# Two Approaches to Estimation of 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. Here we introduce two algorithms that estimate an approximate overcomplete basis quite fast in a high-dimensional space. The first algorithm is based on an assumption that the basis vectors are randomly distributed in the space, and the second on the gaussianization procedure.**

## I. 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 $\mathbf{x} = (x_1, x_2, ..., x_m)^T$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ 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. The columns of $\mathbf{A}$, the $\mathbf{a}_i$, are often called basis vectors.

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 [13], [14], [15]. Such a model is especially interesting when ICA is used for image modeling, because it leads to a decomposition of image windows that is closely related to overcomplete wavelet bases (see [14]). 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 [13]. 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 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. This paper is an extended and improved version of [9].

## II. ASSUMING A PRIOR FOR THE MIXING MATRIX

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], [16], [17].

In the following, we propose two methods that yield bases for overcomplete sparse decompositions of such

0-7803-7278-6/02/$10.00 ©2002 IEEE

data. These methods can, however, also be used on data with subgaussian sources. 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.

### *A. Some properties of an overcomplete ICA basis*

Let us assume that the norms of the basis vectors are set to unity and that the variances of the sources can differ from unity. Note that this does not restrict the distributions of the observed variables. Let us also assume, for simplicity, that the data is prewhitened as a preprocessing step, as in most ICA methods.

Now the dot product between the $i$:th basis vector and the whitened data vector can be written as:

$$\mathbf{a}_i^T \mathbf{z} = \mathbf{a}_i^T \mathbf{A}\mathbf{s} = s_i + \sum_{j \neq i} \mathbf{a}_i^T \mathbf{a}_j s_j \tag{2}$$

The first term is the $i$:th independent component and the second term is approximately gaussian, especially when there is a large number of components in this high-dimensional space, the basis vectors are randomly distributed into the space, and there are no components whose variance is considerably larger than the others. Therefore any good estimate of the mixing matrix should maximize the nongaussianities of these dot products.

On the other hand, simply maximizing the nongaussianities leads to a situation where all the basis vectors point in the most nongaussian direction. However, two random vectors in a high-dimensional space are most likely almost orthogonal, due to quasi-orthogonality. Quasi-orthogonality [10], [11], [12] 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. Therefore any good estimate of the mixing matrix should have quasi-orthogonal columns (i.e. basis vectors $\mathbf{a}_i$). Also, the second term in (2) is gaussian especially when the basis vectors are quasi-orthogonal.

A decomposition with these two properties, i.e. nongaussianity and quasi-orthogonality, seems to capture the essential properties of the decomposition obtained by estimation of the ICA model.

### *B. Bayesian prior for quasi-orthogonality*

We now calculate the probability density for the dot product between two randomly and independently drawn basis vectors: $\mathbf{a}_i^T \mathbf{a}_j$. Assume that the basis vectors are of unit length. In this case these basis vectors can be considered to be points on the surface of an $m$-dimensional unit sphere. The volume of an $m$-dimensional sphere of radius $r$ is

$$V(r) = C_m r^m, \tag{3}$$

where $C_m = \frac{\pi^{\frac{m}{2}}}{\Gamma[\frac{m}{2}+1]}$ is a constant. When you take the portion of the surface of the $m$-dimensional unit sphere that is within an angle of $\alpha$ to a fixed vector and project this onto a hyperplane orthogonal to this vector, you get an $m-1$ dimensional ball of radius $\sin(\alpha)$. When you derivate this with respect to the radius you get the length of the boundary. Therefore the infinitesimal area of a band of width $\mathrm{d}\alpha$ at an angle $\alpha$ on the surface of an $m$-dimensional sphere (whose surface area is scaled to one using $c_m$) is:

$$p_\alpha(\alpha)\mathrm{d}\alpha = c_m \sin^{m-2}(\alpha)\mathrm{d}\alpha \tag{4}$$

Now $c_m = \frac{m-1}{m} \frac{\Gamma[\frac{m}{2}+1]}{\sqrt{\pi}\Gamma[\frac{m-1}{2}+1]}$. By denoting the dot product as $x$, i.e. $\alpha = \arccos(x)$, we get the following probability density:

$$p_x(x) = c_m (1-x^2)^{\frac{m-3}{2}} \tag{5}$$

In this way we get a probability for the mixing matrix $\mathbf{A}$, assuming that all the dot products are independent:

$$p(\mathbf{A}) = \prod_{i<j} c_m (1 - (\mathbf{a}_i^T \mathbf{a}_j)^2)^{\frac{m-3}{2}} \tag{6}$$

The dot products are not quite independent of each other in this space, e.g. if there were $m$ orthogonal vectors in this space, the probability for any of the other $n-m$ vectors to be orthogonal to all these vectors would be zero. In most cases, however, this approximation appears to be good enough.

It can be now easily seen, that a rotation of the coordinate system does not affect this approximation for the probability of $\mathbf{A}$, i.e. $p(\mathbf{WA}) = p(\mathbf{A})$, where $\mathbf{W}$ is orthogonal. One should also note that generally $p(\mathbf{A})$ and $\det(\mathbf{A})$ do not quite behave similarly, even when $\mathbf{A}$ is a square matrix. The determinant goes to zero if there exists a dimension not spanned by the basis vectors, whereas $p(\mathbf{A})$ goes to zero if any two basis vectors point in the same direction. So, for a square $\mathbf{A}$, $p(\mathbf{A}) = 0 \Rightarrow \det(\mathbf{A}) = 0$, but not vice versa.

When the previously mentioned assumptions about $\mathbf{A}$ hold, i.e. the basis vectors $\mathbf{a}_i$ are quasi-orthogonal (randomly distributed) and of unit length, the scaling of the probability made by $\mathbf{A}$ is roughly constant. Therefore the probability for $\mathbf{z}$ given $\mathbf{A}$ can be approximated as follows:

$$p(\mathbf{z}(t)|\mathbf{A}) \approx C \prod_{i=1} p_{y_i}(\mathbf{a}_i^T z(t)) \tag{7}$$

where $C$ is a constant and the variable $y_i$ is the dot product between $\mathbf{a}_i$ and $\mathbf{z}$ as in (2), i.e. sum of $\mathbf{s}_i$ and a gaussian variable. In practice, it seems that the exact form of $p_{y_i}$ is not that important, just as long as it is supergaussian, if $y_i$ is supergaussian, and subgaussian, if $y_i$ is subgaussian.

The posterior probability for the problem can be written as

$$p(\mathbf{A}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{A})p(\mathbf{A})}{p(\mathbf{z})} \tag{8}$$

Here $p(\mathbf{z})$ is constant with respect to $\mathbf{A}$. Note, that $p(\mathbf{A})$ now assigns a higher probability to quasi-orthogonal matrices (when $m > 3$), so that the assumption of quasi-orthogonality of the basis vectors holds and the approximation (7) for $p(\mathbf{z}|\mathbf{A})$ can be used. Finally the following log-probability can be written for the posterior:

$$\begin{aligned}\log p(\mathbf{A}|\mathbf{z}(t), t = 1, ..., T) \approx \\ \sum_t \sum_{i=1}^{n} \log p_{y_i}(\mathbf{a}_i^T \mathbf{z}(t)) \\ + \alpha T \sum_{i<j} \log(1 - (\mathbf{a}_i^T \mathbf{a}_j)^2) + \text{const.}\end{aligned} \tag{9}$$

Here $\alpha$ is a constant that is affected not only by $c_m$, but also by the approximations we have made. We adjust $\alpha$ empirically.

In the following, we maximize this posterior to estimate $\mathbf{A}$, and we denote the maximizing argument as $\hat{\mathbf{A}}$. Note that the difference to maximum likelihood estimation of the classic ICA model (i.e. when $\mathbf{A}$ is square) is that $|\det(\mathbf{A})|$ is replaced by $p(\mathbf{A})$. Note also that previously one of the authors proposed a modification of FastICA to perform a similar estimation by quasi-orthogonality [5], but the quasi-orthogonality measure in the present method has been derived from first principles and it seems that the present method estimates more orthogonal bases. In [9] we approximated $\log(p(\mathbf{A}))$ more heuristically with a power function of the dot products, which also seemed to work quite nicely.

### C. Simulations

First, we tried our method on simulated data. We mixed 40 independent components into a 20 dimensional data space, i.e. $\mathbf{A}$ was a matrix of size $20 \times 40$. Note that in such a case, the second term in (2) contributes about one half of the variance of this dot product (after whitening), so the dot products are not very nongaussian. The sample size was 50000. We generated the ICs from Laplacian and uniform distributions, the first of which is supergaussian and the latter subgaussian.

A general problem in estimating overcomplete bases is that components whose contributions to the data are very small (as measured by the variance of the source) are very difficult to estimate. To avoid this problem, the standard deviations of the sources were uniformly distributed between 0.75 and 1.5. The basis vectors were uniformly distributed on the surface of the 20-dimensional unit sphere.

As a preprocessing step, the data was whitened. We then maximized the posterior in (9) by gradient ascent. For data with Laplacian distributed sources the parameter $\alpha$ was set to the value of 0.35 and $p_{y_i}(y) = \frac{1}{cosh(y)}$, and for uniformly distributed sources the parameter was about 1.5 and the distribution $p_{y_i}(y) = \exp(-y^4)$. The scalings needed in the distributions were implicitly included in $\alpha$.

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. These angles are depicted in Fig. 1. Note that all of these angles are above 60 degrees, which shows good quasi-orthogonality. The probability density shown by the solid line in Fig. 1 for comparison gives the distribution that one would expect for these angles if the estimated basis vectors were distributed randomly in the space. One can see that in fact, the obtained vectors are even more orthogonal than corresponding random vectors. Note that even though we generate the mixing matrix randomly, we then whiten the data, which quasi-orthogonalizes the basis vectors by a small amount.

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. 2. We can see that nearly all the components were quite correctly estimated in both cases.

### D. Experiments on image data

Next we tested our method on image feature extraction. We sampled $12 \times 12$ 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 we estimated 300 components, i.e. a basis more than twice overcomplete. We used the same

0-7803-7278-6/02/$10.00 ©2002 IEEE

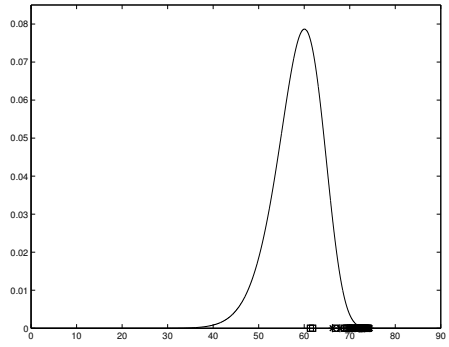

Fig. 1. The quasi-orthogonality of the estimated basis vectors when 40 independent components are mixed into a 20 dimensional space, using the quasi-orthogonalizing prior. Squares: The minimum angles between the estimated basis vectors with Laplacian distributed sources. Asterisks: Uniformly distributed sources. Solid line: Probability density that the minimum angle would have if the vectors were really generated randomly.

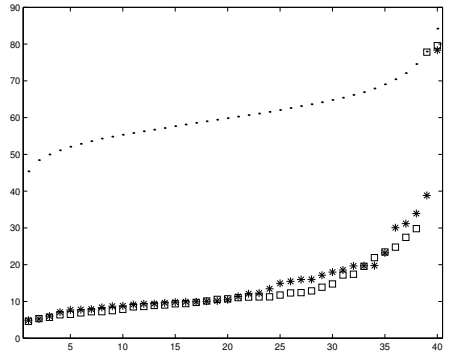

Fig. 2. The angles between the real and matched components, using the quasi-orthogonalization approach. Squares: Laplacian distributed sources. Asterisks: Uniformly distributed sources. Dotted line: The average shape of this graph if the estimates were also drawn randomly.

parameter $\alpha = 0.3$ that we used in the simulations with the Laplacian distributed sources. A supergaussian density was assumed for the independent components by taking $p_{y_i}(y_i) = \frac{1}{\cosh y_i}$.

In Fig. 3, the basis vectors are shown. They are quite similar to what one obtains with ordinary ICA using a supergaussian prior for the independent components. In Fig. 4, we show the angles between the estimated basis vectors in the whitened space; these show that the basis vectors are really quasi-orthogonal.

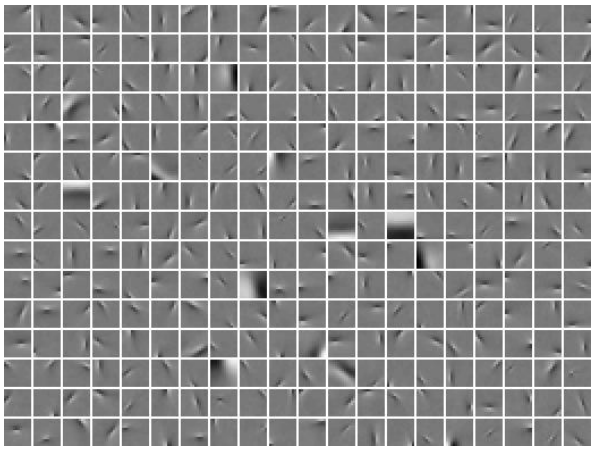

Fig. 3. The basis vectors obtained with the quasi-orthogonalizing prior. The basis vectors are quite similar to those obtained by ordinary ICA, but the basis is more than 2 times overcomplete.

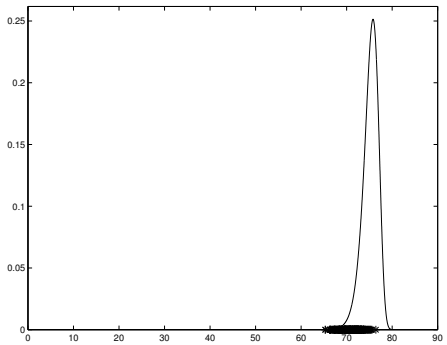

Fig. 4. The minimum angles between the estimated components in the whitened space, using the quasi-orthogonalization approach on image data. Solid line: The distribution this quantity would have, if the vectors were drawn randomly. See caption of Fig. 1.

## III. APPROXIMATE ESTIMATION BY GAUSSIANIZATION

### A. 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. The dot product between the $i$:th basis vector and the data vector is again given by (2). We now wish to estimate the value of the second term in (2) given the value of the dot product. As before, we can assume that this term is approximately gaussian. In a sense the best estimate is then obtained when we monotonically transform the dot product so that it has a gaussian distribution. This is the only transformation that preserves the order of the dot product and produces a gaussian variable.

Assume that we have the $i$:th dot product $y_i = \mathbf{a}_i^T \mathbf{z}$. To gaussianize this direction, we compute the cumulative distribution function, say $F$ of $y_i$. Then we compute for every observation $y_i(t) = \mathbf{a}_i^T \mathbf{z}(t)$ the transform $h(t) = \Phi^{-1}(F(y_i(t)))$, where $\Phi$ 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

$$\mathbf{z}(t) \leftarrow \mathbf{a}_i h(t) + (\mathbf{I} - \mathbf{a}_i \mathbf{a}_i^T)\mathbf{z}(t) \tag{10}$$

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. When using gaussianization, the method does not need to be modified to obtain overcomplete bases. On the other hand, this kind of gaussianization is only applicable in deflationary mode, not in symmetric mode in which the quasi-orthogonalization was used in the preceding section.

### *B. 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 for supergaussian data, and the third power for subgaussian data, 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. 5. All of these angles are above 52 degrees, which shows that we again obtained quite quasi-orthogonal basis vectors. The angles between the original basis vectors and their matched estimates are shown in Fig. 6. Almost all the components were properly estimated with both source distributions. The average error seems to be slightly lower than with the quasi-orthogonalizing prior.

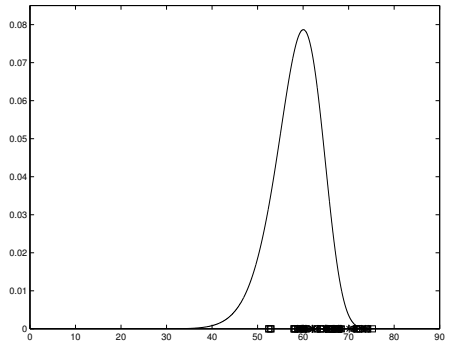


Fig. 5. The quasi-orthogonality of the estimated basis vectors when 40 independent components are mixed into a 20 dimensional space, in the case of the gaussianization method. See caption of Fig. 1.

### *C. 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. 7 we have the obtained basis vectors. These are again similar to those obtained by basic ICA estimation. In Fig. 8 we have the angles between the estimated directions in the whitened space, showing that the basis vectors are quite different from each other.

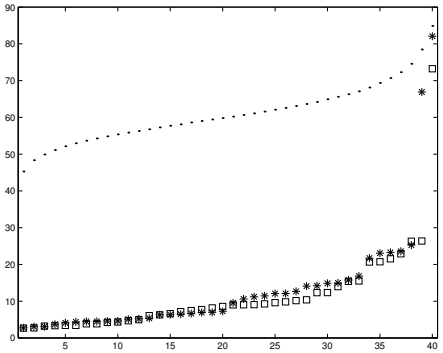

Fig. 6. The angles between the real basis vectors and the matched estimates, for simulated data using the gaussianization procedure. See caption of Fig. 2.

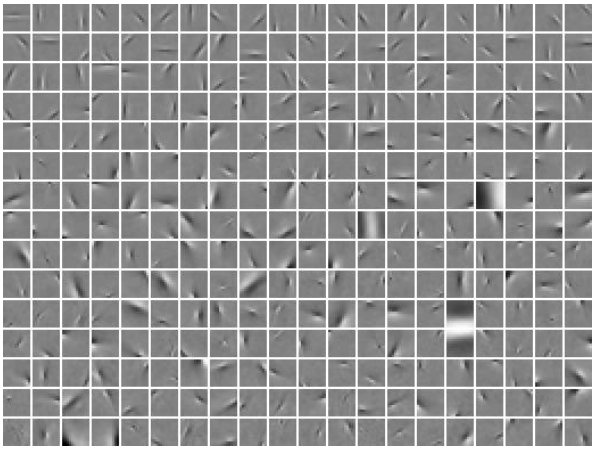

Fig. 7. The image basis vectors obtained using gaussianization. Again, the basis vectors are quite similar to what one obtains with ordinary ICA, but the basis is more than 2 times overcomplete.

## IV. CONCLUSIONS

We introduced two quite fast methods for estimating overcomplete ICA bases in high-dimensional spaces. The first method was based on using a well-chosen prior on the basis vectors, and the second on gaussianization. Simulations and experiments on image data show that the methods work well, thus offering computationally efficient alternatives for overcomplete basis estimation.

### 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.


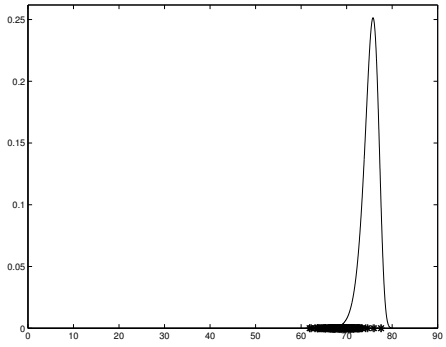

Fig. 8. The angles between the estimated components in the whitened space, for image data and the gaussianization approach. See caption of Fig. 1.


[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] M. Inki and A. Hyvärinen. Two methods for estimating overcomplete independent component bases. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, 2001.

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

[11] 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.

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

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

[14] 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.

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

[16] 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.

[17] 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.