Chapter 1

Independent Components Analysis

*Richard Everson and Stephen J. Roberts**

Independent Components Analysis has recently become an important tool for modelling and understanding empirical datasets. In this chapter we review the theoretical basis of ICA, outline an approach to non-stationary ICA, and describe a number of biomedical case studies. ICA is discussed in the framework of general linear models, which permits comparison with less general methods such as principal components analysis and with flexible models such as neural networks.

1.1 Introduction

Over the last decade in particular there has been much interest in methods of blind source separation and deconvolution (see [1] for a review). The issues involved have traditionally fallen as problems in higher-order statistics. More recently, however, this problem (which for source separation is often referred to as Independent Component Analysis - ICA) has been recast in a neuro-mimetic framework [2] which did much to simplify and popularise the technique. More recently still the ICA problem has been shown to be a maximum-likelihood approach of a latent-variable model [3, 4, 5, 6].

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

1.2 Theory

Independent component analysis considers the set of observations, $\mathbf{x}_t \in \mathbb{R}^{N_o}$, to be produced via a linear mixing of a set of sources, $\mathbf{s}_t \in \mathbb{R}^{N_s}$:

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

This is equivalent to assuming a linear transform from the N_o dimensional observation space to the $N_s \leq N_o$ dimensional source-estimation space of the form:

$$\mathbf{y}_t = \mathbf{W}\mathbf{x}_t \tag{1.2}$$

where $\mathbf{y}_t \in \mathbb{R}^{N_s}$, $\mathbf{x}_t \in \mathbb{R}^{N_o}$, $\mathbf{W} \in \mathbb{R}^{N_s \times N_o}$ and the data samples are indexed by $t = 1, ..., T$. The goal of ICA is to find a matrix $\mathbf{W}$ such that the components of $\mathbf{y}$, y_t^i are maximally independent. This independence is defined in an entropic sense by minimisation of the mutual information $I(\mathbf{y})$ between the components of $\mathbf{y}$, which is equivalent to the Kullback-Leibler (KL) measure between the joint density over $\mathbf{y}$ and the factored density, $\prod_{i=1}^{N_s} p(y^i)$. Maximally independent components y^i therefore maximise the informativeness (in an entropic sense) of this set of estimated sources. Thus ICA seeks a $\mathbf{W}$ that minimises

$$\begin{aligned} I(\mathbf{y}) &= KL(\mathbf{y} \parallel \{y^i\}) \\ &= \int p(\mathbf{y}) \log \left(\frac{p(\mathbf{y})}{\prod_{i=1}^{N_s} p(y^i)} \right) d\mathbf{y} \\ &= -\frac{1}{2} \log |\det \mathbf{W}^T \mathbf{W}| + H(\mathbf{x}) + \sum_{i=1}^{N_s} H(y^i) \end{aligned} \tag{1.3}$$

where $H(\mathbf{x})$ is the data entropy and $H(y^i)$ are the marginal entropies of the latent variable components.

The normalised sample (log) likelihood function of a generative model with independent components is

$$L(\mathbf{x}) \;=\; \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t \mid \mathbf{W}) \;=\; \frac{1}{2} \log |\det \mathbf{W}^T \mathbf{W}| - \sum_{i=1}^{N_s} H(y^i) \tag{1.4}$$

Combining equations (1.3) and (1.4) obtains

$$KL(\mathbf{y} \parallel \{y^i\}) = H(\mathbf{x}) - L(\mathbf{x}) \tag{1.5}$$

Since $H(\mathbf{x})$ is constant, minimising the KL measure (to give maximally independent components) is hence equivalent to maximising the (log) likelihood function of the data [4, 3, 7].

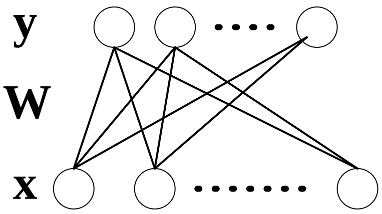

Figure 1.1: **ICA:** *We may regard the ICA model as single-layer neural network. Each estimated source, y^i, may be written as a linear combination of observations (inputs), $\mathbf{x}$, with the 'synaptic' weights forming the un-mixing matrix $\mathbf{W}$. Transfer functions of the 'neurons' are ϕ_i.*

The gradient of L with respect to the elements of $\mathbf{W}$ is needed for the optimisation of L: it is given by

$$\frac{\partial L}{\partial \mathbf{W}_{ij}} = -\mathbf{W}^{-T} - \langle \mathbf{\Phi x} \rangle \tag{1.6}$$

where the components of $\mathbf{\Phi}$ are $\phi_i = \frac{\partial \log p(y^i)}{\partial y^i}$ and $\langle \cdot \rangle$ denotes the sample expectation. Computational implementations commonly choose a particular form for ϕ_i ($\phi = -\tanh$ is a common choice) and the optimum $\mathbf{W}$ found by gradient ascent using (1.6). We mention the important natural gradient modifications to (1.6), which make the learning rule covariant and greatly enhance the learning rate [8, 9, 3].

The component-wise mapping of y^i by $\mathbf{\Phi}$, which follows the linear mapping by $\mathbf{W}$ to the latent space may be implemented by a single-layer neural network.

1.2.1 The decorrelating manifold

In a recent approach to ICA we note that if the components of $\mathbf{y}$ are to be independent, they must be linearly decorrelated. Let us now consider what properties a matrix has if it is to be decorrelating. Linear decorrelation means that the expectation of the product between any pair of different variables is zero i.e.,

$$\langle y_t^l y_t^m \rangle = \delta_{l,m} d_m^2 \tag{1.7}$$

in which y_t^m is the m-th variable (set of data) and d_m^2 is a scale factor corresponding to the power of the m-th component. Hence

$$\mathbf{YY}^T = \mathbf{WXX}^T\mathbf{W}^T = \mathbf{D}^2 \tag{1.8}$$

where the m-th row of $\mathbf{Y}$ is y_t^m, $\mathbf{D}$ is a *diagonal* matrix of scaling factors and $\mathbf{X}$ is the matrix of observations. It is simple to show (see [7]) that if $\mathbf{W}$ is to be a decorrelating matrix then it must be of the form:

$$\mathbf{W} = \mathbf{D}\mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{U}^T \tag{1.9}$$

where the columns of $\mathbf{U}$ are the first N_s principal (eigen) components of $\mathbf{X}$, $\mathbf{\Sigma}$ is the (diagonal) matrix of singular values of $\mathbf{X}$, $\mathbf{D}$ is a diagonal scaling matrix and $\mathbf{Q} \in \mathbb{R}^{N_s \times N_s}$ is a real orthogonal matrix. As is shown in [7] optimisation of the ICA unmixing matrix, $\mathbf{W}$ may then proceed in the $N_s(N_s+1)/2$ dimensional space of decorrelating matrices rather than in the full $N_o N_s$ dimensional space. Furthermore, the resultant matrix is guaranteed to be decorrelating. We optimise using the BFGS pseudo-Newton method [10].

1.2.2 The choice of non-linearity

To fully specify the generative model the marginal densities $p(y^i)$ and hence $\phi_i(y^i)$ must be chosen. However, as the appellation *blind* source separation suggests, they are usually unknown and implementations of ICA are frequently forced to chose ϕ arbitrarily. As mentioned above, a common choice is $\phi(y^i) = -\tanh(y^i)$ which implies that $p(y^i) \propto 1/\cosh(y^i)$. One might suppose that the mis-specification of ϕ would severely handicap ICA's ability to unmix sources whose densities are not of the chosen form. In fact the remarkable ability of ICA to cope with a wide range of sources stems from an additional set of parameters implicit in $\mathbf{W}$ which serve to tune the source density models.

The mutual information $I(\mathbf{y})$ does not depend upon the scale of $\mathbf{y}$, so if $\mathbf{D}$ is a diagonal matrix $I(\mathbf{D}\mathbf{y}) = I(\mathbf{y})$. The unmixing matrix $\mathbf{W}$ may therefore be written as $\mathbf{W} = \mathbf{D}\hat{\mathbf{W}}$, where $\sum_j \hat{\mathbf{W}}_{ij}^2 = 1$, so that $\hat{\mathbf{W}}$ contains no scale information. The "free" parameters $\mathbf{D}_{ii}$ may be incorporated into $\mathbf{\Phi}$ so that, if $\hat{\mathbf{y}} = \hat{\mathbf{W}}\mathbf{x}$, the effective nonlinearities are $\hat{\phi}_i(\hat{y}^i) \equiv \phi(\mathbf{D}_{ii}\hat{y}^i) = -\tanh(\mathbf{D}_{ii}\hat{y}^i)$. Integrating $\phi_i = \partial \log p(y^i)/\partial y^i$ to find the effective marginal density shows that $p(\hat{y}^i) \propto \left\{\cosh(\mathbf{D}_{ii}\hat{y}^i)\right\}^{-1/\mathbf{D}_{ii}}$. When $\mathbf{D}_{ii}$ is large $p(\hat{y}^i)$ approaches a Laplacian density (and $\hat{\phi}$ is effectively a step function), while a Gaussian density is modelled as $\mathbf{D}_{ii} \to 0$ and only the linear part of tanh is effective.

By refining the estimate of ϕ as the calculation proceeds one might expect to be able to estimate a more accurate $\mathbf{W}$ and to be able to separate sources with different densities. A range of source densities, including light-tailed sources, may be modelled with the generalised exponential distribution:

$$p(s^i \,|\, r_i, \beta_i, \mu_i) = \frac{r_i}{2\beta_i \Gamma(1/r_i)} \exp\left\{-\left|\frac{s^i - \mu_i}{\beta_i}\right|^{r_i}\right\} \tag{1.10}$$

The location of the density is set by μ_i and its width by β_i. Clearly p is Gaussian when $r = 2$, Laplacian when $r = 1$, and the uniform distribution is approximated in the limit $r \to \infty$. Rather than learn r, μ and β along with the elements of $\mathbf{W}$, which magnifies the size of the search space, they may be calculated for each y^i at any, and perhaps every, stage of learning. A maximum

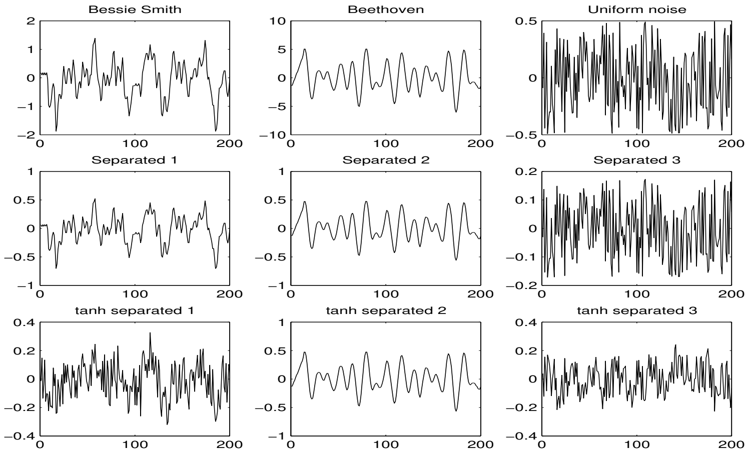

Figure 1.2: **Flexible non-linearity:** *With a fixed (tanh) non-linearity only sources lighter in the tails than a Gaussian can be separated. Uniform noise will hence not be correctly identified as indicated in the poor results of the bottom row. A flexible non-linearity (such as the generalised exponential) achieves excellent results, however (middle row).*

likelihood estimate for r and β may be obtained by solving a *one*-dimensional equation [7].

As an example of the effect the choice of non-linearity may have we show a simple example of separating two sources of music (some blues and some Beethoven) and uniform noise. Fig. 1.2[top row] shows the original sources which are thence mixed. The middle row shows the resultant estimated sources using a flexible non-linearity, the results from which are excellent, whilst the fixed non-linearity (tanh) ICA produces source estimates (bottom row) which are poor.

All results presented in this chapter were obtained using ICA confined to the decorrelating manifold and a generalised-exponential model for the source densities.

1.2.3 Model-order estimation

Although estimating the probable number of sources is possible under a Bayesian interpretation of ICA [11], we are able to take an alternative approach as we evaluate the un-mixing matrix on the decorrelating manifold. All matrices lying on the decorrelating manifold yield models with the same number of sources. We are therefore able to utilise known methods for rank determination of projective matrices (such as $\mathbf{U}$). We adopt the Bayesian approach detailed in [12].

This approach estimates the Bayesian *evidence*† for models of increasing complexity (assumed number of sources). We may choose, for each observation, the model with highest evidence and hence form a histogram of most-likely model orders from the entire data sample. This histogram is then used to select the single, most probable, model order for the entire data sample.

1.3 Non-stationary ICA

Many biomedical applications are non-stationary and the proportions in which sources are mixed drifts with time. Here we present a formulation of ICA that permits non-stationary mixing of the sources; the mixing matrix is thus a (slowly varying) function of time: $\mathbf{A} \equiv \mathbf{A}_t$. We insist, however, that the sources themselves are stationary.

The noiseless, static generative model (equation 1.1) is replaced by a non-stationary mixing and we also model observational noise; thus

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{s}_t + \mathbf{u}_t \tag{1.11}$$

The observational noise is taken to be normally distributed, with zero mean and covariance matrix $\mathbf{R}$. (Attias [13] includes observational noise in a version of static ICA known as Independent Factor Analysis [13].)

The dynamics of $\mathbf{A}_t$ are modelled by a first order Markov process. If we let $\mathbf{a}_t = \text{vec}(\mathbf{A}_t)$ be the $N_o \times N_s$-dimensional vector obtained by stacking the columns of $\mathbf{A}_t$, then $\mathbf{a}_t$ evolves according to

$$\mathbf{a}_t = \mathbf{a}_{t-1} + \mathbf{v}_t \tag{1.12}$$

where ν_t is zero-mean Gaussian noise with covariance $\mathbf{Q}$. The state equation (1.12) and the statistics of $\boldsymbol{\nu}_t$ define the transition density $p(\mathbf{a}_{t+1}|\mathbf{a}_t) = \mathcal{G}(\mathbf{a}_t - \mathbf{a}_{t-1}, \mathbf{Q})$, where $\mathcal{G}(\cdot, \Sigma)$ denotes the Gaussian density function with mean zero and covariance matrix Σ.

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

$$p(\mathbf{s}|\boldsymbol{\theta}) = \prod_{i=1}^{N_s} p(s^i|\boldsymbol{\theta}_i) \tag{1.13}$$

These parameters are taken to be static, but they must be learned as data are observed. As for static ICA, we use generalised exponentials (1.10) ($\boldsymbol{\theta}_i = \{r_i, \beta_i, \mu_i\}$) to model the source densities which are *a priori* unknown.

The problem is now to track $\mathbf{A}_t$ (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)

†The probability of the data set $\mathcal{X}$ given the model $\mathcal{M}$, $P(\mathcal{X}|\mathcal{M})$.

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

$$p(\mathbf{a}_t \,|X_{t-1}) = \int p(\mathbf{a}_t \,|\mathbf{a}_{t-1}) p(\mathbf{a}_{t-1} \,|X_{t-1}) \, d\mathbf{a}_{t-1} \tag{1.14}$$

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

$$p(\mathbf{a}_{t-1} \,|X_{t-1}) = \mathcal{G}(\mathbf{a}_{t-1} - \bar{\mathbf{a}}_t, \mathbf{V}_{t-1}) \tag{1.15}$$

(See [14] for the use of particle filter methods which permit more general state densities.) Since the state equation 1.12) is linear and Gaussian, prediction is then straight-forward:

$$p(\mathbf{a}_t \,|X_{t-1}) = \mathcal{G}(\mathbf{a}_t - \bar{\mathbf{a}}_t, Q + \mathbf{V}_{t-1}) \tag{1.16}$$

On observing datum $\mathbf{x}_t$ the prediction (1.16) may be corrected via Bayes' rule:

$$p(\mathbf{a}_t \,|X_t) \propto p(\mathbf{x}_t \,|\mathbf{a}_{t+1}) p(\mathbf{a}_t \,|X_{t-1}) \tag{1.17}$$

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 (1.11). Since the observational noise is assumed to be Gaussian its density is $p(\mathbf{u}_t) = \mathcal{G}(\mathbf{u}_t, \mathbf{R})$. The likelihood of the observations conditioned on $\mathbf{A}_t$ is given by:

$$p(\mathbf{x}_t \,|\mathbf{A}_t) = \int p(\mathbf{x}_t \,|\mathbf{A}_t, \boldsymbol{\theta}, \mathbf{s}_t) p(\mathbf{s}_t \,|\boldsymbol{\theta}) \, d\mathbf{s}_t \tag{1.18}$$

and since the *sources* are assumed stationary

$$\begin{aligned} p(\mathbf{x}_t \,|\mathbf{A}_t) &= \int p(\mathbf{x}_t \,|\mathbf{A}_t, \mathbf{s}) p(\mathbf{s} \,|\boldsymbol{\theta}) \, d\mathbf{s} \\ &= \int \mathcal{G}(\mathbf{x}_t - \mathbf{A}_t \mathbf{s}, \mathbf{R}) \prod_{i=1}^{N_s} p_i(s^i) \, d\mathbf{s} \end{aligned} \tag{1.19}$$

We emphasise that it is in equation (1.19) that the independence of the sources is modelled by writing the joint source density in factored form.

The prediction (1.14) and correction/update (1.17) 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. After each datum is observed and the state density updated, the maximum *a posteriori* estimate of $\mathbf{A}_t$ is used to estimate the source $\mathbf{s}_t$. Maximum-likelihood estimates of the source parameters $\boldsymbol{\theta}_i = \{r_i, \beta_i, \mu_i\}$ are then determined from sequences $\{s^i_\tau\}^t_{\tau=1}$.

Laplace's approximation can be used to approximate the convolution (1.19) for any fixed $\mathbf{A}_t$ when the observational noise is small; otherwise the integral

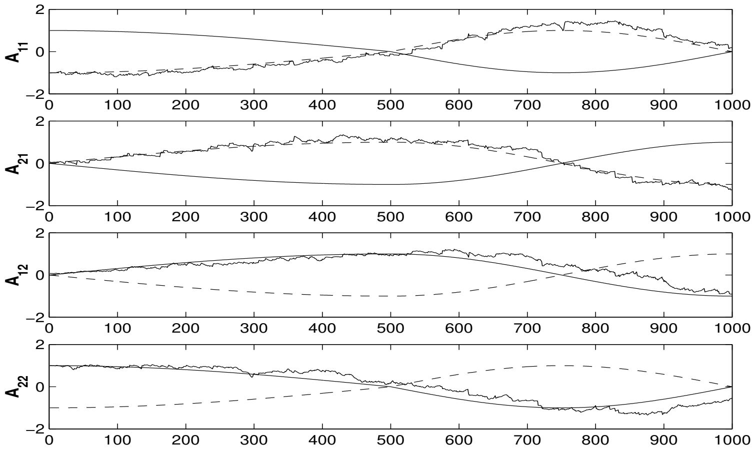

Figure 1.3: **Non-stationary ICA:** *Tracking a mixture of a Laplacian and Gaussian sources.*

can be evaluated by Monte Carlo integration. The corrected pdf $p(\mathbf{a}_t \,|X_t)$ of equation (1.17) is then found by drawing samples, $\mathbf{A}_t|X_t$ from the Gaussian of equation (1.16) and evaluating equation (1.19) 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.

1.3.1 Illustration

We illustrate the method by mixing a Laplacian source and a source with uniform density with a mixing matrix whose components vary sinusoidally with time:

$$\mathbf{A}_t = \begin{pmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{pmatrix} \tag{1.20}$$

Note, however, that the oscillation frequency doubles during the second half of the simulation making it more difficult to track. Fig. 1.3 shows the true mixing matrix and the tracking of it by non-stationary ICA.

Like orthodox ICA, this method cannot distinguish between a column of $\mathbf{A}_t$ and a scaling of the column. Indeed in Fig. 1.3 the algorithm has "latched on" to the negative of the first column of $\mathbf{A}_t$, which is shown dashed. We resolve the scaling ambiguity between the variance of the sources and the scale of the columns of $\mathbf{A}_t$ by insisting that the width of each source is unity; i.e., we ignore the estimated value of β_i (equation 1.10), instead setting $\beta_i = 1 \;\; \forall i$ and allowing all the scale information to reside in the columns of $\mathbf{A}_t$.

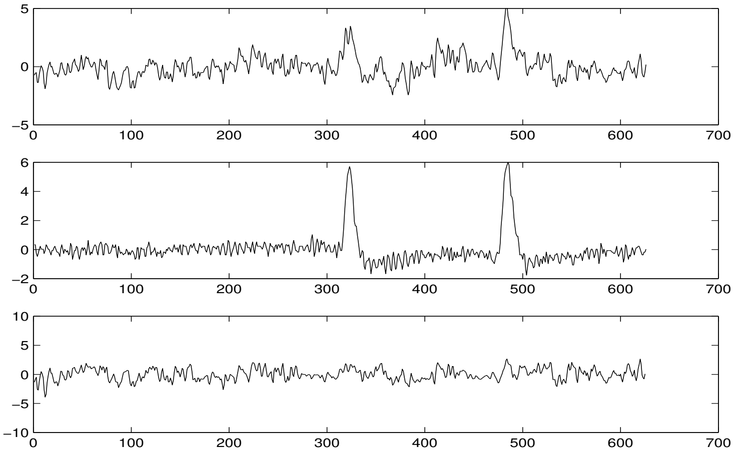

Figure 1.4: **Removal of EOG artefact from EEG:** *EEG channel (C4-A1) [top], EOG trace [middle] and resultant ICA EEG trace [bottom]. The y-axes are all in arbitrary units and the x-axes represent sample numbers at 100Hz.*

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

Estimates of the tracking error are provided by the covariance, $\mathbf{V}_t$, of the state density (1.15). In this case the true $\mathbf{A}_t$ lies within one standard deviation of the estimated $\mathbf{A}_t$ almost all the time.

Note that since the estimate of $\mathbf{A}_t$ uses only observations $\mathbf{x}_1, ..., \mathbf{x}_t$ these results are strictly causal and the method can be used for on-line tracking. If data are analysed retrospectively forward-backward recursions may be used for smoothing rather than filtering [15].

1.4 Applications

We separate the presentation of example results into two subsections. The first details examples in which the assumed number of sources is equal to the number of observations ($N_s = N_o$). This situation is exemplified by artefact-removal applications, in which the interference between multiple observation channels is to be reduced. The second subsection deals with applications in which an unknown number of sources are believed to generate a larger set of observations (i.e. $N_o > N_s$).

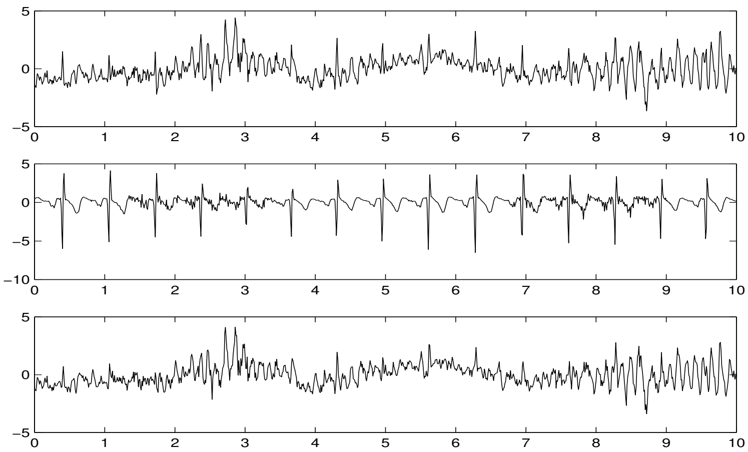

Figure 1.5: **Removal of ECG artefact from EEG:** *EEG channel (C4-A1) [top], ECG trace [middle] and resultant ICA EEG trace [bottom]. Note that the simple linear model of ICA does not adequately reflect the ECG interference in the EEG and hence removal is not complete. The y-axes are all in arbitrary units and the x-axes are in seconds. Sample rate was at 100Hz.*

1.4.1 Source separation

We consider, as examples of ICA with $N_s = N_o$, two cases of artefact removal from the human electroencephalogram (EEG). The first example considers the removal of eye-movement artefact. Eye movements themselves are monitored via the EOG. Fig. 1.4[top & middle] show the EEG[‡] and EOG over some 6 seconds of recording. It is important to note that a linear model, i.e. a superposition, is not inappropriate to describe the eye-movement artefacts in the upper (EEG) trace. Under these circumstances we expect ICA to perform well. Fig. 1.4[bottom] shows the resultant EEG source via ICA. Artefacts from the EOG appear to be well removed.

The second example we include is that of artefacts from heart beats (recorded via the ECG) in the EEG (channel C4-A1 once more). It is clear from the upper two traces of Fig. 1.5 (EEG & ECG) that the linear (superposition) model of ICA is not appropriate. There is, furthermore, a phase shift between the heart-beat events in the two channels. The bottom tracing depicts the resultant EEG source estimated by ICA. Whilst some reduction in the artefact is obtained, it is still clearly visible.

[‡]From the standard '10-20' position C4-A1.

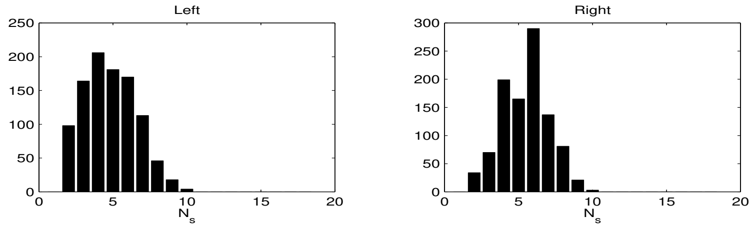

Figure 1.6: **BCI data, model-order histograms:** *Note that $N_s < N_o = 18$ in both cases.*

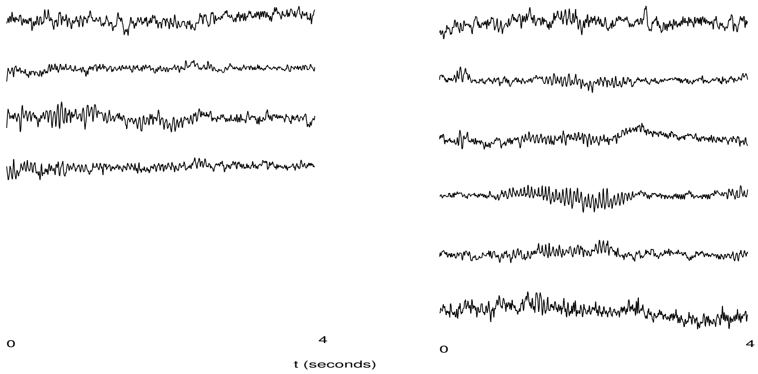

Figure 1.7: **BCI, ICA estimated EEG sources:** *Left and right plots correspond to left and right-hand imagined movements. The movement cue occurred 2 seconds into the block.*

1.4.2 Source number and estimation

Here we consider another EEG case study in which the probable number of 'sources' is unknown but assumed to be less than the number of observations (as is often the case in EEG studies). We present data recorded during a 'brain-computer interface' (BCI) experiment. Recordings were made from 18 channels of EEG from an electrode grid over the primary motor cortex. Subjects were requested to make imaginary left and right-hand finger movements at times given by an external cue [16]. The cue-ing protocol had a countdown to the movement cue in order to allow EEG changes due to movement planning to be detected. Fig. 1.6 shows the model-order histograms for this data. Note that in both cases the estimated number of sources is well below the number of observations ($N_o = 18$). Fig. 1.7 shows the resultant ICA source estimates for left and right-hand imagined movements. The movement cue occurred halfway through the 4 second record. As each source may be regarded as a weighted combination of input channels, each of which is spatially located over the motor cortex, we may map rows of the matrix $\mathbf{W}$ onto a spatial location grid. Fig. 1.8

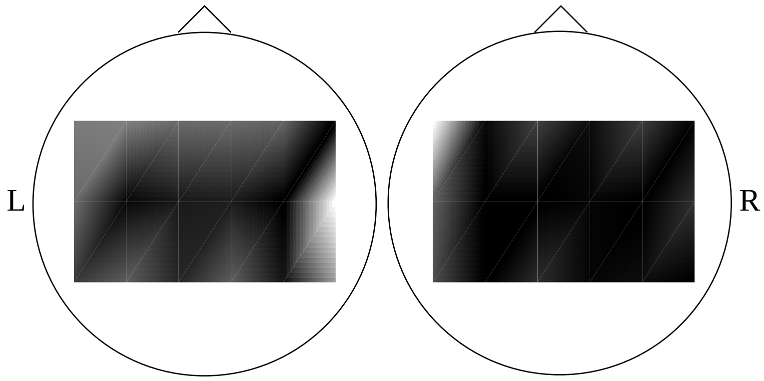

Figure 1.8: **BCI data:** *These plots show the spatial weighting of each original EEG channel in the formation of the highest-power ICA source. Lighter tones represent a higher absolute weighting. Note the contra-lateral location of information in the sources for left (L) and right-hand (R) imagined movements. The nasion is to the top of each plot.*

shows the absolute values of the components of $\mathbf{W}$ associated with the highest-power source estimate for left and right-hand imagined movements. The top of each plot is towards the front of the head and the distributions are smoothed for presentation purposes. Note that higher weighting (lighter tone) is associated with the contra-lateral side for both left and right-hand imagined movements.

1.5 Conclusion

We have described Independent Components Analysis and its application to biomedical data. Particular attention was paid to ensuring that sources are linearly decorrelated and to the problems of flexible source modelling and model order estimation. A new method for tracking non-stationary mixing of stationary sources was introduced.

Orthodox ICA models the observations as being generated by a *linear, noiseless, stationary* mixing of independent sources. Extensions to the ICA formalism now address two of these areas.

An important extension to ICA, due to Attias [13], is dubbed Independent Factor Analysis and models a linear mixing of independent sources with added observational noise.

The state space formulation outlined here permits the tracking of non-stationary mixing matrices. We anticipate that this will become an important tool in the analysis of biomedical data, which are frequently non-stationary, although additional development of the method will be required to handle sources with significant temporal correlations.

The inability of ICA to completely remove nonlinearly mixed heartbeat contamination of EEG data (§1.4.1), highlights the linear nature of ICA. Though nonlinear extensions to ICA have been proposed (and a mixture model appears to be a promising avenue) it is, at present, hard to envisage a completely general approach.

Finally, we point out that all these methods require the dimension of the observation vector to be greater than or equal to the number of sources. Solving the "cocktail party problem" – separating many voices from few observations or deducing global dynamical information from scalar measurements – remains to be solved.

Acknowledgements

We are grateful for discussions with Dirk Husmeier, Will Penny and Iead Rezek. Part of this research was supported by funding from British Aerospace to whom the authors are most grateful.

Bibliography

[1] 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`.

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

[3] 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/`.

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

[5] D.T. Pham. Blind Separation of Instantaneous Mixture of Sources via an Independent Component Analysis. *IEEE Transactions on Signal Processing*, 44(11):2668–2779, 1996.

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

[7] 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`.

[8] 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.

[9] J-F. Cardoso and B. Laheld. Equivarient adaptive source separation. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1996.

[10] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipes in C.* Cambridge University Press, 1991.

[11] S.J. Roberts. Independent Component Analysis: Source Assessment and Separation, a Bayesian Approach. *IEE Proceedings, Vision, Image and Signal Processing*, 1998.

[12] 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 Proceedings - Vision, Image and Signal Processing*, 144(2):116–123, 1997.

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

[14] 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`.

[15] R.M. Everson and S.J. Roberts. Non-stationary Independent Components Analysis. In *Proc. ICANN99.* IEE, 1999.

[16] S.J. Roberts, W. Penny, and I. Rezek. Temporal and Spatial Complexity measures for EEG-based Brain-Computer Interfacing. *Medical and Biological Engineering & Computing*, 1998. In press.