TWO METHODS FOR ESTIMATING OVERCOMPLETE INDEPENDENT COMPONENT BASES

Mika Inki and Aapo Hyvärinen

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland

ABSTRACT

Estimating overcomplete ICA bases is a difficult problem that emerges when using ICA on many kinds of natural data, e.g. image data. Most algorithms are based on approximations of the likelihood, which leads to computationally heavy procedures. Here we introduce two algorithms that are based on heuristic approximations and estimate an approximate overcomplete basis quite fast. The algorithms are based on quasi-orthogonality in high-dimensional spaces, and the gaussianization procedure, respectively.

1. INTRODUCTION

Independent component analysis can be considered to be a fundamental generative model for low-level features of many types of natural data. In ICA the observed data is expressed as a linear transformation of latent variables that are nongaussian and mutually independent. We may express the model as

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_i \mathbf{a}_i s_i \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_m)$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)$ is the vector of the latent variables called the independent components or source signals, and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix.

In the classic case, we assume that the number of independent components equals the number of the observed variables, i.e. $n = m$. Exact conditions for the identifiability of the model were given in [1], and several methods for estimation of the classic ICA model have been proposed in the literature, see [8] for a review.

Recently, a non-classic modification of the model, where it is assumed that the number of independent components is larger than the number of observed variables ($n > m$), has attracted the attention of a number of researchers [12, 13, 14]. Such a model is especially interesting when ICA is used for image modeling, because it leads to decomposition of image windows that is closely related to overcomplete wavelet bases (see [13]). Basically, the larger number of independent components in the model means that we have a larger 'dictionary' from which to construct the representation.

Some methods have already been proposed for estimating the mixing matrix in the ICA model with $n > m$, a problem often called estimation of an overcomplete ICA basis. A drawback with most proposed methods is that they are computationally very demanding. This is basically because the model then becomes a model with missing data. In fact, the evaluation of the likelihood contains an integral and even reasonable approximations of that integral are hard to compute [12]. On the other hand, since these methods are usually applied to data of very high dimensions, it would be very useful to have an estimation method that can cope with very large dimensions with a moderate computational load.

In this paper, we propose two somewhat heuristic methods for approximate estimation of the ICA model with overcomplete bases. The methods are computationally efficient and appear to give good approximations of the optimal estimates.

2. APPROXIMATE ESTIMATION BY QUASI-ORTHOGONALITY

In feature extraction for many kinds of natural data, the ICA model is only a rather coarse approximation. In particular, the number of potential "independent components" seems to be infinite: The set of such components is closer to a continuous manifold that a discrete set. One evidence for this is that in image feature extraction, basic ICA estimation methods give different basis vectors when started with different initial values, and the number of components thus produced does not seem to be limited.

Any basic ICA estimation method for such data gives a rather arbitrary collection of components which are somewhat independent, and have sparse (supergaussian or leptokurtic) marginal distributions. We could argue, therefore, that it is the sparseness that is important, and the exact dependence relations between the components are secondary.

In fact, recent research has revealed important dependencies between the estimated components [6, 7, 15, 16].

In the following, we propose two methods that yield bases for overcomplete sparse decompositions. The method in this section is based on a Bayesian prior on the mixing matrix, and the method in the next section uses a method of gaussianization that has been proposed in projection pursuit literature.

2.1. Sparse approximately uncorrelated decompositions

Let us assume, for simplicity, that the data is prewhitened as a preprocessing step, as in most ICA methods. Then the independent components are simply given by the dot-products of the whitened data vector $\mathbf{z}$ with the basis vectors $\mathbf{a}_i$, barring the noise generated by other components in non-orthogonal directions.

Due to the above considerations, we assume in our approach that what is usually needed, is a collection of basis vectors which has the following two properties.

1. The dot-products $\mathbf{a}_i^T\mathbf{z}$ of the observed data with the basis vectors have sparse (super-Gaussian) marginal distributions.

2. The $\mathbf{a}_i^T\mathbf{z}$ should be approximately uncorrelated ("quasi-uncorrelated"). Equivalently, the vectors $\mathbf{a}_i$ should be approximately orthogonal ("quasi-orthogonal").

A decomposition with these two properties seems to capture the essential properties of the decomposition obtained by estimation of the ICA model. Such decompositions could be called sparse approximately uncorrelated decompositions.

It is clear that it is possible to find highly overcomplete basis sets that have the first property of these two. What is not obvious, however, is that it is possible to find strongly overcomplete decompositions such that the dot-products are approximately uncorrelated. The main point here is that this is possible because of the phenomenon of quasi-orthogonality.

2.2. Bayesian priors for quasi-orthogonality

Quasi-orthogonality [9, 10, 11] is a somewhat counterintuitive phenomenon encountered in very high-dimensional spaces. In a certain sense, there is much more room for vectors in high-dimensional spaces. The point is that in an n-dimensional space, where n is large, it is possible to have (say) $2n$ vectors that are practically orthogonal, i.e. their angles are close to 90 degrees. In fact, when n grows, the angles can be made arbitrarily close to 90 degrees. This must be contrasted with small-dimensional spaces: If, for example, $n = 2$, the even the maximally separated $2n = 4$ vectors exhibit angles of 45 degrees.

Our goal is now to formulate a Bayesian prior for quasi-orthogonality. Such a prior would give high probabilities to mixing matrices with quasi-orthogonal columns. The starting point is to assume that the elements of the basis vectors are drawn randomly, independently from each other. Consider the dot-product between two basis vectors: $\mathbf{a}_i^T\mathbf{a}_j$. Let us normalize this to obtain the quantity

$$d_{ij} = \frac{\mathbf{a}_i^T\mathbf{a}_j}{\|\mathbf{a}_i\|\|\mathbf{a}_j\|} \tag{2}$$

Assume that the dimensions are large. Under the assumption of the independence of the elements of the vectors, the sum $\sum_k a_{ik}a_{j_k}$ has an approximately gaussian distribution, due to the central limit theorem. For large dimensions, the norms in the denominator can be considered to be approximately constant. In the following, we thus approximate the distribution of d_{ij} by a gaussian distribution. Its variance can computed under this independence approximation as $1/n$.

Thus, we could impose a gaussian prior on these dot-products:

$$p(d) = \varphi(d\sqrt{n})\sqrt{n} \tag{3}$$

where φ is the standardized gaussian probability density function. This is used to define the prior for $\mathbf{A}$:

$$p(\mathbf{A}) = \prod_{i\neq j}\varphi(d_{ij}\sqrt{n})\sqrt{n} = \prod_{i\neq j}\varphi(\frac{\mathbf{a}_i^T\mathbf{a}_j}{\|\mathbf{a}_i\|\|\mathbf{a}_j\|}\sqrt{n})\sqrt{n} \tag{4}$$

whose logarithm is given by

$$\log p(\mathbf{A}) = -\frac{n}{2}\sum_{i\neq j}(\frac{\mathbf{a}_i^T\mathbf{a}_j}{\|\mathbf{a}_i\|\|\mathbf{a}_j\|})^2 + \text{const.} \tag{5}$$

This prior assigns higher probabilities to mixing matrices whose columns are quasi-orthogonal.

In practice, this prior can be multiplied by a constant that expresses the strength of the prior. It seems that this strength should be proportional to the number of observations, so that its effect is not diminished for large sample sizes; thus we use a prior strength of the form αT where α is a positive constant that includes the factor $n/2$. Furthermore, in simulations we have noticed that the prior often performs better if the exponential 2 is replaced by a higher number, denoted by λ. Thus, in the following we use the following prior:

$$\log p(\mathbf{A}) = \alpha T\sum_{i\neq j}(\frac{\mathbf{a}_i^T\mathbf{a}_j}{\|\mathbf{a}_i\|\|\mathbf{a}_j\|})^\lambda + \text{const.} \tag{6}$$

One possible reason why higher powers are better can be seen in Fig. 1: The gaussian approximation of the density of the dot-product is good only relatively close to zero; by

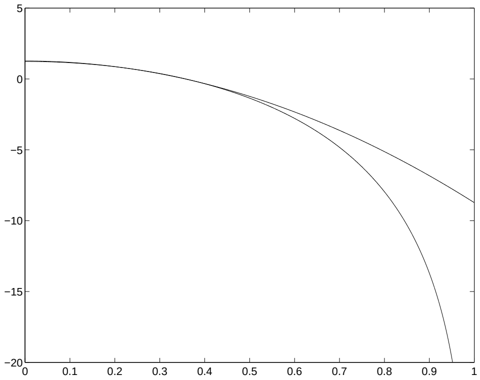

Fig. 1. Logarithms of the actual and the approximated probability densities of the absolute values of the dot product between two random (unit) vectors in a 20-dimensional space. The upper curve corresponds to the gaussian approximation and the lower curve to the actual distribution. The difference for large values may explain why a higher λ than 2 gives better results.

using a higher power, the approximation may be better for larger dot products.

To use the above prior with the ICA likelihood, we do another approximation. Consider the likelihood for ordinary (not overcomplete) ICA:

$$\log L(\mathbf{A}) = \sum_t \sum_{i=1}^{n} \log p_i(\mathbf{w}_i^T \mathbf{z}(t)) + T \log |\det \mathbf{W}|. \quad (7)$$

where the $\mathbf{w}_i$ are the rows of the inverse of $\mathbf{A}$, and $\mathbf{W} = \mathbf{A}^{-1}$. The purpose of the last term involving $\log|\det \mathbf{W}|$ is basically to make to $\mathbf{w}_i$ more or less orthogonal. In fact, this term disappears if the $\mathbf{a}_i$ are constrained orthogonal. Thus, since we are already incorporating the quasi-orthogonality in the prior, we discard this term and simply extend the first term of the likelihood to the overcomplete case. Finally, we arrive at the following expression for the posterior:

$$\log p(\mathbf{A}|\mathbf{z}(t), t = 1, ..., T) = \sum_t \sum_{i=1}^{n} \log p_i(\mathbf{a}_i^T \mathbf{z}(t)) + \alpha T \sum_{i \neq j} \left(\frac{\mathbf{a}_i^T \mathbf{a}_j}{||\mathbf{a}_i|| ||\mathbf{a}_j||}\right)^{\lambda} + \text{const.} \quad (8)$$

In the following, we maximize this posterior to estimate $\mathbf{A}$. Note that previously we proposed a modification of FastICA to perform a similar estimation by quasi-orthogonality [5], but it seems that the present method estimates more orthogonal bases. This may be due to the additional parameters α and λ that can be tuned to fit the data at hand.

2.3. Simulations

First, we tried our method on simulated data. We mixed 40 independent components with Laplacian distributions into a 20 dimensional data space, i.e. $\mathbf{A}$ was a matrix of size 20×40. The sample size was 50000.

A general problem in estimating overcomplete bases is that components whose contributions to the data are very small (as measured by the norm of the corresponding column of $\mathbf{A}$) are very difficult to estimate. To avoid this problem, the columns of the mixing matrix were generated so that the norms of the columns were uniformly distributed between 0.75 and 1.5. Otherwise, the mixing matrix was random.

As a preprocessing step, the data was whitened. We then maximized the posterior in (8) by gradient ascent. The parameter α was fixed to the value of 3. We tried both second and higher powers in the exponent λ of the distribution function. We found that the second power was unable to prevent all the components from converging to the same result. By using the eighth power $\lambda = 8$, the prior worked properly.

To investigate the quasi-orthogonality of the obtained basis vectors, we can look at the minimum angle between one basis vector from the rest. This minimum angle can be calculated from the maximum of the absolute values of the dot products between the basis vector in question and the rest, i.e. from the maximum element of the row (or column) of $\hat{\mathbf{A}}^T\hat{\mathbf{A}}$ corresponding to the basis vector. These angles are depicted in Fig. 2. Note that all of these angles are above 70 degrees, which shows good quasi-orthogonality. The probability density shown by the solid line in Fig. 2 for comparison gives the distribution that one would expect for the elements of $\hat{\mathbf{A}}^T\hat{\mathbf{A}}$ if they were actually distributed randomly in different directions. One can see that in fact, the obtained vectors are even more orthogonal than corresponding random vectors.

The other thing of interest is, of course, how close the estimated basis vectors are to the original basis vectors. This can determined by looking at the absolute value of the elements of $\mathbf{A}^T\hat{\mathbf{A}}$. First we find the element with the largest absolute value from this matrix, remove both the real and the estimated basis vectors corresponding to it, and repeat this until we have a “match” for each basis vector.

The angles (in degrees) between the estimated basis vectors and the matched original basis vectors are shown in Fig. 3. We can see that at least 35 components were quite correctly estimated.

2.4. Experiments on image data

Next we tested our method on image feature extraction. We sampled 12×12 image windows from 13 natural images. We removed the mean from the windows and whitened the thus obtained data vectors. From this 143 dimensional space

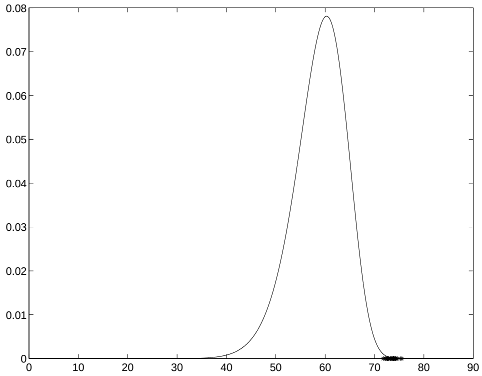

Fig. 2. The separation results when 40 independent components with Laplacian distribution are mixed into a 20 dimensional space, in the quasi-orthogonalization approach. Asterisks: The minimum angles between the estimated basis vectors. Solid line: Probability density that the minimum angle would have if the vectors were really generated randomly.

we estimated 300 components, i.e. a basis more than twice overcomplete. We used the same parameters $\lambda = 8, \alpha = 3$ that we used with the simulations. A supergaussian density was assumed for the independent components by taking $\log p_i(s_i) = \log \cosh s_i$.

In Fig. 4, the basis vectors are shown. They are quite similar to what one obtains with ordinary ICA using a supergaussian prior for the independent components. Note that there are no low-frequency components since we used an exclusively supergaussian prior density for the components. In Fig. 5, we show the distances between the estimated basis vectors in the whitened space; these show that the basis vectors are really quasi-orthogonal.

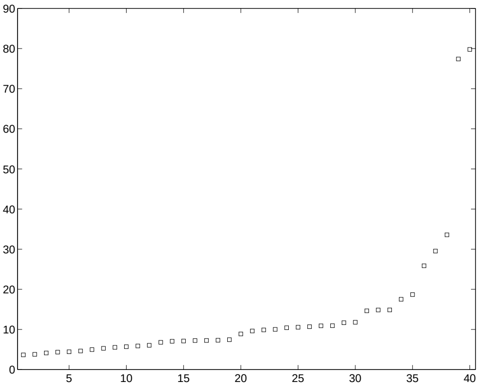

Fig. 3. The distances between the real and matched components, using the quasi-orthogonalization approach.

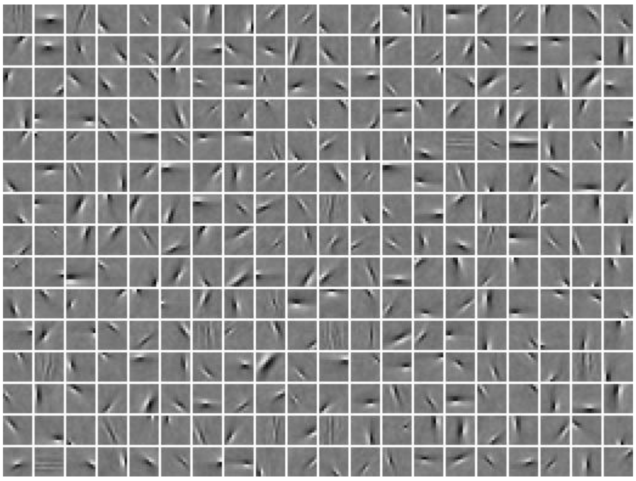

Fig. 4. The basis vectors obtained with the quasi-orthogonalizing prior. The basis vectors are quite similar to those obtained by ordinary ICA, but the basis is more than 2 times overcomplete.

3. APPROXIMATE ESTIMATION BY GAUSSIANIZATION

3.1. Gaussianization vs. orthogonalization

The second method that we propose for approximate estimation of overcomplete ICA bases is based on gaussianization. This idea comes from projection pursuit literature [3]. The point is to replace orthogonalization or quasi-orthogonalization by a nonlinear transform that makes the projections onto already estimated basis vectors gaussian.

We use deflationary estimation of the independent components [2, 4], which means that we first estimate one independent component (typically by maximizing a measure of nongaussianity), then estimate a second component somehow discarding the direction of the first one, and so on, repeating the procedure n times.

The question is then, how to discard the already estimated components. Typically this is done by constraining the search for new independent components to the space that is orthogonal to the already found components; this is more or less equivalent to removing the estimated independent components from the data by linear regression, assuming that the data is prewhitened.

In the gaussianization procedure, we do not remove the components from the data, but we attempt to remove the nongaussianity associated with the component. Assume that we have estimated the i-th component as the linear combination $s_i \underset{i}{=} \mathbf{z}\mathbf{a}^T$. To gaussianize this direction, we compute the cumulative distribution function, say F of s_i. Then we compute for every observation $s_i(t) = \mathbf{a}_i^T \mathbf{z}(t)$ the trans-

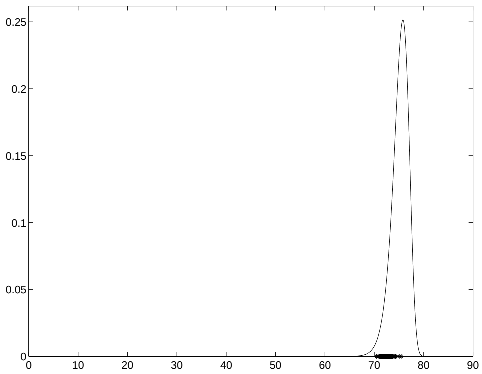

Fig. 5. The distances between the estimated components in the whitened space, using the quasi-orthogonalization approach on image data. See caption of Fig. 2.

form $h(t) = \Phi^{-1}(F(s_i(t)))$, where Φ is the cumulative distribution function of the standardized gaussian distribution. This variable h has a gaussian distribution [3]. To reconstruct the observed $\mathbf{z}(t)$ after this gaussianization, we transform the data back as

$$\mathbf{z}(t) \leftarrow \mathbf{a}_i h(t) + (\mathbf{I} - \mathbf{a}_i \mathbf{a}_i^T)\mathbf{z}(t) \qquad (9)$$

Note that even after m marginal gaussianizations (where m is the dimension of the data) the data is still not distributed according to a joint gaussian distribution: Forcing m marginal distributions to be gaussian does not, in general, make the joint distribution gaussian. In fact, the marginal gaussianizations may interact because the directions are not necessarily orthogonal, so that even the m components that were gaussianized need not have gaussian distributions after the whole process is finished. Compare this with the case of orthogonalization: In orthogonalizing deflation, it is completely impossible to estimate more than m components since one cannot have more than m orthogonal vectors in an m-dimensional space. This is exactly why we had to use quasi-orthogonalization instead of exact orthogonalization in the heuristic extension of the preceding section. In gaussianization, we do not need to modify the method. On the other hand, gaussianization is only applicable in deflationary mode, not in symmetric mode in which the quasi-orthogonalization was used in the preceding section.

3.2. Simulations

We applied our method first on simulated data. The data we used with this approach was identical to that used with the quasi-orthogonalizing prior. The procedure for the estimation was as follows: first we whitened the observed data. Then we estimated one component by using FastICA [4] with the $\tanh$ nonlinearity, and then gaussianized (using the cumulative distribution functions) the component in the direction that FastICA found. Then we estimated another component by FastICA, and so on.

We evaluated the angles between estimated basis vectors in the same manner as with the quasi-orthogonalizing prior. The minimum angles are shown in Fig. 6. All of these angles are above 52 degrees, which shows that we again obtained quite quasi-orthogonal basis vectors. The distances between the original basis vectors and their matched estimates are shown in Fig. 7. Almost all the components were properly estimated.

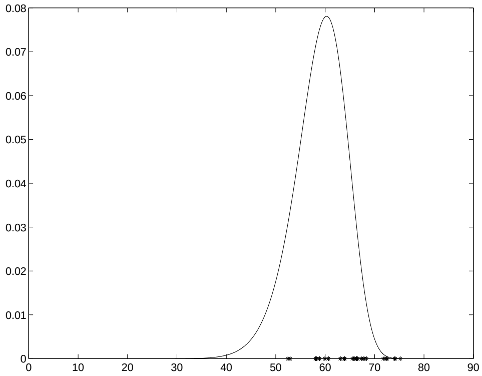

Fig. 6. The separation results when 40 independent components with Laplacian distributions are mixed into a 20 dimensional space, in the case of the gaussianization method. See caption of Fig. 2.

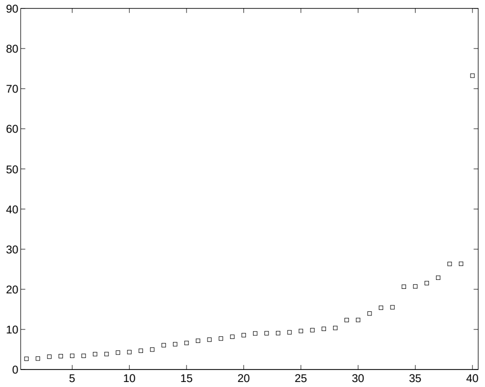

Fig. 7. The distances between the real basis vectors and the matched estimates, for simulated data using the gaussianization procedure.

3.3. Experiments with image data

Finally, we applied our algorithm for image feature extraction. The image data was similar to that used with the quasi-orthogonalizing prior. In Fig. 8 we have the obtained basis vectors. These are again similar to those obtained by basic ICA estimation. We obtained low-frequency components as well, which is in contrast to the quasi-orthogonalization method. In Fig. 9 we have the distances between the estimated directions in the whitened space, showing that the basis vectors are quite diffent from each other.

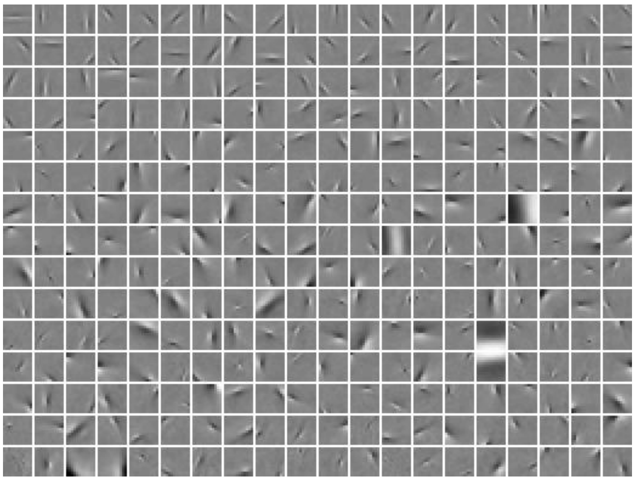

Fig. 8. The image basis vectors obtained using gaussianization. Again, the basis vectors are quite similar to what one obtains with ordinary ICA, but the basis is more than 2 times overcomplete.

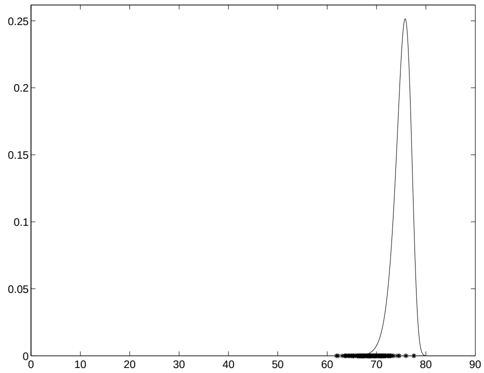

Fig. 9. The distances between the estimated components in the whitened space, for image data and the gaussianization approach. See caption of Fig. 2.

4. CONCLUSION

We introduced two somewhat heuristically motivated methods for estimating overcomplete ICA bases from images. The methods were based on simply extending the estimation principles of basic ICA to the overcomplete case. Simulations and experiments on image data show that the methods work surprisingly well, thus offering computationally efficient alternatives for more statistically principled methods.

5. REFERENCES

[1] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[2] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[3] J.H. Friedman. Exploratory projection pursuit. *J. of the American Statistical Association*, 82(397):249–266, 1987.

[4] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[5] A. Hyvärinen, R. Cristescu, and E. Oja. A fast algorithm for estimating overcomplete ICA bases for image windows. In *Proc. Int. Joint Conf. on Neural Networks*, pages 894–899, Washington, D.C., 1999.

[6] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[7] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7), 2001. in press.

[8] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[9] P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1998.

[10] S. Kaski. Dimensionality reduction by random mapping: fast similarity computation for clustering. In *Proc. Int. Joint Conf. on Neural Networks (IJCNN'98)*, pages 413–418, Anchorage, Alaska, 1998.

[11] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

[12] M. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[13] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[14] P. Pajunen. Blind separation of binary sources with less sensors than sources. In *Proc. Int. Conf. on Neural Networks*, Houston, Texas, 1997.

[15] E. P. Simoncelli and O. Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems 11*, pages 153–159. MIT Press, 1999.

[16] C. Zetzsche and G. Krieger. Nonlinear neurons and high-order statistics: New approaches to human vision and electronic image processing. In B. Rogowitz and T.V. Pappas, editors, *Human Vision and Electronic Imaging IV (Proc. SPIE vol. 3644)*, pages 2–33. SPIE, 1999.