1

ICA: Model order selection and dynamic source models

W.D. Penny and S.J. Roberts, University of Oxford

R.M. Everson, University of Exeter

In this chapter we investigate ICA models in which the number of sources, M, may be less than the number of sensors, N; so-called non-square mixing.

The 'extra' sensor observations are explained as observation noise. This general approach may be called Probabilistic Independent Component Analysis (PICA) by analogy with the Probabilistic Principal Component Analysis (PPCA) model of (12); ICA and PCA don't have observation noise, PICA and PPCA do.

Non-square ICA models give rise to a likelihood model for the data involving an integral which is intractable. In this chapter we build on previous work in which the integral is estimated using a Laplace approximation. By making the further assumption that the unmixing matrix lies on the decorrelating manifold we are able to make a number of simplifications. Firstly, the observation noise can be estimated using PCA methods, and, secondly, optimisation takes place in a space having a much reduced dimensionality; having order M^2 parameters rather than $M \times N$. Again, building on previous work, we derive a model order selection criterion for selecting the appropriate number of sources. This is based on the Laplace approximation as applied to the decorrelating manifold and is compared with PCA model order selection methods.

Standard ICA, if there is such a thing, is not a proper time series model, as each source is considered to be Independent and Identically Distributed (IID). But with dynamic source models, temporal information is used and, as we show, this can lead to much improved source estimation. The second part of this chapter looks at the use of such dynamic source models, where the sources are modelled using a generalised autoregressive (GAR) process. This is the usual autoregressive process but where the noise has a Generalised Exponential (GE) distribution instead of the usual Gaussian.

This chapter consists of six further sections. The first descibres the probability model for non-square ICA and derives the Laplace approximation required to calculate the data likelihood. The second section describes the decorrelating manifold and the third describes ICA and PCA model order selection methods. Section four describes different source models including the GAR process. This includes a description of its own model order criterion for determining the number of taps in the GAR filter. Section five describes results from applying the above methods to the unmixing of music sources and the chapter is concluded in section six.

1.1 A Probabilistic model

The observed variables $\boldsymbol{x}$, of dimension N, are modelled as

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

where $\boldsymbol{e}$ is zero mean Gaussian observation noise having an isotropic covariance matrix with precision β, A is the mixing matrix, and the underlying sources $\boldsymbol{s}$ are statistically independent

$$p(\boldsymbol{s}) = \prod_{i=1}^{M} p(s_i) \tag{1.2}$$

where the sum runs over the M sources. The distribution of the observations conditioned on the mixing matrix and sources is

$$p(\mathbf{x}|A, \mathbf{s}) = \mathcal{N}(\mathbf{x}; A\mathbf{s}, (1/\beta)\boldsymbol{I}) \tag{1.3}$$

where $N(\mathbf{x}; \mu, \Sigma)$ is a normal distribution with mean μ and covariance Σ. The likelihood of a data point is given by

$$p(\mathbf{x}|A) = \int p(\mathbf{x}|A, \mathbf{s})p(\mathbf{s})d\mathbf{s} \tag{1.4}$$

With a non-square ICA model, optimisation takes place in two iterated steps; source estimation and mixing matrix estimation.

1.1.1 Source estimation

The sources can be estimated by noting that their posterior distribution

$$p(\mathbf{s}|A, \mathbf{x}) \propto p(\mathbf{x}|A, \mathbf{s})p(\mathbf{s}) \tag{1.5}$$

is proportional to the 'prior' distribution, $p(\mathbf{s})$, and the source-dependent likelihood $p(\mathbf{x}|A, \mathbf{s})$. An iterative gradient-based scheme exists for estimating the Maximum A Posteriori (MAP) sources, $\mathbf{s}_{MAP}$. This consists of two terms; (i) the gradient of the source-dependent log-likelihood and (ii) the gradient of the log source densities, both of which are given in later parts of this chapter.

Alternatively, the prior can be ignored and the sources set to their Maximum Likelihood (ML) source values, $\mathbf{s}_{ML}$. These are recovered via an unmixing matrix

$$\mathbf{s}_{ML} = W\mathbf{x} \tag{1.6}$$

which is given by the pseudo-inverse of the mixing matrix

$$W = (A^T A)^{-1} A^T \tag{1.7}$$

This unmixing minimises the squared reconstruction error, and therefore maximises the data-likelihood. Computation of the MAP sources will not improve on the ML reconstruction error or the squared error of source estimation, but it will reduce 'cross-talk' between the sources ie. make them more independent. An empirical demonstration of this is given in (1).

1.1.2 Mixing matrix estimation

To compute the likelihood of an observation we must be able to calculate the integral in equation 1.4. If we assume that the distribution over sources is dominated by a single peak, $\hat{\mathbf{s}}$, then the integral can be performed using Laplace's method (4)

$$\int p(\mathbf{x}|A, \mathbf{s}) p(\mathbf{s}) d\mathbf{s} \approx p(\mathbf{x}|A, \hat{\mathbf{s}}) p(\hat{\mathbf{s}}) (2\pi)^{M/2} \det(F)^{-1/2} \tag{1.8}$$

where

$$F = -\left[\frac{d^2 \log p(\mathbf{x}|A, \mathbf{s}) p(\mathbf{s})}{d\mathbf{s}_i d\mathbf{s}_j}\right]_{\mathbf{s}=\hat{\mathbf{s}}} \tag{1.9}$$

In this chapter we use a simplified variant of Laplace's method where the above matrix is replaced by the Hessian (4)

$$H = -\left[\frac{d^2 \log p(\mathbf{x}|A, \mathbf{s})}{d\mathbf{s}_i d\mathbf{s}_j}\right]_{\mathbf{s}=\hat{\mathbf{s}}} \tag{1.10}$$

We have

$$\log p(\mathbf{x}|A, \mathbf{s}) \quad = \quad \frac{N}{2} \log \left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} (\mathbf{x} - A\mathbf{s})^T (\mathbf{x} - A\mathbf{s}) \tag{1.11}$$

giving

$$H = \beta A^T A \tag{1.12}$$

The log-likelihood of an observation, $L \equiv \log p(\mathbf{x}|A)$ is therefore given by

$$\begin{aligned} L &= \frac{N-M}{2}\log\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}(\mathbf{x} - A\hat{\mathbf{s}})^T(\mathbf{x} - A\hat{\mathbf{s}}) \\ &\quad + \log p(\hat{\mathbf{s}}) - \frac{1}{2}\log\det(A^T A) \end{aligned} \tag{1.13}$$

The mixing matrix, A, can be optimised by following the gradient dL/dA using a Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimizer as shown in (11). The noise precision can then be estimated using a fixed point of the above likelihood

$$\frac{1}{\beta} = \frac{1}{N-M} < (\boldsymbol{x} - A\hat{\mathbf{s}})^T(\boldsymbol{x} - A\hat{\mathbf{s}}) > \tag{1.14}$$

where the expectation is taken over all observations.

Fitting a non-square ICA model therefore consists of iterating estimates of the mixing matrix with estimates of the noise precision and the sources. The sources can be estimated either by their MAP or ML values, ie. $\hat{\mathbf{s}} = \mathbf{s}_{MAP}$ or $\hat{\mathbf{s}} = \mathbf{s}_{ML}$, as shown in the previous section. In previous work, (11) has used ML sources.

1.2 The Decorrelating Manifold

In previous work 3 we have constrained the unmixing matrix to be a decorrelating matrix. The motivation for this is that, for sources to be statistically independent, they must be at least linearly decorrelating. Therefore, by ensuring that they are decorrelating, we are at least some way to finding the ICA solution. The corresponding mixing matrix is defined as follows.

If X is an $N \times T$ matrix of zero-mean data vectors, and each entry is normalised by $1/T$, and the Singular Value Decomposition (SVD) of X is given by

$$X = U \Lambda V \tag{1.15}$$

then U contains the principal components of the observation covariance matrix and $\Lambda = [\lambda_1, \lambda_2, ... \lambda_N]$ contains the standard deviations of the corresponding principal components. The mixing matrix is then given by

$$A = U_M \Lambda_M Q^T D^{-1} \tag{1.16}$$

where U_M and Λ_M are the first M columns of U and Λ. The transform is also parameterised by a diagonal scaling matrix D and an orthogonal matrix Q which are both of dimension $M \times M$. The matrices Q and D constitute the ICA transform proper.

1.2.1 Source and noise estimation

The ML source estimates are, as before, given by the pseudo-inverse of the mixing matrix

$$\begin{aligned} W &= (A^T A)^{-1} A^T \\ &= D Q \Lambda_M^{-1} U_M^T \end{aligned} \tag{1.17}$$

operating on the observations

$$\mathbf{s}_{ML} = W\mathbf{x} \tag{1.18}$$

The reconstructed observations are given by

$$A\mathbf{s}_{ML} = U_M U_M^T \mathbf{x} \tag{1.19}$$

which gives an average reconstruction error of

$$E = < (\mathbf{x} - A\mathbf{s}_{ML})^T (\mathbf{x} - A\mathbf{s}_{ML}) > \tag{1.20}$$

$$= Tr[(X - U_M U_M^T X)^T (X - U_M U_M^T X)]$$

$$= Tr[X^T X - X^T U_M U_M^T X] \tag{1.21}$$

By noting that the projection onto the first M principal components is $Y = X^T U_M$, and their covariance is $YY^T = X^T U_M U_M^T X$, E is seen to be the variance of the data not explained by the first M components. Hence

$$E = \sum_{i=M+1}^{N} \lambda_i^2 \tag{1.22}$$

and

$$\frac{1}{\beta} = \frac{1}{N-M} \sum_{i=M+1}^{N} \lambda_i^2 \tag{1.23}$$

Therefore, if ICA is constrained to the decorrelating manifold and ML source estimates are used, the observation noise level is not dependent on D or Q ie. the ICA transform proper. It can therefore be calculated ahead of optimising D and Q and fixed to its calculated value.

1.2.2 *Mixing matrix estimation*

In previous work (3), we showed that for flexible source models (see later), the mixing matrix can be constrained to have rows of length one. For these cases we have $D = I$. The matrix Q is constrained to be orthogonal by writing it as

$$Q = \exp(Z) \tag{1.24}$$

where Z is a skew-symmetric matrix ($Z^T = -Z$) whose non-zero entries z_{ij} are known as Cayley coordinates (6).

By substituting in the average reconstruction error and our chosen form for the mixing matrix, the log-likelihood becomes

$$\begin{aligned} L \quad = \quad & \frac{N-M}{2} \log\left(\frac{\beta}{2\pi e}\right) + \log p(\mathbf{s}_{ML}) \\ & + \log\det(D) - \log\det(Q) - \log\det(\Lambda_M) \end{aligned} \tag{1.25}$$

As Q is an orthonormal matrix, we always have $\det(Q) = 1$. Therefore the only term in the likelihood that depends on Q is the log source density where the dependence is introduced via equations 1.17 and 1.18.

In previous work (3) we show how to compute the derivative of this term and combine it with an expression for dQ/dZ. This then gives the gradient of the likelihood with respect to the Cayley coordinates. Fitting the ICA model therefore corresponds to simply following this gradient using, for example, a BFGS optimiser.

1.3 Model order selection

The optimal number of sources, $\hat{M}$, can be computed by plotting the log-likelihood, $\log p(\mathbf{x}|A)$, as a function of M and choosing the maximum. For most signal processing models, eg. autoregressive models, wavelets or neural networks, model order selection using a maximum likelihood criterion is doomed to failure. This is because as more basis functions are *added* the likelihood increases monotically; the optimal model order is therefore infinite. ICA however, is more like a product model than an additive model, because the sources are independent. As too many sources are postulated, the independence criterion is violated thus reducing the overall likelihood. ICA model order selection using ML is therefore plausible.

For the case of Gaussian sources the ICA model reduces to PCA. We are then able to use PCA model order selection methods such as the Laplace approximation used by (6). By using conjugate priors for the eigenvectors, eigenvalues and noise level and parameterising the eigenvectors using Cayley coordinates (6) shows that the evidence for a PCA model with M sources is

$$p(X|M) \approx p(U) \left(\prod_{j=1}^{M} \lambda_j \right)^{-T/2} \beta^{T(N-M)/2} (2\pi)^{(m+M)/2} |A_Z|^{-1/2} T^{-M/2} \tag{1.26}$$

where $m = NM - M(M+1)/2$ and

$$p(U) = 2^{-M} \prod_{i=1}^{M} \Gamma((N-i+1)/2) \pi^{-(N-i+1)/2} \tag{1.27}$$

where $\Gamma()$ is the Gamma function and

$$|A_Z| = \prod_{i=1}^{M} \prod_{j=i+1}^{N} (\hat{\lambda}_j^{-1} - \hat{\lambda}_i^{-1})(\lambda_i - \lambda_j) T \tag{1.28}$$

and λ_j are the eigenvectors from PCA, and $\hat{\lambda}_j$ are identical except for $j > M$ where $\hat{\lambda}_j = (1/(N-M)) \sum_{j=M+1}^{N} \lambda_j$. Minka's experiments and our own, show this to be a remarkably consistent model order criterion, even with very few data points. Moreover, Minka produces empirical evidence to show that for selected non-Gaussian sources (sound samples with skewed and kurtotic densities), accurate model order selection is still feasible.

1.4 Source Models

1.4.1 Inverse-Cosh Sources

In (2) the source densities are assumed to be inverse-cosh densities

$$p(s_i) = \frac{\cosh^{-1}(s_i)}{Z} \tag{1.29}$$

where Z is a normalising constant. This form arises from the choice of nonlinearity used in the learning algorithm, as discussed in (5). This gives rise to a gradient

$$\frac{d \log p(s_i)}{ds_i} = -\tanh(s_i) \tag{1.30}$$

which is the hyperbolic tangent squashing function used in neural network implementations of ICA models. The normalising constant can be approximated as

$$\log Z = a \log(c+1) + b \tag{1.31}$$

where $a = 0.522$,$b = 0.692$ and $c = 1.397$. A drawback of the IC density, however, is its inability to model sub-Gaussian densities, such as the uniform density.

1.4.2 Generalised Exponential Sources

A more general parametric form which can model super-Gaussian, Gaussian *and* sub-Gaussian forms is the 'Exponential Power Distribution' or 'Generalised Exponential (GE)' density

$$p(s_i) \equiv G(s_i; R_i, \beta_i) = \frac{R_i \beta^{1/R_i}}{2\Gamma(1/R_i)} \exp(-\beta_i |s_i|^{R_i}) \tag{1.32}$$

This density has zero mean, a kurtosis determined by the parameter R_i and a variance which is then determined by $1/\beta_i$. 3 show how to calculate the derivative of the log source density and describe an embedded line search method for estimating $\{R_i, \beta_i\}$.

1.4.3 Generalised Autoregressive Sources

7 have proposed a 'contextual-ICA' algorithm where the sources are conditioned on previous source values. The observations are then generated from an instantaneous mixing of the sources. Their work focuses on using generalized autoregressive (GAR) models for modelling each source. The term 'generalised' is used because the AR models incorporate additive noise which is non-Gaussian. Specifically, Pearlmutter and Parra use an inverse-cosh noise distribution.

Pearlmutter and Parra have shown that contextual-ICA can separate sources which cannot be separated by standard (non-contextual) ICA algorithms. This is because the standard methods utilise only information from the cumulative histograms; temporal information is discarded.

In this chapter we use GAR models with p filter taps and additive noise drawn from a generalised exponential distribution; for $p = 0$ these models therefore reduce to the GE sources described in the previous section. The density model is

$$p(s_i) = G(e_i[t]; R_i, \beta_i) \tag{1.33}$$

where $e_i[t] = s_i[t] - \hat{s}_i[t]$ is the GAR prediction error and $\hat{s}_i[t]$ is the GAR prediction

$$\hat{s}_i[t] = -\sum_{k=1}^{p} c_i(k) s_i[t-k] \tag{1.34}$$

where $c_i(k)$ are the GAR coefficients for the ith source which can collectively be written as a vector $\boldsymbol{c}_i$. The GAR coefficients can be estimated by minimising the error

$$E = \sum_{t=1}^{T} |e_i[t]|^R \tag{1.35}$$

8 derive the corresponding gradients and, again, use BFGS for optimisation. This procedure is embedded within the algorithm for estimating the unmixing matrix, W. The GAR models are re-restimated once for every ten updates of W.

The optimal number of filter taps, $\hat{p}$, can be chosen using a Minimum Description Length (MDL) model order selection criterion. For a data set D and estimated parameters $\hat{\boldsymbol{\theta}}$ of dimension p, the MDL criterion is given by

$$MDL(p) = -\log p(D|\hat{\boldsymbol{\theta}}) + \frac{p}{2} \log T \tag{1.36}$$

where T is the number of data points. For a GAR model this gives

$$MDL_{GAR}(p) = -T \log \left(\frac{R\beta^{(1/R)}}{2\Gamma(1/R)} \right) + \sum_{t=1}^{T} \beta |e_t|^R + \frac{p}{2} \log T \tag{1.37}$$

For $R = 2$, if we ignore terms not involving p or β, this reduces to the well known MDL criterion for an AR model

$$MDL_{AR}(p) = -\frac{T}{2} \log \beta + \frac{p}{2} \log T \tag{1.38}$$

This criterion can be applied to each GAR source in an ICA model, allowing each to have a different number of taps, thus reflecting the dynamic complexity or otherwise, of each source.

1.5 Results

1.5.1 Selecting the number of sources

We now give some results of applying the various model order selection methods for estimating the optimal number of sources. The first method, which we call ICADEC-L, uses the Laplace approximation and constrains the unmixing matrix to be on the decorrelating manifold. This results in the likelihood expression in equation 1.25.

The above measure is compared with PCA model order selection methods. The first of these methods, which we call PCA-L, uses Minka's Laplace approximation described earlier. The second, which we call PCA-MDL, uses an MDL criterion described in (13). The last method, which we call PCA-EV, is the evidence method described in (10).

The methods are applied to four datasets. The first consists of two music sources (the top two in figure 1.1) which are mixed into six observations to which is then added observation noise of variance $1/\beta$. The second data set consists of four music sources (see figure 1.1) which are again mixed into six observations to which we add observation noise. All music sources were normalised to zero mean and unit variance. Fifty data sets of each type are created, where each time, the mixing matrix was set randomly according to a Gaussian distribution. The observation noise sequence was generated afresh each time.

Tables 1 and 2 show the number of times each model selection criterion selected the correct order, for 100 data points and various noise levels. The PCA-L and PCA-MDL criterion appear to offer the best performance, with PCA-L always being slightly better. The ICADEC-L criterion always degrades more rapidly in the presence of noise. The PCA-EV criterion is inconsistent; outperforming all methods on the 4-source task (and actually getting better with increasing noise level), but doing poorly on the 2-source task.

The next two datasets involve EEG sources which were derived as follows. We applied ICADEC to a 22-channel EEG recording, over a time period for which the signal statistics were considered to be stationary (this was found by embedding the ICA model in a hidden Markov process; see 8 for details). The true number of sources underlying this data is unknown, but applying PCA-L gave an answer of 15. We then extracted two data sets; one consisting of 3 sources, shown in Figure 1.2, and one consisting of 10 sources (not shown). All sources were normalised to zero mean and unit variance. These sources were then mixed up to form 20-dimensional observations to which noise was added. Fifty data sets

Table 1.1. *Model order selection with 2 music sources of unit variance. The numbers indicate the percentage of times the correct model order was selected.*

$1/\beta$	PCA-L	PCA-MDL	PCA-EV	ICADEC-L
0.1	100	100	100	100
0.3	94	96	64	86
0.5	96	94	0	62
1.0	78	71	0	8

of each type were created, where each time, the mixing matrix was set randomly according to a Gaussian distribution and the observation noise sequence was generated afresh each time.

The ICADEC-L criterion was applied to only 10 data sets of each type, and at a single noise level ($1/\beta = 0.1$), due to the excessive amount of computation required; the PCA methods perform a single eigen-decomposition of the 20-dimensional space, whereas for ICADEC-L we have to perform 20 separate optimisations.

For the 3-source EEG data set, PCA-L and PCA-MDL achieved 100% correct model order selection at all noise levels ($1/\beta = 0.1, 0.3, 0.5, 1.0$). PCA-EV completely failed at all but the first noise level; as more noise was added it estimated the optimal model order (averaged over the 50 data sets) as 3, 6, 9 and 13 respectively. It therefore mistakenly interprets the extra observation noise as extra sources. ICADEC-L also failed completely on the (limited) data it was applied to, again overestimating the model order. It chose $\hat{M} = 4$ for 9/10 of the data sets and 5 on the remaining one.

The results for the 10-source EEG data set are shown in Table 1.3. Again, PCA-L shows the best performance, closely followed by PCA-MDL. PCA-EV again fails at high noise levels, interpreting the extra noise as extra sources. ICADEC-L performed reasonably well on the limited data it was tried on, getting 70% correct at the first noise level (though the PCA methods get 100% correct).

1.5.2 Comparing source models

Our second set of results compares the different source models; Inverse-Cosh (IC), Generalised Exponential (GE) and Generalised Autoregres-

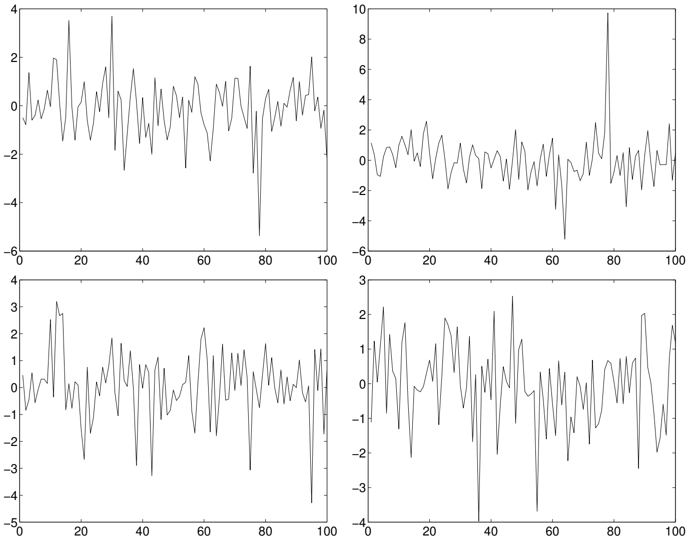

Fig. 1.1. *Music sources used in model order selection experiments. The left two are samples of Beethoven and the right two are samples of Bessie Smith.*

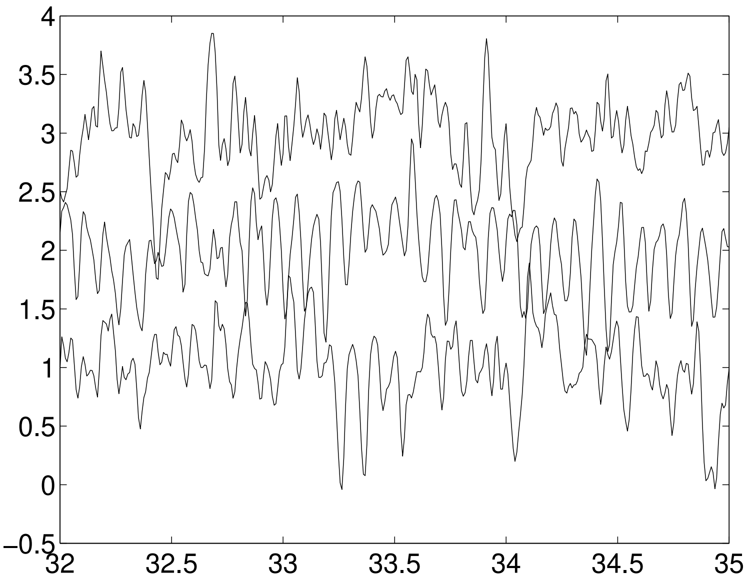

Fig. 1.2. *Three EEG sources.*

Table 1.2. *Model order selection with 4 music sources of unit variance. The numbers indicate the percentage of times the correct model order was selected.*

$1/\beta$	PCA-L	PCA-MDL	PCA-EV	ICADEC-L
0.01	100	96	56	100
0.1	94	94	58	88
0.3	86	82	82	76
0.5	72	70	96	66

Table 1.3. *Model order selection with 10 EEG sources of unit variance. The numbers indicate the percentage of times the correct model order was selected.*

$1/\beta$	PCA-L	PCA-MDL	PCA-EV
0.1	100	100	100
0.3	80	72	50
0.5	58	36	8
1.0	12	4	0

sive (GAR). For the GAR model, application of the MDL criterion to the music sources suggested using a model order of 10. Figures 1.3 and 1.4 show the correlations between true and estimated music sources for the IC model and for the GAR model. This was for a data set containing six observations mixed up from two music sources, as described earlier. The variance of the observation noise was $1/\beta = 0.01$. The corresponding Normalised Mean Squared Errors (NMSE, the squared source estimation error normalised by the variance of the true sources) were 0.0272 for IC, 0.1946 for GE and only 0.0014 for GAR. Table 1.4 shows how unmixing accuracy is dependent on the level of observation noise (we omit the results for the GE model as it is not a good source model for this data set). At low noise levels there is an order of magnitude benefit in using a GAR model, whereas at high noise levels the benefit is modest.

Table 1.4. *Normalised Mean Squared Error of unmixing using Inverse-Cosh (IC) and Generalised Autoregressive (GAR) source models, for various levels of added observation noise, $1/\beta$. The final column shows the relative error. At low noise levels (top-half of table) there is an order of magnitude benefit in using a GAR model, whereas at high noise levels the benefit is modest.*

$1/\beta$	IC	GAR	IC/GAR
0.001	0.0309	0.0013	24
0.005	0.0315	0.0027	12
0.01	0.0328	0.0045	7.3
0.05	0.0412	0.0195	2.1
0.1	0.0577	0.0360	1.6
0.3	0.1111	0.0902	1.2
0.5	0.1587	0.1200	1.3
1.0	0.1807	0.1640	1.1

1.6 Discussion

Of the model order criteria investigated the method of choice is PCA-L, a Bayesian PCA criterion derived by (6). Not only is it accurate, it is also fast. This is important as ICA is increasingly being applied to data sets of a higher dimensionality where, for example in EEG or fMRI analysis, we have tens or hundreds of observations.

The relatively poorer performance of the ICADEC-L criterion is not really surprising as it is, in fact, a maximum likelihood criterion; in its derivation we have not integrated out the mixing matrix or the observation noise. In future, we intend to do this by extending the use of the conjugate priors used in PCA-L to the ICADEC situation; this is a natural extension as, in ICADEC, the mixing matrix component Q is also parameterised using Cayley coordinates.

The use of dynamic source models can, at low noise levels, improve source estimation accuracy by an order of magnitude. In the future, we envisage applying the GAR model order criterion at the same time as mixing matrix estimation. This could be used to allow the overall algorithm to extract only those sources with a high temporal information content. The reduction in the number of sources produced may help to speed-up interpretation of the various ICA components.

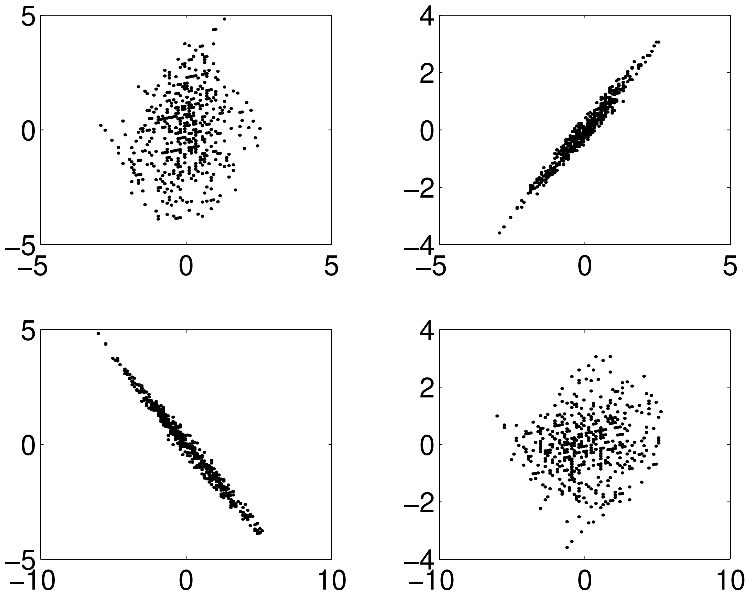

Fig. 1.3. *Source estimation using an Inverse-Cosh source model; plots of true versus estimated sources.*

Bibliography

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

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

R. Everson and S.J. Roberts. Independent Component Analysis: A flexible nonlinearity and decorrelating manifold approach. *Neural Computation*, 11(8), 1999.

R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. Technical Report 254, University of Washington, 1993. http://www.stat.washington.edu/tech.reports/tr254.ps.

D.J.C. Mackay. Maximum likelihood and covariant algorithms for independent component analysis. Technical report, Cavendish Laboratory, University of Cambridge, 1996.

T.P. Minka. Automatic choice of dimensionality for PCA. Technical Report 514, MIT Media Laboratory, Perceptual Computing Section, 2000.

B. A. Pearlmutter and L.C. Parra. Maximum Likelihood Blind Source Separation: A Context-Sensitive Generalization of ICA. In *Advances in Neural Information Processing Systems 9*, pages 613–619. MIT Press, 1997.

W.D. Penny, R.M. Everson, and S.J. Roberts. Hidden Markov Independent Components Analysis. In M. Girolami, editor, *Advances in Independent*

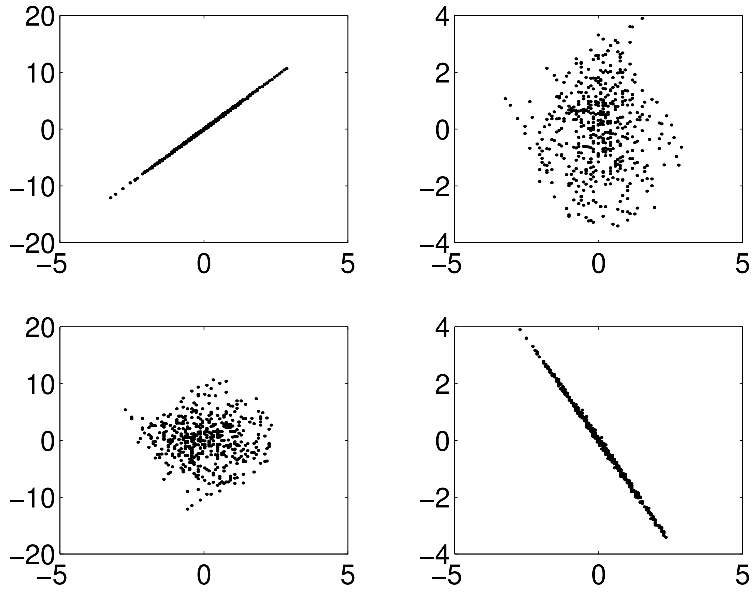

Fig. 1.4. *Source estimation using a GAR source model; plots of true versus estimated sources.*

Component Analysis. Springer, 2000.

W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.

J.J. Rajan and P.J.W. Rayner. Model order selection for the singular value decomposition and the discrete Karhunen-Loeve transform using a Bayesian approach. *IEE Vision, Image and Signal Processing*, 144(2):116–123, 1997.

S.J. Roberts. Independent Component Analysis: Source assessment and Separation, a Bayesian Approach. *IEE Proceedings on Vision, Image and Signal Processing*, 145(3):149–154, 1998.

M. E. Tipping and C. M. Bishop. Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, 11(2):443–482, 1999.

M. Wax and T. Kailath. Detection of Signals by Information Theoretic Criteria. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-32:387–392, 1985.