1 Particle Filters for Non-Stationary ICA

Richard Everson[1] and Stephen Roberts[2]

[1] Department of Computer Science, University of Exeter, Exeter, UK.
R.M.Everson@exeter.ac.uk
[2] Department of Engineering Science, University of Oxford, Oxford. UK
sjrob@robots.ox.ac.uk

1.1 Introduction

Over the last decade in particular there has been much interest in Independent Components Analysis methods (ICA) for blind source separation (BSS) and deconvolution (see [15] for a review). One may think of the blind source separation as the problem of identifying speakers (sources) in a room given only recordings from a number of microphones, each of which records a linear mixture of the sources, whose statistical characteristics are unknown. The casting of this problem in a neuro-mimetic framework [3] has done much to to simplify and popularise the technique. More recently still the ICA solution has been shown to be the maximum-likelihood point of a latent-variable model [17,4,20]

Here we consider the blind source separation problem when the mixing of the sources is non-stationary. Pursuing the speakers in a room analogy, we address the problem of identifying the speakers when they (or equivalently, the microphones) are moving. The problem is cast in terms of a hidden state (the mixing proportions of the sources) which we track using particle filter methods, which permit the tracking of arbitrary state densities.

We first briefly review classical independent componnent analysis. ICA with non-stationary mixing is described in terms of a hidden state model and methods for estimating the sources and the mixing are described. Particle filter techniques are introduced for the modelling of state densities. Finally, we address the non-stationary mixing problem when the sources are independent, but possess temporal correlations.

1.2 Stationary ICA

Classical ICA belongs to the class of latent variable models which model observed data as being generated by the linear mixing of latent sources. We assume that there are M sources whose probability density functions (pdf) are $p_m(s^m)$. Observations, $\boldsymbol{x}_t \in \mathbb{R}^N$, are produced by the instantaneous linear mixing of the sources by A:

$$\boldsymbol{x}_t = A\boldsymbol{s}_t \tag{1.1}$$

In order for the problem to be well posed, the mixing matrix, A, must have at least as many rows as columns ($N \geq M$), so that the dimension of each observation is at least as great as the number of sources. The aim of ICA methods is to recover the latent sources $\hat{\boldsymbol{s}}_t$ by finding W, the (pseudo-) inverse of A:

$$\hat{\boldsymbol{s}}_t = W\boldsymbol{x}_t = WA\boldsymbol{s}_t \tag{1.2}$$

The probability density of the observations is related [19] to the joint source density by:

$$p(\boldsymbol{x}_t \,|\, A) = \frac{p(\boldsymbol{s}_t)}{|\det A|} \tag{1.3}$$

Hence the log likelihood of the observation $\boldsymbol{x}_t$ is

$$\log l = -\log|\det A| + \log p(\boldsymbol{s}_t) \tag{1.4}$$

and the normalised log likelihood of a set of observations $t = 1, ...T$ is therefore

$$\log \mathcal{L} = -\log|\det A| + \frac{1}{T}\sum_{t=1}^{T} \log p(\boldsymbol{s}_t) \tag{1.5}$$

To proceed further some assumptions or prior knowledge about the nature of $p(\boldsymbol{s}_t)$ must be incorporated into the model. Physical considerations may provide strong constraints on the form of $p(\boldsymbol{s}_t)$ (see, for example, Mishkin and Mackay, this volume). The assumption that the sources are linearly decorrelated leads to an $M(M-1)$-dimensional manifold of separating matrices [6]. The further constraint that the sources are normally distributed with zero mean and unit variance leads to the principal components analysis, PCA [12].

Independent components analysis assumes that the sources are independent so that the joint pdf of the sources factorises into the product of marginal densities:

$$p(\boldsymbol{s}_t) = \prod_{m=1}^{M} p(s_t^m) \tag{1.6}$$

Using this factorisation, the normalised log likelihood of a set of T observations is therefore [4,17,20]:

$$\log \mathcal{L} = -\log|\det A| + \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} \log p_m(\hat{s}_t^m) \tag{1.7}$$

The optimum A may then be found by maximisation of $\log \mathcal{L}$ with respect to A, assuming some specific form for the marginal densities $p(s_t^m)$.

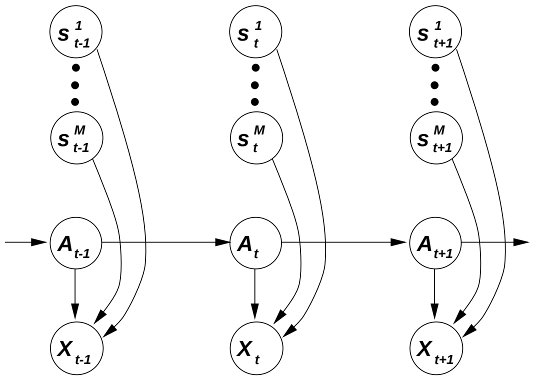

Fig. 1.1. Graphical model describing non-stationary ICA.

Sucessive gradient ascents on $\log l_t$ (equation 1.4) leads to the Bell & Sejnowksi stochastic learning rule for ICA [3], while batch learning is achieved by maximising $\log \mathcal{L}$. Learning rates may be considerably enhanced by modifying the learning rule to make it covariant [1,17]. A common choice is $p(s_t^m) \propto 1/\cosh(s_t^m)$, which leads to a tanh nonlinearity in the learning rule. Althuogh the source model is apparently fixed, scaling of the mixing matrix tunes the model to particular sources [6], and with a tanh nonlinearity platykurtic (heavy tailed) sources can be separated, although not leptokurtic ones. Cardoso [5] has elucidated the conditions under which the true mixing matrix is a stable fixed point of the learning rule.

Adoption of more flexible model for the source densities permits the separation of a wider range of source densities. Attias [2] has used mixtures of Gaussians to model the sources, which permits multi-modal sources and Lee *et al.* [16] switch between sub- and super-Gaussian source models. In the work presented here a generalised exponential model is used.

We emphasise that since the likelihood is unchanged if A is pre-multiplied by a diagonal matrix D or a scaling matrix P, the original scale of the sources cannot be recovered. The separating matrix W is therefore only the inverse of A up to a diagonal scaling and permutation, that is:

$$WA = PD \tag{1.8}$$

1.3 Non-stationary Independent Components Analysis

Figure 1.1 depicts the graphical model describing the conditional independence relations of the non-stationary ICA model. In common with static independent components analysis, we adopt a generative model in which M independent sources are linearly mixed at each instant. Unlike static ICA, however, the mixing matrix A_t is allowed to vary with time. We also assume that the observation $\boldsymbol{x}_t$ is contaminated by normally distributed noise

$\boldsymbol{w}_t \sim \mathcal{N}(0, R)$. Thus

$$\boldsymbol{x}_t = A_t \boldsymbol{s}_t + \boldsymbol{w}_t \tag{1.9}$$

Since the observational noise is assumed to be Gaussian its density is

$$p(\boldsymbol{w}_t) = \mathcal{G}(\boldsymbol{w}_t, R) \tag{1.10}$$

where $\mathcal{G}(\cdot, \Sigma)$ denotes a Gaussian function with zero mean and covariance Σ. The pdf of observations $p(\boldsymbol{x}_t \,|\, A_t)$ is given by

$$p(\boldsymbol{x}_t \,|\, A_t) = \int p(\boldsymbol{x}_t \,|\, A_t, \boldsymbol{\theta}, \boldsymbol{s}_t) p(\boldsymbol{s}_t \,|\, \boldsymbol{\theta}) \, d\boldsymbol{s}_t \tag{1.11}$$

where $\boldsymbol{\theta}$ are parameters pertaining to the particular source model. Since the *sources* are assumed stationary the likelihood of $\boldsymbol{x}_t$ given a particular mixing matrix is

$$p(\boldsymbol{x}_t \,|\, A_t) = \int p(\boldsymbol{x}_t \,|\, A_t, \boldsymbol{s}) p(\boldsymbol{s} \,|\, \boldsymbol{\theta}) \, d\boldsymbol{s} \tag{1.12}$$

As for static ICA, the assumption of the independence of the sources permits the source density to be written in factorised form (equation 1.6) so that the likelihood becomes

$$p(\boldsymbol{x}_t \,|\, A_t) = \int \mathcal{G}(\boldsymbol{x}_t - A_t \boldsymbol{s}, R) \prod_{m=1}^{M} p_m(s^m) \, d\boldsymbol{s} \tag{1.13}$$

We emphasise that it is in equation (1.13) that the independence of the sources is modelled by writing the joint source density in factored form.

The dynamics of A_t are modelled by a first order Markov process, in which the elements of A_t diffuse from one observation time to the next. If we let $\boldsymbol{a}_t = \mathrm{vec}(A_t)$ be the $N \times M$-dimensional vector obtained by stacking the columns of A_t, then $\boldsymbol{a}_t$ evolves according to the state equation:

$$\boldsymbol{a}_{t+1} = F \boldsymbol{a}_t + \boldsymbol{v}_t \tag{1.14}$$

where $\boldsymbol{v}_t$ is zero-mean Gaussian noise with covariance Q, and F is the state transition matrix; in the absence of *a priori* information we take F to be the identity matrix. The state equation (1.14) and the statistics of $\boldsymbol{v}_t$ together define the density $p(\boldsymbol{a}_{t+1} | \boldsymbol{a}_t)$.

A full specification of the hidden state must include the parameter set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_m\}$, $m = 1...M$, which describes the independent source densities. We model the source densities with generalised exponentials, as described in §1.3.1. Since the sources themselves are considered to be stationary, the parameters $\boldsymbol{\theta}$ are taken to be static, but they must be learned as data are observed.

The problem is now to track A_t and to learn $\boldsymbol{\theta}$ as new observations $\boldsymbol{x}_t$ become available. If X_t denotes the collection of observations $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_t\}$, then the goal of filtering methods is to deduce the probability density function of the state $p(\boldsymbol{a}_t \mid X_t)$. This pdf may be found recursively in two stages: prediction and correction. If $p(\boldsymbol{a}_{t-1} \mid X_{t-1})$ is known, the state equation (1.14) and the Markov property that $\boldsymbol{a}_t$ depends only on $\boldsymbol{a}_{t+1}$ permits prediction of the state at time t:

$$p(\boldsymbol{a}_t \mid X_{t-1}) = \int p(\boldsymbol{a}_t \mid \boldsymbol{a}_{t-1}) p(\boldsymbol{a}_{t-1} \mid X_{t-1})\, d\boldsymbol{a}_{t-1} \tag{1.15}$$

The predictive density $p(\boldsymbol{a}_t \mid X_{t-1})$ may be regarded as an estimate of $\boldsymbol{a}_t$ prior to the observation of $\boldsymbol{x}_t$. As the datum $\boldsymbol{x}_t$ is observed, the prediction may be corrected via Bayes' rule

$$p(\boldsymbol{a}_t \mid X_t) = Z^{-1} p(\boldsymbol{x}_t \mid \boldsymbol{a}_t) p(\boldsymbol{a}_t \mid X_{t-1}) \tag{1.16}$$

where the likelihood of the observation given the mixing matrix, $p(\boldsymbol{x}_t \mid \boldsymbol{a}_t)$, is defined by the observation equation (1.9). The normalisation constant Z is known as the innovations probability:

$$Z = p(\boldsymbol{x}_t \mid X_{t-1}) = \int p(\boldsymbol{x}_t \mid \boldsymbol{a}_t) p(\boldsymbol{a}_t \mid X_{t-1})\, d\boldsymbol{a}_t \tag{1.17}$$

The prediction (1.15) and correction/update (1.16) pair of equations may be used to step through the data online, alternately predicting the subsequent state and then correcting the estimate when a new datum is observed. The corrected state at time t is then the basis for the prediction of the state at $t+1$.

1.3.1 Source Model

Before describing the application of particle filters to the prediction and correction equations we briefly describe the source model. In order to be able to separate light-tailed sources a more flexible source model than the traditional $1/\cosh$ density is needed. It is difficult to use a switching model [16] in this context. Attias [2] has used mixtures of Gaussians to model the sources, which permits multi-modal sources; we use generalised exponentials, which provide a good deal of flexibility and do not suffer from the combinatorial complexities associated with mixture models.

Each source density is modelled by

$$p(s^m \mid \boldsymbol{\theta}_m) = z \exp - \left| \frac{s^m - \nu_m}{w_m} \right|^{r_m} \tag{1.18}$$

where the normalising constant is

$$z = \frac{r_m}{2 w_m \Gamma(1/r_m)} \tag{1.19}$$

The density depends upon parameters $\boldsymbol{\theta}_m = \{\mu_m, w_m, r_m\}$. The location of the distribution is set by μ_m, its width by w_m and the weight of its tails is determined by r_m. Clearly p is Gaussian when $r_m = 2$, Laplacian when $r_m = 1$, and the uniform distribution is approximated in the limit $r_m \to \infty$.

Generalised exponential source models in static ICA are able to separate mixtures of Laplacian distributed, Gaussian distributed and uniformly distributed sources—mixtures which methods using a static tanh nonlinearity are unable to separate [6].

1.4 Particle Filters

Practical implementation of the filtering equations (1.15 and 1.16) requires a representation of the predicted and corrected state densities: $p(\boldsymbol{a}_t \,|\, X_{t-1})$ and $p(\boldsymbol{a}_t \,|\, X_t)$. State densities in the Kalman filter [13,11], whose observation equation is linear, are Gaussian and thus easily represented. Nonlinear and non-Gaussian tracking problems must either approximate the state density by a tractable form (usually Gaussian) or use a sampling method. Particle filters, which date back to the Sampling Importance Resampling (SIR) filter of Gordon *et al.* [9], represent the state density $p(\boldsymbol{a}_t \,|\, X_{t-1})$ by a collection or swarm of "particles" each with a probability mass. Each particle's probability mass is modified using the state and observation equations after which a new independent sample is obtained from the posterior $p(\boldsymbol{a}_t \,|\, X_t)$ before proceeding to the next prediction/observation step. Alternatively, particle filters may be viewed as a Monte Carlo integration method for the integrals involved in the state and observation equations. Isard and Blake [10] give a nice introduction to tracking using particle filters.

The SIR algorithm finds a swarm of N_p equally weighted particles which approximate the posterior $p(\boldsymbol{a}_{t-1} \,|\, X_{t-1})$ at time $t-1$. The swarm of particles is regarded as an approximate sample from the true sample. At time t we assume that $\{\boldsymbol{a}_{t-1}^n\}_{n=1}^{N_p}$ is the swarm of N_p particles distributed as an independent sample from $p(\boldsymbol{a}_{t-1} \,|\, X_{t-1})$. Filtering proceeds as follows.

Initialisation

The filter is initialised by drawing N_p samples from the prior $p(\boldsymbol{a}_1)$. In the absence of additional information we choose $p(\boldsymbol{a}_1)$ to be Gaussian distributed with a mean found by performing static ICA on the first, say, 100 observations.

Prediction

Draw N_p samples $\{\boldsymbol{v}_t^1, \boldsymbol{v}_t^2, \ldots \boldsymbol{v}_t^{N_p}\}$ from the state noise density $\mathcal{N}(0, Q)$. Each particle is then propagated through the state equation (1.14) to form a new swarm of particles, $\{\boldsymbol{a}^1_{t|t-1}, ..., \boldsymbol{a}^{N_p}_{t|t-1}\}$, where

$$\boldsymbol{a}^n_{t|t-1} = F\boldsymbol{a}^n_t + \boldsymbol{v}^n_t \tag{1.20}$$

If the particles $\{\boldsymbol{a}^n_{t-1}\}$ are an independent sample from $p(\boldsymbol{a}_{t-1}\,|\,\boldsymbol{x}_{t-1})$ the particles $\boldsymbol{a}^n_{t|t-1}$ are an independent sample from $p(\boldsymbol{a}_t\,|\,X_{t-1})$, so the prediction stage implements equation 1.15.

The prediction stage is rapidly achieved since sampling from the Gaussian distribution is efficient.

Filtering

On the observation of a new datum $\boldsymbol{x}_t$ the prediction represented by the swarm $\{\boldsymbol{a}^n_{t|t-1}\}$ can be corrected. Each particle is weighted by the likelihood of the observation $\boldsymbol{x}_t$ being generated by the mixing matrix represented by $\boldsymbol{a}^n_{t|t-1}$. Thus probability masses q^n_t are assigned by

$$q^n_t = \frac{p(\boldsymbol{x}_t\,|\,\boldsymbol{a}^n_{t|t-1})}{\sum_{k=1}^{N_p} p(\boldsymbol{x}_t\,|\,\boldsymbol{a}^k_{t|t-1})} \tag{1.21}$$

This procedure can be regarded as a discrete approximation to equation 1.16, where the prior $p(\boldsymbol{a}_t\,|\,X_{t-1})$ is approximated by the sample $p(\boldsymbol{a}_t\,|\,X_{t-1})$.

Laplace's approximation can be used to approximate the convolution of equation 1.13 for any fixed A_t when the observational noise is small (see Appendix). Otherwise the integral can be evaluated by Monte Carlo integration, however, since the integral must be evaluated for every particle, Monte Carlo integration is usually prohibitively time consuming.

Resampling

The particles $\boldsymbol{a}^n_{t|t-1}$ and weights q^n_t define a discrete distribution approximating $p(\boldsymbol{a}_t\,|\,\boldsymbol{x}_t)$. These are then resampled with replacement N_p times to form an approximate sample from $p(\boldsymbol{a}_t\,|\,\boldsymbol{x}_t)$, each particle carrying an equal weight. This new sample can now be used as the basis for the next prediction.

Although many variants and improvements to the basic SIR filter exist (see [7] for a review), we have not found the basic SIR filter to be adequate for non-stationary ICA.

1.4.1 Source Recovery

Rather than making strictly Bayesian estimates of the model parameters $\boldsymbol{\theta}^m = \{r_m, w_m, \nu_m\}$, the maximum *a posteriori* (MAP) or mean estimate of A_t is used to estimate $\boldsymbol{s}_t$, after which maximum-likelihood estimates of the parameters are found from sequences $\{s^m_\tau\}^t_{\tau=1}$. Finding maximum-likelihood

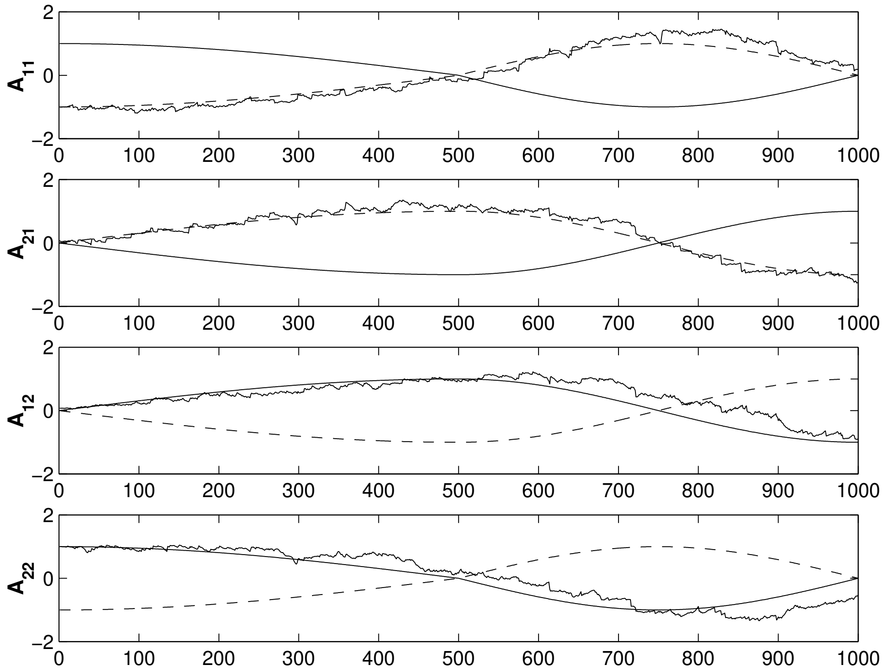

Fig. 1.2. Tracking a mixture of a Laplacian and Gaussian sources.

parameters is readily and robustly accomplished [6]. Each $\boldsymbol{s}_t$ is found by maximising $\log p(\boldsymbol{s}_t \mid \boldsymbol{x}_t, A_t)$, which is equivalent to minimising

$$(\boldsymbol{x}_t - A_t^* \boldsymbol{s}_t)^T R^{-1} (\boldsymbol{x}_t - A_t^* \boldsymbol{s}_t) + \sum_{m=1}^{M} \left| \frac{s_t^m - \nu_m}{w_m} \right|^{r_m} \tag{1.22}$$

where A_t^* is the MAP estimate for A_t. The minimisation can be carried out with a pseudo-Newton method, for example. If the noise variance is small, $\boldsymbol{s}_t \approx A_t^\dagger \boldsymbol{x}_t$, where $A_t^\dagger = (A_t^T A_t)^{-1} A_t^T$ is the pseudo-inverse of A_t, and this estimate provides a good starting guess for a numerical minimisation.

1.5 Illustration

Here we illustrate the method with two examples.

In the first example a Laplacian source ($p(s) \propto e^{-|s|}$) and a source with uniform density are mixed with a mixing matrix whose components vary sinusoidally with time:

$$A_t = \begin{bmatrix} \cos \omega t & \sin \omega t \\ -\sin \omega t & \cos \omega t \end{bmatrix} \tag{1.23}$$

Note, however, that the oscillation frequency doubles during the second half of the simulation making it more difficult to track. Figure 1.2 shows the true mixing matrix and the tracking of it by non-stationary ICA. The solid line marks the mean of the particle swarm. As can be seen from the figure the particle mean tracks the mixing matrix well, although the tracking is poorer in the second half of the simulation where the mixing matrix is changing more rapidly.

Like ICA for stationary mixing, this method cannot distinguish between a column of A_t and a scaling of the column. In Fig. 1.2 the algorithm has "latched on" to the negative of the first column of A_t (shown dashed) which is then tracked for the rest of the simulation.

We resolve the scaling ambiguity between the variance of the sources and the scale of the columns of A_t by insisting that the variance of each source is unity; i.e.,we ignore the estimated value of w_m (equation 1.18), instead setting $w_m = 1$ for all m and allowing all the scale information to reside in the columns of A_t.

To provide an initial estimate of the mixing matrix and source parameters static ICA was run on the first 100 samples. At times $t > 100$ the generalised exponential parameters were re-estimated every 10 observations.

To illustrate the particle filter, every the location of every 50th particle at every 10th timestep is shown in Fig. 1.3. There were 1000 particles in total. The solid line here shows the true mixing matrix.

Estimates of the tracking error are provided by the covariance of the state density, which is found from the particle swarm. In this case the true A_t lies within one standard deviation of the estimated A_t almost all the time (see Fig. 1.3). We remark that it appears to be more difficult to track the columns associated with light-tailed sources than heavy-tailed sources. We note, furthermore, that the Gaussian case appears to be most difficult. In Fig. 1.2, A_{11} and A_{21} mix the Laplacian source, and the uniform source is mixed by A_{12} and A_{22} which are tracked less well, especially during the second half of the simulation. We suspect that the difficulty in tracking columns associated with nearly Gaussian sources is due to the ambiguity between a Gaussian source and the observational noise which is assumed to be Gaussian.

It is easy to envisage situations in which the mixing matrix might briefly become singular. For example, if the microphones are positioned so that each receives the same proportions of each speaker the columns of A_t are linearly dependent and A_t is singular. In this situation A_t cannot be inverted and source estimates (equation 1.22) are very poor. To cope with this we monitor the condition number of A_t; when it is large, implying that A_t is close to singular, the source estimates are discarded for the purposes of inferring the source model parameters, $\{r_m, w_m, \mu_m\}$.

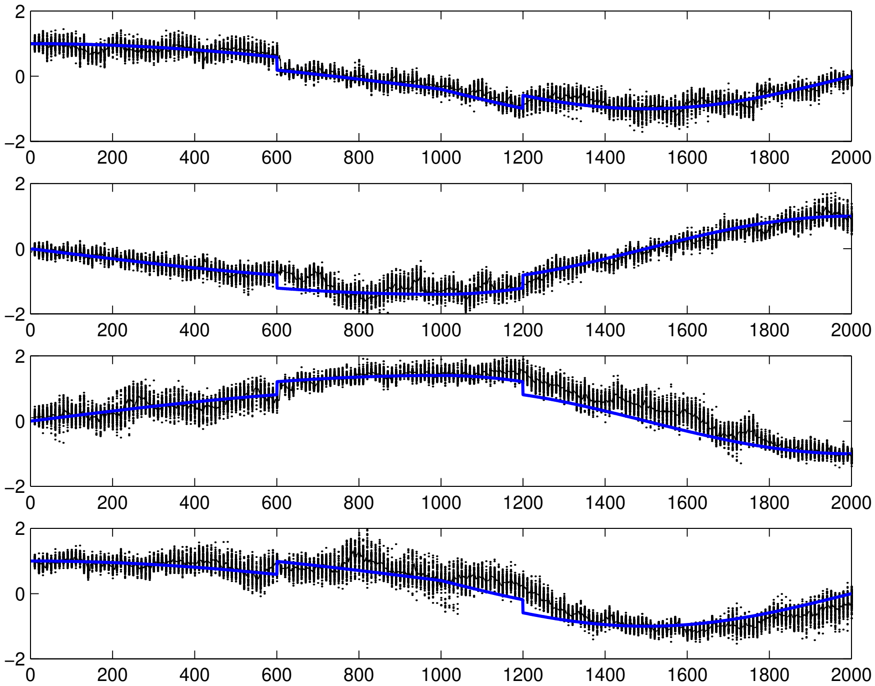

Fig. 1.3. The swarm of particles tracking the mixing matrix for a mixture of a Laplacian and a Gaussian source. Thick solid lines mark the true mixing matrix; thin solid lines mark the mean of the particle swarm. Every 50th particle is plotted every 10th timestep.

In Fig. 1.4 we show non-stationary ICA applied to Laplacian and uniform sources mixed with the matrices

$$A_t = \begin{bmatrix} \cos 2\omega t & \sin \omega t \\ -\sin 2\omega t & \cos \omega t \end{bmatrix} \tag{1.24}$$

where ω is chosen so that A_{1000} is singular. Clearly the mixing matrix is tracked through the singularity, although not so closely as when A_t is well conditioned. Figure 1.5 shows the condition number of the MAP A_t. The normalising constant $Z = p(\boldsymbol{x}_t \,|\, X_{t-1})$ in the correction equation (1.16) is known as the innovations probability (equation 1.17) and measures the degree to which a new datum fits the dynamic model learned by the tracker. Discrete changes of state are signalled by low innovations probability. The innovations probability is approximated by

$$p(\boldsymbol{x}_t \,|\, X_{t-1}) \approx \sum_{n=1}^{N_p} p(\boldsymbol{x}_t \,|\, \boldsymbol{a}^n_{t|t-1}) \tag{1.25}$$

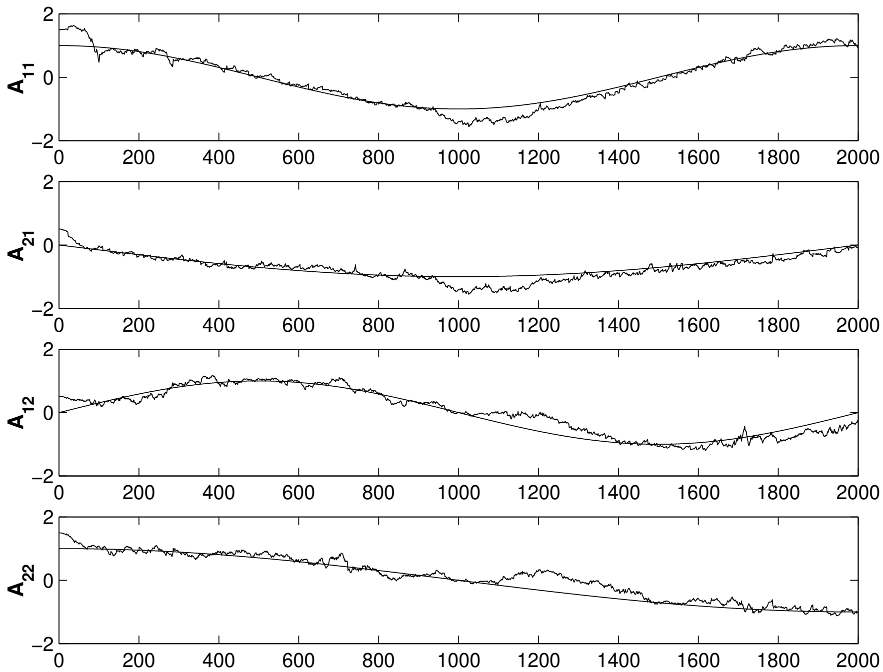

Fig. 1.4. Tracking through a singularity. The mixing matrix is singular at $t = 1000$.

when the state density is represented by particles. Figure 1.5 shows the innovations probability for the mixing shown in Fig. 1.4: the presence of the singularity is clearly reflected.

Note also that the simulation shown in Figure 1.4 was deliberately initialised fairly close to, but not exactly at the true A_1. The "latching on" of the tracker to the correct mixing matrix in the first 100 observations is evident in the figure.

1.6 Smoothing

The filtering methods presented estimate the mixing matrix as $p(A_t | X_t)$. They are therefore strictly causal and can be used for online tracking. If the data are analysed retrospectively future observations $(\boldsymbol{x}_\tau, \tau > t)$ may be used to refine the estimate of A_t. The Markov structure of the generative model permits the pdf $p(\boldsymbol{a}_t | X_T)$ to be found from a forward pass through the data, followed by a backward sweep in which the influence of future observations on $\boldsymbol{a}_t$ is evaluated. See, for example, [8] for a detailed exposition of forward-backward recursions.

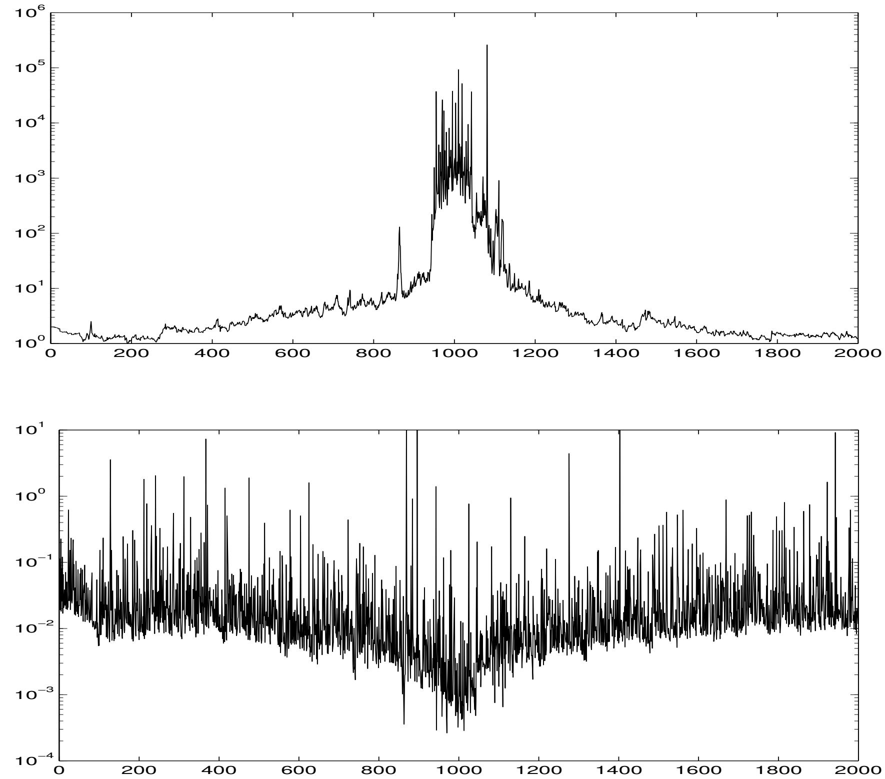

Fig. 1.5. Top: Condition number of the MAP estimate of A_t. At $t = 1000$ the true mixing matrix is singular. Matrices with condition numbers greater than 10 were not used for estimating the source parameters. **Bottom:** Innovations probability $p(\boldsymbol{x}_t \,|\, X_{t-1})$.

In the forward pass the joint probability

$$p(\boldsymbol{a}_t, \boldsymbol{x}_1, ...\boldsymbol{x}_t) \equiv \alpha_t = \int \alpha_{t-1} p(\boldsymbol{a}_t \,|\, \boldsymbol{a}_{t-1})\, p(\boldsymbol{x}_t \,|\, \boldsymbol{a}_t)\, d\boldsymbol{a}_{t-1} \tag{1.26}$$

is recursively evaluated. In the backward sweep the condtional probability

$$p(\boldsymbol{x}_{t+1}, ..., \boldsymbol{x}_T \,|\, \boldsymbol{a}_t) \equiv \beta_t = \int \beta_{t+1}\, p(\boldsymbol{a}_{t+1} \,|\, \boldsymbol{a}_t) p(\boldsymbol{x}_{t+1} \,|\, \boldsymbol{a}_{t+1})\, d\boldsymbol{a}_{t+1} \tag{1.27}$$

is found. Finally the two are combined to produce a smoothed non-causal estimate of the mixing matrix:

$$p(\boldsymbol{a}_t \,|\, \boldsymbol{x}_1, ..., \boldsymbol{x}_T) \;\propto\; \alpha_t \beta_t \tag{1.28}$$

The forwad density α_t and the backward density can each be approximated by a swarm of particles. However, in order to combine them to produce

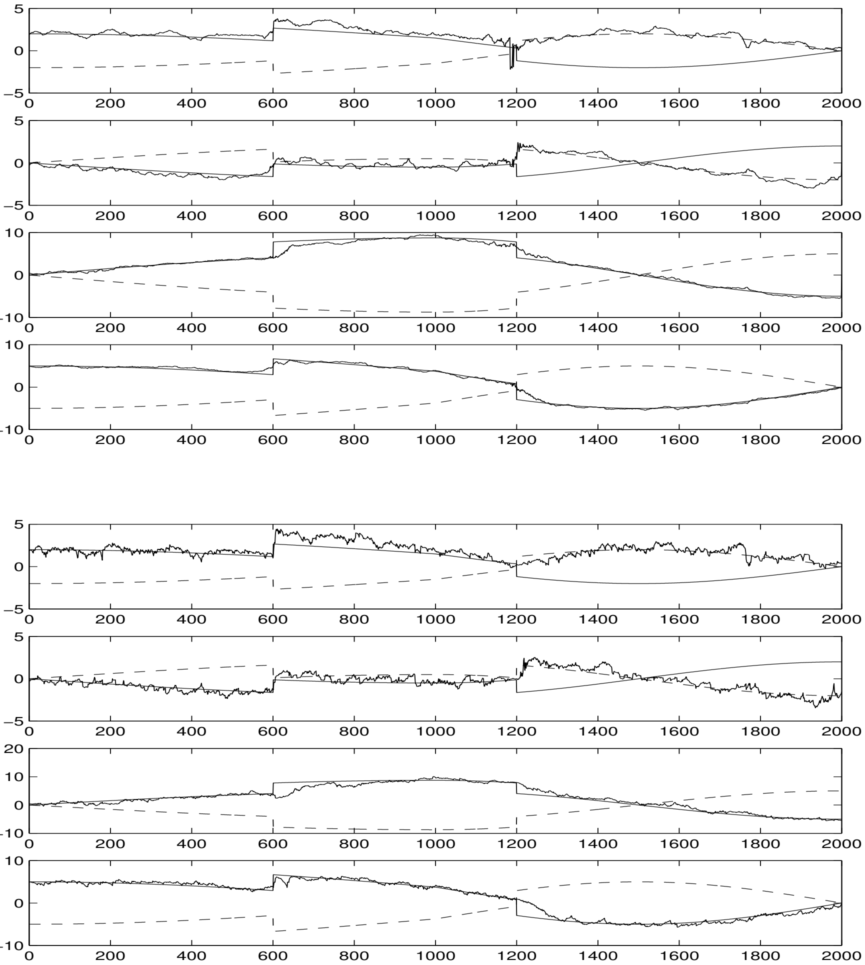

Fig. 1.6. Top: Retrospective tracking with forward-backard recursions. **Bottom:** Online filtering of the same data. Dashed lines show the negative of the mixing matrix elements.

the smoothed estimate (equation 1.28) necessitates storing the entire history of the particles during the forward sweep [14]. The storage problems inherent in this method can be somewhat alleviated by ignoring the influence of observations outside some window around the current observation. An alternative, which we have adopted is to calculate the mean and covariance of the particle swarm approximating α_t at each t on the forward sweep. These are then combined with a Gaussian approximation to β_t from the backward sweep. The Gaussian approximation makes the calculation of equation 1.28

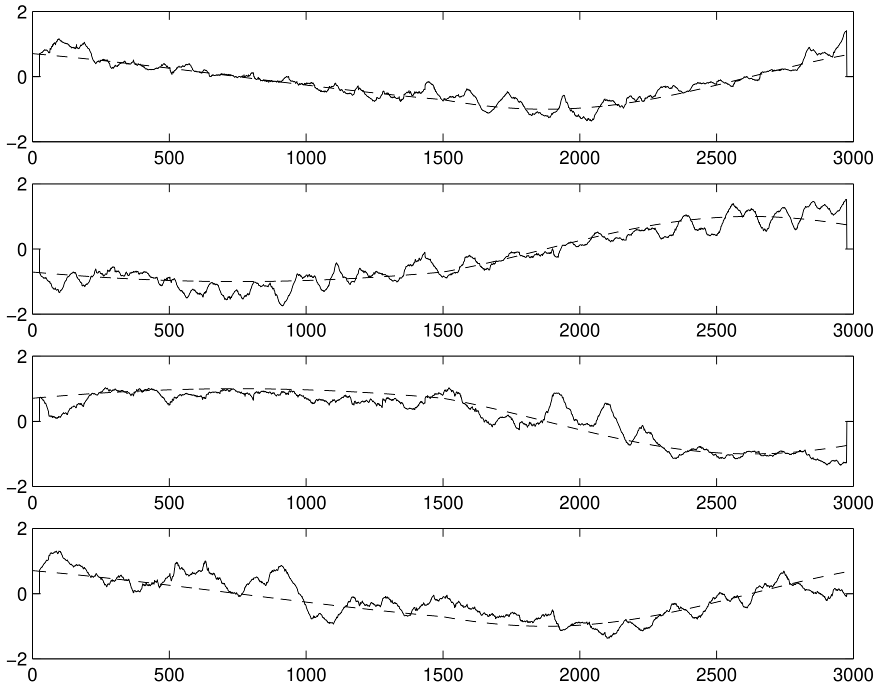

Fig. 1.7. Tracking temporally correlated sources. **Top:** Elements of the mixing matrix during tracking of temporally correlated sources. Dashed lines show the true mixing matrix elements.

particularly simple and only two quantities, the mean and covariance of α_t need be stored.

Figure 1.6 illustrates tracking by both smoothing and causal filtering. As before the elements of the mixing matrix vary sinusoidally with time except for discontinous jumps at $t = 600$ and 1200. Both the filtering and forward-backward recursions track the mixing matrix; however the smoothed estimate is less noisy and more accurate, particularly at the discontinuities. Note also that following the discontinuity at $t = 1200$ the negative of the first columm of A_t is tracked.

1.7 Temporal Correlations

The graphical model in Figure 1.1 assumes that successive samples from each source are independent, so that the sources are stochastic. When temporal correlations in the sources are present the model must be modified to include the conditional dependence of s_t^m on s_{t-1}^m. In this case the hidden state is now comprised of $\boldsymbol{a}_t$ and the states of the sources $\boldsymbol{s}_t$, and predictions and

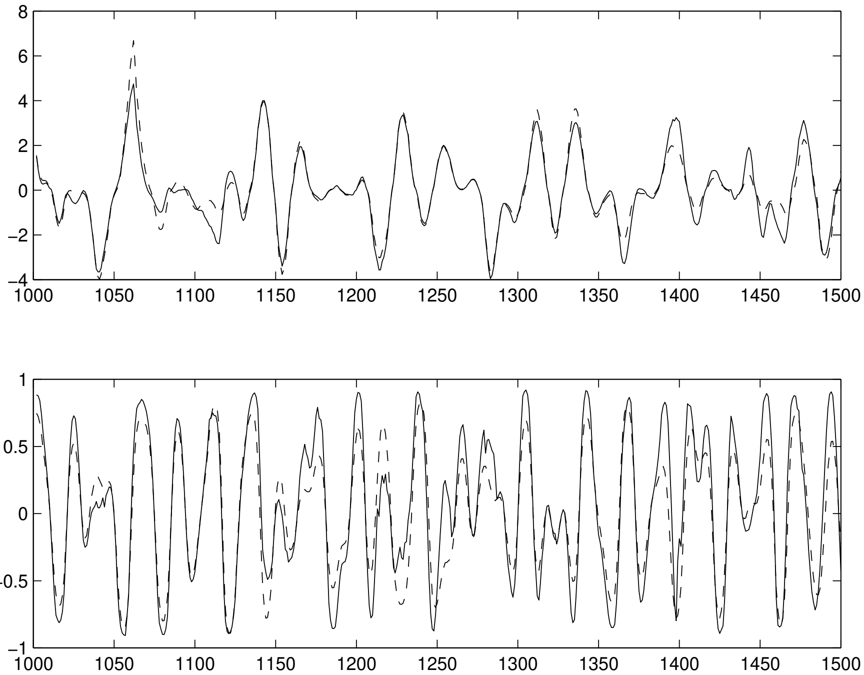

Fig. 1.8. Recovered sources and the true sources (dashed) for times 1000 - 1500 for the tracking shown in Fig. 1.7

corrections for the full state should be made. Since the sources are independent, predictions for the each source and $\boldsymbol{a}_t$ may be made independently and the system is a factorial hidden Markov model [8].

A number of source predictors have been implemented, including the Kalman filter, AR models and Gaussian mixture models. However, we find that in all these cases the combined tracker is unstable. The instability arises because the change in observation from $\boldsymbol{x}_{t+1}$ to $\boldsymbol{x}_t$ cannot be unambiguously assigned to either a change in the mixing matrix or a change in the sources. Small errors in the prediction of the sources induce errors in the mixing matrix estimates, which in turn lead to errors in subsequent source predictions; these errors are then incorporated into the predictive model for the sources and further (worse) errors in the prediction are made. This problem is not present in the stochastic case because the source model is much more tightly constrained.

Under the assumption that the sources evolve on a rapid timescale compared with the mixing matrix, the effect of temporal correlations in the sources may be removed by averaging over a sliding window. That is, the likelihood $p(\boldsymbol{x}_t \,|\, A_t)$ used in the correction step (equation 1.16) is replaced by

$$\left\{ \prod_{\tau=-L}^{L} p(\boldsymbol{x}_{t+\tau}|A_{t+\tau}) \right\}^{\frac{1}{2L+1}} \tag{1.29}$$

The length of the window $2L+1$ is chosen to be of a typical timescale of the sources. Tracking using the averaged likelihood is computationally expensive because at each t the $p(\boldsymbol{x}_{t+\tau}|A_{t+\tau})$ must be evaluated for each τ in the sliding window. An alternative method of destroying the source temporal correlations is to replace the likelihood $p(\boldsymbol{x}_t \,|\, A_t)$ with $p(\boldsymbol{x}_{t+\tau}|A_{t+\tau})$ with τ chosen at random from within the sliding window ($-L \le \tau \le L$). This is no more expensive than using $p(\boldsymbol{x}_t \,|\, A_t)$ and effectively destroys the source correlations.

Figures 1.7 and 1.8 illustrates the tracking of a mixing matrix with temporally correlated sources. The window length was $L = 50$. Tracking is not as accurate as in the stochastic case, however the mixing matrix is followed and the sources are recovered well.

If more precise knowledge about the sources is available it might be incorporated into the source evolution equations in the form of a Bayesian prior. With sufficient information to constrain the sources this might permit the tracking and separation of temporally correlated sources without recourse to temporal averaging.

1.8 Conclusion

We have presented a method for independent components analysis when the mixing proportions are non-stationary. The Sampling Importance Resampling particle filter permits easy tracking of the state density even though the density is non-Gaussian and the observation equation is nonlinear. The method is strictly causal and can be used for online tracking (or "filtering"). If data are analysed retrospectively forward-backward recursions may be used for smoothing rather than filtering. The mixing of temporally correlated sources may be tracked by averaging or sampling from within a sliding window.

In common with most tracking methods, the state noise covariance Q and the observational noise covariance R are parameters which must be set. Although we have not addressed the issue here, it is straight-forward, though laborious, to obtain maximum-likelihood estimates for them using the EM method [8]. It would also be possible to estimate the state mixing matrix F in the same manner.

Although we have modelled the source densities here with generalised exponentials, which permits the separation of a wide range of sources, it is possible to both generalise or restrict the source model. More complicated (possibly multi-modal) densities may be represented by a mixture of Gaussians. On the other hand, if all the sources are restricted to be Gaussian the method becomes a tracking factor analyser. In the zero noise limit the method performs non-stationary principal components analysis.

Acknowledgement

We gratefully acknowledge funding from British Aerospace plc.

1.9 Appendix: Laplace's approximation for the likelihood

Here we use Laplace's approximation to evaluate the convolution integral (1.13) which gives the likelihood of an observation $\boldsymbol{x}$ given the mixing matrix A:

$$p(\boldsymbol{x}\,|\,A) = \int \mathcal{G}(\boldsymbol{x} - A\boldsymbol{s}, R) \prod_{m=1}^{M} p_m(s^m)\, d\boldsymbol{s} \tag{1.30}$$

The source densities are taken to be generalised exponentials (1.18).

Laplace's approximation [18] for the N-dimensional integral

$$I = \int f(\boldsymbol{y}) \exp\{-h(\boldsymbol{y})\}\, d\boldsymbol{y} \tag{1.31}$$

where $h(\boldsymbol{y})$ is sharply peaked, is given by

$$I \approx \int f(\hat{\boldsymbol{y}}) \exp\{-h(\hat{\boldsymbol{y}}) - (\boldsymbol{y} - \hat{\boldsymbol{y}})^T H(\boldsymbol{y} - \hat{\boldsymbol{y}})/2\}\, d\boldsymbol{y} \tag{1.32}$$

Here $H = \frac{\partial^2 h}{\partial \boldsymbol{y}_i \boldsymbol{y}_j}$ is the Hessian and $\hat{\boldsymbol{y}}$ is the $\boldsymbol{y}$ that minimises h. Thus

$$\begin{align} I &\approx f(\hat{\boldsymbol{y}}) \exp -h(\hat{\boldsymbol{y}}) \sqrt{\det 2\pi H^{-1}} \tag{1.33} \\ &= \frac{(2\pi)^{\frac{N}{2}}}{\sqrt{\det H}} f(\hat{\boldsymbol{y}}) \exp -h(\hat{\boldsymbol{y}}) \tag{1.34} \end{align}$$

If the observational noise is small compared to the amplitude of the sources we regard the Gaussian term as $\exp\{-h\}$ and the generalised exponential and normalising factors as f. Then $h(\boldsymbol{s}) = (A\boldsymbol{s} - \boldsymbol{x})^T R^{-1}(A\boldsymbol{x} - \boldsymbol{x})/2$ is minimised at $\hat{\boldsymbol{s}} = (A^T C^{-1} A)^{-1} A^T R^{-1} \boldsymbol{x}$ and the Hessian is $H = A^T R^{-1} A$. Consequently the likelihood is approximated as

$$p(\boldsymbol{x}\,|\,A) \approx \rho \exp\{-(A\hat{\boldsymbol{s}} - \boldsymbol{x})^T R^{-1}(A\hat{\boldsymbol{s}} - \boldsymbol{x})/2\}\ \exp - \sum \left| \frac{\hat{s}_i - \nu_i}{w_i} \right|^{R_i} \tag{1.35}$$

where the normalising factor ρ is

$$\rho = \frac{Z}{\sqrt{\det(A^T R^{-1} A)\det(R)}} \tag{1.36}$$

Comparison with Monte-Carlo integration shows that this approximation is good over a wide range of noise covariances. Evaluation of the likelihood using (1.35) is much more efficient than Monte-Carlo integration or other numerical quadratures.

References

1. S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, Cambridge MA, 1996. MIT Press.
2. H. Attias. Independent factor analysis. *Neural Computation*, 11(5):803–852, 1999.
3. A.J. Bell and T.J. Sejnowski. An information maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.
4. J-F. Cardoso. Infomax and Maximum Likelihood for Blind Separation. *IEEE Sig. Proc. Letters*, 4(4):112–114, 1997.
5. J-F. Cardoso. On the stability of source separation algorithms. In T. Constantinides, S.-Y. Kung, M. Niranjan, and E. Wilson, editors, *Neural Networks for Signal Processing VIII*, pages 13–22. IEEE Signal Processing Society, IEEE, 1998.
6. R.M. Everson and S.J. Roberts. ICA: A flexible non-linearity and decorrelating manifold approach. *Neural Computation*, 11(8), 1999. Available from `http://www.dcs.ex.ac.uk/academics/reverson`.
7. P. Fearnhead. *Sequential Monte Carlo methods in filter theory.* PhD thesis, Merton College, University of Oxford, 1999.
8. Z. Ghahramani. Learning Dynamic Bayesian Networks. In C.L. Giles and M. Gori, editors, *Adaptive Processing of Temporal Information*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 1999.
9. N. Gordon, D. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.
10. M. Isard and A. Blake. Contour tracking by stochastic density propagation of conditional density. In *Proc. European Conf. Computer Vision*, pages 343–356, Cambridge, UK, 1996.
11. A.H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, 1973.
12. I.T. Joliffe. *Principal Component Analysis.* Springer-Verlag, 1986.
13. R. Kalman and R. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering, Trans. ASME Series D*, 83(95-108), 1961.
14. G. Kitigawa. Monte Carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

15. T-W. Lee, M. Girolami, A.J. Bell, and T.J. Sejnowski. A Unifying Information-theoretic Framework for Independent Component Analysis. *International Journal on Mathematical and Computer Modeling*, 1998. (In press). Available from `http://www.cnl.salk.edu/~tewon/Public/mcm.ps.gz`.
16. T-W. Lee, M. Girolami, and T.J. Sejnowski. Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources. *Neural Computation*, 11:417–441, 1999.
17. D.J.C. MacKay. Maximum Likelihood and Covariant Algorithms for Independent Component Analysis. Technical report, University of Cambridge, December 1996. Available from `http://wol.ra.phy.cam.ac.uk/mackay/`.
18. J.J.K. O' Ruanaidth and W.J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.
19. A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 1991.
20. B. Pearlmutter and L. Parra. A Context-Sensitive Generalization of ICA. In *International Conference on Neural Information Processing*, 1996.