??, ??, 1–10 (??)

Blind Source Separation for Non-stationary Mixing

RICHARD EVERSON AND STEPHEN ROBERTS

R.M.Everson@exeter.ac.uk, sjrob@robots.ox.ac.uk

Department of Computer Science,
University of Exeter, Exeter, UK
and
Department of Engineering Science,
University of Oxford, Oxford. UK

Received ??; Revised ??

Editors: ??

Abstract. Blind source separation attempts to recover independent sources which have been linearly mixed to produce observations. We consider blind source separation with non-stationary mixing, but stationary sources. The linear mixing of the independent sources sources is modelled as evolving according to a Markov process, and a method for tracking the mixing and simultaneously inferring the sources is presented. Observational noise is included in the model. The technique may be used for online filtering or retrospective smoothing. The tracking of mixtures of temporally correlated is examined and sampling from within a sliding window is shown to be effective for destroying temporal correlations. The method is illustrated with numerical examples.

Keywords: Blind source separation, independent component analysis, non-stationary, particle filters

1. Introduction

Over the last decade in particular there has been much interest in methods of blind source separation (BSS) and deconvolution (see [11] 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 (which is often referred to as Independent Component Analysis – ICA) in a neuromimetic 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 [13, 4, 14]

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 dynamic methods similar to the Kalman filter.

We first briefly review classical ICA and describe a source model which permits the separation of light-tailed (leptokurtic) sources as well as heavy tailed sources, which the standard ICA model implicitly assumes. 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. Finally we address the non-stationary mixing problem when the sources are independent, but possess temporal correlations.

2. Stationary ICA

Classical ICA assumes that there are M independent sources whose probability density functions are $p_m(s^m)$. Observations, $\mathbf{x}_t \in \mathbb{R}^N$, are produced by the instantaneous linear mixing of the sources by A:

$$\mathbf{x}_t = A\mathbf{s}_t \tag{1}$$

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 $\hat{\mathbf{s}}_t$ by finding W, the (pseudo-) inverse of A:

$$\hat{\mathbf{s}}_t = W\mathbf{x}_t = WA\mathbf{s}_t \tag{2}$$

The assumption that the sources are independent means that the joint probability density funcion (pdf) of the sources factorises into the product of marginal densities:

$$p(\mathbf{s}_t) = \prod_{m=1}^{M} p(s_t^m) \tag{3}$$

Using this factorisation, the (pseudo) likelihood of the observation $\mathbf{x}_t$ is [4, 13, 14]:

$$\log l = -\log|\det A| - \sum_{m=1}^{M} \log p_m(\hat{s}_t^m) \tag{4}$$

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

$$\log \mathcal{L} = -\log|\det A| - \frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} \log p_m(\hat{s}_t^m) \tag{5}$$

The optimum A may then be found by maximisation of $\log \mathcal{L}$ with respect to A, assuming some specific form for $p(\hat{s}_t^m)$. Sucessive gradient ascents on $\log l$ 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, 13].

Since the likelihood is unchanged if A is premultiplied 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:

$$WA = PD \tag{6}$$

In order to maximise the likelihood some assumptions about the form of the source pdfs $p(s_t^m)$ must be made, even though they are *a priori* unknown. 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.

2.1. Generalised Exponentials

By adopting a more flexible model for the source densities one might be able to separate 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.* [12] switch between sub- and super-Gaussian source models.

In order to be able to separate light-tailed sources we have used the generalised exponential density:

$$p(s^m|\boldsymbol{\theta}_m) = z \exp - \left| \frac{s^m - \nu_m}{w_m} \right|^{r_m} \tag{7}$$

where the normalising constant is

$$z = \frac{r_m}{2w_m\Gamma(1/r_m)} \tag{8}$$

and 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 $r_m \to \infty$.

Rather than learn $\{\mu_m, w_m, r_m\}$ along with the elements of the separating matrix W, which magnifies the size of the search space, they may be calculated from the sequences $\{s_t^m\}$ $(t = 1, ...T)$ at any, and perhaps every, stage of learning. The location parameter is well estimated by the sample mean and the maximum likelihood estimate for r_m and w_m may be obtained by solving a *one*-dimensional equation [6].

We have used the generalised exponentials in a quasi-Newton (BFGS [15]) ICA algorithm. At each stage of the optimisation the parameters $\{\mu_m, w_m, r_m\}$ describing the distribution of the mth separated variable were found, permitting the calculation of $\log \mathcal{L}$ and its gradient. This algorithm is able to separate a mixture of a Laplacian source, a Gaussian source and a uniformly distributed source. Algorithms using a static tanh nonlinearity are unable to separate this mixture. Further details are given in [6].

3. Non-stationary Blind Source Separation

Figure 1 shows the graphical model describing the conditional independence relations of the non-stationary BSS model. In common with static blind source separation, we adopt a generative model in which M independent sources are linearly mixed at each instant. Unlike static BSS, however, the mixing matrix A_t is allowed to vary with time. We also assume that the observation $\mathbf{x}_t$ is contaminated by normally distributed noise $\mathbf{w}_t \sim \mathcal{N}(0, R)$. Thus

$$\mathbf{x}_t = A_t \mathbf{s}_t + \mathbf{w}_t \tag{9}$$

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 $\mathbf{a}_t = \mathrm{vec}(A_t)$ be the $N \times M$-dimensional vector obtained by stacking the columns of A_t, then $\mathbf{a}_t$ evolves according to

$$\mathbf{a}_{t+1} = F\mathbf{a}_t + \mathbf{v}_t \tag{10}$$

where $\mathbf{v}_t$ 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 (10) and the statistics of $\mathbf{v}_t$ together define the density $p(\mathbf{a}_{t+1}|\mathbf{a}_t)$.

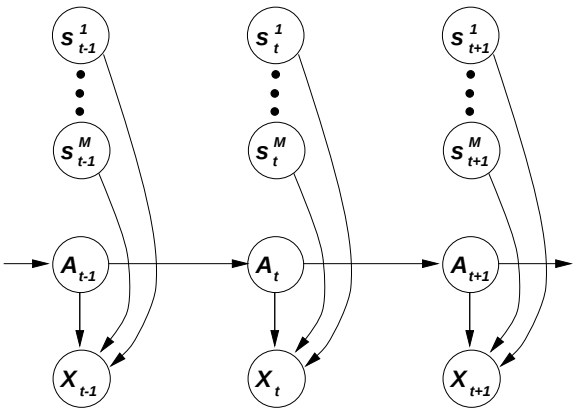

Fig. 1. Graphical model describing non-stationary BSS

A full specification of the state must include the parameter set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_m\}$, $m = 1...M$, which describes the independent source densities:

$$p(\mathbf{s}|\boldsymbol{\theta}) = \prod_{m=1}^{M} p(s^m|\boldsymbol{\theta}_m) \tag{11}$$

We model the source densities with generalised exponentials, as described in §2.1. 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 $\mathbf{x}_t$ become available. If X_t denotes the collection of observations $\{\mathbf{x}_1, ..., \mathbf{x}_t\}$, then the goal of filtering methods is to deduce the probability density function of the state $p(\mathbf{a}_t|X_t)$. This pdf may be found recursively in two stages: prediction and correction. If $p(\mathbf{a}_{t-1}|X_{t-1})$ is known, the state equation (10) and the Markov property that $\mathbf{a}_t$ depends only on $\mathbf{a}_{t+1}$ permits prediction of the state at time t:

$$p(\mathbf{a}_t|X_{t-1}) = \int p(\mathbf{a}_t|\mathbf{a}_{t-1}) p(\mathbf{a}_{t-1}|X_{t-1})\, d\mathbf{a}_{t-1} \tag{12}$$

The predictive density $p(\mathbf{a}_t|X_{t-1})$ may be regarded as an estimate of $\mathbf{a}_t$ prior to the observation of $\mathbf{x}_t$. As the datum $\mathbf{x}_t$ is observed, the

prediction may be corrected via Bayes' rule

$$p(\mathbf{a}_t \,|\, X_t) = Z^{-1} p(\mathbf{x}_t \,|\, \mathbf{a}_t) p(\mathbf{a}_t \,|\, X_{t-1}) \tag{13}$$

where the likelihood of the observation given the mixing matrix, $p(\mathbf{x}_t \,|\, \mathbf{a}_t)$, is defined by the observation equation (9). The normalisation constant Z is known as the innovations probability:

$$\begin{aligned} Z &= p(\mathbf{x}_t \,|\, X_{t-1}) \\ &= \int p(\mathbf{x}_t \,|\, \mathbf{a}_t) p(\mathbf{a}_t \,|\, X_{t-1}) \, d\mathbf{a}_t \end{aligned} \tag{14}$$

The prediction (12) and correction/update (13) 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 arrives.

3.1. Prediction

Since the state equation is linear and Gaussian the state transition density is

$$p(\mathbf{a}_t \,|\, \mathbf{a}_{t-1}) = \mathcal{G}(\mathbf{a}_t - F\mathbf{a}_{t-1}, Q) \tag{15}$$

where $\mathcal{G}(\cdot, \Sigma)$ denotes the Gaussian density function with mean zero and covariance matrix Σ.

We represent the prior density $p(\mathbf{a}_{t-1} \,|\, X_{t-1})$ as a Gaussian:

$$p(\mathbf{a}_{t-1} \,|\, X_{t-1}) = \mathcal{G}(\mathbf{a}_{t-1} - \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}) \tag{16}$$

Prediction is then straight-forward:

$$\begin{aligned} &p(\mathbf{a}_t \,|\, X_{t-1}) = \\ &\qquad \mathcal{G}(\mathbf{a}_t - F\boldsymbol{\mu}_{t-1}, Q + F\Sigma_{t-1}F^T) \end{aligned} \tag{17}$$

3.2. Correction

On the observation of a new datum $\mathbf{x}_t$ the prediction (17) can be corrected. Since the observational noise is assumed to be Gaussian its density is

$$p(\mathbf{w}_t) = \mathcal{G}(\mathbf{w}_t, R) \tag{18}$$

The pdf of observations $p(\mathbf{x}_t \,|\, A_t)$ is given by

$$p(\mathbf{x}_t \,|\, A_t) = \int p(\mathbf{x}_t \,|\, A_t, \boldsymbol{\theta}, \mathbf{s}_t) p(\mathbf{s}_t \,|\, \boldsymbol{\theta}) \, d\mathbf{s}_t \tag{19}$$

and since the *sources* are assumed stationary

$$\begin{aligned} p(\mathbf{x}_t \,|\, A_t) &= \int p(\mathbf{x}_t \,|\, A_t, \mathbf{s}) p(\mathbf{s} \,|\, \boldsymbol{\theta}) \, d\mathbf{s} \\ &= \int \mathcal{G}(\mathbf{x}_t - A_t\mathbf{s}, R) \prod_{m=1}^{M} p_m(s^m) \, d\mathbf{s} \end{aligned} \tag{20}$$

We emphasise that it is in equation (20) that the independence of the sources is modelled by writing the joint source density in factored form.

Laplace's approximation can be used to approximate the convolution (20) for any fixed A_t when the observational noise is small; otherwise the integral can be evaluated by Monte Carlo integration. The corrected pdf $p(\mathbf{a}_t \,|\, X_t)$ of equation (13) is then found by drawing samples, $A_t | X_t$ from the Gaussian of equation (17) and evaluating equation (20) for each sample.

The mean and covariance of the corrected $p(\mathbf{a}_t \,|\, X_t)$ are found from the samples and the density approximated once again by a Gaussian before the next prediction is made.

Rather than representing the state densities as Gaussians at each stage more flexibility may be obtained with particle filter techniques [9, 10]. In these methods the state density is represented by a collection of "particles," each with a probability mass. Each particle's probability is modified using the state and observation equations, after which a new independent sample is obtained using sampling importance resampling before proceding to the next prediction/observation step. Though computationally more expensive than the Gaussian representation, these methods permit arbitrary observational noise distributions to be modelled and more complicated, possibly multi-modal, state densities. The application of particle filter methods to non-stationary ICA is described elsewhere [7].

3.3. 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) estimate of A_t is used to estimate $\mathbf{s}_t$, after which maximum-likelihood estimates of the parameters are found from sequences $\{s_\tau^m\}_{\tau=1}^t$. Finding maximum-likelihood parameters is readily and robustly accomplished [6]. Each $\mathbf{s}_t$ is found by maximising

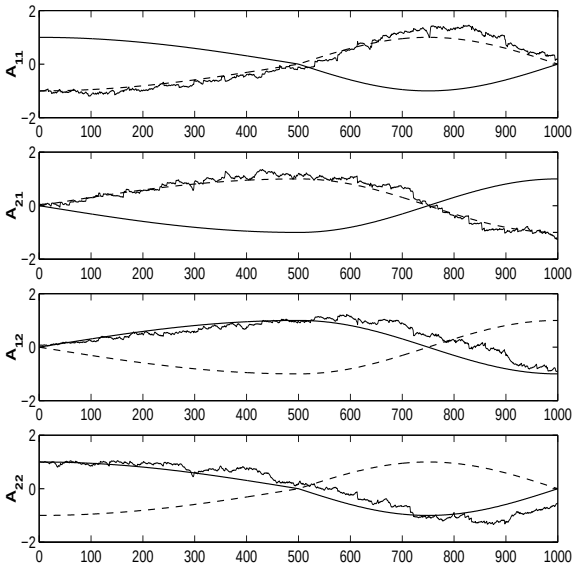

Fig. 2. Tracking a mixture of a Laplacian and Gaussian sources.

$\log p(\mathbf{s}_t|\mathbf{x}_t, A_t)$, which is equivalent to minimising

$$(\mathbf{x}_t - A_t^*\mathbf{s}_t)^T R^{-1}(\mathbf{x}_t - A_t^*\mathbf{s}_t) + \sum_{m=1}^{M} \left| \frac{s_t^m}{w_m} \right|^{r_m} \tag{21}$$

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, $\mathbf{s}_t \approx A_t^\dagger \mathbf{x}_t$, where $A_t^\dagger = (A_t^T A_t)^{-1} A_t^T$ is the pseudo-inverse of A_t.

4. 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:

$$A_t = \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix} \tag{22}$$

Note, however, that the oscillation frequency doubles during the second half of the simulation making it more difficult to track. Figure 2 shows the true mixing matrix and the tracking of it by non-stationary ICA.

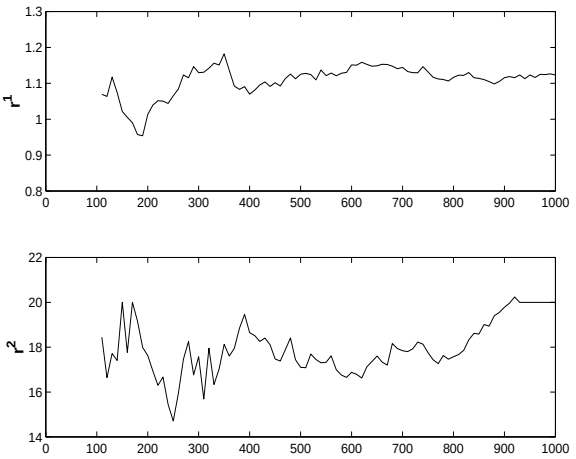

Fig. 3. Online estimates of the generalised exponential parameters r_m during the tracking shown in figure 2.

Like BSS for stationary mixing, this method cannot distinguish between a column of A_t and a scaling of the column. In figure 2 the algorithm has "latched on" to the negative of the first column of A_t, which is shown dashed. 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 7), 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. Figure 3 shows that the estimated source parameters converge to close to their correct values of 1 for the Laplacian source and "large" for the uniform source.

Estimates of the tracking error are provided by the covariance, Σ_t, of the state density (equation 16). In this case the true A_t lies within one standard deviation of the estimated A_t almost all the time. 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 figure 2, A_{11} and A_{21} mix the

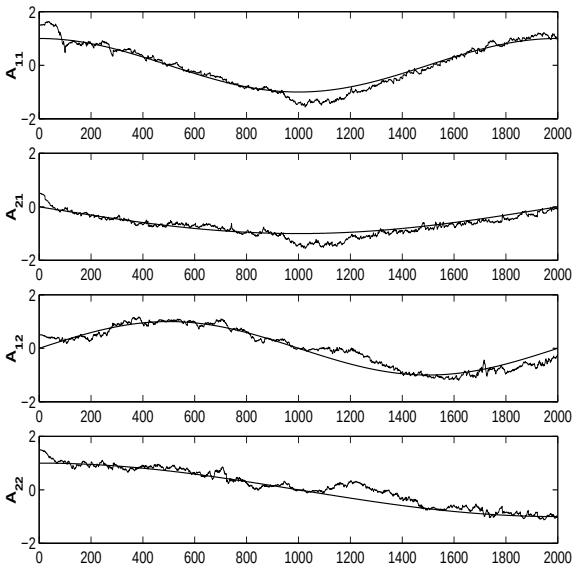

Fig. 4. Tracking through a singularity. The mixing matrix is singular at $t = 1000$.

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 21) 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, $\{r_m, w_m, \mu_m\}$.

In figure 4 we show non-stationary BSS applied to Laplacian and uniform sources mixed with the matrices

$$A_t = \begin{bmatrix} \cos 2\omega t & \sin \omega t \\ -\sin 2\omega t & \cos \omega t \end{bmatrix} \qquad (23)$$

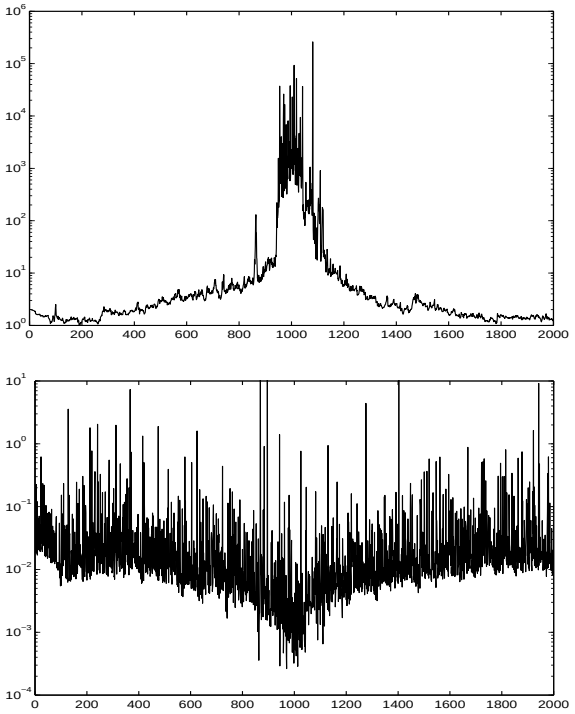

Fig. 5. **Top:** Condition number of the MAP estimate of A_t. At $t = 1000$ the true mixing matrix is singular. Matrices with condition numbers greater than 10 were not used for estimating the source parameters. **Bottom:** Innovations probability $p(\mathbf{x}_t \,|\, X_{t-1})$.

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 5 shows the condition number of the MAP A_t. The normalising constant $Z = p(\mathbf{x}_t \,|\, X_{t-1})$ in the prediction equation (17) is known as the innovations probability 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. Figure 5 also shows the innovations probability for the mixing shown in figure 4: the presence of the singularity is clearly reflected.

Note also that the simulation shown in Figure 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.

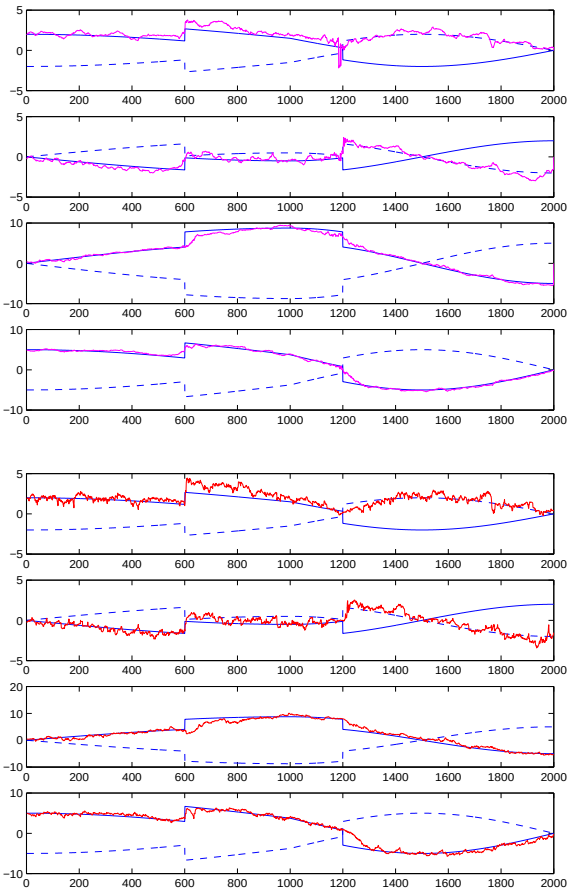

Fig. 6. **Top:** Retrospective tracking with forward-backard recursions. **Bottom:** Online filtering of the same data. Dashed lines show the negative of the mixing matrix elements.

5. 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 ($\mathbf{x}_\tau$, $\tau > t$) may be used to refine the estimate of A_t. The Markov structure of the generative model permits the pdf $p(\mathbf{a}_t|X_T)$ to be found from a forward pass through the data, followed by a backward sweep in which the influence of future observations on $\mathbf{a}_t$ is evaluated. See, for example, [8] for a detailed exposition of forward-backward recursions.

In the forward pass the joint probability

$$p(\mathbf{a}_t, \mathbf{x}_1, ...\mathbf{x}_t) = \alpha_t$$
$$= \int \alpha_{t-1} p(\mathbf{a}_t|\mathbf{a}_{t-1})\, p(\mathbf{x}_t|\mathbf{a}_t)\, d\mathbf{a}_{t-1} \quad (24)$$

is recursively evaluated. In the backward sweep the condtional probability

$$p(\mathbf{x}_{t+1}, ..., \mathbf{x}_T|\mathbf{a}_t) = \beta_t$$
$$= \int \beta_{t+1}\, p(\mathbf{a}_{t+1}|\mathbf{a}_t) p(\mathbf{x}_{t+1}|\mathbf{a}_{t+1})\, d\mathbf{a}_{t+1} \quad (25)$$

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

$$p(\mathbf{a}_t|\mathbf{x}_1, ..., \mathbf{x}_T) \;\propto\; \alpha_t\beta_t \quad (26)$$

If α_t and β_t are each approximated by Gaussians it is nescessary to save only the means and covariance matrices

Figure 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 the following the discontinuity at $t = 1200$ the negative of the first columm of A_t is tracked.

6. Temporal Correlations

The graphical model in Figure 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 $\mathbf{a}_t$ and the states of the sources $\mathbf{s}_t$, and predictions and corrections for the full state should be made. Since the sources are independent, predictions for the each source and $\mathbf{a}_t$ 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, the fundamental indeterminacy of the source scales ren-

ders the combined tracker unstable. The instability arises because the change in observation from $\mathbf{x}_{t+1}$ to $\mathbf{x}_t$ 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 $p(\mathbf{x}_t|A_t)$ used in the correction step (equation 13) is replaced by

$$\left\{\prod_{\tau=-L}^{L} p(\mathbf{x}_{t+\tau}|A_{t+\tau})\right\}^{\frac{1}{2L+1}} \tag{27}$$

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 $p(\mathbf{x}_{t+\tau}|A_{t+\tau})$ must be evaluated for each τ in the sliding window. An alternative method of destroying the source temporal correlations is to replace the likelihood $p(\mathbf{x}_t|A_t)$ with $p(\mathbf{x}_{t+\tau}|A_{t+\tau})$ with τ chosen at random from within the sliding window ($-L \leq \tau \leq L$). This is no more expensive than using $p(\mathbf{x}_t|A_t)$ and effectively destroys the source correlations.

Figure 7 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.

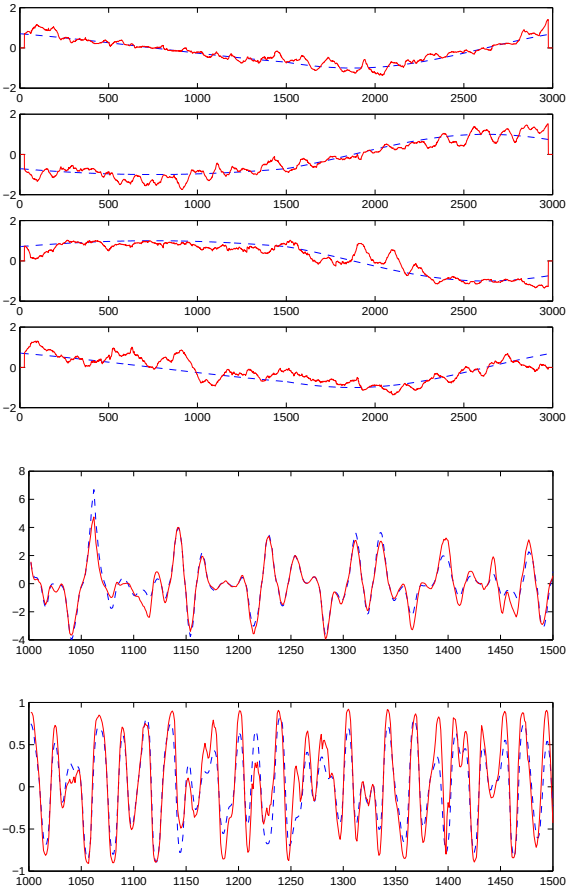

Fig. 7. Tracking temporally correlated sources. **Top:** Elements of the mixing matrix during tracking. Dashed lines show the true mixing matrix elements. **Bottom:** Recovered sources and the true sources (dashed) for times 1000 - 1500.

7. Conclusion

We have presented a method for blind source separation when the mixing proportions are non-stationary. 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 component analysis.

Acknowledgement

We gratefully acknowledge partial funding from British Aerospace plc.

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 nonlinearity and decorrelating manifold approach. *Neural Computation*, 11(8), 1999. Available from `http://www.dcs.ex.ac.uk/academics/reverson`.
7. R.M. Everson and S.J. Roberts. Particle filters for Non-stationary Independent Components Analysis. Technical Report TR99-6, Imperial College, 1999. Available from `http://www.ee.ic.ac.uk/research/neural/everson`.
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. 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`.
12. 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.
13. 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/`.
14. B. Pearlmutter and L. Parra. A Context-Sensitive Generalization of ICA. In *International Conference on Neural Information Processing*, 1996.
15. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 2 edition, 1992.

Richard Everson graduated with a degree in physics from Cambridge University in 1983. He worked in fluid mechanics at Brown and Yale Universities on fluid mechanics and data analysis problems until moving to Rockefeller University, New York, to work on optical imaging and modelling of the visual cortex. After working at Imperial College, University of London, he moved to the Department of Computer Science at Exeter University in 1999. Current research interests include data analysis and pattern recognition; quantitative analysis of brain function, particularly from cortical optical imaging data and EEG; modelling of cortical architecture; Bayesian methods and signal and image processing.

Stephen Roberts graduated from Oxford University with a degree in physics in 1987. After working in industry he returned to Oxford and obtained his DPhil in 1991. He was lecturer in Engineering Science at St. Hugh's College, Oxford, prior to his appointment as lecturer in the Department of Electrical & Electronic Engineering at Imperial College, Univer-

sity of London, in 1994. In 1999 he was appointed a University Lecturer in Information Engineering at Oxford. His research interests include data analysis, information theory, neural networks, scale space methods, Bayesian methods, image and signal processing, machine learning and artificial intelligence.