Independent Component Analysis: A flexible nonlinearity and decorrelating manifold approach

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

December 22, 1998

Abstract

Independent Components Analysis finds a linear transformation to variables which are maximally statistically independent. We examine ICA and algorithms for finding the best transformation from the point of view of maximising the likelihood of the data. In particular, we discuss the way in which scaling of the unmixing matrix permits a "static" nonlinearity to adapt to various marginal densities. We demonstrate a new algorithm that uses generalised exponential functions to model the marginal densities and is able to separate densities with light tails.

We characterise the manifold of decorrelating matrices and show that it 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, and as an example use the algorithm to find independent component basis vectors for an ensemble of portraits.

1 Introduction

Finding a natural cooordinate system is an essential first step in the analysis of empirical data. Principal components analysis (PCA) is often used to find a basis set which is determined by the dataset itself. The principal components are orthogonal and projections of the data onto them are *linearly* decorrelated, which can be ensured by considering the *second order* statistical properties of the data. Independent components analysis (ICA), which has enjoyed recent theoretical (Bell and Sejnowski 1995; Cardoso and Laheld 1996; Cardoso 1997; Pham 1996; Lee et al. 1998) and empirical (Makeig et al. 1996; Makeig et al. 1997) attention, aims at a loftier goal: it seeks a linear transformation to coordinates in which the data are maximally statistically independent, not merely decorrelated. 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. In this paper, we seek to better understand and improve the technique. To this end we explicitly calculate the likelihood landscape in the space of all unmixing matrices and examine the way in which the maximum likelihood basis is achieved. The likelihood landscape is used to show how conventional algorithms for ICA which use fixed nonlinearities are able to adapt to a range of source densities by scaling the unmixed variables. We have implemented an ICA algorithm which can separate leptokurtic (i.e., heavy-tailed) and platykurtic (i.e., light-tailed) sources, by modelling marginal densities with the family of generalized exponential densities. We examine ICA in the

[0] `$ Id: notes.tex,v 1.15 1998/12/22 14:28:56 rme Exp rme $`

context of decorrelating transformations, and derive an algorithm which operates on the manifold of decorrelating matrices. As an illustration of our algorithm we apply it to the "Rogues Gallery" – an ensemble of portraits (Sirovich and Sirovich 1989) – in order to find the independent components basis vectors for the ensemble.

2 Background

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

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

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

$$I(\mathbf{a}) = \int p(\mathbf{a}) \log \frac{p(\mathbf{a})}{\prod_k p_k(a_k)} d\mathbf{a} \tag{2}$$

When the joint probability $p(\mathbf{a})$ can be factored into the product of the marginal densities $p_k(a_k)$, the various components of $\mathbf{a}$ are statistically independent and the mutual information is zero. ICA thus finds a *factorial coding* (Barlow 1961) for the observations.

The model we have in mind is that the observations were generated by the noiseless linear mixing of K independent sources $s_k(t)$, so that

$$\mathbf{x} = M\mathbf{s}. \tag{3}$$

The matrix W is thus to be regarded as the (pseudo) inverse of the *mixing matrix*, M. Thus successful estimation of W constitutes *blind* source separation. It should be noted, however, that it may not be possible to find a factorial coding with a linear change of variables, in which case there will be some remaining statistical dependence between the a_k.

ICA has been brought to the fore by Bell & Sejnowski's (1995) neuro-mimetic formulation, which we now briefly summarise. For simplicity, we keep to the standard assumption that $K = N$.

Bell & Sejnowski introduce a nonlinear, component-wise mapping $\mathbf{y} = \mathbf{g}(\mathbf{a})$, $y_k = g_k(a_k)$ into a space in which the marginal densities are uniform. The linear transformation followed by the nonlinear map may be accomplished by a single layer neural network in which the elements of W are the weights and the K neurons have transfer functions g_k.

Since the mutual information is constant under invertible, component-wise changes of variables, $I(\mathbf{a}) = I(\mathbf{y})$, and since the g_k are, in theory at least, chosen to generate uniform marginal densities, $p_k(y_k)$, 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 approach to the maximum entropy, and (if $\mathbf{g}(\mathbf{a})$ is chosen correctly) the minimum mutual information, requires that the gradient of H with respect to the elements of W:

$$\frac{\partial H}{\partial W_{ij}} = \frac{\partial \log|\det W|}{\partial W_{ij}} + \sum_k \left\langle \frac{\partial}{\partial W_{ij}} \log g_k'(a_k) \right\rangle = W^{-\mathsf{T}} + \left\langle \mathbf{z}\mathbf{x}^{\mathsf{T}} \right\rangle \tag{5}$$

where $z_i = \phi_i(a_i) = g_i''/g_i'$ and $\langle \cdot \rangle$ denote expectations. If a gradient-ascent method is applied, the estimates of W are then 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 an online stochastic gradient ascent to maximum entropy. Various modifications of this scheme, such as MacKay's covariant algorithm (MacKay 1996) and Amari's natural gradient scheme (Amari et al. 1996) enhance the convergence rate, but the basic ingredients remain the same.

If one sacrifices the plausibility of a biological interpretation of the ICA algorithm, much more efficient optimisation of the unmixing matrix is possible. In particular, quasi-Newton methods, such as the BFGS scheme (Press et al. 1992), which approximate the Hessian $\partial^2 H/\partial W_{ij} W_{lm}$, can speed up finding the unmixing matrix by at least an order of magnitude.

3 Likelihood Landscape

Cardoso (1997) and MacKay (1996) have each shown that the neuro-mimetic formulation is equivalent to a maximum likelihood approach. MacKay in particular shows that the log likelihood for a single observation $\mathbf{x}(t)$ is

$$\log P(\mathbf{x}(t)|W) = \log|\det W| + \sum_k \log p_k(a_k(t)). \tag{6}$$

The normalised log likelihood for the entire set of observations is therefore

$$\log \mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} \log P(\mathbf{x}(t)|W) = \log|\det W| - \sum_k H_k(a_k), \tag{7}$$

where

$$H_k(a_k) = \frac{1}{T}\sum_{t=1}^{T} \log p_k(a_k(t)) \approx -\int p_k(a_k)\log p_k(a_k) da_k \tag{8}$$

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

Note also that the mutual information (2) is given by

$$I(\mathbf{a}) = \int p(\mathbf{a})\log p(\mathbf{a}) d\mathbf{a} + \sum_k H_k(a_k) \tag{9}$$

$$= -H(\mathbf{a}) + \sum_k H_k(a_k) \tag{10}$$

and since $H(\mathbf{a}) = \log|\det W| - H(\mathbf{x})$ (Papoulis 1991), the likelihood is related to the mutual information by

$$I(\mathbf{a}) = H(\mathbf{x}) - \log\mathcal{L}. \tag{11}$$

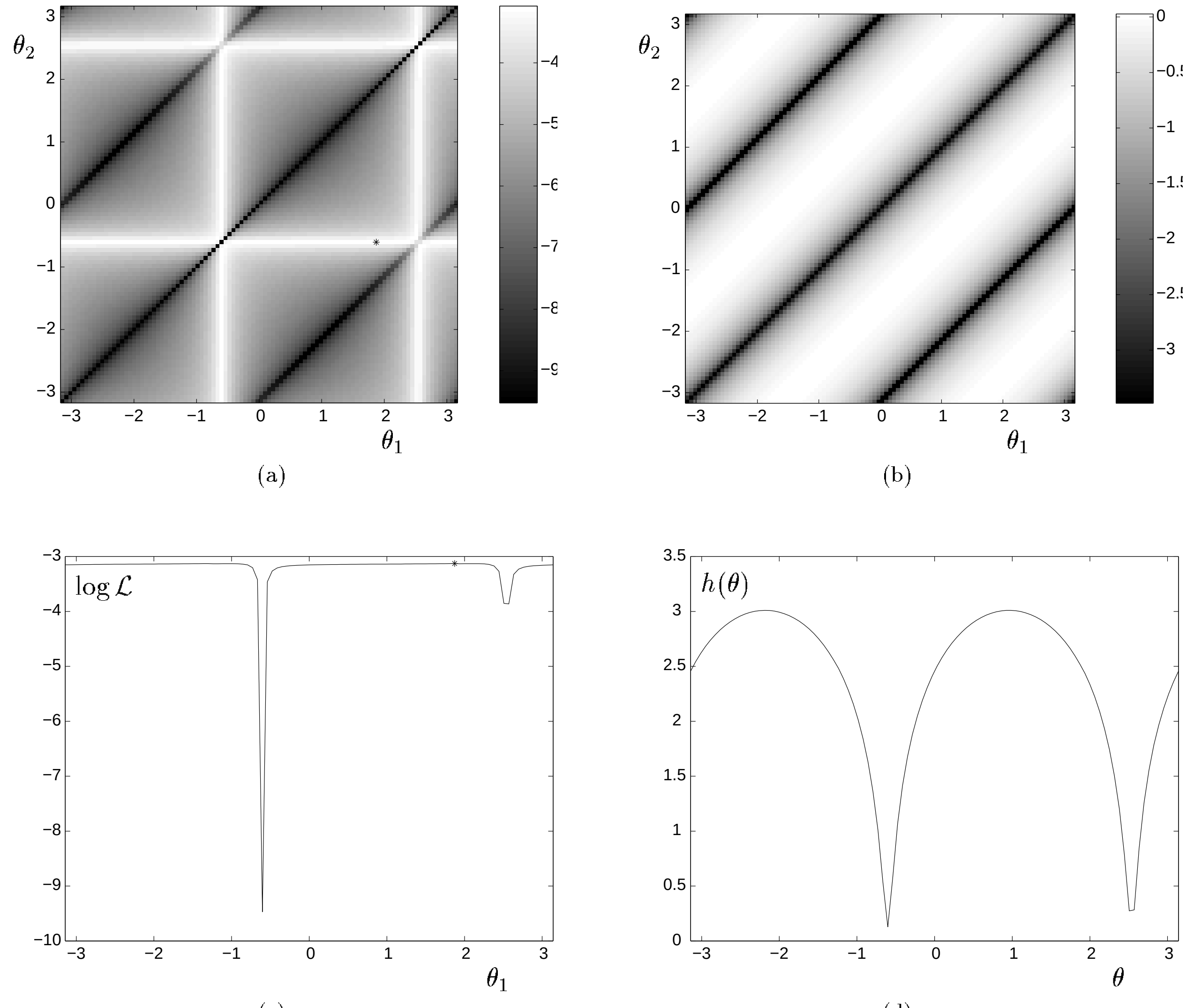
Thus the mutual information is a constant, $H(\mathbf{x})$, minus the log likelihood, so that hills in the log likelihood are valleys in the mutual information.

The mutual information $I(\mathbf{a})$ is invariant under rescaling of $\mathbf{a}$, so if D is a diagonal matrix, $I(D\mathbf{a}) = I(\mathbf{a})$. Since the entropy $H(\mathbf{x})$ is constant, equation 11 shows that the likelihood does not depend upon the scaling of the rows of W. 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,

$$W = \begin{pmatrix} \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \end{pmatrix}, \tag{12}$$

and the likelihood plotted as a function of θ_1 and θ_2. Figure 1 shows the log likelihood for the



(a)

(b)

(c)

(d)

Figure 1: *Likelihood landscape for a mixture of a Laplacian 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:** $\log|\det W(\theta_1, \theta_2)|$. **c:** *Log likelihood along the "ridge"* $\theta_2 = const.$*, passing through the maximum likelihood.* **d:** *The marginal entropy,* $H_k(a_k) = h(\theta_k)$.

mixture of a Gaussian source and a Laplacian ($p(s) \propto e^{-|s|}$) source with $M = \begin{pmatrix} 2 & 1 \\ 3 & 1 \end{pmatrix}$. Also plotted are the constituent components of $\log \mathcal{L}$: namely $\log |\det W|$ and H_k. Here the entropies were calculated by modelling the marginal densities with a generalised exponential (see below), but histogramming the $a(t)$ and numerical quadrature gives very similar, though coarser, results.

Several features deserve comment.

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.

Symmetries. Clearly $\log \mathcal{L}$ is doubly periodic in θ_1 and θ_2. Additional symmetries are conferred by the facts that

1. $\log |\det W|$ is symmetric in the line $\theta_1 = \theta_2$.

2. The likelihood is unchanged under permutation of the coordinates (here θ_1 and θ_2). In this example $H_k(a_k)$ depends only on the angle, and not on the particular k; that is, $H_k(a_k)$ may be written as $h(\theta_k)$ for some function h, which depends, of course, on the data, $\mathbf{x}(t)$. Consequently

$$\log \mathcal{L} = \log |\det W| - \sum_k h(\theta_k). \tag{13}$$

$h(\theta)$ is graphed in Figure 1d for the Gaussian/Laplacian example.

Analogous symmetries are retained in higher dimensional examples.

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 W lies on a ridge with steep sides and a flat top. Figure 1c 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 picture gives a slightly distorted view of the likelihood landscape faced by learning algorithms because they generally work in terms of the full matrix W, rather than with the row-normalised form.

3.1 Mixture of images

As a more realistic example, Figure 2 shows the likelihood landscape for a pair of images mixed with the mixing matrix $M = \begin{pmatrix} 0.7 & 0.3 \\ 0.55 & 0.45 \end{pmatrix}$. Since the distributions of pixel values in the images are certainly not unimodal the marginal entropies were calculated by histogramming the a_k and numerical quadrature.

The overall likelihood is remarkably similar in structure to the Laplacian-Gaussian mixture shown above. The principal difference is that the top of the ridge is now bumpy and gradient-based algorithms may get stuck at a local maximum as illustrated in Figure 3. The imperfect unmixing by the matrix at a local maximum is evident as the ghost of Einstein haunting the house. Unmixing by the maximum likelihood matrix is not quite perfect (the maximum likelihood unmixing matrix is not quite M^{-1}) because the source images are not in fact independent, indeed the correlation matrix $\langle \mathbf{s}\mathbf{s}^{\mathsf{T}} \rangle = \begin{pmatrix} 1.0000 & -0.2354 \\ -0.2354 & 1.0000 \end{pmatrix}$.

4 Choice of squashing function

The algorithm, as outlined above, leaves open the choice of the "squashing functions" g_k, whose function is to map the transformed variables, a_k, into a space in which their marginal densities are

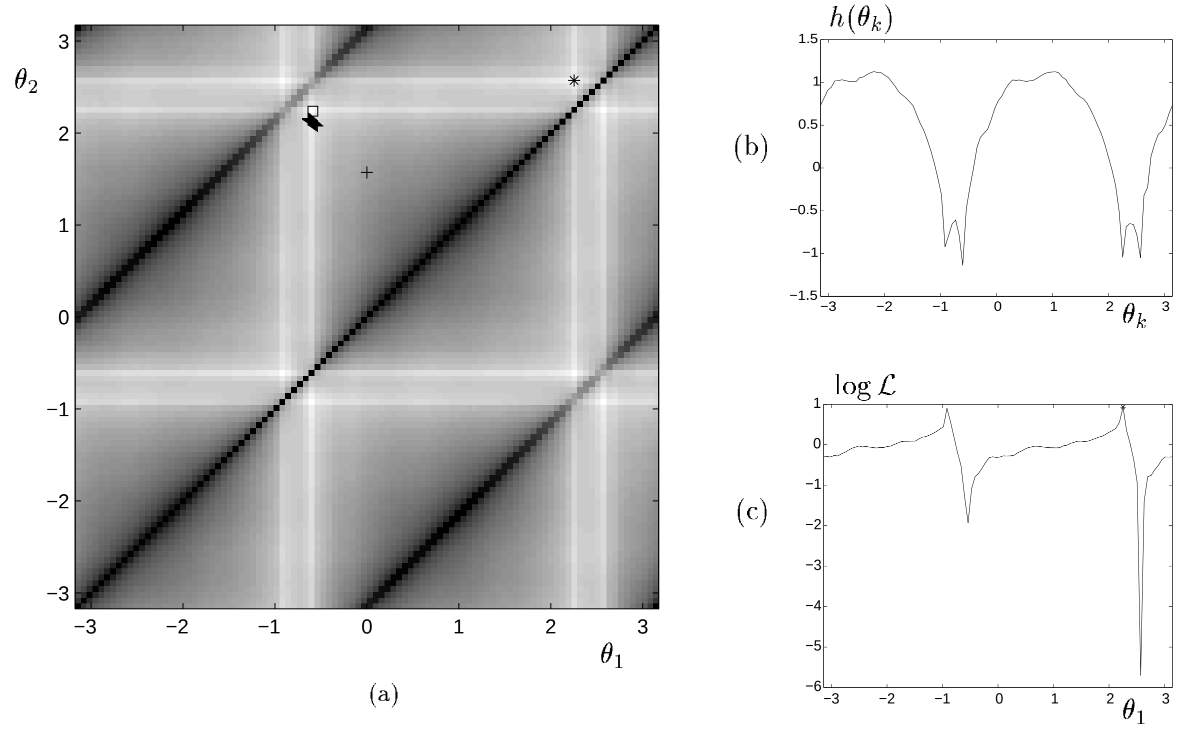


Figure 2: *Likelihood landscape for a mixture of two images.* **a:** *Log likelihood,* $\log L$*, plotted as a function of* θ_1 *and* θ_2*. Dark gray indicates low likelihood matrices and white indicates high likelihood matrices. The maximum likelihood matrix is indicated by the* $*$*; and* M^{-1} *is indicated by the square. Note that symmetry means that an equivalent maximum likelihood matrix is almost coincident with* M^{-1}*. Crosses indicate the trajectory of estimates of* W *by the relative gradient algorithm, starting with* $W_0 = I$*.* **b:** *The marginal entropy,* $H_k(a_k) = h(\theta_k)$*.* **c:** *Log likelihood along the "ridge"* $\theta_2 = const.$*, passing through the maximum likelihood.*

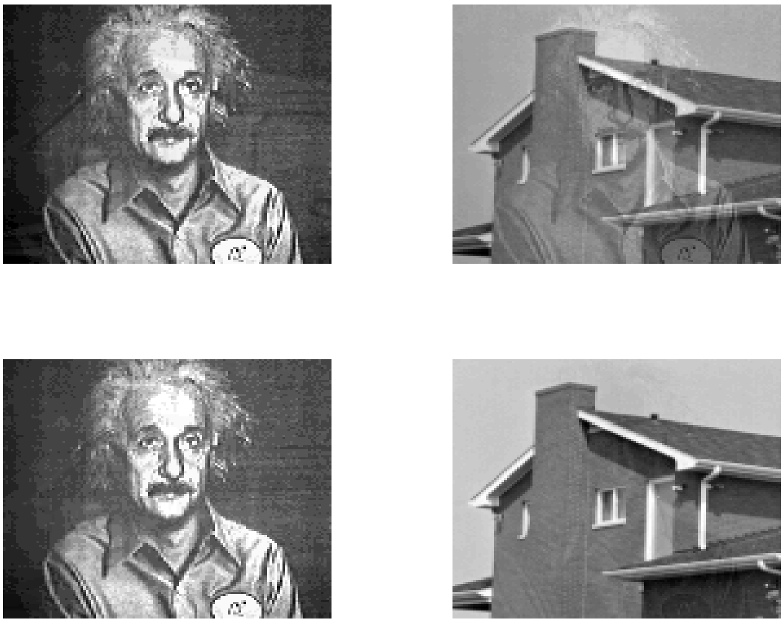


Figure 3: *Unmixing of images by maximum likelihood and sub-optimal matrices.* ***Top row:*** *Images unmixed by a matrix at a local maximum (*$\log \mathcal{L} = -0.0611$*). The trajectory followed by the relative gradient algorithm that arrived at this local maximum is shown in figure 2.* ***Bottom row:*** *Images unmixed by the maximum likelihood matrix (*$\log \mathcal{L} = 0.9252$*).*

uniform. It should be pointed out that what is actually needed are the functions $\phi_k(a_k)$ rather than the g_k themselves.

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

$$g(a) = P(a) = \int_{-\infty}^{a} p(x)dx, \tag{14}$$

where the subscripts k have been dropped for ease of notation. Combining (14) and $\phi(a) = g''/g'$ gives alternative forms for the ideal ϕ:

$$\phi(a) = \frac{\partial p}{\partial P} = \frac{\partial \log p}{\partial a} = \frac{p'(a)}{p(a)} \tag{15}$$

In practice, however, the marginal densities are not known. Bell & Sejnowski recognised this and investigated a number of forms for g and hence ϕ. Current folklore maintains (and MacKay (1996) gives a partial proof) that so long as the marginal densities are heavy tailed (platykurtic) almost any squashing function that is the cumulative density function of a positive kurtosis density will do, and the generalised sigmoidal function and the negative hyperbolic tangent are common choices. Solving (15) with $\phi(a) = -\tanh(a)$ shows that using the hyperbolic tangent is equivalent to assuming $p(a) = 1/(\pi \cosh(a))$.

4.1 Learning the nonlinearity

Multiplication of W by a diagonal matrix D does not change the mutual information between the unmixed variables, that is $I(DW\mathbf{x}) = I(W\mathbf{x})$. It therefore appears that the scaling of the rows of W is irrelevant. However, the mutual information does depend upon D if it is calculated using marginal densities that are not the true source densities. This is precisely the case faced by learning algorithms using an *a priori* fixed marginal density, e.g., $p(a) = 1/(\pi \cosh(a))$ as implied by choosing $\phi = -\tanh$. As figure 4 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 -3.9975, which is to be compared with the true maximum likelihood of -3.1178. 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 M^{-1}, figure 4b shows the log likelihood of DW^* as a function of D. The maximum log likelihood of -3.1881 is achieved for $D = \begin{pmatrix} 1.67 & 0 \\ 0 & 5.094 \end{pmatrix}$.

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. Let $W = D\hat{W}$, where D is diagonal and $\hat{W}$ is row-normalised and let $\mathbf{a} = D\hat{\mathbf{a}} = D\hat{W}\mathbf{x}$, so that $\hat{\mathbf{a}}$ are the unmixed variables produced by the row-normalised unmixing matrix. The nonlinearity is thus $\phi(a_k) = \phi(D_{kk}\hat{a}_k) \equiv \hat{\phi}(\hat{a}_k)$. If $\phi(a_k) = -\tanh(a_k)$, then $\hat{\phi}(\hat{a}_k) = -\tanh(D_{kk}\hat{a}_k)$. The marginal density modelled by $\hat{\phi}$ (for the row-normalised unmixing matrix) is discovered by solving (15) for p, which yields $p(\hat{a}) \propto 1/[\cosh(D_{mm}\hat{a}_k)]^{1/D_{mm}}$. A range of densities is therefore parameterised by D_{kk} : as $D_{kk} \to 0$, $p(\hat{a}_k)$ approximates a Gaussian density, while for large D_{kk} the nonlinearity $\hat{\phi}$ is suited to a Laplacian density. Figure 4d shows the convergence of D as W is located for the Laplacian/Gaussian mixture using the relative gradient algorithm. The component for which $D \to\approx 1.67$ is the unmixed Gaussian component, while the component for which $D \to\approx 5$ is the Laplacian component.

An observation of Cardoso (1997) shows what the correct scaling is. Suppose that W is a scaled version of the (not row-normalised) maximum likelihood unmixing matrix: $W = DM^{-1}$. The the gradient of the likelihood (5) is

$$\frac{\partial \mathcal{L}}{\partial W} = (D^{-1} + \left\langle \Phi(DM^{-1}\mathbf{x})\mathbf{s}^{\mathsf{T}} \right\rangle) M^{\mathsf{T}} \tag{16}$$

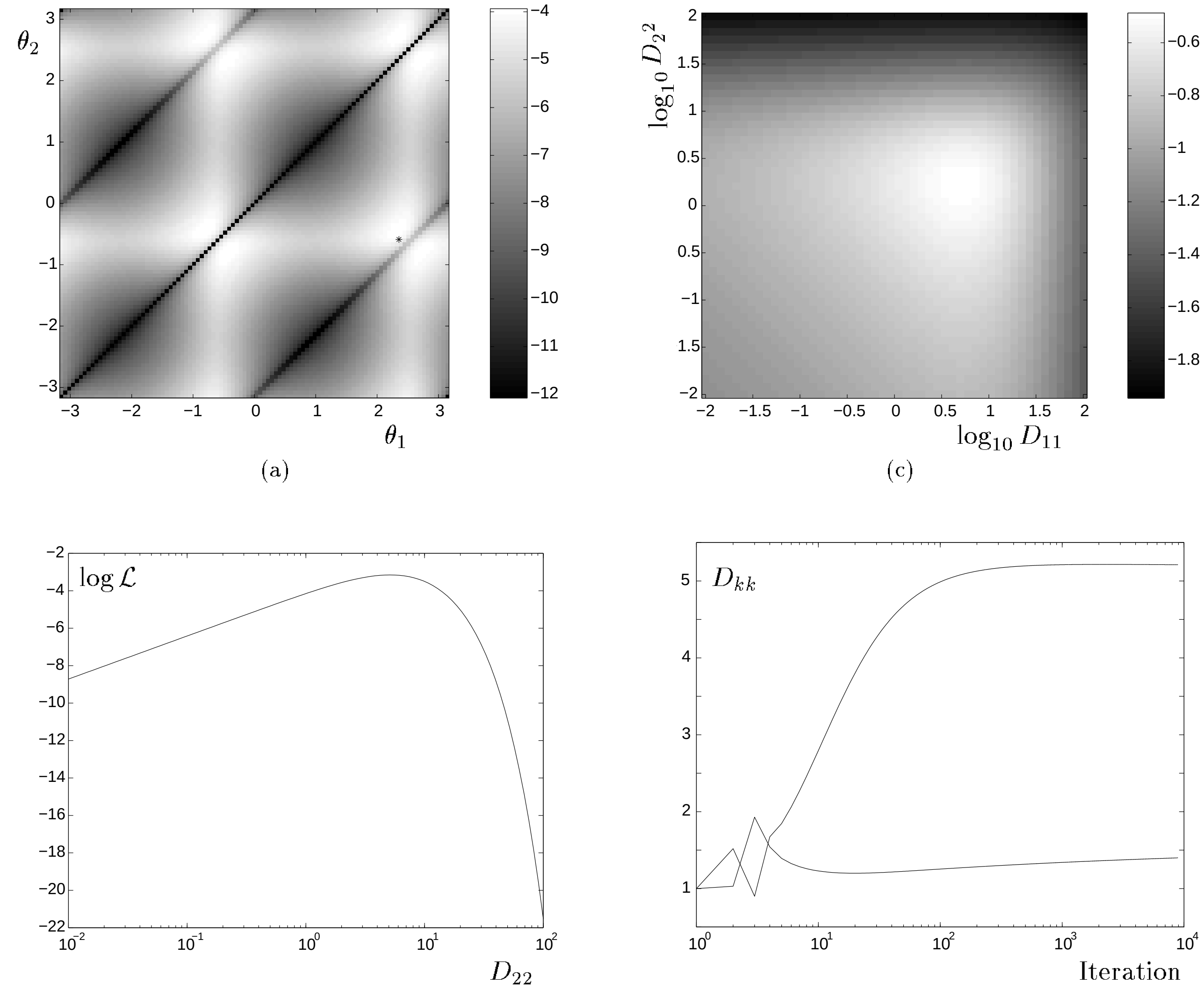
$$= (D^{-1} + \left\langle \Phi(D\mathbf{s})\mathbf{s}^{\mathsf{T}} \right\rangle) M^{\mathsf{T}}, \tag{17}$$

where $\mathbf{\Phi}(\mathbf{a}) = (\phi(a_1), ..., \phi(a_K))^{\mathsf{T}}$. Since the sources are independent and ϕ is a monotone function $\langle \phi(D_i s_i) s_j \rangle = 0$ for $i \neq j$ and the likelihood is maximum for the scaling factors given by $\langle \phi(D_k s_k) s_k D_k \rangle = -1$.

The manner in which nonlinearity is learned can be seen by noting that 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 direct substitution in (15) shows that $\phi(a) = -a$ is the correct nonlinearity for a Gaussian distribution. On the other hand, the unmixed Laplacian component sees the nonlinearity more like a step function, which is appropriate for a Laplacian density.

Densities with tails lighter than Gaussian require a ϕ with positive slope at the origin 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, covariant and BFGS variations of the ICA algorithm all fail to separate a mixture of a uniform source, a Gaussian source and a Laplacian source.

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 "single" nonlinearity, and their inability to unmix light-tailed sources (such as uniform densities).



(a)

(c)

(b)

(d)

Figure 4: *Likelihood for a Laplacian-Gaussian mixture, assuming $1/\cosh$ sources.* **a:** *Normalised log likelihood plotted in the space of two-dimensional, row-normalised unmixing matrices. 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_{kk}, the gray scale is $-\log_{10}(|\log L|)$.* **c:** *Section, $D_{11} = const.$, through the maximum in (a).* **d:** *Convergence of the diagonal elements of D, as W is found by the relative gradient algorithm.*

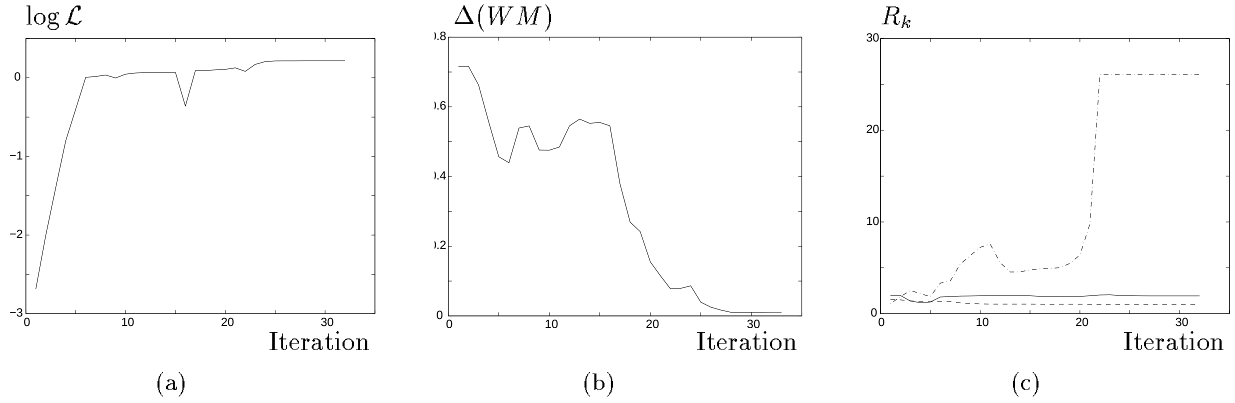


Figure 5: *Separating a mixture of uniform, Gaussian and Laplacian sources.* **a:** *Likelihood* $\log \mathcal{L}$ *of the unmixing matrix plotted against iteration.* **b:** *Fidelity of the unmixing* $\Delta(WM)$ *plotted against iteration.* **c:** *Estimates of* R_k *for the unmixed variables. R describes the power of the generalised exponential: the two lower curves converge to approximately 1 and 2, describing the separated Laplacian and Gaussian components, while the upper curve (limited to 25 for numerical reasons) describes the unmixed uniform source.*

4.2 Generalised Exponentials

By refining the estimate of the marginal densities 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, and especially light-tailed, densities. An alternative approach advanced by (Lee et al. 1998) is to switch (according to the kurtosis of the estimated source) between fixed $-\tanh(\cdot)$ and $+\tanh(\cdot)$ nonlinearities. We have investigated a number of methods of estimating $\phi(a)$ from the T instances of $a(t)$, $t = 1...T$. Briefly, we find that non-parametric methods using the cumulative density or kernel density estimators (Wand and Jones 1995) are too noisy to permit the differentiation required to obtain $\phi = p'/p$.

MacKay (1996) has suggested generalising the usual $\phi(a) = -\tanh(a)$ to use a gain β; that is $\phi(a) = -\tanh(\beta a)$. As discussed in §4.1, scaling the rows of W effectively incorporates a gain into the nonlinearity and permits it to model a range of heavy-tailed densities. To provide a little more flexibility than the hyperbolic tangent with gain, we have used the generalised exponential distribution:

$$p(a|\beta, R) = \frac{R\beta^{1/R}}{2\Gamma(1/R)} \exp\{-\beta|a|^R\}. \tag{18}$$

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 $R = 2$, Laplacian when $R = 1$, and the uniform distribution is approximated in the limit $R \to \infty$. This parametric model, like the hyperbolic tangent, assumes that the marginal densities are unimodal and symmetric about the mean.

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_k at any, and perhaps every, stage of learning. Formulae for maximum likelihood estimators of β and R are given in the Appendix.

4.2.1 Example

We have implemented an adaptive ICA algorithm using the generalised exponential to model the marginal densities. Schemes based on the relative gradient algorithm and the BFGS method have been used, but the quasi-Newton scheme is much more efficient and we discuss that here.

The BFGS scheme minimises $-\log \mathcal{L}$ (see equation 7). At each stage of the minimisation the parameters R_k and β_k, describing the distribution of the kth unmixed variable, were calculated. With these on hand $-\log \mathcal{L}$ can be calculated from the marginal entropies (8) and the gradient found from

$$-\frac{\partial \log \mathcal{L}}{\partial W} = -W^{-\mathsf{T}} - \left\langle \mathbf{z}\mathbf{x}^{\mathsf{T}} \right\rangle, \tag{19}$$

where $z_k = \phi(a_k|\beta_k, R_k)$ is evaluated using the generalised exponential. Note that (19) assumes that R and β are fixed and independent of W, though in fact they depend on W_{ij} because they are evaluated from $a_k(t) = \sum_m W_{km}x_m(t)$. In practice, this leads to small errors in the gradient (largest at the beginning of the optimisation, before R and β have reached their final values), to which the quasi-Newton scheme is tolerant.

Two measures were used to assess the scheme's performance: first, the log likelihood (7) was calculated; the second measures how well W approximates M^{-1}. Recall that changes of scale in the a_k and permutation of the order of the unmixed variables do not affect the mutual information, so rather than $WM = I$ we expect $WM = PD$ 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, diag(A). Consequently the error in W may be assessed by

$$\Delta(MW) = \Delta(WM) = \min_P \frac{||WMP - \text{diag}(WMP)||}{||WM||}, \tag{20}$$

where the minimum is taken over all permutation matrices, P. Of course, when the sources are independent $\Delta(WM)$ should be zero, though when they are not independent the maximum likelihood unmixing matrix may not correspond to $\Delta(WM) = 0$.

Figure 5 shows the progress of the scheme in separating a Laplacian source, $s_1(t)$, a uniformly distributed source, $s_2(t)$, and a Gaussian source, $s_3(t)$, mixed with

$$M = \begin{pmatrix} 0.2519 & 0.0513 & 0.0771 \\ 0.5174 & 0.6309 & 0.4572 \\ 0.1225 & 0.6074 & 0.4971 \end{pmatrix}. \tag{21}$$

There were $T = 1000$ observations. The log likelihood and $\Delta(WM)$ show that the generalised exponential adaptive algorithm (unlike the $\phi = -\tanh$) succeeds in separating the sources.

5 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:

$$\langle a_n a_k \rangle \approx \frac{1}{T}\sum_{t=1}^{T} a_k(t)a_n(t) = \frac{1}{T}(a_k, a_n)_t = \delta_{mn}d_n^2, \tag{22}$$

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 may write

$$AA^T = D^2 \tag{23}$$

where A is the matrix whose kth row is $a_k(t)$ 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 uncorrelated.

Equation (23) comprises $K(K-1)/2$ relations which must be satisfied if W is to be a decorrelating matrix. (There are only $K(K-1)/2$ relations rather than K^2 because (1) AA^T is symmetric, so demanding that $[D^2]_{ij} = 0$ $(i \neq j)$ is equivalent to requiring that $[D^2]_{ji} = 0$; and (2) the diagonal elements of D are not specified: we are only demanding that cross-correlations are zero.)

Clearly there are many decorrelating matrices, of which the ICA unmixing matrix is just one, and we mention a few others below. The decorrelating matrices comprise an $K(K+1)/2$-dimensional manifold in the KN-dimensional space of possible unmixing matrices, and we may seek the ICA unmixing matrix on this manifold.

If W is a decorrelating matrix, we have

$$AA^{\mathsf{T}} = WXX^{\mathsf{T}}W^{\mathsf{T}} = D^2 \tag{24}$$

and if none of the rows of A is identically zero,

$$D^{-1}WXX^{\mathsf{T}}W^{\mathsf{T}}D^{-1} = I_K \tag{25}$$

Now, if $Q \in \mathbb{R}^{K\times K}$ is a real orthogonal matrix and $\hat{D}$ another diagonal matrix,

$$\hat{D}QD^{-1}WXX^{\mathsf{T}}W^{\mathsf{T}}D^{-1}Q^{\mathsf{T}}\hat{D} = \hat{D}^2 \tag{26}$$

so $\hat{D}QD^{-1}W$ is also a decorrelating matrix.

Note that 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

$$X = U\Sigma V^{\mathsf{T}} \tag{27}$$

be a singular value decomposition of the data matrix $X = [\mathbf{x}(1), \mathbf{x}(2), ...\mathbf{x}(T)]$; $U \in \mathbb{R}^{K\times K}$ and $V \in \mathbb{R}^{T\times T}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{K\times T}$ 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

$$W = DQW_0 = DQ\Sigma^{-1}U^{\mathsf{T}}. \tag{28}$$

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 $1/T$.

The manifold of decorrelating matrices is seen to be $K(K+1)/2$-dimensional: it is the Cartesian product of the K-dimensional manifold $\mathcal{D}$ of scaling matrices and the $(K-1)K/2$-dimensional manifold of orthogonal matrices $\mathcal{Q}$. Explicit coordinates on $\mathcal{Q}$ are given by

$$Q = e^S \tag{29}$$

where S is an anti-symmetric matrix $(S^{\mathsf{T}} = -S)$. Each of the above-diagonal elements of S may be used as a coordinate for $\mathcal{Q}$.

Particularly well-known decorrelating matrices (Bell and Sejnowski 1997; Penev and Atick 1996) are:

PCA $Q = I$ and $D = \Sigma$. In this case W simply produces the principal components representation. The columns of U form a new orthogonal basis for the data and the mean squared projection onto the kth coordinates is σ_k^2/T. The PCA solution holds a special position among decorrelating transforms because it simultaneously finds orthonormal bases for both the row (V) and column (U) spaces of X. Viewed in these bases, the data is decomposed into a sum of products which are *linearly* decorrelated in both space and time. The demand by ICA of in-

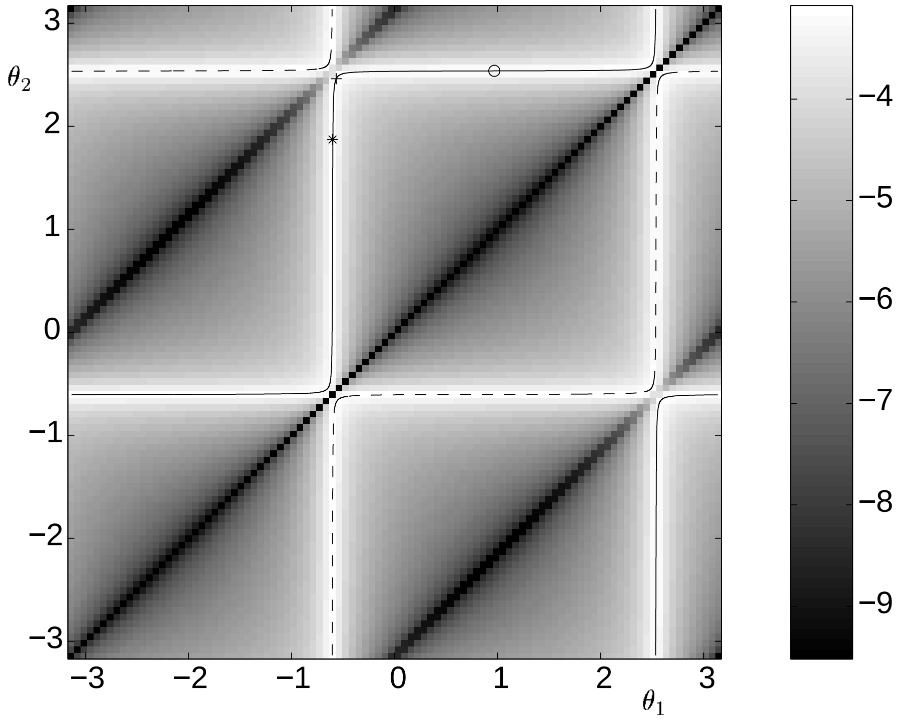


Figure 6: *The manifold of (row-normalised) decorrelating matrices plotted on the likelihood function for the mixture of Gaussian and Laplacian sources. Leaves of the manifold corresponding to* $\det Q = \pm 1$ *are as solid and dashed lines respectively. The symbols mark the locations of decorrelating matrices corresponding to PCA (*$\circ$*), ZCA (*$+$*) and ICA (*$*$*).*

	$\Delta(WM)$	$\log \mathcal{L}$
PCA	0.3599	-3.1385
ZCA	0.3198	-3.1502
ICA	0.1073	-3.1210

Table 1: *PCA, ZCA and ICA errors in inverting the mixing matrix (equation 20) and* $\log \mathcal{L}$*, the log likelihood of the unmixing matrix (equation 7)*

dependence in time, rather than just linear decorrelation, can only be achieved by sacrificing orthogonality between the elements of the spatial basis, i.e., the rows of W.

ZCA $Q = U$ and $D = TI$. Bell and Sejnowski (1997) call decorrelation with the symmetrical decorrelating matrix, $W^{\mathsf{T}} = W$, the zero-phase components analysis. Unlike PCA, whose basis functions are global, ZCA basis functions are local and whiten each row of WX so that it has unit variance.

ICA In the sense that it is neither local or global, ICA is intermediate between ZCA and PCA. No general analytic form for Q and D can be given, and the optimum Q must be sought by minimising equation (2) (the value of D is immaterial since $I(D\mathbf{a}) = I(\mathbf{a})$). It is important to note that if the optimal W is found within the space of decorrelating matrices it may not minimise the mutual information, which also depends on higher moments, as well as some other W which does not yield an exact linear decorrelation.

When $K = 2$ the manifold of decorrelating matrices is 3-dimensional, since two parameters are required to specify D and a single angle parameterises Q. Since multiplication by a diagonal matrix

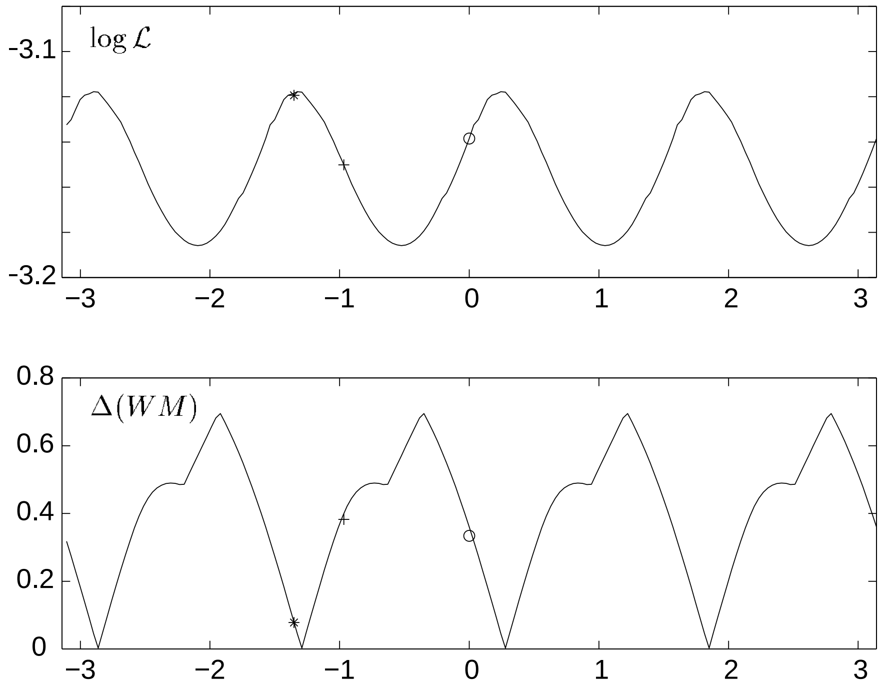


Figure 7: *Likelihood and errors in inverting the mixing matrix as the* $\det Q = +1$ *leaf of the decorrelating manifold is traversed. The symbols mark the locations of decorrelating matrices corresponding to PCA (○), ZCA (+) and ICA (∗).*

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 – this has been done for the Gaussian/Laplacian example in figure 6. Also plotted on the figure are the locations of the orthogonal matrices corresponding to the PCA and ZCA decorrelating matrices. The manifold consists of two non-intersecting leaves, corresponding to $\det Q = \pm 1$, which run close to the tops of the ridges in the likelihood. Figure 7 shows the likelihood and errors in inverting the mixing matrix as the $\det Q = +1$ leaf is traversed. Table 5 gives the likelihoods and errors in inverting the mixing matrix for the PCA, ZCA and ICA.

In general the decorrelating manifold does not exactly coincide with the top of the likelihood ridges, though numerical computations suggest that it is usually close. When the sources are Gaussians the decorrelating manifold and the likelihood ridge are identical, but all decorrelating matrices (PCA, ICA, ZCA, etc) have the same (maximum) likelihood and the top of the ridge is flat.

This characterisation of the decorrelating matrices does not assume that the number of observation sequences N is equal to the assumed number of sources K, and it is interesting to observe that if $K < N$ the reduction in dimension from $\mathbf{x}$ to $\mathbf{a}$ is accomplished by $\Sigma^{-1} U_K^\mathsf{T} \in \mathbb{R}^{K \times N}$, where U_K consists of the first K columns of U. This is the transformation onto the decorrelating manifold and is the same irrespective of whether the final result is PCA, ZCA or ICA. It should be noted that the transformation onto the decorrelating manifold is a projection, and data represented by the low power (high index) principal components is discarded by projecting onto the manifold. It might therefore appear that the projection could erroneously discard low variance principal components that nonetheless correspond to (low power) independent components. Proper selection of the model order, K, involves deciding how many linearly mixed components can be distinguished from noise, which can be done on the basis of the (linear) covariance matrix (Everson and Roberts 1998). The number of relevant independent components can therefore be determined before projecting onto the decorrelating manifold and so any directions which are discarded should correspond to noise. We emphasise that with sufficient data, the maximum likelihood unmixing matrix lies on the decorrelating manifold and will be located by algorithms confined to the manifold.

An important characteristic of the PCA basis (the columns of U) is that it minimises reconstruction error. A vector $\mathbf{x}$ is approximated by $\tilde{\mathbf{x}}$ projecting $\mathbf{x}$ onto the first K columns of U

$$\tilde{\mathbf{x}} = U_K U_K^{\mathsf{T}} \mathbf{x}, \tag{30}$$

where U_K denotes the first K columns of U. The mean squared approximation error

$$\epsilon_K^{(PCA)} = \left\langle \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 \right\rangle \tag{31}$$

is minimised amongst all linear bases by the PCA basis for any K. Indeed the PCA decomposition is easily derived by minimising this error functional with the additional constraint that the columns of U_K are orthonormal. It is a surprising fact that this minimum reconstruction error property is shared by all the decorrelating matrices, and in particular by the (non-orthogonal) ICA basis which is formed by the rows of W. This is easily seen by noting that the approximation in terms of K ICA basis functions is

$$\tilde{\mathbf{x}} = W^{\dagger} W \mathbf{x}, \tag{32}$$

where the pseudo-inverse of W is

$$W^{\dagger} = U_K \Sigma Q^{\mathsf{T}} D^{-1}. \tag{33}$$

The approximation error is therefore

$$\begin{aligned} \epsilon_K^{(ICA)} &= \left\langle \|\mathbf{x} - W^{\dagger} W \mathbf{x}\|^2 \right\rangle & (34) \\ &= \left\langle \|\mathbf{x} - (U_K \Sigma Q^{\mathsf{T}} D^{-1})(D Q \Sigma^{-1} U_K^{\mathsf{T}})\mathbf{x}\|^2 \right\rangle & (35) \\ &= \left\langle \|\mathbf{x} - U_K U_K^{\mathsf{T}} \mathbf{x}\|^2 \right\rangle & (36) \\ &= \epsilon_K^{(PCA)} & (37) \end{aligned}$$

Penev and Atick (1996) have also noticed this property in connection with local feature analysis.

5.1 Algorithms

Here we examine algorithms which seek to minimise the mutual information using an unmixing matrix W which is drawn from the class of linearly decorrelating matrices $\mathcal{D} \times \mathcal{Q}$ and therefore has the form of equation (23). Since $I(D\mathbf{a}) = I(\mathbf{a})$ for any diagonal D, at first sight it appears that we may choose $D = I_K$. However, as the discussion in §4.1 points out, the elements of D serve as adjustable parameters tuning a model marginal density to the densities generated by the a_k. 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 of decorrelating matrices.

A straightforward method is to use one of the popular minimisation schemes (Bell and Sejnowski 1995; Amari et al. 1996; MacKay 1996) to take one or several steps towards the minimum and then to replace the current estimate of W with the nearest decorrelating matrix.

Finding the nearest decorrelating matrix requires finding the D and Q that minimise

$$\|W - DQW_0\|^2 \tag{38}$$

When $D = I_K$ (i.e., when an adaptive ϕ is being used) this is a simple case of the matrix Procrustes problem (Golub and Loan 1983; Horn and Johnson 1985). The minimising Q is the orthogonal polar factor of WW_0^T. That is, if $WW_0^\mathsf{T} = YSZ^\mathsf{T}$ is a SVD of W, then $Q = YZ^\mathsf{T}$. When $D \neq I_K$, (38) must be minimised numerically to find D and Q (Everson 1998).

This scheme permits estimates of W to leave the decorrelating manifold, because derivatives are taken in the full space of $K \times K$ matrices. It might be anticipated that a more efficient algorithm would be one which constrains W to remain on the manifold of decorrelating matrices, and we now examine algorithms which enforce this constraint.

5.1.1 Optimising on the decorrelating manifold

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 (29); however, with fixed nonlinearities it is essential to allow D to vary. In this case the optimum is sought in terms of the $(K-1)K/2$ above-diagonal elements of S and the K elements of D.

Optimisation schemes perform best if the independent variables have approximately equal magnitude. To ensure the correct scaling we write

$$D = \Sigma\tilde{D} \tag{39}$$

and optimise the likelihood with respect to the elements of $\tilde{D}$ (which are $O(1)$) along with S_{pq}. An initial pre-processing step is to transform the data into the whitened PCA coordinates; thus

$$\hat{X} = \Sigma^{-1}U^\mathsf{T}X. \tag{40}$$

The normalised log likelihood is

$$\log\mathcal{L}(\hat{X}|\tilde{D}, Q) = \log|\det\Sigma\tilde{D}Q| + \left\langle \sum_k \log p_k(a_k(t)) \right\rangle. \tag{41}$$

The gradient of $\log\mathcal{L}$ with respect to $\tilde{D}$ is

$$\frac{\partial\log\mathcal{L}}{\partial\tilde{D}_i} = \tilde{D}_i^{-1} + \left\langle \phi_i(a_i)\sigma_i\sum_j Q_{ij}\hat{\mathbf{x}}_j(t) \right\rangle \tag{42}$$

and

$$\frac{\partial\log\mathcal{L}}{\partial Q_{ij}} = \langle\phi_i(a_i)D_i\hat{\mathbf{x}}_j(t)\rangle = Z_{ij} \tag{43}$$

Using the parameterisation (29), equation (43), and

$$2\frac{\partial Q_{ij}}{\partial S_{pq}} = Q_{ip}\delta_{qj} - Q_{iq}\delta_{pj} + Q_{qj}\delta_{pi} - Q_{pj}\delta_{qi}, \tag{44}$$

the gradient of $\log\mathcal{L}$ with respect to the above-diagonal elements S_{pq} $(p < q \leq K)$ of the anti-symmetric matrix is given by:

$$\frac{\partial\log\mathcal{L}}{\partial S_{pq}} = -Q_{mp}Z_{mq} + Q_{mq}Z_{mp} - Q_{qm}Z_{pm} + Q_{pm}Z_{qm} \tag{45}$$

(summation on repeated indices). With the gradient on hand, gradient descent or, more efficiently, quasi-Newton schemes may be used.

When the nonlinearity is adapted to the unmixed marginal densities, one simply sets $\tilde{D} = I_K$ in (41) and (43), and the optimisation is conducted on in the $K(K-1)/2$-dimensional manifold $\mathcal{Q}$.

Clearly, a natural starting guess for W is the PCA unmixing matrix given by $S = 0$, $\tilde{D} = I_K$.

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 $K(K-1)/2$-dimensional (or if fixed nonlinearities are used, $K(K+1)/2$-dimensional) space of decorrelating matrices rather than in the full K^2-dimensional space of order K matrices. For large problems and especially if the Hessian matrix is being used this provides considerable computational savings in locating the optimum matrix.

3. The scaling matrix D, which does not provide any additional information, is effectively 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. Of course with sufficient data, a necessary condition for independence is linear decorrelation and the optimum ICA matrix will lie on the decorrelating manifold. Nonetheless, the decorrelating matrix is generally very close to the optimum matrix and provides a good starting point from which to find it.

6 Rogues Gallery

The hypothesis that human faces are composed from an admixture of a small number of canonical or basis faces was first examined by Sirovich and Kirby (1987, 1990). It has inspired much research in the pattern recognition (Atick et al. 1995) and psychological (O'Toole et al. 1991; O'Toole et al. 1991) communities. Much of this work has focused on *eigenfaces*, which are the principal components of an ensemble of faces and are therefore mutually orthogonal. As an application of our adaptive ICA algorithm on the decorrelating manifold we have computed the independent components for an ensemble faces – dubbed the "Rogues Gallery" by Sirovich and Sirovich (1989). The model we have in mind is that a particular face, $\mathbf{x}$, is an admixture of K basis functions, the coefficients of the admixture being drawn from K independent sources $\mathbf{s}$. If the ensemble of faces is subjected to ICA the rows of the unmixing matrix are estimates of the basis functions, which (unlike the eigenfaces) need not be orthogonal.

There were 143 clean-shaven, male Caucasian faces in the original ensemble, but the ensemble was augmented by the reflection of each face in its midline to make 286 faces in all (Kirby and Sirovich 1990). The mean face was subtracted from each face of the ensemble before ICA. Independent components were estimated using a quasi-Newton scheme on the decorrelating manifold with generalised exponential modelling of the source densities. Since the likelihood surface has many local maxima, the optimisation was run repeatedly ($K+1$ times for K assumed sources) each run starting from a different (randomly chosen) initial decorrelating matrix. One of the initial conditions always included the PCA unmixing matrix and it was found that this initial matrix always lead to the ICA unmixing matrix with the highest likelihood. It was also always the case that the ICA unmixing matrix had a higher likelihood than the PCA unmixing matrix. We remark that an adaptive optimisation scheme using our generalised exponential approach was essential: several of the unmixed variables had densities with tails lighter than Gaussian.

Principal components are naturally ordered by the associated singular value, σ_k, which measures standard deviation of projection of the data onto the kth principal component: $\sigma_k^2 = \left\langle (\mathbf{u}_k^\mathsf{T}\mathbf{x})^2 \right\rangle$. In an analogous manner we may order the ICA basis vectors by the scaling matrix D. Hereafter we assume that the unmixing matrix is row-normalised, and denote the ICA basis vectors by $\mathbf{w}_k$.

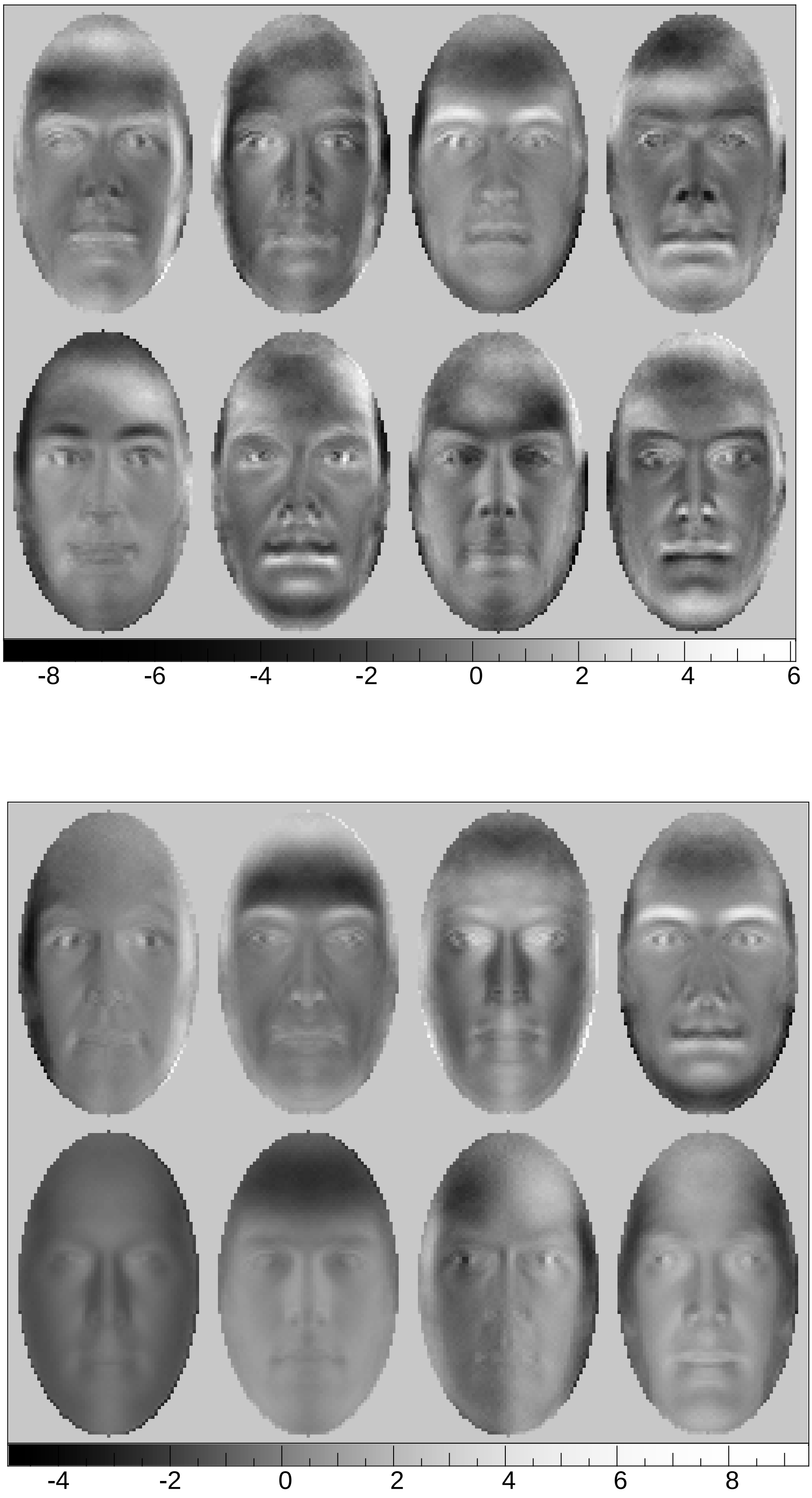


Figure 8: *Independent basis functions and principal components of faces.* **a:** *The first 8 independent component basis faces,* $\mathbf{w}_k$ *from an* $K = 20$ *ICA of the faces ensemble.* **b:** *The first 8 principal components,* $\mathbf{u}_k$*, from the same ensemble.*

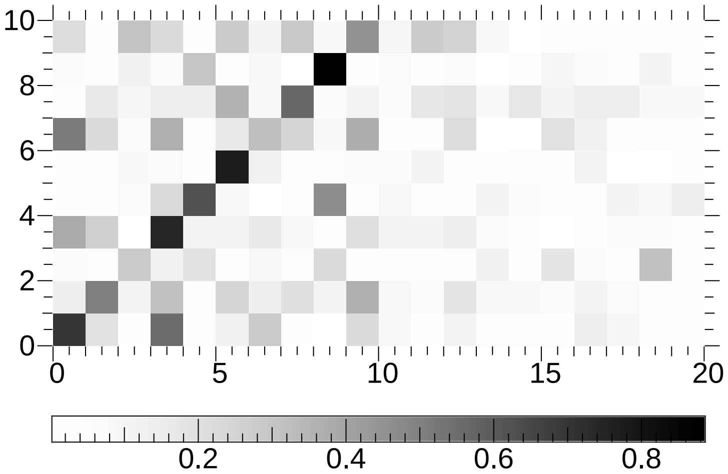


Figure 9: *The matrix* $|W^{(20)}W^{(10)^\mathsf{T}}|$ *showing the inner products between the independent components basis vectors for* $K = 10$ *and* $K = 20$ *assumed sources.*

Then $D_k^2 = \langle (\mathbf{w}_k^\mathsf{T}\mathbf{x})^2 \rangle$ measures the mean squared projection of the data onto the kth normalised ICA basis vector. We then order the $\mathbf{w}_k$ according to D_k, starting with the largest.

Figure 8 shows the first eight ICA basis vectors together with the eight principal components from the same dataset. The $\mathbf{w}_k$ were calculated with $K = 20$, i.e., it was assumed that there are 20 significant ICA basis vectors. Although independent components have to be recalculated for each K, we have found a fair degree of concurrence between the basis vectors calculated for different K. For example, figure 9a shows the matrix of inner products between the basis vectors calculated with $K = 10$ and $K = 20$; i.e., $|W^{(20)}W^{(10)^\mathsf{T}}|$. The large elements on or close to the diagonal and the small elements in the lower half of the matrix indicate that the basis vectors retain their identities as the assumed number of sources increases.

As the figure shows, the ICA basis vectors have more locally concentrated power than the principal components. Power is concentrated around sharp gradients or edges in the images, in concurrence with Bell and Sejnowski's (1997) observation that the ICA basis functions are edge detectors. As Bartlett, Lades, and Sejnowski (1998) have found, this property may make the ICA basis vectors useful feature detectors since the edges are literal features! We also note that, unlike the $\mathbf{u}_k$, the $\mathbf{w}_k$ are not forced to be symmetric or anti-symmetric in the vertical midline. There is a tendency for the midline to be a line of symmetry and we anticipate that with a sufficiently large ensemble, the w_k would acquire exact symmetry.

As the assumed number of sources is increased the lower-powered independent component basis vectors approach the principal components. This is illustrated in figure 10, which shows the matrix of inner products between the $\mathbf{w}_k$ from a $K = 20$ source model and the first 20 principal components. For k greater than about 12 the angle between the principal components $\mathbf{u}_k$ and $\mathbf{w}_k$ is small. Bartlett, Lades, and Sejnowski (1998) have calculated independent component basis vectors for a different ensemble of faces, and do not report this tendency for the independent components to resemble the principal components. However, their unmixing matrix is not guaranteed to be decorrelating. It is possible that our algorithm is getting stuck at local likelihood maxima close to the PCA unmixing matrix, however, initialising the optimisation at randomly chosen positions on the decorrelating manifold failed to find W with a greater likelihood than those presented here.

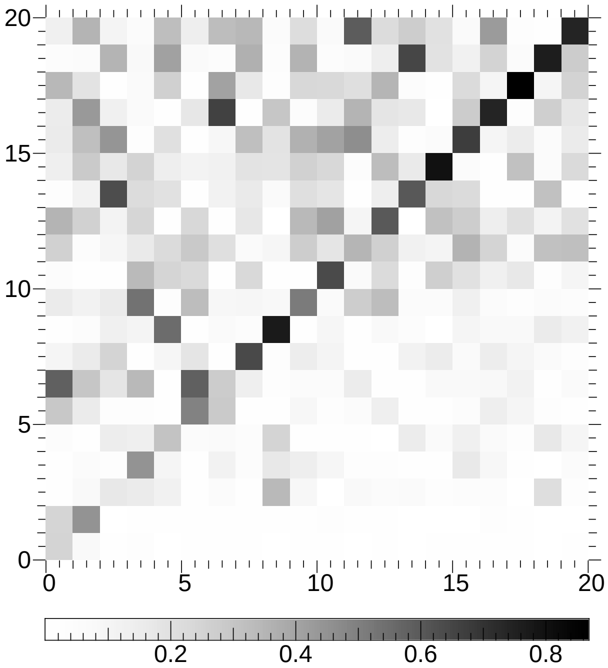


Figure 10: *The matrix* $|W^{(20)}U_{20}^{\mathsf{T}}|$ *showing the inner products between the independent components basis vectors* ($K = 20$) *and the first 20 principal components.*

We suspect that the proximity of the later ICA basis vectors to the principal components is due to the fact that the independent components are constrained to lie on the decorrelating manifold, the noisy condition (and relatively small size ($T = 286$)) of our ensemble, factors which also prevent meaningful estimates of the true number of independent sources.

7 Summary and Conclusion

We have used the likelihood landscape as a numerical tool to better understand independent components analysis and the manner in which gradient-based algorithms work. In particular we have tried to make plain the role that scaling of the unmixing matrix plays in adapting a "static" nonlinearity to the nonlinearities required to unmix sources with differing marginal densities. To cope with light-tailed densities we have demonstrated a scheme that uses generalised exponential functions to model the marginal densities. Despite the success of this scheme in separating a mixture of Gaussian, Laplacian and uniform sources, additional work is required to model sources which are heavily skewed or which have multi-modal densities.

Numerical experiments show that the manifold of decorrelating matrices lies close to the ridges of high-likelihood unmixing matrices in the space of all unmixing matrices. We have shown how to find the optimum ICA matrix on the manifold of decorrelating matrices and we have used the algorithm to find independent component basis vectors for a rogues gallery. Seeking the ICA unmixing matrix on the decorrelating manifold naturally incorporates the case in which there are more observations, N, than sources, K. Selection of the correct number of sources, especially with few data, can be difficult particularly as ICA does not model observational noise (but see Attias (1998)), however the model order may be selected *before* projection onto the decorrelating manifold. In common with other authors, we note that real cocktail party problem – separating many voices from few observations – remains to be solved (for machines).

Finally, independent components analysis depends on minimising the mutual information between the unmixed variables, which is identical to minimising the Kullback-Leibler divergence between the between the joint density $p(\mathbf{a})$ and the product of the marginal densities $\prod_k p_k(a_k)$. The Kullback-Leibler divergence is one of many measures of disparity between densities (see, for example, Basseville (1989)) and one might well consider using a different one. Particularly attractive is the Hellinger distance which is a metric and not just a divergence. When an unmixing matrix which makes the mutual information zero can be found, the Hellinger distance is also zero. However, when some residual dependence between the unmixed variables remains these various divergences will vary in their estimate of the best unmixing matrix.

Acknowledgements

We are grateful for discussions with Will Penny and Iead Rezek, and we thank Michael Kirby and Larry Sirovich for supplying the rogues gallery ensemble. Part of this research was supported by funding from British Aerospace to whom the authors are most grateful. We also acknowledge helpful comments given by two anonymous referees.

A Appendix

Here we give formulae for estimating the generalised exponential (18) parameters β and R from T observations $a(t)$. The normalised log likelihood is

$$L = \log R + \frac{1}{R}\log\beta - \log 2 - \log\Gamma(1/R) - \beta{\sum}'|a_t|^R \tag{46}$$

where $\sum' \equiv T^{-1}\sum_{t=1}^{T}$. The derivative of the L with respect to β is

$$\frac{\partial L}{\partial\beta} = \frac{1}{R\beta} - {\sum}'|a_t|^R. \tag{47}$$

Setting this equal to zero gives β in terms of R and we can solve the one-dimensional problem $\frac{dL}{dR} = 0$ to find the maximum likelihood parameters.

$$\frac{dL}{dR} = \frac{\partial L}{\partial R} + \frac{\partial L}{\partial\beta}\frac{\partial\beta}{\partial R} \tag{48}$$

but the second term is zero if the solution is sought along the curve defined by $\frac{\partial L}{\partial\beta} = 0$. It is straight-forward to find

$$\frac{\partial L}{\partial R} = \frac{1}{R} - \frac{1}{R^2}\log\beta + \frac{1}{R^2}\psi(1/R) - \beta{\sum}'|a_t|^R\log|a_t| \tag{49}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. Since there is only one finite R for which $\frac{dL}{dR}$ is zero, this is readily and robustly accomplished. We remark that the domain of attraction for a Newton's method is quite small, and Newton's method offers only a slight advantage over straightforward bisection.

References

Amari, S., A. Cichocki, and H. Yang (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, and M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, Cambridge MA, pp. 757–763. MIT Press.

Atick, J., P. Griffin, and A. Redlich (1995). Statistical Approach to Shape from Shading: Reconstruction of 3D Face Surfaces from Single 2D Images. *Neural Computation* (6), 1321–1340.

Attias, H. (1998). Independent factor analysis. *Neural Computation. In press.*

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication.* MIT Press.

Bartlett, M., H. Lades, and T. Sejnowski (1998, January). Independent components representations from face recognition. In *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology: Conference on Human Vision and Electronic Imaging III*, San Jose, California. SPIE. *In press.* Available from `http://www.cnl.salk.edu/~marni`.

Basseville, M. (1989). Distance measures for signal processing and pattern recognition. *Signal Processing 18*, 349–369.

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

Bell, A. and T. Sejnowski (1997). The “Independent Components” of Natural Scenes are Edge Filters. *Vision Research 37*(23), 3327–3338.

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

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

Everson, R. (1998). Orthogonal, but not orthonormal, Procrustes problems. *Advances in Computational Mathematics.* (Submitted). Available from `http://www.ee.ic.ac.uk/research/neural/everson`.

Everson, R. and S. Roberts (1998). Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Trans. Sig. Proc.*. (*Submitted.*) Available from `http://www.ee.ic.ac.uk/research/neural/everson`.

Golub, G. and C. V. Loan (1983). *Matrix Computations.* Oxford: North Oxford Academic.

Horn, R. and C. Johnson (1985). *Matrix Analysis.* Cambridge University Press.

Kirby, M. and L. Sirovich (1990). Application of the Karhunen-Loève Procedure for the Characterization of Human Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(1), 103–108.

Lee, T.-W., M. Girolami, A. Bell, and T. Sejnowski (1998). A Unifying Information-theoretic Framework for Independent Component Analysis. *International Journal on Mathematical and Computer Modeling.* (In press). Available from `http://www.cnl.salk.edu/~tewon/Public/mcm.ps.gz`.

Lee, T.-W., M. Girolami, and T. Sejnowski (1998). Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources. *Neural Computation. In press.* Available from `http://www.cnl.salk.edu/~tewon`.

MacKay, D. (1996, December). Maximum Likelihood and Covariant Algorithms for Independent Component Analysis. Technical report, University of Cambridge. Available from `http://wol.ra.phy.cam.ac.uk/mackay/`.

Makeig, S., , T.-P. Jung, A. Bell, D. Ghahremani, and T. Sejnowski (1997). Transiently Time-locked fMRI Activations revealed by independent components analysis. *Proceedings of the National Academy of Sciences 95*, 803–810.

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

O'Toole, A., H. Abdi, K. Deffenbacher, and J. Bartlett (1991). Classifying Faces by Race and Sex Using an Autoassiciative Memory Trained for Recognition. In K. Hammond and D. Gentner (Eds.), *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 847–851.

O'Toole, A., K. Deffenbacher, H. Abdi, and J. Bartlett (1991). Simulating the "Other-Race Effect" As a Problem in Perceptual Learning. *Connection Science 3*(2), 163–178.

Papoulis, A. (1991). *Probability, Random Variables and Stochastic Processes.* McGraw-Hill.

Penev, P. and J. Atick (1996). Local Feature Analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems 7*(3), 477–500.

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

Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (1992). *Numerical Recipes in C* (2 ed.). Cambridge: Cambridge University Press.

Sirovich, L. and M. Kirby (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America 4A*(3), 519–524.

Sirovich, L. and C. Sirovich (1989). Low dimensional description of complicated phenomena. *Contemporary Mathematics 99*, 277–305.

Wand, M. and M. Jones (1995). *Kernel Smoothing.* Number 60 in Monographs on statistics and applied probability. London: Chapman and Hall.