A flexible non-linearity and decorrelating manifold approach to ICA

Richard Everson and Stephen Roberts*
Department of Electrical and Electronic Engineering,
Imperial College of Science, Technology & Medicine,
London. UK.
{r.everson,s.j.roberts}@ic.ac.uk

Abstract

Independent Components Analysis finds a linear transformation to variables which are maximally statistically independent. We examine ICA from the point of view of maximising the likelihood of the data. We elucidate how scaling of the unmixing matrix permits a "static" nonlinearity to adapt to various marginal densities. We demonstrate a new algorithm that uses generalised exponentials functions to model the marginal densities and is able to separate densities with light tails.

We characterise decorrelating matrices and numerically show that the manifold of decorrelating matrices lies along the ridges of high-likelihood unmixing matrices in the space of all unmixing matrices. We show how to find the optimum ICA matrix on the manifold of decorrelating matrices.

1 Introduction

Independent components analysis, which has enjoyed recent theoretical [2, 7, 4, 6] and empirical (e.g., [8]) attention, seeks a linear transformation to coordinates in which the data are maximally statistically independent and not just linearly decorrelated, as would be obtained with principal components analysis. Viewed from another perspective, ICA is a method of separating independent sources which have been linearly mixed to produce the data.

Despite its recent popularity, aspects of the ICA algorithms are still poorly understood. Here we seek to better understand and improve the technique.

Consider a set of T observations, $\mathbf{x}(t) \in \mathbb{R}^N$. Independent components analysis seeks a linear transformation $W \in \mathbb{R}^{M \times N}$ to a new set of variables,

$$\mathbf{a} = W\mathbf{x}, \tag{1}$$

in which the components of $\mathbf{a}$, $a_m(t)$, are maximally independent in a statistical sense. The degree of independence is measured by the mutual information

*We gratefully acknowledge partial support from British Aerospace.

between the components of $\mathbf{a}$:

$$I(\mathbf{a}) = \int p(\mathbf{a}) \log \frac{p(\mathbf{a})}{\prod_m p_m(a_m)} d\mathbf{a}. \tag{2}$$

When the joint probability $p(\mathbf{a})$ can be factored into the product of the marginal densities $p_m(a_m)$, the various components of $\mathbf{a}$ are statistically independent and the mutual information is zero.

If the observations were generated by the noiseless linear mixing of M independent sources $s_m(t)$, so that

$$\mathbf{x} = \mathcal{M}\mathbf{s}, \tag{3}$$

then W is to be regarded as the (pseudo) inverse of the *mixing matrix*, $\mathcal{M}$. Thus successful estimation of W constitutes *blind* source separation.

ICA has been brought to the fore by Bell & Sejnowski's neuro-mimetic formulation [2], which we briefly summarise. For simplicity, we keep to the standard assumption that $M = N$. Bell & Sejnowski introduce a non-linear, component-wise mapping, $\mathbf{y} = \mathbf{g}(\mathbf{a})$, $y_m = g_m(a_m)$ into a space in which the marginal densities are uniform. The linear transformation followed by the non-linear map may be accomplished by a single layer neural network in which the elements of W are the weights and the M neurons have transfer functions g_m.

Since the mutual information is constant under one-to-one, component-wise changes of variables, $I(\mathbf{a}) = I(\mathbf{y})$, and since the g_m are, in theory at least, chosen to generate uniform marginal densities, $p_m(y_m)$, the mutual information $I(\mathbf{y})$ is equal to the negative of the entropy of $\mathbf{y}$:

$$I(\mathbf{y}) = -H(\mathbf{y}) = \int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}. \tag{4}$$

Any gradient-based optimisation approach to the maximum entropy, and therefore the minimum mutual information, requires the gradient of H with respect to the elements of W:

$$\frac{\partial H}{\partial W_{ij}} = \frac{\partial \log |W|}{\partial W_{ij}} + \sum_m \left\langle \frac{\partial}{\partial W_{ij}} \log g'_m(a_m) \right\rangle = W^{-\mathsf{T}} + \left\langle \mathbf{z}\mathbf{x}^{\mathsf{T}} \right\rangle, \tag{5}$$

where $z_i = \phi_i(a_i) = g''_i/g'_i$. If a gradient-ascent method is applied, the estimates of W are updated according to $\Delta W = \nu \partial H / \partial W$ for some learning rate ν. Bell & Sejnowski drop the expectation operator in order to perform a stochastic gradient descent to the minimum mutual information. Various modifications, such as MacKay's covariant algorithm [7] and Amari's natural gradient scheme [1], enhance the convergence rate, but the basic ingredients remain the same. Much more efficient optimisation is possible using, for example, quasi-Newton methods (e.g., BFGS [9]) if the biological plausibility of a biological implementation is sacrificed.

2 Likelihood Landscape

Cardoso [4] and MacKay [7] have each shown that the neuro-mimetic formulation is equivalent to a maximum likelihood approach. The normalised log likelihood for the entire set of observations $\mathbf{x}(t), t = 1...T$, is

$$\log \mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \log P(\mathbf{x}(t)|W) = \log|\det W| - \sum_{m} H_m(a_m), \tag{6}$$

where

$$H_m(a_m) = \frac{1}{T} \sum_{t=1}^{T} \log p_m(a_m(t)) \approx - \int p_m(a_m) \log p_m(a_m) da_m \tag{7}$$

is an estimate of the marginal entropy of the mth unmixed variable.

Note also that the mutual information is given by

$$I(\mathbf{a}) = \int p(\mathbf{a}) \log p(\mathbf{a}) d\mathbf{a} + \sum_{m} H_m(a_m) = H(\mathbf{x}) - \log \mathcal{L}. \tag{8}$$

Thus the mutual information is a constant, $H(\mathbf{x})$, minus the log likelihood, so that hills in the log likelihood are valleys in the mutual information.

Equation (6) requires some sort of normalisation since multiplication of W by a diagonal matrix does not change the mutual information, but would permit us to make the likelihood take on arbitrary values. We therefore choose to normalise W so that the sum of the squares of the elements in each row is unity: $\sum_j W_{ij}^2 = 1 \ \ \forall i$.

When only two sources are mixed, the row normalised W may be parameterised by two angles: $W = \begin{pmatrix} \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \end{pmatrix}$ and the likelihood plotted as a function of θ_1 and θ_2. Figure 1 shows the log likelihood for the mixture of a Gaussian and a bi-exponential source with $\mathcal{M} = [2, 1; 3, 1]$. Several features deserve comment.

1. Singularities: Rows of W are linearly dependent when $\theta_1 = \theta_2 + n\pi$, so $\log|\det W|$ and hence $\log \mathcal{L}$ are singular.

2. Symmetries: Clearly $\log \mathcal{L}$ is doubly periodic in θ_1 and θ_2. Additional symmetry is conferred by the fact that the likelihood unchanged under permutation of the coordinates (here θ_1 and θ_2). Analogous symmetries are retained in higher dimensional examples.

3. Ridges: The maximum likelihood is achieved for several (θ_1, θ_2) related by symmetry, one instance of which is marked by a star in the figure. The maximum likelihood lies on a ridge with steep sides and a flat top. Figure 1b shows a section along the ridge. The rapid convergence of ICA algorithms is probably due to the ease in ascending the sides of the ridge; arriving at the very best solution requires a lot of extra work. Note however, that this is a slightly distorted view of the landscape faced by learning algorithms because they generally work in terms of the full matrix W, rather

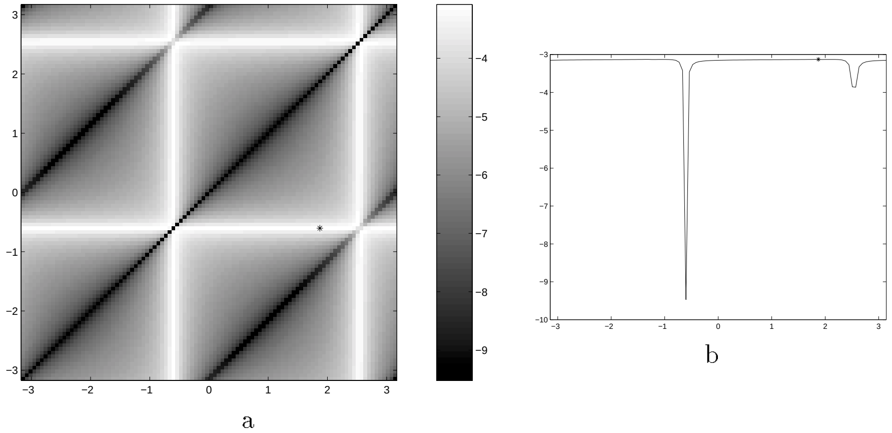

Figure 1: *Likelihood landscape for a mixture of a bi-exponential and Gaussian sources.* ***a:*** *Log likelihood,* $\log \mathcal{L}$*, plotted as a function of* θ_1 *and* θ_2*. Dark gray indicates low likelihood matrices and white indicates high likelihood matrices. The maximum likelihood matrix (i.e., the ICA unmixing matrix) is indicated by the* $*$*.* ***b:*** $\log \mathcal{L}$ *along the "ridge"* $\theta_2 = const.$*, passing through the maximum likelihood.*

than the with row-normalised form.

4. Local maxima: The structure of the likelihood for more realistic examples is very similar to figure 1. The principal difference is that the top of the ridges are frequently bumpy (because the marginal densities are not strictly unimodal) and gradient-based algorithms may stick at a local maximum.

3 Choice of squashing function

The algorithm, as outlined above, leaves open the choice of the "squashing functions" g_m, which map the transformed variables a_m into a space in which their marginal densities are uniform. In fact, what is actually needed are the functions $\phi_m(a_m)$ rather than the g_m themselves.

If the marginal densities are known it is, in theory, a simple matter to find the appropriate g_m, since the function which is the cumulative marginal density is the map into a space in which the density is uniform. That is

$$g(a) = P(a) = \int_{-\infty}^{a} p(x)dx, \tag{9}$$

where the subscripts m have been dropped for ease of notation. Substituting (9) in $\phi(a) = g''/g'$ gives alternative forms for the ideal ϕ:

$$\phi(a) = \frac{\partial p}{\partial P} = \frac{\partial \log p}{\partial a} = \frac{p'(a)}{p(a)}. \tag{10}$$

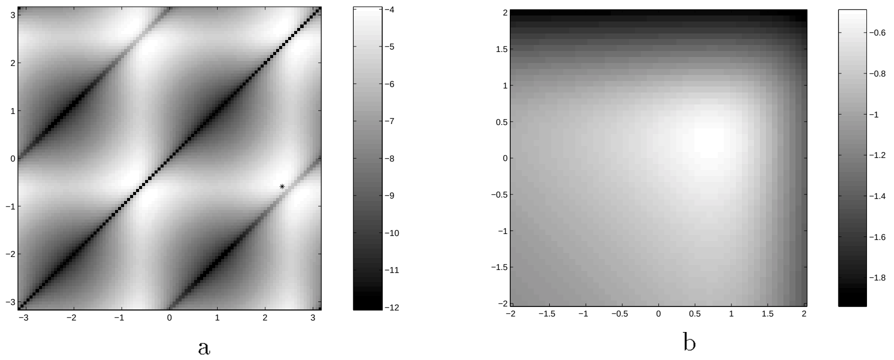

Figure 2: *Likelihood for a bi-exponential/Gaussian mixture, assuming* $1/\cosh$ *sources.* **a:** *Normalised log likelihood plotted in the space of two-dimensional, row-normalised unmixing matrices.* $\mathcal{M}^{-1}$ *is marked with a star.* **b:** *The log likelihood plotted as a function of the elements* D_{11} *and* D_{22} *of a diagonal matrix multiplying the row-normalised maximum likelihood matrix. Since the log likelihood becomes very large and negative for large* D_{mm}*, the gray scale is* $-\log_{10}(|\log \mathcal{L}|)$*. The axes are labelled with* $\log_{10} D_{mm}$*.*

In practice, however, the marginal densities are not known, and a number of forms for g and hence ϕ have been investigated [2]. Current folklore maintains that so long as the marginal densities are heavy tailed (platykurtic) almost any function that squashes will do, and the generalised sigmoidal function and the negative hyperbolic tangent are common choices. Solving (10) with $\phi(a) = -\tanh(a)$ shows that using the hyperbolic tangent is equivalent to assuming $p(a) = 1/(\pi \cosh(a))$.

3.1 Learning the nonlinearity

In section 2 it was argued that, by multiplying the unmixing matrix by a diagonal matrix D, the likelihood could be made to take on arbitrary values without changing the mutual information between the unmixed variables. The estimated mutual information is not, however, independent of D if the unmixed densities appearing in (6) are not the true source densities. This is precisely the case faced by learning algorithms using an *a priori* fixed marginal density. As figure 2 shows, the likelihood landscape for *row-normalised* unmixing matrices using $p(a) = 1/(\pi \cosh(a))$ is similar in form to the likelihood shown in figure 1, though the ridges are not so sharp and the maximum likelihood is only -3.99, which is to be compared with the true maximum likelihood of -3.12. Multiplying the row-normalised mixing matrix by a diagonal matrix, D, opens up the possibility of better fitting the unmixed densities to $1/(\pi \cosh(a))$. Choosing W^* to be the row-normalised $\mathcal{M}^{-1}$, figure 2b shows the log likelihood of DW^* as a function of D: it clearly achieves a maximum for a particular choice of D.

In fact, by adjusting the overall scaling of each row of W, ICA algorithms are "learning the nonlinearity". We may think of the diagonal terms being incorporated into the nonlinearity as adjustable parameters, which are learned along with the row-normalised unmixing matrix.

Numerical experiments show that during learning the weights are adjusted so that the variance of the unmixed Gaussian component is small, while the width of the unmixed exponential component remains relatively large. This means that the Gaussian component really only "feels" the linear part of $-\tanh$ close to the origin and (10) shows that $\phi(a) = -a$ is the correct nonlinearity for a Gaussian distribution. On the other hand, the unmixed bi-exponential component sees the nonlinearity more like a step function, which is appropriate for a bi-exponential density.

Densities with tails lighter than Gaussian (i.e., $p\frac{\partial \log p}{\partial p} < 1$) require a sub-linear ϕ, and it might be expected that the $-\tanh(a)$ nonlinearity would be unable to cope with such marginal densities. Indeed, with $\phi = -\tanh$, the relative gradient [1], covariant [7] and BFGS variations of the ICA algorithm all fail to separate a mixture of uniform, Gaussian and bi-exponential sources.

This point of view gives a partial explanation for the spectacular ability of ICA algorithms to separate sources with different heavy-tailed densities using a "static" non-linearity, and their inability to unmix light-tailed sources (such as uniform densities).

3.2 Adapting ϕ

By refining the estimate of ϕ as the calculation proceeds one might expect to be able to estimate a more accurate W and to be able to separate sources with different densities. Unfortunately numerical experiments show that non-parametric methods, such as kernel density estimators, yield densities which are too noisy to permit the numerical differentiation required to find ϕ.

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

$$p(a|\beta, R) = \frac{R\beta^{1/R}}{2\Gamma(1/R)} \exp\{-\beta|a|^R\}. \tag{11}$$

The width of the distribution is set by $1/\beta$, while the weight of its tails is determined by R. Clearly p is Gaussian when $R = 2$, bi-exponential when $R = 1$, and the uniform distribution is approximated in the limit $R \to \infty$.

Rather than learn R and β along with the elements of W, which magnifies the size of the search space, they may be calculated for each a_m at any, and perhaps every, stage of learning. A maximum likelihood estimate for R and β may be obtained by solving a *one*-dimensional equation [5].

We have used the generalised exponentials in a quasi-Newton (BFGS [9]) ICA algorithm. At each stage of the minimisation the parameters R_m and β_m, describing the distribution of the mth unmixed variable were found, permitting the calculation of $-\log \mathcal{L}$ and its gradient. This algorithm is able to separate a mixture of a bi-exponential 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 [5], and we give an example on real data below.

4 Decorrelating matrices

If an unmixing matrix can be found, the unmixed variables are, by definition, independent. One consequence is that the cross-correlation between any pair of unmixed variables is zero:

$$\langle a_n a_m \rangle \approx \frac{1}{T}\sum_{t=1}^{T} a_m(t)a_n(t) = \frac{1}{T}(a_m, a_n)_t = \delta_{mn} d_n^2, \tag{12}$$

where $(\cdot, \cdot)_t$ denotes the inner product with respect to t, and d_n is a scale factor. Since all the unmixed variables are pairwise decorrelated we have

$$AA^{\mathsf{T}} = WXX^{\mathsf{T}}W^{\mathsf{T}} = D^2, \tag{13}$$

where A is the matrix whose mth row is $a_m(t)$ and D is a diagonal matrix of scaling factors. We will say that a decorrelating matrix for data X is a matrix which, when applied to X, leaves the rows of A decorrelated. Equation (13) comprises $M(M-1)/2$ relations which must be satisfied if W is to be a decorrelating matrix. Clearly there are many decorrelating matrices, of which the ICA unmixing matrix is just one.

If $Q \in \mathbb{R}^{M\times M}$ is an orthogonal matrix and $\hat{D}$ another diagonal matrix,

$$\hat{D}QD^{-1}WXX^{\mathsf{T}}W^{\mathsf{T}}D^{-1}Q^{\mathsf{T}}\hat{D} = \hat{D}^2 \tag{14}$$

so $\hat{D}QD^{-1}W$ is also a decorrelating matrix. The matrix $D^{-1}W$ not only decorrelates, but makes the rows of A orthonormal. It is straightforward to produce a matrix which does this. Let $X = U\Sigma V^{\mathsf{T}}$ be a singular value decomposition of the data matrix $X = [\mathbf{x}(1), \mathbf{x}(2), ...\mathbf{x}(T)]$; $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{T\times T}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{M\times T}$ is a matrix with singular values, $\sigma_i > 0$, arranged along the leading diagonal and zeros elsewhere. Then let $W_0 = \Sigma^{-1}U^{\mathsf{T}}$. Clearly the rows of $W_0X = V^{\mathsf{T}}$ are orthonormal, so the class of decorrelating matrices is characterised as

$$W = DQW_0 = DQ\Sigma^{-1}U^{\mathsf{T}}. \tag{15}$$

The columns of U are the familiar principal components of principal components analysis and $\Sigma^{-1}U^{\mathsf{T}}X$ is the PCA representation of the data X, but normalised or "whitened" so that the variance of the data projected onto each principal component is unity.

Particularly well-known decorrelating matrices [3] are: **PCA:** $Q = I$ and $D = \Sigma$. In this case W simply produces the principal components representation. **ZCA:** $Q = U$ and $D = I$. Bell and Sejnowski [3] call decorrelation with the symmetrical decorrelating matrix $W^{\mathsf{T}} = W$ the zero-phase components analysis. **ICA:** No general analytic form for Q and D can be given, and the optimum Q must be sought by minimising equation (2).

The characterisation of W (15) shows that the decorrelating manifold is $M(M+1)/2$-dimensional: the Cartesian product of the M-dimensional

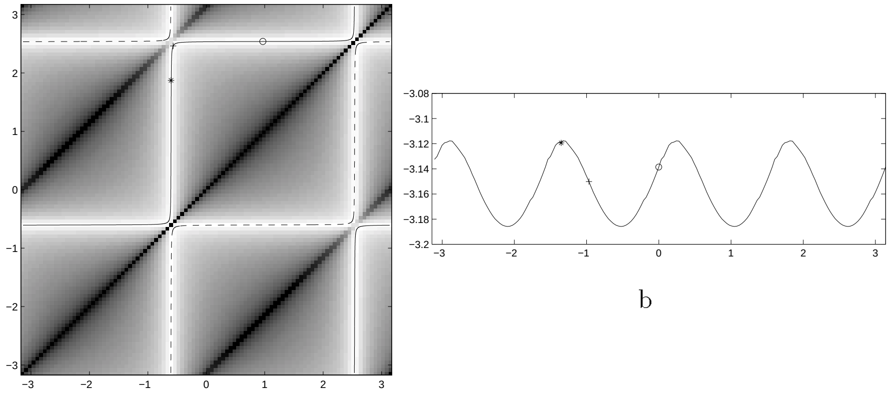

Figure 3: *Decorrelating manifold.* **a:** *The manifold of (row-normalised) decorrelating matrices plotted on the likelihood function for the mixture of Gaussian and bi-exponential sources. Leaves of the manifold corresponding to* $\det Q = \pm 1$ *are as solid and dashed lines respectively. The symbols mark the locations of decorrelating matrices corresponding to PCA* ($\circ$)*, ZCA* ($+$) *and ICA* ($*$). **b:** *Likelihood as the* $\det Q = +1$ *leaf of the decorrelating manifold is traversed.*

manifold $\mathcal{D}$ of scaling matrices and the $(M-1)M/2$-dimensional manifold of orthogonal matrices $\mathcal{Q}$. When $M = 2$ the manifold of decorrelating matrices is 3-dimensional. Since multiplication by D does not change the decorrelation, D is relatively unimportant and the manifold of row-normalised decorrelating matrices (which lies in $\mathcal{D} \times \mathcal{Q}$) may be plotted on the likelihood landscape, as shown in figure 3 for the Gaussian/bi-exponential mixture. The manifold consists of two non-intersecting leaves, corresponding to $\det Q = \pm 1$, which run along the tops of the ridges in the likelihood.

4.1 ICA on the decorrelating manifold

Explicit coordinates on $\mathcal{Q}$ are given by the matrix exponential of an anti-symmetric matrix S:

$$Q = e^{S}; \qquad S^{\mathsf{T}} = -S. \tag{16}$$

The above-diagonal elements of S may be used as coordinates for $\mathcal{Q}$ and the diagonal elements of D are coordinates for $\mathcal{D}$. Thus $-\log \mathcal{L}(D, Q)$ may be minimised by gradient descent or, more efficiently, by BFGS schemes [5].

Since $I(D\mathbf{a}) = I(\mathbf{a})$ for any diagonal D, at first sight it appears that we may choose $D = I_M$. However, as the discussion in section 3.1 points out, the elements of D serve as adjustable parameters tuning a model marginal density to the densities generated by the a_m. If a "fixed" nonlinearity is to be used, it is therefore crucial to permit D to vary and to seek W on the full manifold $\mathcal{D} \times \mathcal{Q}$ of decorrelating matrices. When the marginal densities are modelled with an adaptive nonlinearity D may be held constant and the unmixing matrix sought on $\mathcal{Q}$, using the parameterisation (16). In either case a good starting point is the PCA unmixing matrix.

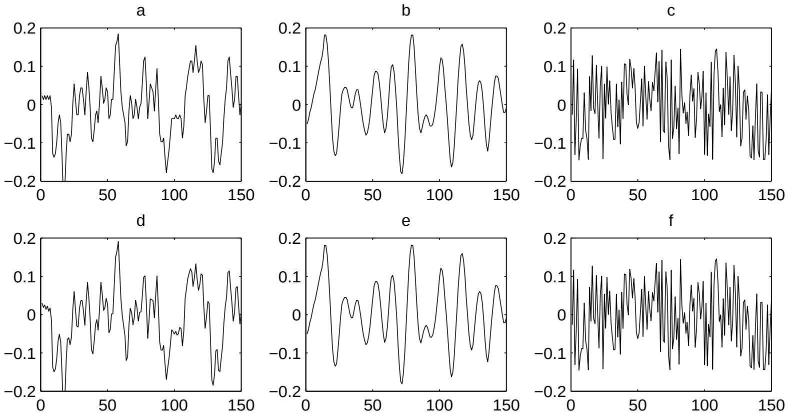

Figure 4: *Separation of two music sources and uniform noise. Top row: 150 samples of the original sources, $s_m(t)$. Bottom row: the unmixed variables, $a_m(t)$. (f_{samp} = 11.3kHz). To facilitate comparison both sources and the unmixed variables have been normalised to unit variance.*

Finding the ICA unmixing matrix on the manifold of decorrelating matrices has a number of advantages. (**1**) The unmixing matrix is guaranteed to be linearly decorrelating. (**2**) The optimum unmixing matrix is sought in the $M(M-1)/2$-dimensional (or with fixed nonlinearities, $M(M+1)/2$-dimensional) space of decorrelating matrices rather than in the full MN-dimensional space. For large problems and especially if the Hessian matrix is being used this provides considerable computational savings. (**3**) The scaling matrix D, which does not provide any additional information, is removed from the numerical solution.

A potentially serious disadvantage is that with small amounts of data the optimum matrix on $\mathcal{Q}$ may not coincide with the maximum likelihood ICA solution, because an unmixing matrix which does not produce exactly linear decorrelation may more effectively minimise the mutual information. Nonetheless, the decorrelating matrix is generally very close to the optimum matrix and provides a good starting point from which to find it.

5 Illustration

In conclusion we illustrate the adaptive nonlinearity and decorrelating manifold approach by applying it to a mixture of three sources: uniform noise and two fragments of music (a Beethoven string quartet and an old recording of a blues ballad). The music sources each had unit variance and the noise was distributed between ±0.5; the elements of $\mathcal{M}$ were chosen at random in $[0,1]$. Figure 4 shows 150 samples of the sources together with the estimated sources found by the adaptive, decorrelating manifold algorithm. It is clear that the algorithm has done a good job in separating the sources: the noisy blues recording is estimated together with its noise (plots a and d), while the string quartet is uncontaminated (plots b and e). To the ear the recovered sources are indistinguishable from the originals, and in particular there is no trace of music in the unmixed noise.

Changes of scale in the a_m and permutation of the order of the unmixed variables do not affect the mutual information, so rather than $W\mathcal{M} = I$ we expect $W\mathcal{M} = PD$ for some diagonal matrix D and permutation matrix P. Under the Frobenius norm, the nearest diagonal matrix to any given matrix A is just its diagonal elements, $\mathrm{diag}(A)$. Consequently the error in W may be quantitatively assessed with

$$\Delta(\mathcal{M}W) = \Delta(W\mathcal{M}) = \min_P \|W\mathcal{M}P - \mathrm{diag}(W\mathcal{M}P)\| / \|W\mathcal{M}\|, \qquad (17)$$

where the minimum is taken over all permutation matrices, P. For perfect unmixing $\Delta(W\mathcal{M}) = 0$. The adaptive, decorrelating manifold algorithm finds W with $\Delta(W\mathcal{M}) = 0.0383$.

The marginal densities of the music sources are both heavy tailed, but the uniform noise has lighter tails than Gaussian. We find that the adaptive algorithm is essential for separating the light tailed source: the algorithms given in [1, 7] and the decorrelating manifold algorithm with $-\tanh$ nonlinearity all audibly fail to separate the sources and all have $\Delta(W\mathcal{M}) > 0.6$.

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, *NIPS*96*, volume 8, pages 757–763, Cambridge MA, 1996. MIT Press.

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

[3] A.J. Bell and T.J. Sejnowski. The "Independent Components" of Natural Scenes are Edge Filters. *Vision Research*, 37(23):3327–3338, 1997.

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

[5] R.M. Everson and S.J. Roberts. ICA: A flexible non-linearity and decorrelating manifold approach. *Neural Comp.*, 1998. (*Submitted.*) Available from `http://www.ee.ic.ac.uk/research/neural/everson`.

[6] 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.

[7] 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/`.

[8] S. Makeig, A. Bell, T-P Jung, and T.J. Sejnowski. Independent Component Analysis of Electroencephalographic Data. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge MA, USA 1996.

[9] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, 2 edition, 1992.