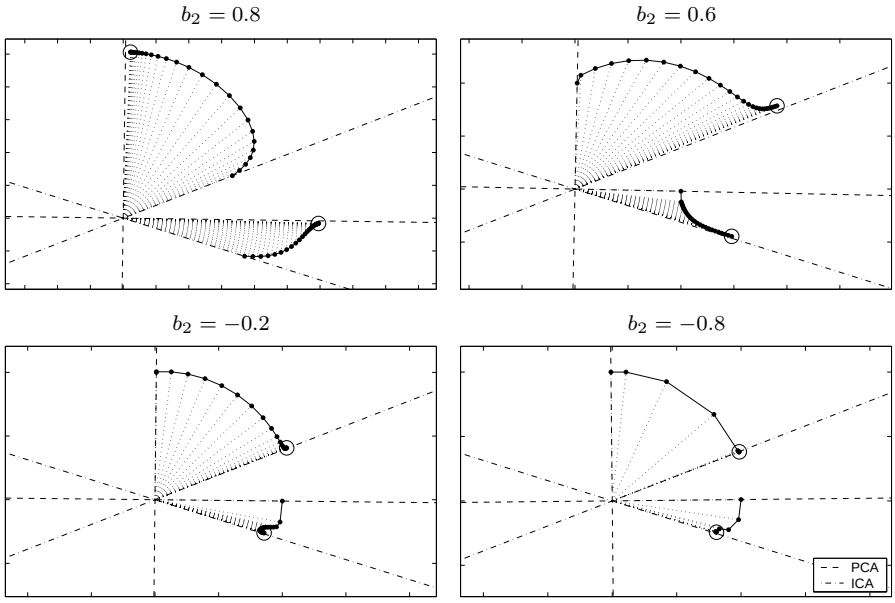


Figure 4. ICA model with temporally correlated sources. The results achieved with the unrestricted $q(\mathbf{s}(t))$ for the same data sets as in Figure 3. The current estimates of the columns of $\mathbf{A}$ are plotted after every 100 iterations for $b_2 = -0.2, -0.8$, every 1000 iterations for $b_2 = 0.6$ and every 5000 iterations for $b_2 = 0.8$, the intervals between the ticks on all axes are equal. The rotation of the solution is much slower in the case when the souce dynamics is just slightly different ($b_1 = 0.8$, $b_2 = 0.6$).

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 factorial approximation: The ICA solution is found except for the case where $b_1 = b_2$. In that case, the model converged to the PCA solution despite initialisation in the ICA solution.

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 is equal (i.e. $b_1 = b_2$), 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. In the presented experiments, the rotation in this case is defined by the factorial form of $q(\mathbf{A})$ which yields the principal component solution as explained in Section 4.3.

6. Non-Gaussian source models

In Sections 3–5, the rotational invariance was removed by introducing linear dynamics with diagonal matrix $\mathbf{B}$. A more common way to fix the rotation is to model the sources by a non-Gaussian distribution. In this section, we extend the analysis to the case of non-Gaussian source models. We consider two different non-Gaussian models: a simple model for super-Gaussian sources and the most commonly used mixture-of-Gaussians (MoG) model.

6.1. Super-Gaussian source model

If the distribution of the sources is known to be symmetric and super-Gaussian (which corresponds to positive kurtosis), an easy way to model the source distribution is using a Gaussian distribution

$$s_j(t) \sim \mathcal{N}\left(s_j(t) \,|\, 0, \sigma_j^2(t)\right) \tag{15}$$

with zero mean and variance $\sigma_j^2(t)$ changing with time (see examples of sources generated according to this model in Figure 5). We model the changing variance $\sigma_j^2(t)$ using a variance node, another time-dependent Gaussian variable $u_j(t)$ (Valpola et al., 2004), which yields the density model presented in Appendix B.1.

For this ICA model, the optimal unrestricted posterior approximation $q(\mathbf{s}(t))$ can be shown to be Gaussian with the covaraince of a form similar to (11):

$$\Sigma_{\mathbf{s}(t),\mathrm{opt}} = \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \Sigma_m(t)^{-1} \right\rangle^{-1}$$

where $\Sigma_m(t)$ is the time-dependent diagonal covariance of the source prior. And again, the expectation $\left\langle \mathbf{s}(t)\mathbf{s}^T(t) \right\rangle$ appears in the optimal covariance for $q(\mathbf{A}_{i,:})$ just as in (14). Therefore, the same effect of the posterior approximation is expected for this source model as well.

We studied this model experimentally on 10-dimensional mixtures of two super-Gaussian sources generated according to the Gaussian model (15) with the changing variance

$$\sigma_j^2(t) = e^{-u_j(t)}\,, \qquad u_j(t) \sim N(0, (2\nu)^2)\,. \tag{16}$$

The variance parameter ν was varied over the range $[0.6,\ 1]$ in order to control the non-Gaussianity of the generated sources and, therefore, the strength of evidence in support of the ICA solution present in the data (see Figure 5). The same mixing matrix $\mathbf{A}$ as in Section 5 was used.

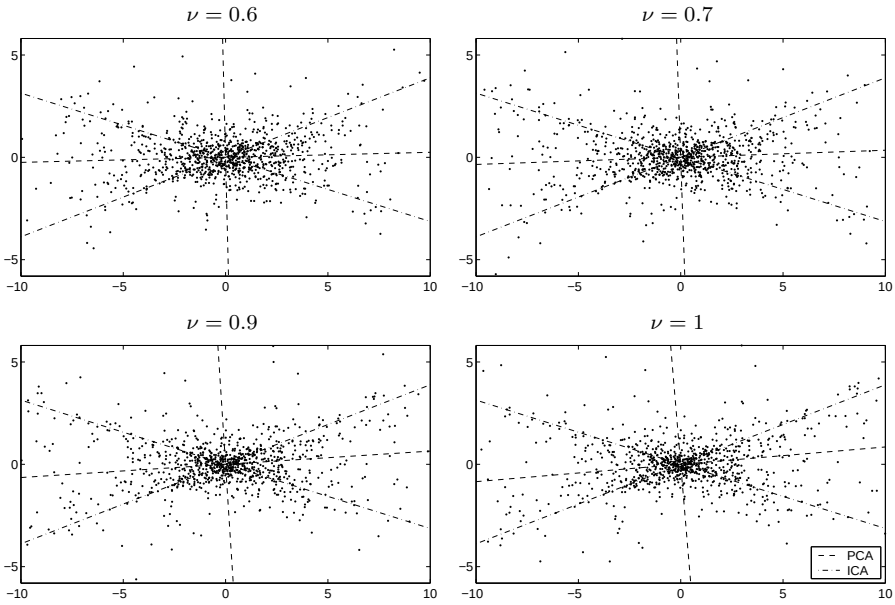


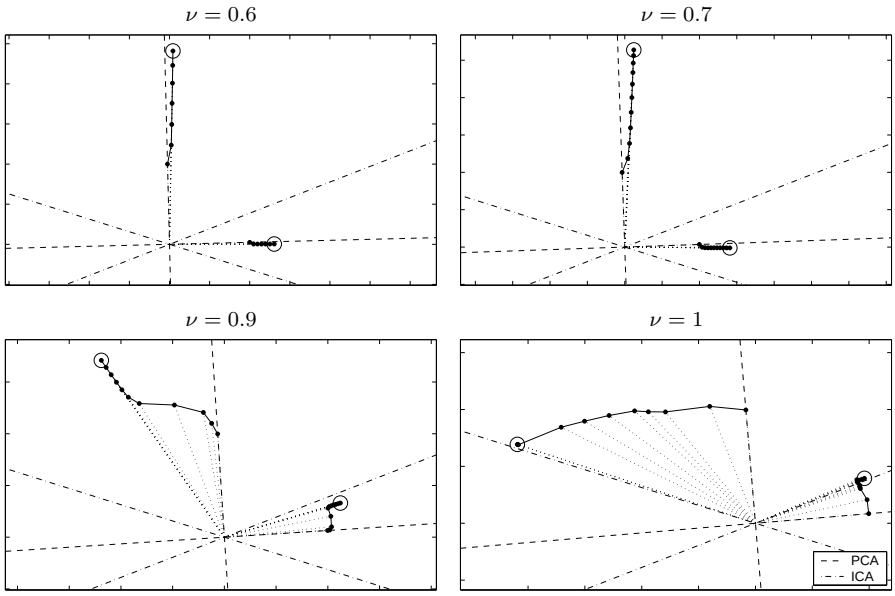
Figure 5. The noisy linear mixtures of two super-Gaussian sources for different values of the variance parameter ν. The samples are plotted in the subspace spanned by the columns of the mixing matrix. The PCA and ICA directions are also shown on the plots.

The model with the factorial approximation $q(\mathbf{s}(t))$ was implemented using the building blocks software (Valpola et al., 2003) based on the learning rules presented in (Valpola et al., 2001). The results of the simulations for two different initialisations are presented in Figure 6. The same effect clearly appears in this model as well. An interesting result of these experiments is the existence of two local minima for mediate ν: one with nearly orthogonal mixing vectors (and close to PCA for small values of ν) and the other one close to ICA.

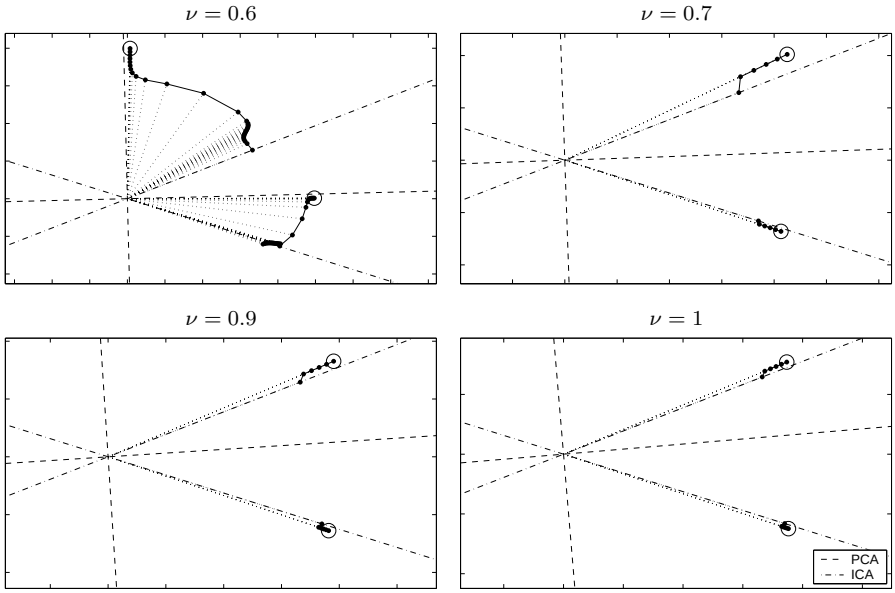
The model with the unrestricted $q(\mathbf{s}(t))$ was implemented using Matlab (Ilin, 2004) based on the learning rules presented in Appendix B.2. In the experiments with the unrestricted $q(\mathbf{s}(t))$, the correct ICA solution was found for all the four data sets (the results are not presented here).

6.2. Mixture-of-Gaussians model for sources

We now study the same effect of the posterior approximation for the mixture-of-Gaussians source model which is most commonly used in variational Bayesian ICA (Attias, 1999; Lappalainen, 1999; Miskin and MacKay, 2000; Choudrey et al., 2000; Valpola, 2000; Chan et al., 2002).



(a) Initialisation in PCA



(b) Initialisation in ICA

Figure 6. ICA with the super-Gaussian source model and the factorial $q(\mathbf{s}(t))$. The track of the columns of $\mathbf{A}$ during learning for four data sets with $\nu \in \{0.6, 0.7, 0.9, 1.0\}$. The current estimates of the columns of $\mathbf{A}$ are plotted after every 5000 iterations for the PCA initialisation and every 10000 iterations for the ICA initialisation. The final solution is circled. The intervals between the ticks on all axes are equal.

The optimal unrestricted posterior $q(\mathbf{s}(t))$ for this model would be a mixture of Gaussians, typically with a large number of mixture components: Whereas the prior mixture $p(\mathbf{s}(t))$ can be expressed as a product of simple mixtures $p(s_j(t))$ with K_j components each, modelling posterior correlations means that each and every multivariate Gaussian has to be modelled separately in the posterior. Therefore, the optimal source posterior is

$$\begin{aligned} q_{\text{opt}}(\mathbf{s}(t)) &= \sum_{\boldsymbol{\lambda}} q(\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}) q(\mathbf{s}(t) \,|\, \boldsymbol{\lambda}(t) = \boldsymbol{\lambda}) \\ &= \sum_{\boldsymbol{\lambda}} q(\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}) \mathcal{N}\left(\mathbf{s}(t) \,|\, \mu_{\mathbf{s}(t),\boldsymbol{\lambda}}, \Sigma_{\mathbf{s},\boldsymbol{\lambda}}\right) \end{aligned} \tag{17}$$

where $\boldsymbol{\lambda}$ is a vector whose components $\lambda_j \in \{1, \ldots, K_j\}$ define the Gaussians chosen for sources s_j. There are $\prod_j K_j$ choices for $\boldsymbol{\lambda}$ and therefore the source posterior is a mixture of $\prod_j K_j$ Gaussians. The sum $\sum_{\boldsymbol{\lambda}}$ means $\sum_{\lambda_1=1}^{K_1} \cdots \sum_{\lambda_m=1}^{K_m}$ and the covariance matrices of the mixture components are as follows:

$$\Sigma_{\mathbf{s},\boldsymbol{\lambda}} = \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \Sigma_{m,\boldsymbol{\lambda}} \right\rangle^{-1} . \tag{18}$$

Here, $\Sigma_{m,\boldsymbol{\lambda}}$ is the diagonal covariance of the conditional source prior $p(\mathbf{s}(t)|\boldsymbol{\lambda}(t) = \boldsymbol{\lambda})$. Note that the mixture covariances $\Sigma_{\mathbf{s},\boldsymbol{\lambda}}$ are same for all $\mathbf{s}(t)$.

As follows from (18), using the factorial approximation $q(\mathbf{s}(t) \,|\, \boldsymbol{\lambda}(t) = \boldsymbol{\lambda}) = \prod_j q(s_j(t) \,|\, \boldsymbol{\lambda}(t) = \boldsymbol{\lambda})$ yields the same orthogonalising effect for the mixing matrix $\mathbf{A}$ as in the models already considered in this article.

The approach proposed in (Miskin and MacKay, 2000) uses a simpler (and therefore coarser) approximation of the cost function (4). The MoG prior $p(s_j(t))$ for every source sample is approximated by only one Gaussian whose parameters are calculated using a set of coefficients λ_{t,j,k_j} (see the description of this model in Appendix C). This yields a Gaussian posterior $q(\mathbf{s}(t))$ whose optimal full covariance is

$$\Sigma_{\mathbf{s}(t),\text{opt}} = \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \text{diag}\left(\begin{bmatrix} \vdots \\ \sum_{k_j=1}^{K_j} \lambda_{t,j,k_j} \sigma_{j,k_j}^{-2} \\ \vdots \end{bmatrix} \right) \right\rangle^{-1} \tag{19}$$

where σ_{j,k_j}^2 is the variance of the k_j mixture component in the prior for s_j. The covariance $\Sigma_{\mathbf{s}(t),\text{opt}}$ is again similar to (11), and becomes diagonal if and only if $\mathbf{A}$ has orthogonal columns w.r.t. the inverse noise covariance Σ_n^{-1}. And, of course, the same optimal $q(\mathbf{A}_{i,:})$ like in (14) appears in the MoG models as well.

In the experiments with the MoG model for the sources, we used 10-dimensional mixtures of two super-Gaussian sources generated according to (15)-(16) with the parameter ν varying over the range [0.6, 1.2].

The simpler models with the Gaussian $q(\mathbf{s}(t))$ and covariance (19) were implemented in Matlab (Ilin, 2004) according to the learning rules presented in Appendix C.2. The fully factorial MoG posterior (17) was implemented using the building blocks presented in (Valpola et al., 2001) and the MoG block presented in (Harva, 2004). The number of mixture components for each of the two sources was set to three in all experiments.

Figure 7 shows the simulation results achieved with the factorial Gaussian $q(\mathbf{s}(t))$: The same orthogonalising effect is clearly demonstrated experimentally. Note that these results are very similar to the ones presented in Section 6.1 for the super-Gaussian source model (see Figure 6 for comparison). Qualitatively same results were obtained with the factorial MoG $q(\mathbf{s}(t))$ as well.

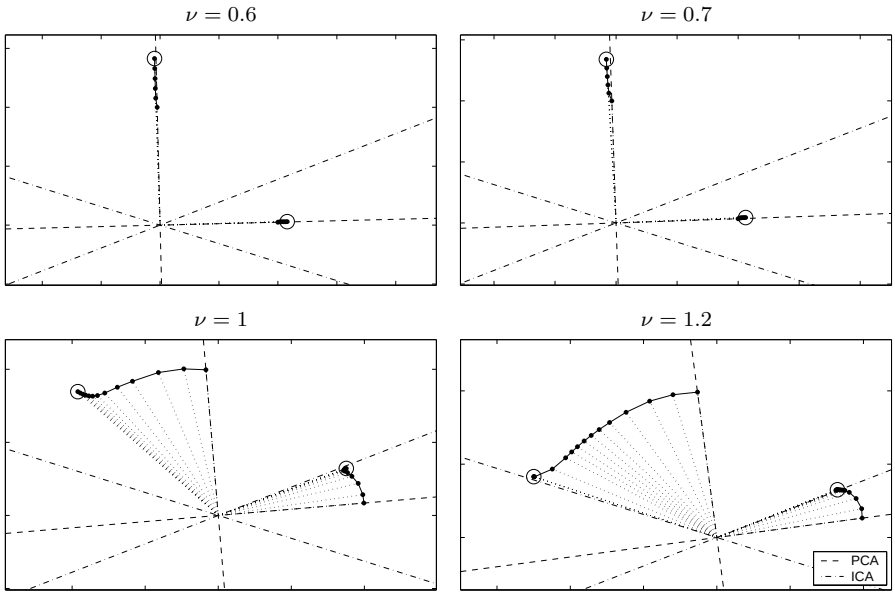
The experiments also showed that modelling the posterior correlations in $q(\mathbf{s}(t))$ helps remove the orthogonalising effect: In the case of the Gaussian $q(\mathbf{s}(t))$ with the full covariance matrix (19), the correct ICA solution was found for all the four data sets.

7. Discussion

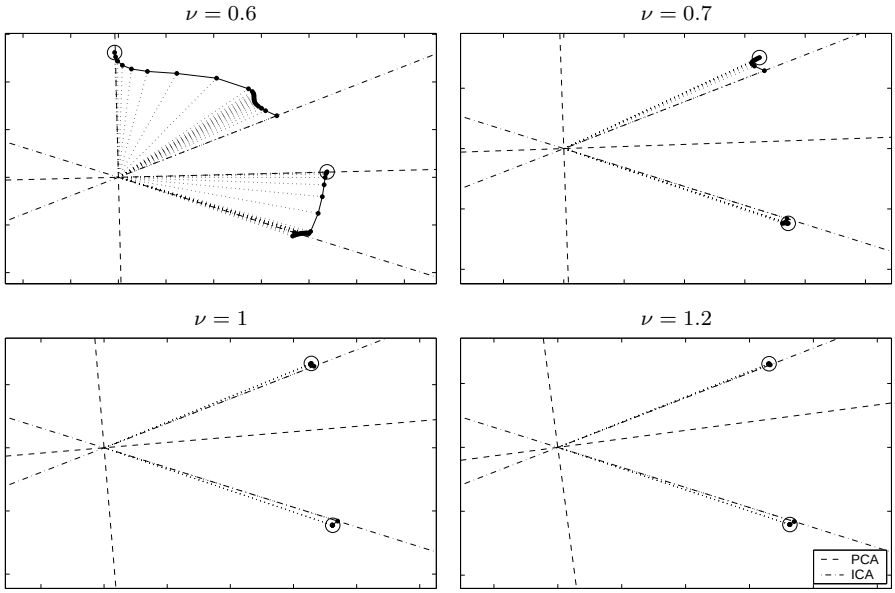
As we have seen, the form of the posterior approximation can strongly affect the result found by ensemble learning. We based the analysis on linear models with either temporally correlated or non-Gaussian sources for the sake of simplicity. The situation is slightly more complicated with nonlinear mixtures 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 nonlinear cases as well. In nonlinear mixtures, 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 feed-forward mappings with element-wise 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



(a) Initialisation in PCA



(b) Initialisation in ICA

Figure 7. ICA with the MoG source model and the factorial $q(\mathbf{s}(t))$. The track of the columns of $\mathbf{A}$ during learning for four data sets with $\nu \in \{0.6, 0.7, 1.0, 1.2\}$. The current estimates of the columns of $\mathbf{A}$ are plotted after every 2000 iterations for the PCA initialisation and every 1000 iterations for the ICA initialisation. The final solution is circled. The intervals between the ticks on all axes are equal.

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.

Acknowledgments

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. The authors would like to thank Markus Harva for his contribution to the code for the factorial MoG posterior.

References

Attias, H.: 1999, 'Independent Factor Analysis'. *Neural Computation* **11**(4), 803–851.

Barber, D. and C. Bishop: 1998, 'Ensemble Learning for Multi-Layer Networks'. In: M. Jordan, M. Kearns, and S. Solla (eds.): *Advances in Neural Information Processing Systems 10*. Cambridge, MA, USA: The MIT Press, pp. 395–401.

Belouchrani, A., K. A. Meraim, J.-F. Cardoso, and E. Moulines: 1997, 'A blind source separation technique based on second order statistics'. *IEEE Trans. on Signal Processing* **45**(2), 434–44.

Chan, K., T. Lee, and T. J. Sejnowski: 2002, 'Variational Learning of Clusters of Undercomplete Nonsymmetric Independent Components'. *Journal of Machine Learning Research* **3**, 99–114.

Chan, K., T. Lee, and T. J. Sejnowski: 2003, 'Variational Bayesian Learning of ICA with Missing Data'. *Neural Computation* **15**(8), 1991–2011.

Choudrey, R., W. Penny, and S. Roberts: 2000, 'An Ensemble Learning Approach to Independent Component Analysis'. In: *Proc. of the IEEE Workshop on Neural Networks for Signal Processing, Sydney, Australia, December 2000*. pp. 435–444, IEEE Press.

Ghahramani, Z. and M. Beal: 2001, 'Propagation Algorithms for Variational Bayesian Learning'. In: T. Leen, T. Dietterich, and V. Tresp (eds.): *Advances in Neural Information Processing Systems 13*. Cambridge, MA, USA, pp. 507–513, The MIT Press.

Ghahramani, Z. and G. E. Hinton: 2000, 'Variational Learning for Switching State-Space Models'. *Neural Computation* **12**(4), 963–996.

Harva, M.: 2004, 'Hierarchical Variance Models of Image Sequences'. Master's thesis, Helsinki University of Technology, Espoo.

Hyvärinen, A., J. Karhunen, and E. Oja: 2001, *Independent Component Analysis*. J. Wiley.

Ilin, A.: 2004, 'Matlab code for variational Bayesian learning of linear ICA models'. *http://www.cis.hut.fi/projects/bayes/software/npl2004/*.

Ilin, A. and H. Valpola: 2003, 'On the Effect of the Form of the Posterior Approximation in Variational Learning of ICA Models'. In: *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*. Nara, Japan, pp. 915–920.

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

Miskin, J. and D. J. C. MacKay: 2000, 'Ensemble Learning for Blind Image Separation and Deconvolution'. In: M. Girolami (ed.): *Advances in Independent Component Analysis*. Springer-Verlag, pp. 123–141.

Tong, L., V. Soo, R. Liu, and Y. Huang: 1990, 'Amuse: a new blind identification algorithm'. In: *Proc. ISCAS*. New Orleans, USA.

Valpola, H.: 2000, '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.

Valpola, H., M. Harva, and J. Karhunen: 2004, 'Hierarchical Models of Variance Sources'. *Signal Processing* **84**(2), 267–282.

Valpola, H., A. Honkela, M. Harva, A. Ilin, T. Raiko, and T. Östman: 2003, 'Bayes Blocks software library'. *http://www.cis.hut.fi/projects/bayes/software/*.

Valpola, H. and J. Karhunen: 2002, 'An Unsupervised Ensemble Learning Method for Nonlinear Dynamic State-Space Models'. *Neural Computation* **14**(11), 2647–2692.

Valpola, H., T. Raiko, and J. Karhunen: 2001, 'Building Blocks for Hierarchical Latent Variable Models'. In: *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*. San Diego, USA, pp. 710–715.

Ziehe, A., K.-R. Mller, G. Nolte, B.-M. Mackert, and G. Cuiro: 1998, 'Artifact Reduction in Magnetoneurography Based on Time-Delayed Second Order Correlations'. Technical Report 31, GMD - Forschungszentrum Informationstechnik GmbH.

Appendix

A. ICA with temporally correlated sources

A.1. 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; $\text{diag}(\mathbf{v})$ denotes a diagonal matrix with the elements of vector $\mathbf{v}$ on its main diagonal; and the exponential function $e^{-\mathbf{v}}$ is applied component-wise to the elements of its vector argument $\mathbf{v}$.

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

The prior for the (hyper)parameters $\mathbf{A}$, $\boldsymbol{\theta}_{\text{rest}}$:

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

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

A.2. The learning rules

The parameters of $q(\mathbf{v}_x)$, $q(\mathbf{v}_s)$, $q(m_{v_x})$, $q(v_{v_x})$, $q(m_{v_s})$, $q(v_{v_s})$ and factorial $q(\mathbf{s}_t)$ are updated using the rules presented in (Valpola et al.,

2001). The only difference is calculating the variance $\widetilde{f}_i$ of the function $f_i = \mathbf{A}_{i,:}\mathbf{s}$ when updating $q(v_{x,i})$:

$$\widetilde{f}_i = \langle \mathbf{A}_{i,:} \rangle \Sigma_{\mathbf{s}} \langle \mathbf{A}_{i,:} \rangle^T + \sum_{j=1}^{m} \widetilde{a}_{ij} \left\langle s_j^2 \right\rangle$$

where $\Sigma_{\mathbf{s}}$ is the posterior covariance of $\mathbf{s}$.

The following recursive learning rules for $\mathbf{s}_t = \mathbf{s}(t)$, $\mathbf{A}$, $\mathbf{B}$ are obtained as a result of using conjugate priors.

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

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

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

   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.

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

$$\begin{aligned} q(\mathbf{A}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \mathcal{N}\left(a_{ij} | \overline{a}_{ij}, \widetilde{a}_{ij}\right) \\ \widetilde{a}_{ij}^{-1} &= \langle \alpha_j \rangle + \langle e^{v_{x,i}} \rangle \sum_{t=1}^{N} \left\langle s_{t,j}^2 \right\rangle \\ \overline{a}_{ij} &= \widetilde{a}_{ij} \langle e^{v_{x,i}} \rangle \sum_{t=1}^{N} [x_{t,i} \langle s_{t,j} \rangle - \sum_{k \neq j} \langle a_{ik} \rangle \langle s_{t,k} s_{t,j} \rangle] \end{aligned}$$

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

$$\begin{aligned} q(\mathbf{B}) &= \prod_{j=1}^{m} \mathcal{N}\left(b_j | \overline{b}_j, \widetilde{b}_j\right) \\ \widetilde{b}_j^{-1} &= \langle \beta_j \rangle + \langle e^{v_{s,j}} \rangle \sum_{t=2}^{N} \left\langle s_{t-1,j}^2 \right\rangle \\ \overline{b}_j &= \widetilde{b}_j \langle e^{v_{s,j}} \rangle \sum_{t=2}^{N} \langle s_{t,j} s_{t-1,j} \rangle \end{aligned}$$

B. ICA with super-Gaussian source model

B.1. The density model

The ICA model with super-Gaussian sources (see Section 6.1) has the same density model as in Appendix A.1 except for the prior for the sources $\mathbf{S}$:

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

where $\Sigma_m(t) = \text{diag}(e^{-\mathbf{u}(t)})$.

The prior for the hyperparameters corresponding to the source model:

$$\begin{aligned}
p(\{\mathbf{u}(t)\}_{t=1}^{N}|\mathbf{m}_u, \mathbf{v}_u) &= \prod_{t=1}^{N} \mathcal{N}\left(\mathbf{u}(t)|\mathbf{m}_u, \text{diag}(e^{-\mathbf{v}_u})\right) \\
p(\mathbf{m}_u|m_{m_u}, v_{m_u}) &= \prod_{j=1}^{m} \mathcal{N}\left(m_{u,j}|m_{m_u}, e^{-v_{m_u}}\right) \\
p(\mathbf{v}_u|m_{v_u}, v_{v_u}) &= \prod_{j=1}^{m} \mathcal{N}\left(v_{u,j}|m_{v_u}, e^{-v_{v_u}}\right)
\end{aligned}$$

$$m_{m_u}, v_{m_u}, m_{v_u}, v_{v_u} \sim \mathcal{N}(0, \gamma)$$

B.2. The learning rules

The parameters of $q(m_{v_x})$, $q(v_{v_x})$, $q(\mathbf{u}_t)$, $q(\mathbf{m}_u)$, $q(\mathbf{v}_u)$, $q(m_{m_u})$, $q(v_{m_u})$, $q(m_{v_u})$, $q(v_{v_u})$ and factorial $q(\mathbf{s}_t)$ are updated using the rules presented in (Valpola et al., 2001). The update rules for the parameters $\mathbf{v}_x$, $\mathbf{A}$ are the same as in Appendix A. The update rules for unrestricted $q(\mathbf{s}_t)$ are obtained as a result of using conjugate prior:

$$\begin{aligned}
q(\mathbf{s}_t) &= \mathcal{N}\left(\mathbf{s}_t|\overline{\mathbf{s}}_t, \Sigma_{\mathbf{s}_t}\right) \\
\Sigma_{\mathbf{s}_t} &= \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{A} + \Sigma_m(t)^{-1} \right\rangle^{-1} \\
\overline{\mathbf{s}}_t &= \Sigma_{\mathbf{s}_t} \left\langle \mathbf{A}^T \Sigma_n^{-1} \mathbf{x}_t \right\rangle
\end{aligned}$$

C. ICA with MoG source model

C.1. The density model

The ICA model with the MoG source prior (see Section 6.2) has the same density model as in Appendix A.1 except for the prior for the sources $\mathbf{S}$:

$$p(\mathbf{S}|\boldsymbol{\theta}_{\text{rest}}) = \prod_{t=1}^{N}\prod_{j=1}^{m}\sum_{k=1}^{K_j}\pi_{j,k}\mathcal{N}\left(s_{t,j}|m_{j,k}, e^{-v_{j,k}}\right)$$

The prior for the hyperparameters corresponding to the source model:

$$\begin{aligned}
p(\{\pi_{j,k}\}_{k=1}^{K_j}|\mathbf{c}) &= \text{Dirichlet}(\{\pi_{j,k}\}_{k=1}^{K_j}|\{c_{j,k}\}_{k=1}^{K_j}) \\
p(\{m_{j,k}\}_{k=1}^{K_j}|m_{m,j}, v_{m,j}) &= \prod_{k=1}^{K_j}\mathcal{N}\left(m_{j,k}|m_{m,j}, e^{-v_{m,j}}\right) \\
p(\{v_{j,k}\}_{k=1}^{K_j}|m_{v,j}, v_{v,j}) &= \prod_{k=1}^{K_j}\mathcal{N}\left(v_{j,k}|m_{v,j}, e^{-v_{v,j}}\right) \\
p(m_{m,j}|m_{m_m}, v_{m_m}) &= \mathcal{N}\left(m_{m,j}|m_{m_m}, e^{-v_{m_m}}\right) \\
p(v_{m,j}|m_{v_m}, v_{v_m}) &= \mathcal{N}\left(v_{m,j}|m_{v_m}, e^{-v_{v_m}}\right) \\
p(m_{v,j}|m_{m_v}, v_{m_v}) &= \mathcal{N}\left(m_{v,j}|m_{m_v}, e^{-v_{m_v}}\right) \\
p(v_{v,j}|m_{v_v}, v_{v_v}) &= \mathcal{N}\left(v_{v,j}|m_{v_v}, e^{-v_{v_v}}\right) \\
j &= 1, \ldots, m
\end{aligned}$$

$$m_{m_m}, v_{m_m}, m_{v_m}, v_{v_m}, m_{m_v}, v_{m_v}, m_{v_v}, v_{v_v} \sim \mathcal{N}(0, \gamma)$$

A set of coefficients $\lambda_{t,j,k}$ simplifying the cost function is used like in (Miskin and MacKay, 2000; Chan et al., 2002):

$$-\log p(s_{t,j}|\boldsymbol{\theta}_{\text{rest}}) \leq \sum_{k=1}^{K_j}\lambda_{t,j,k}\log\frac{\pi_{j,k}\mathcal{N}\left(s_{t,j}|m_{j,k}, e^{-v_{j,k}}\right)}{\lambda_{t,j,k}}$$

C.2. The learning rules

The update rules for the parameters $\mathbf{v}_x$, m_{v_x}, v_{v_x}, $\mathbf{A}$ are the same as in Appendix A. The parameters of $q(m_{m,j})$, $q(v_{m,j})$, $q(m_{v,j})$, $q(v_{v,j})$, $q(m_{m_m})$, $q(v_{m_m})$, $q(m_{v_m})$, $q(v_{v_m})$, $q(m_{m_v})$, $q(v_{m_v})$, $q(m_{v_v})$, $q(v_{v_v})$ are updated using the rules presented in (Valpola et al., 2001). The rest of the update rules are as follows:

1. The update rule for $\lambda_{t,j,k}$ is

$$\lambda_{t,j,k} \propto \exp\{\langle \log \pi_{j,k} + \log \mathcal{N}\left(s_{t,j}|m_{j,k}, e^{-v_{j,k}}\right)\rangle\}$$

with the normalisation conditions

$$\sum_{k=1}^{K_j} \lambda_{t,j,k} = 1\,.$$

2. The parameters of $q(m_{j,k})$, $q(v_{j,k})$ are updated similarily to the rules presented in (Valpola et al., 2001) with the exception that the gradients from the children $s_{t,j}$ are weighed by the coefficients $\lambda_{t,j,k}$.

3. The update rule for $q(\{\pi_{j,k}\}_{k=1}^{K_j})$ is obtained as a result of using conjugate prior:

$$q(\{\pi_{j,k}\}_{k=1}^{K_j}) = \text{Dirichlet}(\{\pi_{j,k}\}_{k=1}^{K_j}|\{c_{j,k} + \textstyle\sum_{t=1}^{N} \lambda_{t,j,k}\}_{k=1}^{K_j})$$

4. The update rules for $q(\mathbf{s}_t)$ with full covariance:

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

$$\Sigma_{\mathbf{s}_t} = \left\langle \mathbf{A}^T\Sigma_n^{-1}\mathbf{A} + \text{diag}\left(\begin{bmatrix} \vdots \\ \sum_{k_j=1}^{K_j} \lambda_{t,j,k_j} e^{v_{j,k_j}} \\ \vdots \end{bmatrix}\right)\right\rangle^{-1}$$

$$\overline{\mathbf{s}}_t = \Sigma_{\mathbf{s}_t}\left\langle \mathbf{A}^T\Sigma_n^{-1}\mathbf{x}_t + \begin{bmatrix} \vdots \\ \sum_{k_j=1}^{K_j} \lambda_{t,j,k_j} e^{v_{j,k_j}} m_{j,k_j} \\ \vdots \end{bmatrix}\right\rangle$$

The update rules for $q(\mathbf{s}_t)$ with diagonal covariance:

$$q(\mathbf{s}_t) = \prod_{j=1}^{m} \mathcal{N}\left(s_{t,j}|\overline{s}_{t,j}, \widetilde{s}_{t,j}\right)$$

$$\widetilde{s}_{t,j}^{-1} = \sum_{k=1}^{K_j} \lambda_{t,j,k}\left\langle e^{v_{j,k}}\right\rangle + \sum_{i=1}^{n}\left\langle e^{v_{x,i}}\right\rangle\left\langle a_{ij}^2\right\rangle$$

$$\overline{s}_{t,j} = \widetilde{s}_{t,j}\left(\sum_{k=1}^{K_j} \lambda_{t,j,k}\left\langle e^{v_{j,k}} m_{j,k}\right\rangle + \right.$$

$$\left. + \sum_{i=1}^{n}\left\langle e^{v_{x,i}}\right\rangle\left\langle a_{ij}\right\rangle\left[x_{t,i} - \sum_{k\neq j}\left\langle a_{ik}\right\rangle\left\langle s_{t,k}\right\rangle\right]\right)$$