Non-stationary Independent Component Analysis

Richard Everson and Stephen J. Roberts

Department of Electrical & Electronic Engineering,
Imperial College, London SW7 2BT.
{r.everson,s.j.roberts}@ic.ac.uk

Abstract

Blind source separation with non-stationary mixing, but stationary sources is considered. The linear mixing of the independent sources is modelled as evolving according to a first order 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 is illustrated with numerical examples.

1 Introduction

Over the last decade in particular there has been much interest in methods of blind source separation and deconvolution (see [9] 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 for source separation is often referred to as Independent Component Analysis - ICA) in a neuro-mimetic framework [2] 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 [11, 3, 12]

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.

2 Theory

In common with static ICA we assume a generative model in which there are M independent sources whose probability density functions are $p_m(s^m)$. The sources are mixed by a matrix A_t which, unlike static ICA, is allowed to vary with time. The observation $\mathbf{x}_t \in \mathbb{R}^N$ at time t $(t = 1, \ldots T)$ is a linear mixture to which is added observational noise, $\mathbf{w}_t$:

$$\mathbf{x}_t = A_t \mathbf{s}_t + \mathbf{w}_t \tag{1}$$

The mixing matrix 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 observational noise is taken to be normally distributed, with zero mean and covariance matrix R.

Orthodox ICA takes the mixing matrix to be constant in time and assumes that the observations are noise-free. An unmixing matrix W (the inverse of A, up to multiplication by a diagonal matrix and a permutation matrix) can be found by minimising the mutual information between the unmixed sources, $\hat{\mathbf{s}}_t = W\mathbf{x}_t$. Attias [1] has significantly developed ICA by introducing Independent Factor Analysis, which includes observational noise and so bears the same relation to ICA as factor analysis does to PCA.

Here we permit A_t to vary with time. The dynamics of A_t are modelled by a first order Markov process. If we let $\mathbf{a}_t = \mathrm{vec}(A_t)$ be the $N \times M$-dimensional vector obtained by

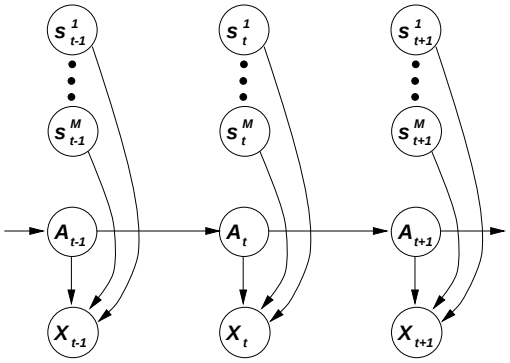

Figure 1: *Graphical model describing non-stationary ICA*

stacking the columns of A_t, then $\mathbf{a}_t$ evolves according to

$$\mathbf{a}_{t+1} = F\,\mathbf{a}_t + \mathbf{v}_t \tag{2}$$

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

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

$$p(\mathbf{s}\,|\boldsymbol{\theta}) = \prod_{m=1}^{M} p(s^m|\boldsymbol{\theta}_m) \tag{3}$$

These parameters are taken to be static, but they must be learned as data are observed. A full description of the source model is deferred to section 2.1. Figure 1 illustrates the graphical model describing the conditional independence relations of the non-stationary ICA model.

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 (pdf) 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 (2) gives a prediction of the state at time t:

$$p(\mathbf{a}_t\,|X_{t-1}) = \int p(\mathbf{a}_t\,|\mathbf{a}_{t-1})p(\mathbf{a}_{t-1}\,|X_{t-1})\,d\mathbf{a}_{t-1} \tag{4}$$

As the datum $\mathbf{x}_t$ is observed, the prediction may be corrected via Bayes' rule

$$p(\mathbf{a}_t\,|X_t) = Z^{-1}p(\mathbf{x}_t\,|\mathbf{a}_t)p(\mathbf{a}_t\,|X_{t-1}) \tag{5}$$

where the normalisation constant Z is known as the innovations probability:

$$\begin{aligned} Z &= p(\mathbf{x}_t\,|X_{t-1}) \\ &= \int p(\mathbf{x}_t\,|\mathbf{a}_t)p(\mathbf{a}_t\,|X_{t-1})\,d\mathbf{a}_t \end{aligned} \tag{6}$$

The likelihood of the datum $\mathbf{x}_t$ given the mixing matrix $\mathbf{a}_t$ is $p(\mathbf{x}_t\,|\mathbf{a}_t)$ which is defined by the observation equation (1).

The prediction (4) and correction/update (5) pair of equations may be used to step through the data, alternately predicting the subsequent state and then correcting the estimate when a new datum arrives.

2.1 Source Model

The source densities are *a priori* unknown and must be modelled. Orthodox ICA uses an apparently fixed source model, although scaling of the mixing matrix tunes the model to particular sources [5, 4]. To separate sources with tails lighter than Gaussian a more flexible source model must be used. Lee *et al* [10] switch between sub- and super-Gaussian source models, and Attias has used mixtures of Gaussians [1] which permit multi-modal sources. We have found generalised exponentials to be effective for separating sub and super-Gaussian sources [5] and we use them to model the sources here. Thus we model each source density as

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

where the normalising constant is

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

The location of the density is set by ν_m and its width by w_m. When $r_m = 1$ the source has a Laplacian density; when $r_m = 2$ the generalised exponential is a Gaussian, and for large r_m the pdf approaches a uniform density.

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 [5]. Each $\mathbf{s}_t$ is found by maximising $\log p(\mathbf{s}_t\,|\,\mathbf{x}_t, A_t)$, which is equivalent to minimising

$$(\mathbf{x}_t - A_t^*\mathbf{s}_t)^T R^{-1}(\mathbf{x}_t - A_t^*\mathbf{s}_t) + \sum_{m=1}^{M}\left|\frac{s_t^m}{w_m}\right|^{r_m} \tag{9}$$

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.

2.2 Prediction

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

$$p(\mathbf{a}_t\,|\,\mathbf{a}_{t-1}) = \mathcal{G}(\mathbf{a}_t - F\,\mathbf{a}_{t-1}, Q) \tag{10}$$

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:

$$p(\mathbf{a}_{t-1}\,|\,X_{t-1}) = \mathcal{G}(\mathbf{a}_{t-1} - \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}) \tag{11}$$

Prediction is then straight-forward:

$$p(\mathbf{a}_t\,|\,X_{t-1}) = \mathcal{G}(\mathbf{a}_t - F\boldsymbol{\mu}_{t-1}, Q + F\Sigma_{t-1}F^T) \tag{12}$$

2.3 Correction

On the observation of a new datum $\mathbf{x}_t$ the prediction (12) can be corrected. Since the observational noise is assumed to be Gaussian its density is

$$p(\mathbf{w}_t) = \mathcal{G}(\mathbf{w}_t, R) \tag{13}$$

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

$$p(\mathbf{x}_t\,|\,A_t) = \int p(\mathbf{x}_t\,|\,A_t, \boldsymbol{\theta}, \mathbf{s}_t) p(\mathbf{s}_t\,|\,\boldsymbol{\theta})\, d\mathbf{s}_t \tag{14}$$

and since the *sources* are assumed stationary

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

$$= \int \mathcal{G}(\mathbf{x}_t - A_t\mathbf{s}, R) \prod_{m=1}^{M} p_m(s^m)\, d\mathbf{s} \tag{15}$$

We emphasise that it is in equation (15) 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 (15) 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 (5) is then found by drawing samples, $A_t|X_t$ from the Gaussian of equation (12) and evaluating equation (15) 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 (see, for example, [8]). 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 multimodal, state densities. The application of particle filter methods to non-stationary ICA is described elsewhere [6].

3 Results

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:

$$A_t = \begin{bmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{bmatrix} \tag{16}$$

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.

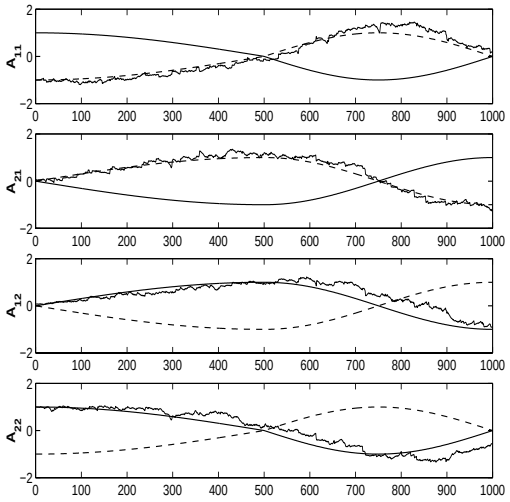

Figure 2: *Tracking a mixture of a Laplacian and Gaussian sources.*

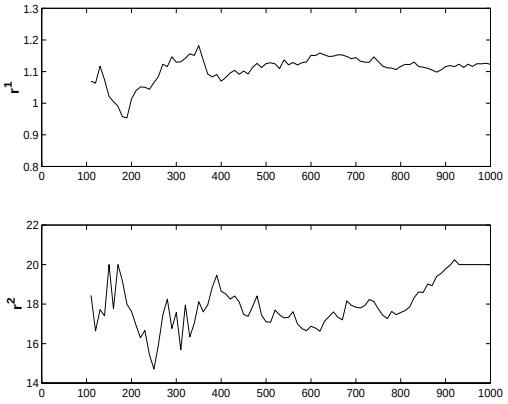

Figure 3: *Online estimates of the generalised exponential parameters r_m during the tracking shown in figure 2.*

Like orthodox ICA, 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 $w_m = 1 \ \ \forall m$ 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" (truncated at 20) for the uniform source.

Estimates of the tracking error are provided by the covariance, Σ_t, of the state density (equation 11). 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, while the columns pertaining to Gaussian sources are hardest to follow. In figure 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 in 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 become briefly 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 9) 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 ICA applied to Laplacian and uniform sources mixed with the matrices

$$A_t = \begin{bmatrix} \cos 2\omega t & \sin \omega t \\ -\sin 2\omega t & \cos \omega t \end{bmatrix} \qquad (17)$$

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 (12) 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 sig-

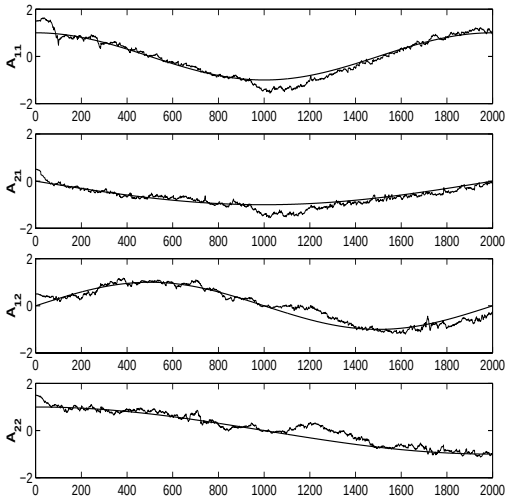

Figure 4: *Tracking through a singularity. The mixing matrix is singular at* $t = 1000$.

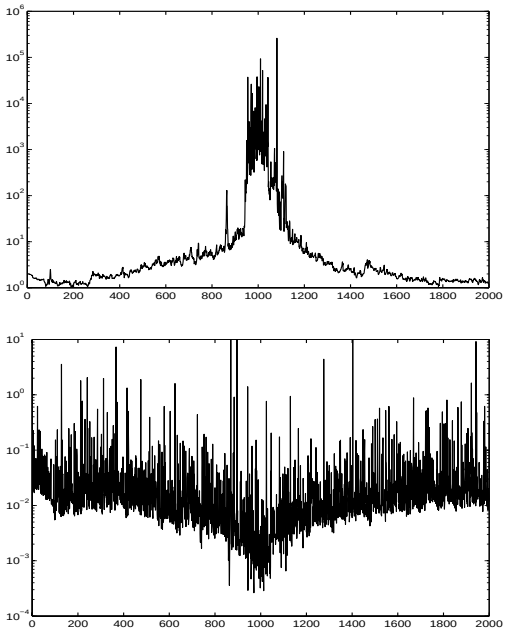

Figure 5: ***Top:*** *Condition number of the MAP estimate of* A_t. *At* $t = 1000$ *the true mixing matrix is singular. Matrices with condition numbers greater than 10 were not used for estimating the source parameters.* ***Bottom:*** *Innovations probability* $p(\mathbf{x}_t \,|\, X_{t-1})$.

nalled 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.

4 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, [7] for a detailed exposition of forward-backward recursions.

Figure 6 illustrates tracking both by 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.

5 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.

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 [7].

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

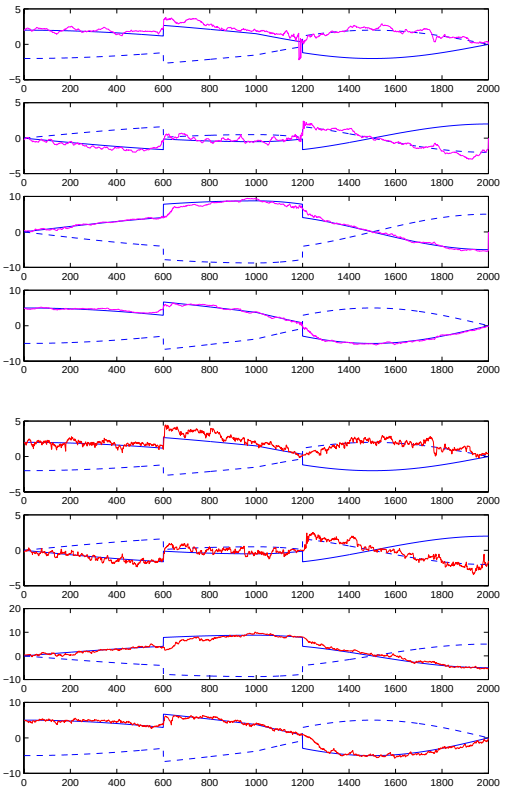

Figure 6: ***Top:*** *Retrospective tracking with forward-backard recursions.* ***Bottom:*** *Online filtering of the same data. Dashed lines show the negative of the mixing matrix elements.*

limit the method performs non-stationary principal component analysis.

Finally we remark that current work concentrates on tracking the sources themselves. This is important when successive samples from each (independent) sources are not independent [12].

Acknowledgement

We gratefully acknowledge partial funding from British Aerospace plc.

References

[1] H. Attias. Independent factor analysis. *Neural Computation*, 1998. *In press.*

[2] A.J. Bell and T.J. Sejnowski. An information maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] J-F. Cardoso. Infomax and Maximum Likelihood for Blind Separation. *IEEE Sig. Proc. Letters*, 4(4):112–114, 1997.

[4] J-F. Cardoso. On the stability of source separation algorithms. In *Neural Networks for Signal Processing VIII*, 1998.

[5] R.M. Everson and S.J. Roberts. ICA: A flexible non-linearity and decorrelating manifold approach. *Neural Computation*, 1999. (*To appear.*) Available from `http://www.ee.ic.ac.uk/research/neural/everson`.

[6] R.M. Everson and S.J. Roberts. Particle filters for Non-stationary Indpendent Components Analysis. Technical Report TR99-6, Imperial College, 1999. Available from `http://www.ee.ic.ac.uk/research/neural/everson`.

[7] 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.

[8] 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.

[9] 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*, 1999.

[10] 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.

[11] 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/`.

[12] B. Pearlmutter and L. Parra. A Context-Sensitive Generalization of ICA. In *International Conference on Neural Information Processing*, 1996.