FLEXIBLE BAYESIAN INDEPENDENT COMPONENT ANALYSIS FOR BLIND SOURCE SEPARATION

R.A. Choudrey and S.J. Roberts

University of Oxford
Robotics Research
Oxford, U.K.

ABSTRACT

Independent Component Analysis (ICA) is an important tool for extracting structure from data. ICA is traditionally performed under a maximum likelihood scheme in a latent variable model and in the absence of noise. Although extensively utilised, maximum likelihood estimation has well known drawbacks such as overfitting and sensitivity to local-maxima. In this paper, we propose a Bayesian learning scheme using the *variational* paradigm to learn the parameters of the model, estimate the source densities, and - together with *Automatic Relevance Determination* (ARD) - to infer the number of latent dimensions. We illustrate our method by separating a noisy mixture of images, estimating the noise and correctly inferring the true number of sources.

1. INTRODUCTION

Independent Component Analysis (ICA) seeks to extract salient features and structure from a dataset which is assumed to be a linear mixture of independent underlying (hidden) features. The goal of ICA is to 'unmix' the dataset and recover these features.

ICA has traditionally been performed in the noise-less limit [1], [2], with noise often being dealt with as an extra source. More recently, however, Attias [3] extended ICA and incorporated full covariance noise into the ICA framework. The model, dubbed by Attias as Independent Factor Analysis (IFA), was subsequently learnt through a maximum likelihood EM algorithm.

Lappalainen introduced an *ensemble learning* formalism (a special case of the variational framework) for ICA in [4], where the posterior over the 'ensemble' of hidden variables and parameters is approximated. A similar method is used in [5] but where a richer variety of functional forms for the priors is used. In addition to this, an extra distribution is placed over the variances in the mixing matrix prior in an attempt to automatically determine the number of hidden sources, a practice known as Automatic Relevance Determination (ARD) [6]. Crucially, however, the source model used in [5] is kept fixed and only the parameters of the sensor model are learnt. If the source model is not known, or not chosen correctly, an incorrect and ill-fitting model will be learnt. This is relaxed in [7], but only unimodal source densities are modelled, greatly restricting its flexibility.

In line with [3], we choose a fully-adaptable factorial Mixture of Gaussians (MoG) as our source model allowing us to recover arbitrary source densities. We further extend this formalism by bringing the model into the Bayesian sphere, allowing us to incorporate prior knowledge of the problem domain while avoiding over-fitting. We also employ ARD to infer the number of latent dimensions as part of the learning process. To overcome the heavy computational load associated with Bayesian learning, we use the variational framework to make assumptions about the posterior and thus allow tractability of the Bayesian model.

2. THE MODEL

In common with ICA in the literature, we choose a generative model to work with. The observed variables, $\mathbf{y}$, of dimension S are modelled as a linear combination of statistically independent latent variables, $\boldsymbol{x}$, of dimension L with added Gaussian noise

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{u} \tag{1}$$

where $\boldsymbol{H}$ is an $S \times L$ *mixing matrix* and $\boldsymbol{u}$ is S-dimensional additive noise. In signal processing nomenclature, S is the number of (observed) sensors and L is the number of latent (hidden) sources.

The noise is assumed to be Gaussian, with zero mean and diagonal precision matrix $\boldsymbol{\Lambda}$.The probability of observing data vector $\boldsymbol{y}^n$ is then given by

$$p(\boldsymbol{y}^n|\boldsymbol{x}^n, \boldsymbol{H}, \boldsymbol{\Lambda}) = \left|\det(\frac{1}{2\pi}\boldsymbol{\Lambda})\right|^{\frac{1}{2}} \exp[-E_D] \tag{2}$$

where

$$E_D = \frac{1}{2}(\boldsymbol{y}^n - \boldsymbol{H}\boldsymbol{x}^n)^T\boldsymbol{\Lambda}(\boldsymbol{y}^n - \boldsymbol{H}\boldsymbol{x}^n) \tag{3}$$

Since the sources $\boldsymbol{x} = \{x_1, .., x_i, .., x_L\}$ are mutually independent, the distribution over $\boldsymbol{x}$ for data point n can be written as

$$p(\boldsymbol{x}^n) = \prod_{i=1}^{L} p(x_i^n) \tag{4}$$

where the product runs over the L sources.

In ICA, one attempts to uncover the hidden source signals that give rise to a set of observed sensor signals. In principle, this is achieved by calculating the posterior over the latent variables (sources) given the observed variables (sensor signals) and the model

$$p(\boldsymbol{x}^n|\boldsymbol{y}^n, \mathcal{M}) = \frac{p(\boldsymbol{y}^n|\boldsymbol{x}^n, \mathcal{M})p(\boldsymbol{x}^n|\mathcal{M})}{p(\boldsymbol{y}^n|\mathcal{M})} \tag{5}$$

where $p(\boldsymbol{x}^n|\mathcal{M})$ is the source model and $p(\boldsymbol{y}^n|\mathcal{M})$ is a normalising factor often called the marginal likelihood, or *evidence* for model $\mathcal{M}$.

2.1. Source Model

The choice of a flexible and mathematically attractive (tractable) source model is crucial if a wide variety of source distributions are to be modelled; in particular, the source model should be capable of encompassing both super- and sub-Gaussian distributions (distributions with positive and negative kurtosis respectively) and complex multi-modal distributions.

One such distribution is a factorised mixture of Gaussians with L factors (i.e. sources) and m_i components per source:

$$\begin{aligned} p(\boldsymbol{x}^n|\boldsymbol{\theta}) &= \prod_{i=1}^{L}\sum_{q_i=1}^{m_i} p(q_i^n = q_i|\boldsymbol{\pi}_i)p(x_i^n|q_i,\boldsymbol{\mu}_{i,q_i},\beta_{i,q_i}) \\ &= \prod_{i=1}^{L}\sum_{q_i=1}^{m_i} \pi_{i,q_i}\mathcal{N}(x_i^n;\boldsymbol{\mu}_{i,q_i},\beta_{i,q_i}) \end{aligned} \quad (6)$$

where the mixing proportions $\pi_{i,q_i} = p(q_i^n = q_i|\boldsymbol{\pi}_i)$, the prior probability of choosing component q_i of the i^{th} source. q_i^n is a variable indicating which component of the i^{th} source is chosen for generating x_i^n and takes on values of $\{q_i = 1,\ldots,q_i = m_i\}$. The mean and precision of component q_i in source i are $\boldsymbol{\mu}_{i,q_i}$ and β_{i,q_i} respectively. The parameters of source i are $\boldsymbol{\theta}_i = \{\boldsymbol{\pi}_i,\boldsymbol{\mu}_i,\boldsymbol{\beta}_i\}$ where bold face indicates the vector of m_i parameters. The complete parameter set of the source model is $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_L\}$.

The complete collection of possible source states is denoted $\boldsymbol{q} = \{\boldsymbol{q}_1,\boldsymbol{q}_2,\ldots,\boldsymbol{q}_{\boldsymbol{m}}\}$ and runs over all $\boldsymbol{m} = \prod_i m_i$ possible combinations of source states. The probability of state $\boldsymbol{q}^n$ being chosen and generating source vector $\boldsymbol{x}^n$ is

$$\begin{aligned} p(\boldsymbol{x}^n,\boldsymbol{q}^n|\boldsymbol{\theta}) &= \prod_{i=1}^{L} p(q_i^n = q_i|\boldsymbol{\pi}_i)p(x_i^n|q_i,\boldsymbol{\mu}_{i,q_i},\beta_{i,q_i}) \\ &= p(\boldsymbol{q}^n|\boldsymbol{\pi})p(\boldsymbol{x}^n|\boldsymbol{q}^n,\boldsymbol{\theta}) \end{aligned} \quad (7)$$

where $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\ldots,\boldsymbol{\pi}_L\}$. Note that the product of L 1-dimensional MoGs in (6) is equivalent to a single MoG in L-dimensional space with $\boldsymbol{m}$ states.

The likelihood of the IID data $\boldsymbol{D} = \{\boldsymbol{y}^1,\boldsymbol{y}^2,\ldots,\boldsymbol{y}^N\}$ given the model parameters $\boldsymbol{\Theta} = \{\boldsymbol{H},\boldsymbol{\Lambda},\boldsymbol{\theta}\}$ can now be written as

$$p(\boldsymbol{D}|\boldsymbol{\Theta}) = \prod_{n=1}^{N}\sum_{\boldsymbol{q}=1}^{\boldsymbol{m}}\int p(\boldsymbol{y}^n,\boldsymbol{x}^n,\boldsymbol{q}^n|\boldsymbol{\Theta})d\boldsymbol{x} \quad (8)$$

where $d\boldsymbol{x} = \prod_i dx_i$.

The parameters, $\boldsymbol{\Theta}$, of the model can be learnt through a maximum likelihood approach such as the Expectation-Maximisation (EM) algorithm ([8], [9]) (see [3] for a comprehensive derivation of the EM algorithm with regard to ICA/IFA). The resultant values can then be used to reconstruct the sources via (5).

3. BAYESIAN INFERENCE AND VARIATIONAL LEARNING

The maximum likelihood approach to learning the parameters of the model is well documented (see [10], [11], [3] for an introduction), as are the pitfalls. We choose to take the Bayesian approach and infer the posterior distributions over parameters $\{\boldsymbol{H},\boldsymbol{\Lambda},\boldsymbol{\theta}\}$ and hidden variables $\{\boldsymbol{x},\boldsymbol{q}\}$. First, we will state the prior distributions over the hidden variables and model parameters.

3.1. The Priors

Because of source independence, it follows that the distribution over the MoG component indicator variables, $p(\boldsymbol{q}|\boldsymbol{\pi})$, is a product over all π_{i,q_i}^n where i indexes the sources and n the data. The prior over the source model (MoG) parameters is a product of priors over $\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\beta}$. The prior over the mixing proportions, $\boldsymbol{\pi}_i$, for the i^{th} source is a symmetric Dirichlet with hyper-parameter λ_{i0}. The prior over each MoG mean, μ_{i,q_i}, is a Gaussian with mean m_{i0} and precision τ_{i0} and the prior over the associated precision, β_{i,q_i}, is a Gamma with width and scale hyper-parameters b_{i0} and c_{i0} respectively.

The prior over the sensor noise precision, $\boldsymbol{\Lambda}$, is a product of Gamma distributions for each diagonal element, Λ_j, with width and scale hyper-parameters b_{Λ_j} and c_{Λ_j}. The prior over each element of the mixing matrix, H_{ji} is a zero-mean Gaussian with precision α_i for each column. By monitoring the evolution of the precisions $\boldsymbol{\alpha}$, the relevance of each source may be determined (ARD). If α_i is large, column i of $\boldsymbol{H}$ will be close to zero, indicating source i is irrelevant. Finally, the prior over each α_i is a Gamma($b_{\alpha_i},c_{\alpha_i}$).

Bayesian inference in such a model is computationally intensive and often intractable. An important and efficient tool in approximating posterior distributions is the *variational method* (see [12] for an excellent tutorial). In particular, we take the *variational Bayes* approach detailed in [13].

3.2. Variational Bayesian Learning

In the variational Bayes framework, the objective function to be maximised is the *negative free energy*, F

$$F = \langle\log p(\boldsymbol{D},\boldsymbol{W})\rangle_{p'(\boldsymbol{W})} + \mathcal{H}\left[p'(\boldsymbol{W})\right] \quad (9)$$

where $\boldsymbol{W} = \{\boldsymbol{\Lambda},\boldsymbol{H},\boldsymbol{\alpha},\boldsymbol{x},\boldsymbol{q},\boldsymbol{\theta}\}$. The first term in (9) is the expectation of the joint density of hidden and observed variables with respect to an approximating posterior $p'(\boldsymbol{W})$. The second term is the entropy of $p'(\boldsymbol{W})$. The negative free energy forms a strict lower bound on the evidence, $p(\boldsymbol{D}|\mathcal{M})$, of the model, with the difference being the Kullback-Leibler (KL) divergence between the true and approximating posteriors. Maximising this function is equivalent to minimising the KL divergence between the true and approximate posteriors. A wide variety of models and assumptions can be compared and contrasted by calculating the free energy of each model. The higher the free energy, the higher the likelihood of the data under that model, and, therefore, the better that model is at 'explaining' the data.

By choosing $p'(\boldsymbol{W})$ such that it factorises, terms in each hidden variable can be maximised individually. We choose the following factorisation

$$p'(\boldsymbol{W}) = p'(\boldsymbol{\Lambda})p'(\boldsymbol{H})p'(\boldsymbol{\alpha})p'(\boldsymbol{x}|\boldsymbol{q})p'(\boldsymbol{q})p'(\boldsymbol{\theta}) \quad (10)$$

where $p'(\boldsymbol{\theta}) = p'(\boldsymbol{\pi})p'(\boldsymbol{\mu})p'(\boldsymbol{\beta})$ and $p'(a|b)$ is the approximating density of $p(a|b,\boldsymbol{D})$. The term $p'(\boldsymbol{x}|\boldsymbol{q})$ in (10) implies a mixture posterior source density for source i

$$p'(x_i^n) = \sum_{q_i=1}^{m_i} p'(q_i^n = q_i)p'(x_i|q_i) \quad (11)$$

$$\doteq \sum_{q_i=1}^{m_i} \gamma_{i,qi}^n\mathcal{N}(x_i^n;\hat{\mu}_{i,q_i}^n,\hat{\beta}_{i,q_i}^n) \quad (12)$$

where we have stipulated a MoG density for reasons explained later.

We will also stipulate that the posteriors over the sources factorise such that

$$p'(\boldsymbol{x}) = \prod_{i=1}^{L} p'(x_i) \quad \text{and therefore} \quad p'(\boldsymbol{H}) = \prod_{i=1}^{L} p'(\boldsymbol{H}_i) \quad (13)$$

where $\boldsymbol{H}_i$ is the i^{th} column of the mixing matrix, $\boldsymbol{H}$. This additional factorisation allows efficient scaling of computation with the number of hidden sources, with little loss of accuracy [7].

By substituting $p(\boldsymbol{D}, \boldsymbol{W})$ and (10) into (9), we obtain expressions for the negative free energy, F, of our model. One may now proceed by specifying functional forms of each of the approximating posteriors and using these in (9) as shown by [4]. As shown in [14], however, there is no need to specify functional forms for (all) the approximating posteriors as they 'fall-out' of the maximisation process, helped by the factorised form of $p'(\boldsymbol{W})$. The optimal form for each posterior is simply given by

$$p'(W_k) \propto p(W_k) \exp\left[\langle \log p(\boldsymbol{D}, \boldsymbol{W}) \rangle_{\prod_{l \neq k} p'(W_l)}\right] \quad (14)$$

where the index k refers to the k^{th} parameter in $\boldsymbol{W}$.

This can be fully applied if $p'(\boldsymbol{x}) = p'(\boldsymbol{x}|\boldsymbol{q})$, allowing free-form optimisation giving the ensemble learning algorithms presented in [15]. This factorisation gives a Gaussian posterior over $\boldsymbol{x}$. Using (10), however, requires a functional form for (11) to allow expectations of the data likelihood, (2), under the mixture density (11) to be taken. To allow flexibility (and conjugacy with the prior), we specify (12). The posterior expectations are then given by

$$\langle x_i{}^n \rangle = \sum_{q_i=1}^{m_i} p'(q_i^n = q_i) \langle x_i{}^n | q_i \rangle \quad (15)$$

$$\langle x_i^{n\,2} \rangle = \sum_{q_i=1}^{m_i} p'(q_i^n = q_i) \langle x_i^{n\,2} | q_i \rangle \quad (16)$$

where

$$p'(q_i^n = q_i) = \gamma_{i,qi}^n \quad (17)$$

$$\langle x_i{}^n | q_i^n \rangle = \hat{\mu}_{i,q_i}^n \quad (18)$$

$$\langle x_i^{n\,2} | q_i^n \rangle = (\hat{\mu}_{i,q_i}^n)^2 + \frac{1}{\hat{\beta}_{i,q_i}^n} \quad (19)$$

The energy F is maximised in a similar way to the full free-form approach, except $p'(\boldsymbol{x}, \boldsymbol{q})$ which is found by differentiating this term and optimising w.r.t. the variational parameters in (12). The update equations for these parameters are very similar to those found in [3] (albeit with expectations of arguments rather than point estimates).

All the derived posteriors require solving a set of coupled hyper-parameter update equations. In practice, this is best achieved by first cycling through $p'(\boldsymbol{x}), p'(\boldsymbol{H}), p'(\boldsymbol{\alpha})$ until convergence. These values are then passed to $p'(\boldsymbol{\theta})$, whose constituent updates are cycled until convergence. The hyper-parameters for $p'(\boldsymbol{\Lambda})$ are updated, then the whole process is repeated until convergence.

Once trained, the model can be used to reconstruct hidden source signals (to within a scaling and permutation) given a data-set by calculating $\langle q_i \rangle$ and $\langle x_i \rangle$ under their respective posteriors over the whole data-set, and given the (now fixed) model parameter posteriors.

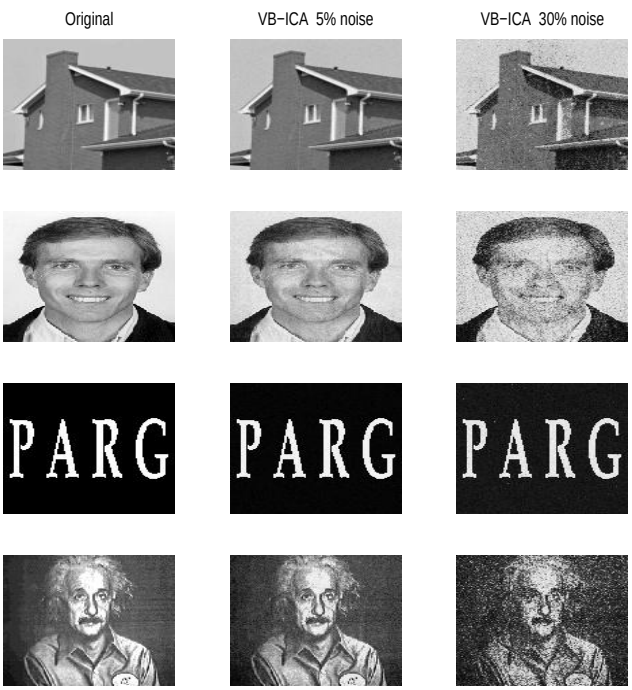

Fig. 1. The source images and their reconstructions using VB-ICA

4. RESULTS

Four 127x127 images with very different densities (see figures 1 and 2) were mixed to produce eight sensor signals giving a 8x16129 data matrix. This was duplicated with zero-mean isotropic noise added with a variance of 0.1 (approximately 5pct noise) to one set while noise with a variance of 2 (approximately 30pct noise) was added to the other. 1000 data-vectors were drawn at random from each set (the same random vectors from both). Two sets of models ranging from latent dimensionality 1-8, and with 5 components per MoG, were trained on the two datasets using the variational Bayes ICA (VB-ICA) algorithm. The variables $\boldsymbol{x}$, $\boldsymbol{H}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$ were initialised using SVD while the MoG parameters were initialised using k-means clustering on $\boldsymbol{x}$. We then ran the models until the negative free energy converged (to within 0.01 pct) or until a maximum of 200 iterations was reached. We then used the trained models to unmix the whole 8x16129 data-matrix and reassemble the original pictures. Due to the indeterminancy introduced by the noise, ICA can only ever recover a scaled permutation of the original sources. In the results that follow, the reconstructions have been re-ordered and scaled by -1 (if necessary) to aid comparison.

4.1. Blind source seperation

Figure 1 compares the reconstructed images with the originals for the 4-source models. Both low and high noise data matrices are unmixed convincingly well. All images are very well seperated, albeit with the high-noise reconstructions exhibiting some noise. The low-noise reconstructions are particularly impressive, with little or no cross-talk. The reasons for such good recovery are evident in figure 2. The first column shows a histogram of the original images. The low-noise models capture the shape of the densities very accurately, while the high-noise pdfs have captured the general

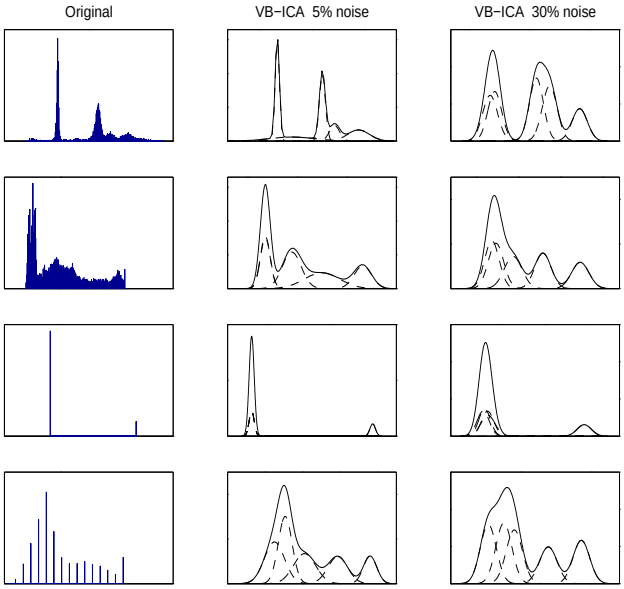

Fig. 2. The source densities and their reconstructions

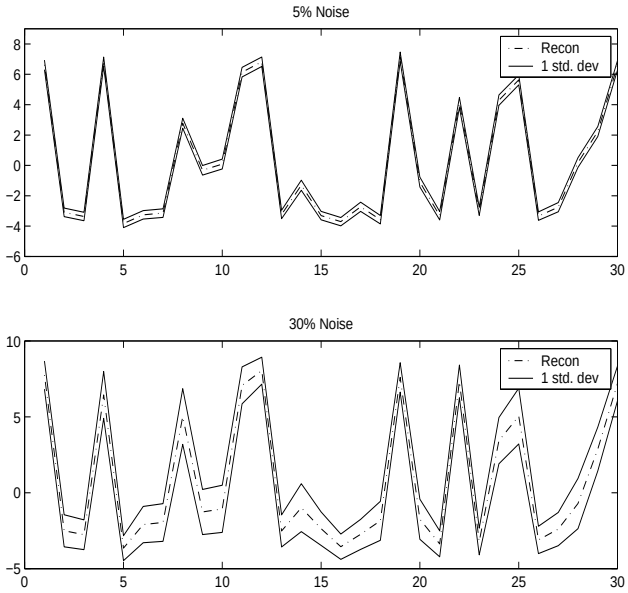

Fig. 3. Example reconstruction with errors

shapes even though they have modelled some of the noise. In particular, note the final row; the recovered densities are 'smoothed' versions of the quantised original. This is an effect of using 5 MoG components to try and capture the 15 sharp peaks of the original. If the number of components is increased, the MoGs start to capture the inherent quantisation. This increases the quality of the reconstruction (in this case 'Einstein') marginally, although what effect this has on generality is a point for further research.

An important aspect of Bayesian reasoning is a model's ability to quantify its confidence in its results. For example, figure 3 plots 30 sample points from the reconstruction of source 2 ('S. J. Roberts'). The dashed line is the reconstruction ($\langle x_2 \rangle$), while the solid line contours 1 standard deviation according to $p'(x_2)$. The difference in uncertainty between the low- and high-noise source reconstructions is clear to see, with an average precision of 10.93 for the low-noise estimate, and 0.67 for the high-noise source signal. The quantification of uncertainty in parameter and hidden variable values can be extracted for *any* variable in the model. This extra information is obviously very important if and when results are assessed, especially if results are to be used for any decision-making or training of some other network.

4.2. Model selection

We found we could perform model selection in a number of ways - using ARD, monitoring the estimated noise precision and observing the negative free energy. The major benefit of ARD is that multiple models of varying dimensionality need not be trained; any unnecessary sources are automatically killed. Figure 4.2 shows the recovered source signals for the training data from a model with 8 sources. Figure 4.2a illustrates how ARD has suppressed the spurious sources in the low-noise case, inferring the correct latent dimensionality of 4. The precisions (expectations), $\boldsymbol{\alpha}$, are plotted in figure 5. The first four columns of the mixing matrix have precisions very close to zero. The prior on $\boldsymbol{H}_{j,1:4}$ is therefore wide enough to allow the columns to take values significantly different to zero (the prior mean), in effect allowing information to be channeled from the data to the source MoGs via these columns. The last four coefficients are of the order of 10^3, tightly constraining the prior on $\boldsymbol{H}_{j,5:8}$ around zero, effectively blocking any information flow from the data to sources 5-8.

In the high-noise case, ARD breaks down. The significant level of noise tricks ARD into thinking an extra source is needed to explain the extra data variance. The fifth coefficient in figure 5 is sufficiently close to zero, allowing some of the noise through to an extra source. Although the true latent dimensionality is not found in this case, one benefit of this noise channeling is to reduce the amount of noise absorbed by the 'useful' sources than would otherwise be the case (something noted in the rms error between the true and reconstructed images).

We also noticed how the estimated noise variance was a good indicator for the true number of sources. Figure 6 shows how the noise estimation varies across model orders. We have plotted the ratio of estimated to true noise so that a value of 1 indicates a perfect estimation. The noise estimation absorbs extra variance when not enough sources are trained. As soon as there are enough sources to explain the variance (in our case, 4), the noise precision jumps to a value closer to the true precision. This is particularly marked for the low-noise case, where there is approximately a six-fold change. Note how the noise precision is actually overestimated. No ICA algorithm is perfect - some of the noise will always be absorbed by the sources, leaving less noise for $\boldsymbol{\Lambda}$ to explain away. For the high-noise case, the peak is much less prounounced, making it harder to judge the true underlying dimensionality.

As expected, the free-energy was the best indicator for both the low- and high-noise data. Figure 7 plots the negative free energy across model orders. The low-noise curve clearly peaks at model order 4. The high-noise curve is much shallower, but is still clearly maximum at 4 sources. In fact, if plotted on its own scale, the high-noise curve has a similar shape to the low-noise curve. The

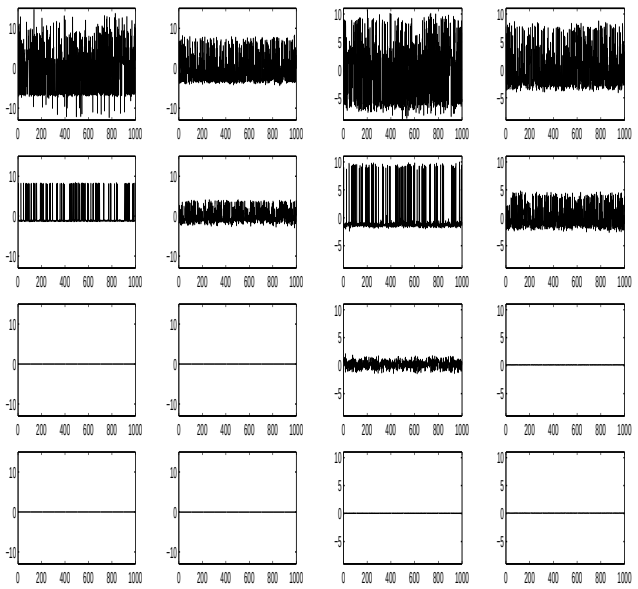

Fig. 4. Reconstructed source signals from the training data

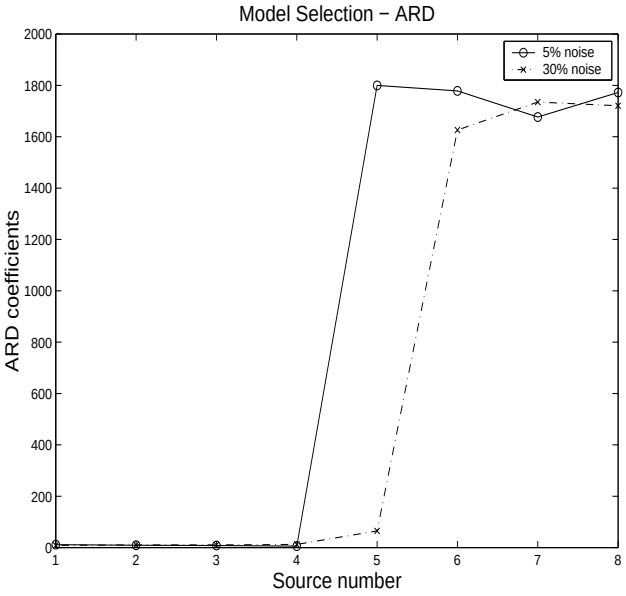

Fig. 5. ARD coefficients as a function of source number

'one-scale' plot nicely illustrates how the energy for the high-noise data is much lower than the low-noise data - i.e. the model is more confident in its results when trained on the cleaner data.

Both the noise and energy monitoring methods require the training of multiple models to infer the model order, although the free-energy is very robust in the presence of noise. ARD, on the other hand, needs only a single model (full-rank i.e. No. sources = No. sensors) in order to infer the latent dimensionality. Unfortunately, it can be tripped up if the training data is too noisy. Which strategy one chooses depends on one's circumstances. If the data dimensionality is large, the training of a full-rank model may be computationally prohibiting, especialy if a relatively small number of sources is expected. ARD is not as robust as the free-energy in the presence of noise, so the best strategy maybe to train a full-rank model to get a rough estimate, then train a few models either side of this estimate to zero in on the best one. Ultimately, this a point of further research and investigation.

5. DISCUSSION

In this paper, we have presented a method for Bayesian ICA which allows potentially any (stationary) source density to be modelled efficiently and accurately. We have demonstrated this by unmixing a noisy mixture of images with very different and complex pdfs, something not possible by fixed or uni-modal source models. This Bayesian formalism trains robust models which can be interrogated to quantify uncertainties in the patterns found and the parameters learned. We have shown that the true number of sources can be inferred as part of the learning process using Automatic Relevance Determination, highlighting both its merits and demerits. We have also shown how monitoring the estimated noise and free-energy across model orders also picks out the true latent dimensionality, with the free-energy being particularly robust.

The VB-ICA algorithm can easily be extended to learn positive-only bases and mixing using methods introduced in [7], something that will be presented elsewhere. As well as the Blind Source Seperation problem, VB-ICA can be used for any ICA related task, such as adaptive speech filtering, speech signal coding, biomedical signal processing, image compression, text modeling, financial data analysis and many other diverse applications. The ability to capture accurate representations of the latent densities allows a great flexibility and power. Coupled with the inherent model-order selection, we believe VB-ICA can be used for a wide variety of intelligent pattern recognition.

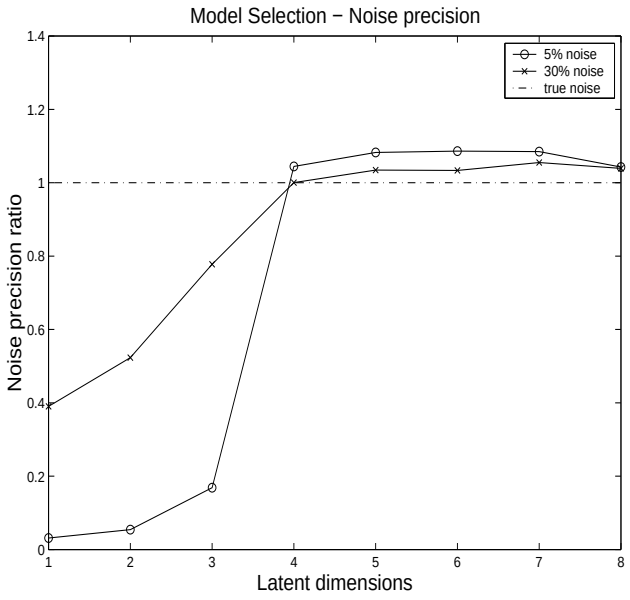

Fig. 6. Estimated noise precision s a function of latent dimensionality

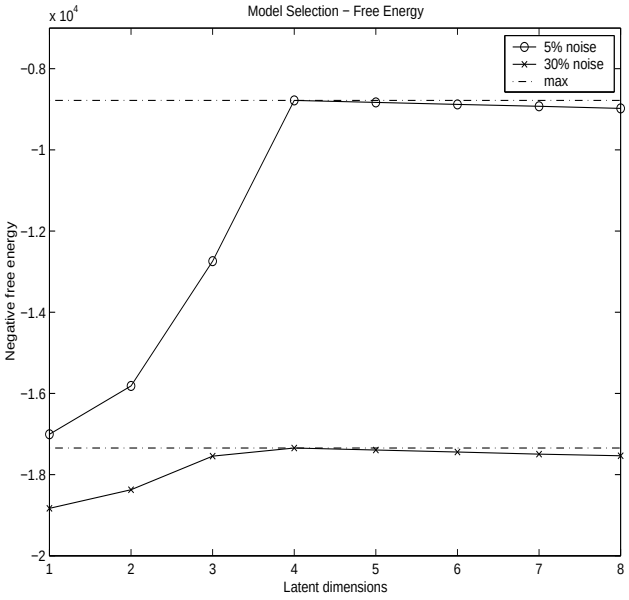

Fig. 7. Negative free energy as a function of latent dimensionality

6. REFERENCES

[1] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[2] T. W. Lee, M. Girolami, A. J Bell, and T. J. Sejnowski, "A unifying information-theoretic framework for independent component analysis," *International Journal of Mathematical and Computer Modelling (In press)*, 1998.

[3] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1998.

[4] H Lappalainen, "Ensemble learning for independent component analysis," in *Proceedings of the First International Workshop on Independent Component Analyis*, 1999, pp. 7–12.

[5] C. M. Bishop and N. D. Lawrence, "Variational bayesian independent component analysis," Tech. Rep., Computer Laboratory, University of Cambridge, 2000.

[6] D. J. C. MacKay, "Probable networks and plausible predictions - a review of practical bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, pp. 469–505, 1995.

[7] J. W. Miskin and D. J. C. MacKay, "Ensemble learning for blind source seperation," in *ICA: Principles and Practice*, S. J. Roberts and R. M. Everson, Eds., chapter 7. Cambridge University Press, 2001.

[8] A. P. Dempster, N. M. Laird, and Rubin D. B., "Maximum liklihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.

[9] R. D. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. The MIT Press, Cambridge, Massachusetts, 1999.

[10] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ica," in *ICONIPS '96*, 1996, pp. 151–157.

[11] Cardoso, "Infomax and maximum liklihood for blind source separation," *IEEE Letters on Signal Processing*, vol. 4, pp. 112–114, 1997.

[12] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, Ed. The MIT Press, Cambridge, Massachusetts, 1999.

[13] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[14] D. J. C. MacKay, "Developments in probabilistic modelling with neural networks - ensemble learning," in *Proceedings of the third Annual Symposium on Neural Networks*, Nijmagen, The Netherlands, 1995, pp. 191–198, Springer.

[15] R. Choudrey, W.D. Penny, and S.J. Roberts, "An ensemble learning approach to independent component analysis," in *Proceedings of Neural Networks for Signal Processing X, Sydney, December 2000*. IEEE Signal Processing Society, December 2000.