

Estimating sub- and super-Gaussian densities  
using ICA and exponential power distributions  
with applications to natural images.

Michael S. Lewicki

`lewicki@cnbc.cmu.edu`

Computer Science Department and  
Center for the Neural Basis of Cognition

Carnegie Mellon University

Mellon Institute 115

4400 Fifth Avenue

Pittsburgh, PA 15213

Last modified: Oct 22, 1999

## Abstract

Exponential power distributions (Box and Tiao, 1973) provide a general method for modeling non-Gaussian statistical structure of univariate distributions that have the form  $p(x) \propto \exp(-|x|^q)$ . By inferring  $q$ , a wide class of statistical distributions can be characterized including uniform, Gaussian, Laplacian, and other so-called sub- and super-Gaussian densities. Using this distribution in independent component analysis (ICA), we show that the exponential power distribution can be used to infer the optimal degree of non-Gaussian statistical structure for multivariate densities. We also show that this class of distributions provides a near exact fit to the source distributions of natural images and demonstrate that this leads to better estimated coding efficiency.

## 1 Introduction

In signal processing and pattern classification, the performance of a method is often determined by how well it can model the underlying statistical distribution of the data. One recent example of this is independent component analysis (ICA). The success of ICA on problems such as blind source separation and signal analysis results directly from its ability to model non-Gaussian statistical structure.

ICA assumes that a multidimensional data signal is composed of a linear superposition of independent sources

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

where  $\mathbf{x}$  is the  $L$ -dimensional data signal (also called the mixtures),  $\mathbf{s}$  is the  $L$ -dimensional source vector, and  $A$  is the  $L \times L$  mixing matrix. The objective is to model the statistical structure of the data, which requires inference of both the unknown mixing matrix and the unknown sources. If the source distributions are assumed to be Gaussian, this technique is equivalent to principal component analysis (PCA). PCA assumes the data to be distributed according to a multivariate Gaussian, which implies that the mixing matrix, the so-called “independent components,” are orthogonal. In contrast, ICA places no restrictions on the mixing matrix and assumes that the source distributions are non-Gaussian.

In many applications of ICA, the form of the source distribution (or equivalently the “non-linearity”) is fixed. More recent work has extended these results so that the form of distribution can also be inferred from the data for example, using Gaussian mixtures (Attias, 1998) or mixtures of sub-Gaussian

and super-Gaussian densities (Lee et al., 1999a).

Below we present a simple extension of ICA in which the source densities are modeled using exponential power distributions. These distributions use a single parameter to describe the deviation from the standard normal and can capture a wide range of distributions from uniform to near-delta functions. We show that this class of distributions is particularly good for capturing the statistical structure of natural images and demonstrate that the codes learned for both whitened and unwhitened natural images yield better estimated coding efficiency than previous models.

## 2 Exponential power distributions

The exponential power distribution<sup>1</sup> is used to model distributions that deviate from normality. In its simplest form, this distribution is

$$p(x) \propto \exp\left(-\frac{1}{2}|x|^q\right). \quad (2)$$

By varying the exponent  $q$ , it is possible to describe Gaussian, platykurtic, and leptokurtic distributions. Using  $q = 2/(1 + \beta)$ , Box and Tiao (1973) expressed this distribution in the following general form

$$p(x|\mu, \sigma, \beta) = \frac{\omega(\beta)}{\sigma} \exp\left[-c(\beta) \left|\frac{x - \mu}{\sigma}\right|^{2/(1+\beta)}\right], \quad -\infty < x < \infty, \quad (3)$$

where

$$c(\beta) = \left[\frac{\Gamma[\frac{3}{2}(1 + \beta)]}{\Gamma[\frac{1}{2}(1 + \beta)]}\right]^{1/(1+\beta)} \quad (4)$$

and

$$\omega(\beta) = \frac{\Gamma[\frac{3}{2}(1 + \beta)]^{1/2}}{(1 + \beta)\Gamma[\frac{1}{2}(1 + \beta)]^{3/2}}, \quad \sigma > 0, \quad -\infty < x < \infty, \quad \beta > -1. \quad (5)$$

In this form, the data's mean and standard deviation are given by  $\mu$  and  $\sigma$ , respectively. The parameter  $\beta$  is a measure of kurtosis and controls the distribution's deviation from normality.<sup>2</sup> When  $\beta = 0$ , the

<sup>1</sup>also called a generalized Laplacian or generalized Gaussian.

<sup>2</sup>Box and Tiao (1973) considered the case for  $\beta$  over the range  $[-1, 1]$ , but the distributions are also valid for the more general case for  $\beta > 1$ .

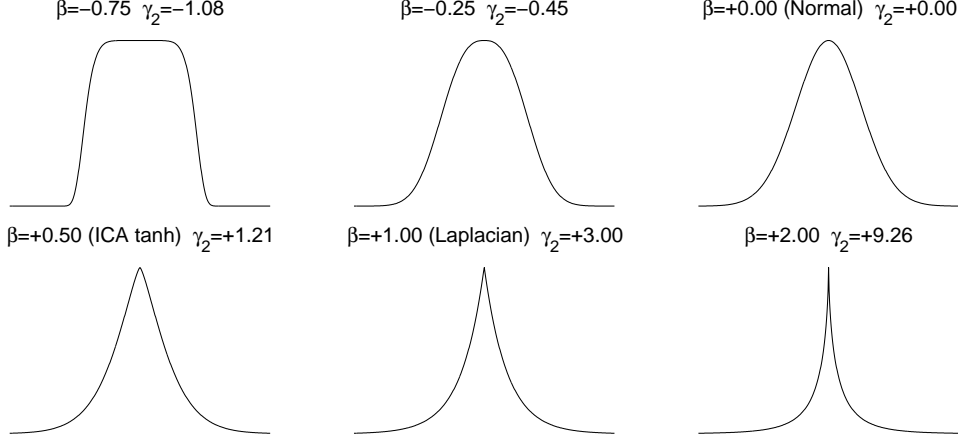


Figure 1: Exponential power distributions for various values  $\beta$ . The parameter  $\beta$  is also a measure of the distribution's kurtosis, and varies with the standard kurtosis measure,  $\gamma_2$  (eq.6).

distribution is the standard normal; it is a Laplacian (or double exponential) for  $\beta = 1$ . As  $\beta \rightarrow -1$ , the distribution becomes uniform over the unit interval. As  $\beta \rightarrow \infty$ , the distribution a delta function at zero. The parameter  $\beta$  can also be converted to the standard kurtosis measure  $\gamma_2 = E(x - \mu)^4 / \sigma^4 - 3$ . For the exponential power distribution, this relation is

$$\gamma_2 = \frac{\Gamma[\frac{5}{2}(1 + \beta)]\Gamma[\frac{1}{2}(1 + \beta)]}{\Gamma[\frac{5}{2}(1 + \beta)]^2} - 3. \quad (6)$$

Figure 2 shows examples of the exponential power distribution for various values of  $\beta$  and the corresponding values of  $\gamma_2$ . In addition to the standard normal and the Laplacian, also shown is the distribution labeled “ICA tanh” which shows the best-fitting exponential power distribution to the implied prior distribution under the widely-used tanh non-linearity in independent component analysis (ICA) (See section ?? below). The form of this prior is

$$p(x) = \theta \cosh(\beta x)^{-\theta/\beta} / Z, \quad (7)$$

where  $Z = 2\pi^{-1/2}\Gamma(1 + \theta/2\beta)/\Gamma((\theta + \beta)/2\beta)$ . The fit by the exponential power distribution (for  $\theta = \beta = 1$ ) is almost exact, differing only by having a slightly sharper peak, and yielding a Kullback-Leibler divergence of 0.0007 for the optimum of  $\beta = 0.495$  and  $\sigma = 1.525$ .

### 3 Estimating $\beta$

For the purposes of independent component analysis, it is natural to assume zero mean and unit variance. The problem then becomes to estimate the value of  $\beta$  from the data. This can be accomplished simply finding the maximum posteriori value of  $\beta$ . The posterior distribution of  $\beta$  given the observations  $\mathbf{x} = \{x_1, \dots, x_N\}$  is

$$p(\beta|\mathbf{x}) \propto p(\mathbf{x}|\beta)p(\beta), \quad (8)$$

where the data likelihood is

$$p(\mathbf{x}|\beta) = \prod_n \omega(\beta) \exp \left[ -c(\beta)|x_n|^{2/(1+\beta)} \right], \quad (9)$$

and  $p(\beta)$  defines the prior distribution for  $\beta$ . Because  $\beta > -1$ , it is convenient to use  $p(\beta) \sim \text{Gamma}(1+\beta|a, b)$ . Choosing the values  $a = 2$  and  $b = 2$  gives a broad prior distribution with a 95% density range of  $[-0.5, 10.5]$ , which is sufficient for our purposes here. See Box and Tiao (1973) for further discussion on inference with the exponential power distribution.

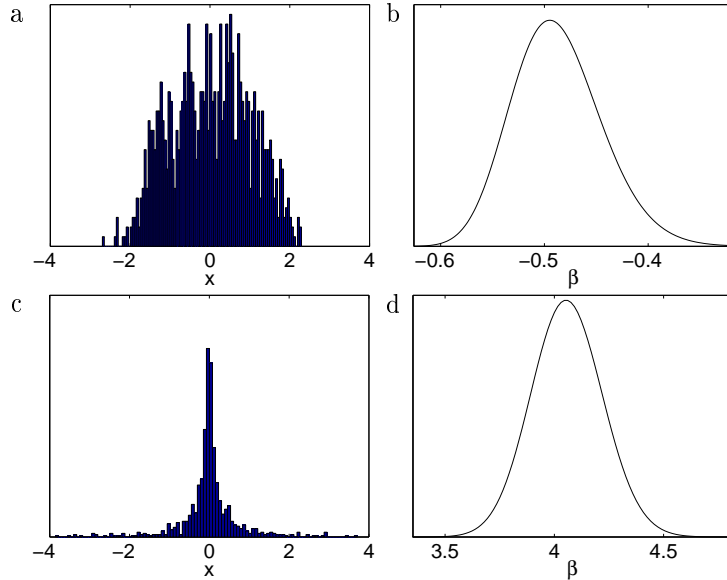


Figure 2: Posterior distributions for test data. The histograms (a and c) show samples (1000 in each set) drawn from two different exponential power distributions,  $\beta = -0.5$  (a) and  $\beta = 4$  (c), with zero mean and unit variance. The adjacent plots (b and d) show the corresponding posterior distributions (eq.8) for  $\beta$ .

## 4 Application to independent component analysis

The objective of ICA is to infer both the unknown sources and the unknown mixing matrix from the data signal. This problem, also called blind source separation, can be formulated explicitly as one of density estimation (Pearlmutter and Parra, 1996; MacKay, 1996; Cardoso, 1997). The data likelihood is derived by marginalizing over the sources

$$p(\mathbf{x}|\mathbf{A}) = \int p(\mathbf{x}|\mathbf{s}, \mathbf{A})p(\mathbf{s}) d\mathbf{s}. \quad (10)$$

Because there is a unique expression for the data in terms of the sources,  $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$ , the conditional likelihood is a delta function

$$p(\mathbf{x}|\mathbf{s}, \mathbf{A}) = \delta(\mathbf{x} - \mathbf{A}\mathbf{s}). \quad (11)$$

In this case, the expression for the data likelihood is

$$p(\mathbf{x}|\mathbf{A}) = \frac{p(\mathbf{s})}{|\det \mathbf{A}|}. \quad (12)$$

Performing gradient ascent on this expression gives a rule for learning the mixture matrix,  $\mathbf{A}$

$$\Delta \mathbf{A} \propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{x}|\mathbf{A}) = -\mathbf{A}(\mathbf{z}\mathbf{s}^T - \mathbf{I}). \quad (13)$$

where the prefactor  $\mathbf{A}\mathbf{A}^T$  is used to obtain the natural gradient solution (Amari et al., 1996) which gives an ascent direction that is insensitive to rescalings of the data. The vector  $\mathbf{z}$  is a function of the prior and is defined by  $\mathbf{z} = (\log p(\mathbf{s}))'$ . If for  $p(\mathbf{s})$  we use the exponential power distribution (eq.3)

$$z_i = -\theta |s_i - \mu_i|^{q-1} q c \sigma_i^{-q}, \quad (14)$$

where  $\theta = \text{sign}(s_i - \mu_i)$ ,  $q = 2/(1 + \beta_i)$ , and  $c = [\Gamma(3/q)/\Gamma(1/q)]^{q/2}$ . Details of the learning rule derivation are given the appendix.

Figure 2 shows examples of the fitting two dimensional distributions with the ICA-exponential power model. The values of  $\beta$  were (re)estimated periodically during learning by maximizing eq.8.

Figure 3 shows examples of the fitting two dimensional distributions with the ICA-exponential power model. The values of  $\beta$  were (re)estimated periodically during learning by maximizing the posterior (eq.8).

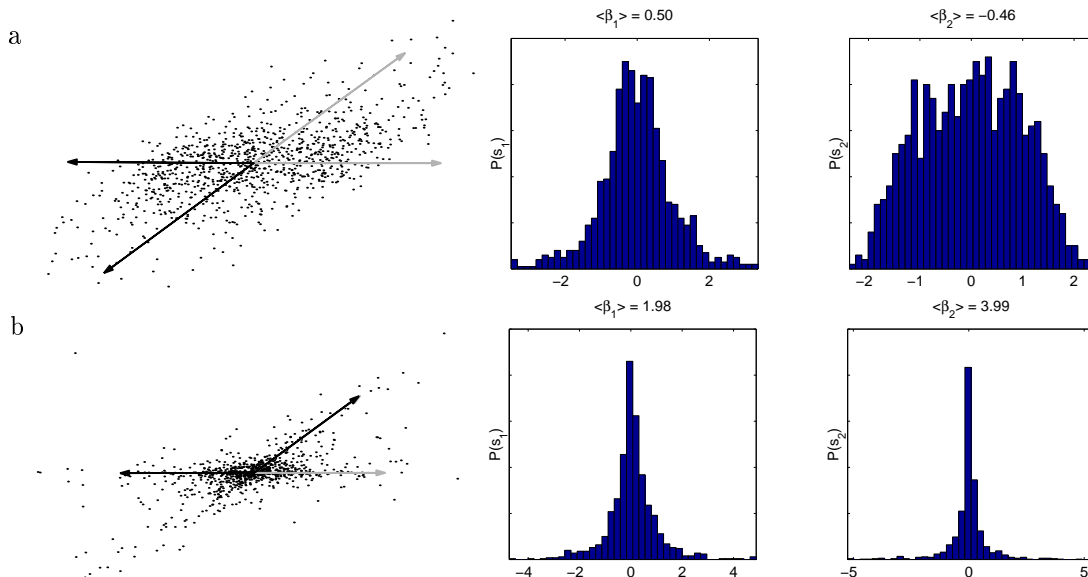


Figure 3: Fitting independent components of 2D distributions using exponential power source priors. The scatter plots on the left show the data distributions. The superimposed gray vectors were used to generate the distributions. The black vectors show the learned basis vectors. For clarity, the vectors have been rescaled. The histograms on the right show the distribution of coefficients and the inferred values of the exponential power parameter  $\beta$ . (a) The first example shows a mixture of super- and sub-Gaussian sources in which the true values of  $\beta$  were 0.5 and -0.5. (b) The second example shows a mixture of two super-Gaussian sources in which the true  $\beta$  values were equal to 2.0 and 4.0.

## 5 Learning codes for natural images

One application where the assumption of non-Gaussian distributions has proved critical is in the learning of codes (or independent components or basis functions) for natural images (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Lewicki and Olshausen, 1999). In previous work, the prior distribution assumed for the coefficients was assumed to be fixed. With the above results, we can let the data determine the best prior out the of class of exponential power distributions. This distribution has been shown to accurately model the statistics of wavelet coefficients for images and have been found to have values of  $\beta$  in the range  $[1, 1.3]$  (Mallat, 1989; Simoncelli and Adelson, 1996; Buccigrossi and Simoncelli, 1997).

The training data consisted of  $12 \times 12$  image patches randomly sampled from the ten  $512 \times 512$  images in the dataset of Olshausen and Field (1996). The patches were repeatedly resampled throughout training

to avoid reuse of any one set of patches. To speed convergence, the bases were initialized to the identity matrix, but with one basis set to unity to capture DC variation in the image structure. The basis matrix was adapted using eq.13 on blocks of 100 patterns. The gradient stepsize was initialized to 0.01 and gradually reduced to 0.0001 after 10,000 iterations. The added overhead compared to standard ICA is negligible. Training took about 4 hours on a 300Mhz Pentium II.

The bases were trained on two versions of the natural image data set, a “whitened” version, identical to that used in Olshausen and Field (1996), and “raw” version. The whitened data was filtered using a combined low-pass/whitening filter with frequency response  $L(f) = f \exp(-f/f_0)^n$  with  $f_0 = 200$  cycles/picture and  $n = 4$ . The “raw” data set used the same parameters, but with the frequency response  $L(f) = \exp(-f/f_0)^n$ . This low-pass filtered the original images to remove artifacts of the sampling grid. See Olshausen and Field (1997) for a more detailed discussion of these issues.

Figure 4 shows subsets of the learned basis functions on the two data sets and the inferred values of  $\beta$  versus the basis function norms. The  $\beta$  values for both image data sets lie in the range of 1 to 2.5 indicating that the source densities show a high degree of sparseness (i.e. greater kurtosis) and none fall close the Gaussian range. It is also evident that there is a slight trend for the sparseness to increase with the value of the norm, i.e. with spatial frequency. Not plotted are  $\beta$  values for the DC components, which were 0.82 for the whitened images and -0.70 (i.e. close to uniform) for the raw images.

Figure 5 shows that the fitted exponential power source densities closely matches the observed source distributions. The estimated coding efficiency (using the methods described in Lewicki and Olshausen (1999)) for the whitened basis was  $3.64 \pm 0.12$  bits per pixel for 10 samples of 1000 patterns each at 7 bits of precision. This compares to an estimate of 4.7 bits per pixel using the standard ICA model (Lee et al., 1999b). Using the same parameters, a value  $2.19 \pm 0.03$  was estimated for the raw basis, which is considerably lower, because whitening removes much of the redundancy. This bound could be further improved if additive noise were included in the model (Lewicki and Olshausen, 1999).

Estimated Bits Per Pixel for 7 Bits of Precision			
coefficient prior	(equivalent) exp. power $\beta$	dataset	
		whitened	raw
Gaussian	(0.00)		
cosh (ICA tanh)	( $\approx 0.39$ )	$4.07 \pm 0.12$	
Laplacian	(1.00)	$3.89 \pm 0.17$	$2.24 \pm 0.04$
exp. power	[-0.7, 2.5]	$3.64 \pm 0.12$	$2.19 \pm 0.03$

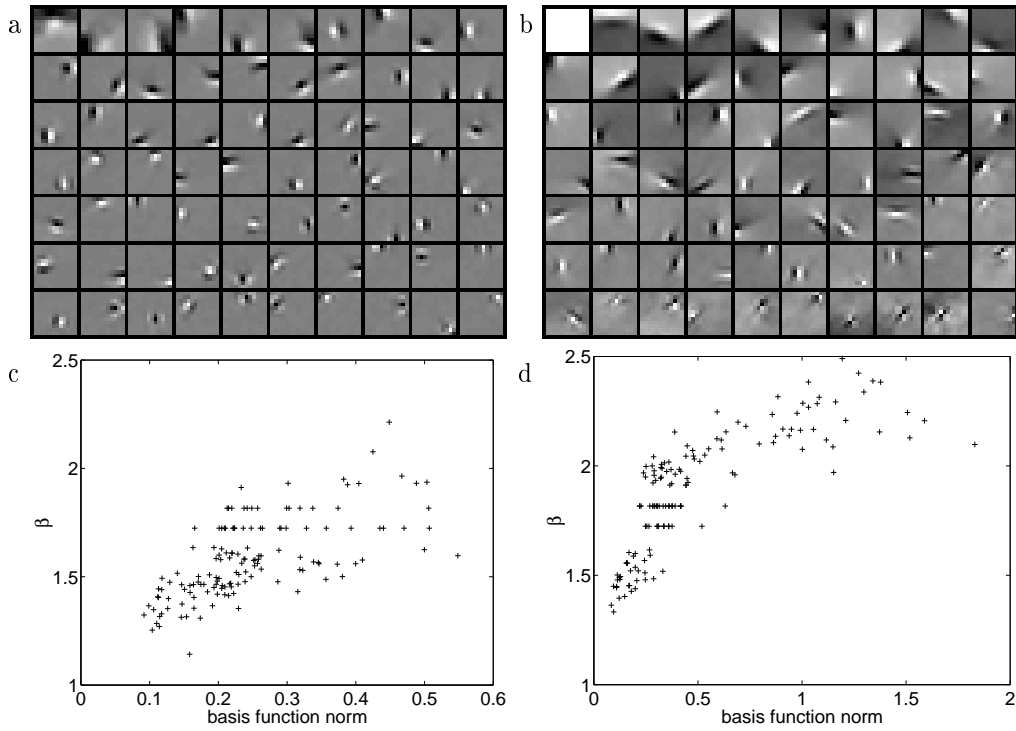


Figure 4: Learning codes of natural images by density estimation. The upper graphs show a sample of the learned basis functions in decreasing order of the  $L^2$  norm for whitened images (a) and for the raw images (b). The scatter plots show the inferred values of  $\beta$  versus the basis function norms.

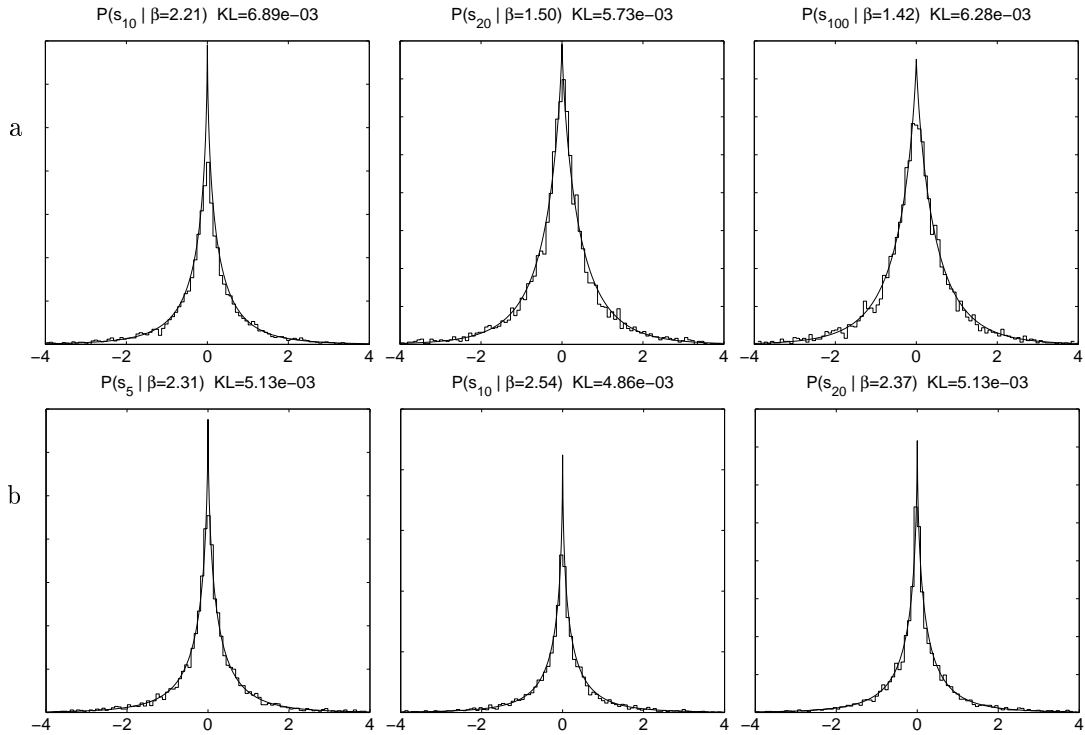


Figure 5: Fits of some inferred source densities (smooth curves) compared to observed distributions (step curves) for the whitened images (a) and raw images (b).

## 6 Discussion

Using exponential power densities with ICA improves the accuracy of the density estimation. These models allow direct estimate of the deviation from normality and include a wide range of sub- and super-Gaussian densities and seem particularly well suited for describing the statistics of natural images.

This distribution is also the statistical interpretation of the  $L^p$  norm loss function. Non-Gaussian loss functions have been advocated for robust estimation and for obtaining sparse representations (Box and Tiao, 1973; Mallat and Zhang, 1993; Chen et al., 1996) using overcomplete bases. Often the exponent is chosen arbitrarily, but here we have used probabilistic modeling to choose the exponent value that best fits the statistical distribution of the data.

More accurate estimates of the underlying density yields a number of benefits including improved coding efficiency for compression algorithms, better blind separation of a greater variety of sources, and better performance for denoising algorithms. Incorporating such multivariate density estimates into mixture models (Lee et al., 1999b) could lead to better classification of non-Gaussian clusters, such as those containing outliers.

## Appendix

### Derivation of the Learning Rule

We first review some useful notation and results for matrix calculus (Anderson, 1984; Dwyer, 1967; Fang and Zhang, 1990). A convenient notation for the derivative of one vector with respect to another is

$$\frac{\partial \mathbf{z}^T}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}^T}{\partial \mathbf{x}} \frac{\partial \mathbf{z}^T}{\partial \mathbf{y}}. \quad (15)$$

In expressing matrix derivatives, it is useful to define index vectors and matrices. Let  $\mathbf{e}_i(n) = (0, \dots, 0, 1, 0, \dots)^T$  be an  $n \times 1$  vector with a one at the  $i^{\text{th}}$  position and  $\mathbf{E}_{ij}(m, n) = \mathbf{e}_i(m) \mathbf{e}_j^T(n)$ . These can sometimes be more simply written as  $\mathbf{e}_i$  and  $\mathbf{E}_{ij}$ . Then  $\mathbf{e}_i^T \mathbf{e}_j = \delta_{ij}$ , where  $\delta_{ij} = 1$  if  $i = j$  and 0 otherwise. We state the following useful relations.

$$\mathbf{E}_{ij}(m, n) \mathbf{x}_{m \times 1} = \mathbf{e}_i(m) x_j, \quad (16)$$

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{A} \mathbf{X}^{-1} \mathbf{B} = -\mathbf{A} \mathbf{X}^{-1} \mathbf{E}_{ij} \mathbf{X}^{-1} \mathbf{B}, \quad (17)$$

and

$$\frac{\partial}{\partial \mathbf{X}} \log |\mathbf{X}| = \mathbf{X}^{-T}. \quad (18)$$

We now proceed with the derivation of the learning rule. The likelihood of the data given the model is

$$p(\mathbf{x}|\mathbf{A}) = \frac{p(\mathbf{s})}{|\det \mathbf{A}|}, \quad (19)$$

where  $\mathbf{s} = \mathbf{A}^{-1} \mathbf{x}$ . For purposes of optimizing the model parameters, it is convenient to use the derivative of the log likelihood

$$\frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{x}|\mathbf{A}) = \frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{s}) - \frac{\partial}{\partial \mathbf{A}} \log |\det \mathbf{A}| \quad (20)$$

We first consider the derivative of  $p(\mathbf{s})$  w.r.t a single element  $a_{ij}$  of the matrix  $\mathbf{A}$

$$\frac{\partial}{\partial a_{ij}} \log p(\mathbf{s}) = \frac{\partial \mathbf{s}^T}{\partial a_{ij}} \mathbf{z}, \quad (21)$$

where  $\mathbf{z} = \partial \log p(\mathbf{s}) / \partial \mathbf{s}$ . Then

$$\frac{\partial \mathbf{s}^T}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}} [\mathbf{A}^{-1} \mathbf{x}]^T \quad (22)$$

$$= [\mathbf{A}^{-1} \mathbf{E}_{ij} \mathbf{A}^{-1} \mathbf{x}]^T \quad (23)$$

$$= -\mathbf{s}^T \mathbf{E}_{ij}^T \mathbf{A}^{-T} \quad (24)$$

$$= -(\mathbf{E}_{ij} \mathbf{s})^T \mathbf{A}^{-T} \quad (25)$$

$$= -s_j \mathbf{e}_i^T \mathbf{A}^{-T} \quad (26)$$

$$= -s_j (\mathbf{A}_i^{-1})^T, \quad (27)$$

where  $\mathbf{A}_i^{-1}$  is the  $i^{\text{th}}$  column of  $\mathbf{A}^{-1}$ .

Using the above results we then have

$$\frac{\partial}{\partial a_{ij}} \log p(\mathbf{s}) = -s_j (\mathbf{A}_i^{-1})^T \mathbf{z} \quad (28)$$

$$= -s_j \sum_k [\mathbf{A}_i^{-1}]_k z_k \quad (29)$$

$$= -s_j [\mathbf{A}^{-T} \mathbf{z}]_i . \quad (30)$$

The derivative in terms of the whole matrix  $\mathbf{A}$  can then be expressed as

$$\frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{s}) = -\mathbf{A}^{-T} \mathbf{z} \mathbf{s}^T . \quad (31)$$

Combining results yields

$$\frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{x}|\mathbf{A}) = -\mathbf{A}^{-T} \mathbf{z} \mathbf{s}^T - \mathbf{A}^{-T} . \quad (32)$$

We can premultiply by  $\mathbf{A} \mathbf{A}^T$  to obtain the natural gradient solution (Amari et al., 1996)

$$\mathbf{A} \mathbf{A}^T \frac{\partial}{\partial \mathbf{A}} \log p(\mathbf{x}|\mathbf{A}) = -\mathbf{A} (\mathbf{z} \mathbf{s}^T - \mathbf{I}) . \quad (33)$$

## References

- Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural and Information Processing Systems*, volume 8, pages 757–763, San Mateo. Morgan Kaufmann.
- Anderson, T. W. (1958,1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York. First and Second Edition.
- Attias, H. (1998). Blind separation of noisy mixtures: An EM algorithm for independent factor analysis. *Neural Computation*. Submitted.
- Bell, A. J. and Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Res.*, 37(23):3327–3338.

- Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.
- Buccigrossi, R. W. and Simoncelli, E. P. (1997). Image compression via joint statistical characterization in the wavelet domain. Technical Report 414, Univ. Pennsylvania.
- Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4:109–111.
- Chen, S., Donoho, D. L., and Saunders, M. A. (1996). Atomic decomposition by basis pursuit. Technical report, Dept. Stat., Stanford Univ., Stanford, CA.
- Dwyer, P. S. (1967). Some applications of matrix derivatives in multivariate analysis. *Am. Stat. Assoc. Journal*, 62:607–625.
- Fang, K. T. and Zhang, Y. T. (1990). *Generalized Multivariate Analysis*. Springer-Verlag, Berlin.
- Lee, T.-W., Girolami, M., and Sejnowski, T. J. (1999a). Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources. *Neural Computation*, 11(2):409–433.
- Lee, T.-W., Lewicki, M. S., and Sejnowski, T. J. (1999b). ICA mixture models for unsupervised classification of non-Gaussian sources and automatic context switching in blind signal separation. *IEEE PAMI*. submitted.
- Lewicki, M. S. and Olshausen, B. A. (1999). A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. of Am. A: Optics, Image Science, and Vision*. in press.
- MacKay, D. J. C. (1996). Maximum likelihood and covariant algorithms for independent component analysis. University of Cambridge, Cavendish Laboratory. Available at <ftp://w01.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz>.
- Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE PAMI*, 11:674–693.
- Mallat, S. G. and Zhang, Z. F. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

- Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, 381:607–609.
- Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, 37(23).
- Pearlmutter, B. A. and Parra, L. C. (1996). A context-sensitive generalization of ICA. In *International Conference on Neural Information Processing*, pages 151–157.
- Simoncelli, E. P. and Adelson, E. H. (1996). Noise removal via Bayesian wavelet coring. In *3rd IEEE Int'l Conf on Image Processing, Lausanne Switzerland*.