ELSEVIER

Neurocomputing 49 (2002) 151–162

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Imposing sparsity on the mixing matrix in independent component analysis

Aapo Hyvärinen*, Karthikesh Raju

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

Received 15 February 2001; accepted 22 October 2001

Abstract

In independent component analysis, prior information on the distributions of the independent components is often used; some weak information may in fact be necessary for successful estimation. In contrast, prior information on the mixing matrix is usually not used. This is because it is considered that the estimation should be completely blind as to the form of the mixing matrix. Nevertheless, it could be possible to find forms of prior information that are sufficiently general to be useful in a wide range of applications. In this paper, we argue that prior information on the sparsity of the mixing matrix could be a constraint general enough to merit attention. In a biological interpretation, sparseness of mixing matrix means sparse connectivity of the neural network. We show that the computational implementation of such sparsifying priors on the mixing matrix is very simple since in many cases they can be expressed as conjugate priors. The property of being conjugate priors means that essentially the same algorithm can be used as in ordinary ICA. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Independent component analysis; Sparsity; Prior information

1. Introduction

Independent component analysis (ICA) [16] is a statistical model where the observed data is expressed as a linear transformation of latent variables that are non-Gaussian and mutually independent. The classic version of the model can be expressed as

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

* Corresponding author. Tel.: +358-9-451-3278; fax: +358-9-451-3277.
E-mail address: aapo.hyvarinen@hut.fi (A. Hyvärinen).

where $\mathbf{x}=(x_1,x_2,\ldots,x_n)^T$ is the vector of observed random variables, $\mathrm{s}=(s_1,s_2,\ldots,s_n)^T$ is the vector of the independent latent variables (the "independent components"), and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The problem is then to estimate both the mixing matrix $\mathbf{A}$ and the realizations of the latent variables $\mathbf{s}_i$, using observations of $\mathbf{x}$ alone. Exact conditions for the identifiability of the model were given in [7]; the most fundamental is that the independent components s_i must be non-Gaussian [7]. A considerable amount of research has been recently conducted on the estimation of this model, see e.g. [1,2,4–6,8,12].

We thus have some prior knowledge on the distribution of the independent components: they are assumed to be non-Gaussian. Non-Gaussian variables can be roughly divided into two groups: super-Gaussian and sub-Gaussian variables, although slightly different definitions exist. In many cases, it is further assumed that we know the nature of the independent components: whether they are sub-Gaussian or super-Gaussian. This is the case, for example, in image feature extraction [19,3,9], in which the components are usually assumed to be super-Gaussian, or sparse. This is not an arbitrary assumption, but a simple consequence of the fact that the independent components estimated from image data are super-Gaussian with few exceptions.

On the other hand, no prior knowledge on the mixing matrix [15] is used in the basic ICA model. This has the advantage of giving the model great generality. In many application areas, however, information on the form of the mixing matrix is available. Using prior information on the mixing matrix is likely to give better estimates of the matrix for a given number of data points. This is of great importance in situations where the computational costs of ICA estimation are so high that they severely restrict the amount of data that can be used, as well as in situations where the amount of data is restricted due to the nature of the application.

This situation can be compared to that found in regression, where overlearning is a very general phenomenon. The classical way of avoiding overlearning in regression, i.e. overfitting, is to use of regularizing priors, which typically penalize regression functions that have large curvatures, i.e. lots of "wiggles". This makes it possible to use regression methods even when the number of parameters in the model is very large compared to the number of observed data points. In the extreme theoretical case, the number of parameters in infinite, but the model can still be estimated from finite amounts of data by using prior information. Thus suitable priors can reduce overlearning [14].

One example of using prior knowledge that predates modern ICA methods is the literature of beamforming (see the discussion in [5]), where a very specific form of the mixing matrix is represented by a small number of parameters. In investigations on application of ICA to magnetoencephalography [22], it has been found that the independent components can be modelled by the classic dipole model, an information that could be used to constrain the form of the mixing coefficients [17]. The problem with these methods is, however, that they may be applicable to a few data sets only, and lose the generality that is one of the main factors in the current flood of interest in ICA.

In this paper, we introduce a form of prior information on the mixing matrix that is both general enough to be used in many applications and strong enough to increase the performance of ICA estimation. First we investigate the possibility of using two simple classes of priors for the mixing matrix $\mathbf{A}$: Jeffreys' prior and quadratic priors. We come to the conclusion that these two classes are not very useful in ICA. Then we introduce the concept of sparse priors. These are priors that enforce a sparse structure on the mixing matrix. In other words, the prior penalizes mixing matrices with a large number of significantly non-zero entries. Thus this form of prior is similar to the prior knowledge on the sparseness of the independent components. In fact, due to this similarity, sparse priors are so-called conjugate priors, which implies that estimation using this kind of priors is particularly easy: Ordinary ICA methods can be simply adapted to using such priors. Sparse priors are particularly useful in image feature extraction, where a link to sparsely connected networks can be made.

Preliminary results were reported in [11].

2. Background: Jeffreys' and quadratic priors

In the following, we assume that the estimator $\mathbf{B}$ of the inverse of the mixing matrix $\mathbf{A}$ is constrained so that the estimates of the independent components $\mathbf{y}=\mathbf{Bx}$ are *white*, i.e. decorrelated and of unit variance: $E\{\mathbf{y}\mathbf{y}^T\}=\mathbf{I}$. This restriction facilitates greatly the analysis. For its justification, see e.g. [7,12]. We concentrate here on formulating priors for $\mathbf{B}=\mathbf{A}^{-1}$. Completely analogue results hold for priors on $\mathbf{A}$.

2.1. Jeffreys' prior

The classical prior in Bayesian inference is Jeffreys' prior. It is considered a maximally uninformative prior, which already indicates that it is probably not useful for our purpose.

Indeed, it was shown in [20] that Jeffreys' prior has the form:

$$p(\mathbf{B}) \propto |\det \boldsymbol{B}^{-1}|. \tag{2}$$

Now, the constraint of whiteness of the $\mathbf{y}=\mathbf{Bx}$ means that $\mathbf{B}$ can be expressed as $\mathbf{B}=\mathbf{UV}$, where $\mathbf{V}$ is a constant matrix, and $\mathbf{U}$ is restricted to be orthogonal. But we have $\det \mathbf{B} = \det \mathbf{U} \det \mathbf{V} = \det \mathbf{V}$, which implies that Jeffreys's prior is constant in the space of allowed estimators (i.e. decorrelating $\mathbf{B}$). Thus we see that Jeffreys' prior has no effect on the estimator, and therefore cannot reduce overlearning.

2.2. Quadratic priors

In regression, the use of quadratic regularizing priors is very common. It would be tempting to try to use the same idea in the context of ICA. Especially in feature extraction, we could require the columns of $\mathbf{A}$, i.e. the features, to be smooth in the same sense as smoothness is required of regression functions. In other words, we could consider every column of $\mathbf{A}$ as a discrete approximation of a smooth function, and choose a prior that imposes smoothness for the underlying continuous function. Similar arguments hold for priors defined on the rows of $\mathbf{B}$, i.e. the filters corresponding to the features.

The simplest class of regularizing priors is given by quadratic priors. We will show here, however, that such quadratic regularizers, at least the simple class that we define below, do not change the estimator.

Consider priors that are of the form

$$\log p(\mathbf{B}) = \sum_{i=1}^{n} \mathbf{b}_i^{\mathrm{T}} \mathbf{M} \mathbf{b}_i + const., \tag{3}$$

where the $\mathbf{b}_i^{\mathrm{T}}$ are the rows of $\mathbf{B} = \mathbf{A}^{-1}$, and $\mathbf{M}$ is a matrix that defines the quadratic prior. For example, for $\mathbf{M} = \mathbf{I}$ we have a "weight decay" prior $\log p(\mathbf{B}) = \sum_i \|\mathbf{b}_i\|^2$. Alternatively, we could include in $\mathbf{M}$ some differential operators so that the prior would measure the "smoothnesses" of the $\mathbf{b}_i$, in the sense explained above. The prior can be manipulated algebraically to yield

$$\sum_{i=1}^{n} \mathbf{b}_i^{\mathrm{T}} \mathbf{M} \mathbf{b}_i = \sum_{i=1}^{n} \mathrm{tr}(\mathbf{M} \mathbf{b}_i \mathbf{b}_i^{\mathrm{T}}) = \mathrm{tr}(\mathbf{M} \mathbf{B}^{\mathrm{T}} \mathbf{B}). \tag{4}$$

Quadratic priors have little significance in ICA estimation, however. To see this, let us constrain the estimates of the independent components to be white as above. This means that we have

$$E\{\mathbf{y}\mathbf{y}^{\mathrm{T}}\} = E\{\mathbf{B}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\} = \mathbf{B}\mathbf{C}\mathbf{B}^{\mathrm{T}} = \mathbf{I} \tag{5}$$

in the space of allowed estimates, which gives after some algebraic manipulations $\mathbf{B}^{\mathrm{T}}\mathbf{B} = \mathbf{C}^{-1}$. Now we see that

$$\sum_{i=1}^{n} \mathbf{b}_i^{\mathrm{T}} \mathbf{M} \mathbf{b}_i = \mathrm{tr}(\mathbf{M}\mathbf{C}^{-1}) = const. \tag{6}$$

In other words, the quadratic prior is constant. The same result can be proven for a quadratic prior on $\mathbf{A}$. Thus, quadratic priors seem to be of little interest in ICA.

3. Sparse priors on the mixing matrix

3.1. Motivation

A much more satisfactory class of priors is given by what we call sparse priors. This means that the prior information says that most of the elements of each row of $\mathbf{B}$ are zero. The motivation for considering sparse priors is both empirical and algorithmic.

Empirically, it has been observed in feature extraction of images that the obtained filter tend to be localized in space. This implies that the distribution of the elements b_{ij} of the filter $\mathbf{b}_i$ tends to be sparse, i.e. most elements are practically zero. A similar phenomenon can be seen in analysis of magnetoencephalography, where each source signal is usually captured by a limited number of sensors. This is due to the spatial localization of the sources and the sensors.

The algorithmic appeal of sparsifying priors, on the other hand, is based on the fact that sparse priors can be made to be conjugate priors. This is a special class of priors, and means that estimation of the model using this prior requires only very simple modifications in ordinary ICA algorithms.

Another motivation for sparse priors is their neural interpretation. Biological neural networks are known to be sparsely connected, i.e. only a small proportion of all possible connections between neurons are actually used. This is exactly what sparse priors model. This interpretation is especially interesting when ICA is used in modelling of the visual cortex [3,10,19].

3.2. Measuring sparsity of mixing matrix

Sparsity of a random variable, say s, can be measured by expectations of the form $E\{G(s)\}$, where G is a non-quadratic function, for example the following

$$G(s) = -|s|. \tag{7}$$

The use of such measures requires that the variance of s is normalized to a fixed value, and its mean is zero.

Let us assume that the data $\mathbf{x}$ is *whitened* as a preprocessing step (this assumption will be discussed in detail below). This means that the data is linearly transformed into $\mathbf{z} = \mathbf{V}\mathbf{x}$ so that the covariance of the whitened data equals identity: $E\{\mathbf{z}\mathbf{z}^{\mathrm{T}}\} = \mathbf{I}$. Denote by $\mathbf{W}$ the separating matrix applied to the whitened data.

Now, constraining the estimates $\mathbf{y} = \mathbf{W}\mathbf{z}$ of the independent components to be white implies that $\mathbf{W}$ is orthogonal, which implies that the sum of the squares of the elements $\sum_j w_{ij}$ is equal to one for every i. The elements of each row of $\mathbf{W}$ can be then considered a realization of a random variable of zero mean and unit variance. This means we could measure the sparsities of the rows of $\mathbf{W}$ using a sparsity measure of the form (7).

Thus, we can define a sparse prior of the form

$$\log p(\mathbf{W}) = \sum_{i=1}^{n} \sum_{j=1}^{n} G(w_{ij}) + const., \tag{8}$$

where G is the logarithm of some super-Gaussian density function (up to some additive constant), and again $\mathbf{w}_i^{\mathrm{T}} = (w_{i1}, \ldots, w_{in})$ are the rows of $\mathbf{A}^{-1}$. The function G in (7) is such a log-density, so we see that we have here a measure of sparsity of the $\mathbf{w}_i$.

The prior in (8) has the nice property of being a conjugate prior. Let us assume that the *independent components are super-Gaussian*, and for simplicity, let us further assume that they have identical distributions, with log-density G. Now we can take that same log-density as the log-prior density G in (8). Then we can write the prior in the form

$$\log p(\mathbf{W}) = \sum_{i=1}^{n} \sum_{j=1}^{n} G(\mathbf{w}_i^{\mathrm{T}} \mathbf{e}_j) + const., \tag{9}$$

where we denote the canonical basis vectors by $\mathbf{e}_i$, i.e. the ith element of $\mathbf{e}_i$ is equal to one, and all the others are zero.

Assume now that we have T whitened observations $\mathbf{z}(t)$, $t = 1,\ldots,T$. The likelihood of $\mathbf{W}$ that is constrained to be orthogonal is simply given by [13]

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{i=1}^{n} G(\mathbf{w}_i^{\mathrm{T}} \mathbf{z}(t)) + const. \tag{10}$$

Thus the posterior distribution has the form:

$$\log p(\mathbf{W}|\mathbf{z}) = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} G(\mathbf{w}_i^{\mathrm{T}} \mathbf{z}(t)) + \sum_{j=1}^{n} G(\mathbf{w}_i^{\mathrm{T}} \mathbf{e}_j) \right] + const. \tag{11}$$

This form shows that the posterior distribution has the same form as the prior distribution (and, in fact, the original likelihood). Priors with this property are called conjugate priors in Bayesian theory. The usefulness of conjugate priors resides in the property that the prior can be considered to correspond to a “virtual” sample. The posterior distribution in (11) has the same form as the likelihood of a sample of size $T + n$ which consists of both the observed $\mathbf{z}(t)$ and the canonical basis vectors $\mathbf{e}_i$. In other words, the posterior in (11) is the likelihood of the augmented (whitened) data sample

$$\mathbf{z}^*(t) = \begin{cases} \mathbf{z}(t) & \text{if } 1 \leqslant t \leqslant T \\ \mathbf{e}_{t-T} & \text{if } T < t \leqslant T + n. \end{cases} \tag{12}$$

Thus, using conjugate priors has the additional benefit that we can use exactly the same algorithm for maximization of the posterior as in ordinary maximum likelihood estimation of ICA. All we need to do is to add this virtual sample to the data; the virtual sample is of same size n as the dimension of the data.

Note that we could also have sub-Gaussian (anti-sparse) conjugate priors just in the same way as sparse priors. The property of being conjugate is more general than what we use in this paper. However, the utility of non-sparse conjugate priors is not as obvious as the utility of sparse priors.

3.3. Modifying prior strength

The conjugate priors given above can be generalized by considering a family of super-Gaussian priors given by

$$\log p(\mathbf{W}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha G(\mathbf{w}_i^{\mathrm{T}} \mathbf{e}_j) + const. \tag{13}$$

Using this kind of prior means that the virtual sample points are weighted by some parameter α. This parameter expresses the degree of belief that we have in the prior. A large α means that the belief in the prior is strong. Also, the parameter α could be different for different i, but this seems less useful here. The posterior distribution has then the form:

$$\log p(\mathbf{W}|\mathbf{z}) = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} G(\mathbf{w}_i^{\mathrm{T}} \mathbf{z}(t)) + \sum_{j=1}^{n} \alpha G(\mathbf{w}_i^{\mathrm{T}} \mathbf{e}_j) \right] + const. \tag{14}$$

The above expression can be further simplified in the case where the assumed density of the independent components is Laplacian, i.e. $G(y) = -|y|$. In this case, the α can multiply the $\mathbf{e}_j$ themselves:

$$\log p(\mathbf{W}|\mathbf{z}) = \sum_{i=1}^{n} \left[\sum_{t=1}^{T} |\mathbf{w}_i^{\mathrm{T}} \mathbf{z}(t)| - \sum_{j=1}^{n} |\mathbf{w}_i^{\mathrm{T}}(\alpha \mathbf{e}_j)| \right] + const., \tag{15}$$

which is simpler than (14) from the algorithmic viewpoint: It amounts to the addition of just n virtual data vectors of the form $\alpha\mathbf{e}_j$ to the data. This avoids all complications due to the differential weighting of sample points in (14), and ensures that any conventional ICA algorithm can be used by simply adding the virtual sample to the data. In fact, the Laplacian prior is most often used in ordinary ICA algorithms, sometimes in the form of the log cosh function that can be considered as a smoother approximation of the absolute value function.

The estimation of a suitable α is a further problem that could presumably be solved by Bayesian methods. In this paper, we simply try different values for α's (see Section 5) to find a good one.

3.4. Priors and whitening

Above, we assumed that the data is preprocessed by whitening. It should be noted that the effect of the sparse prior is dependent on the whitening matrix. This is because sparseness is imposed on the separating matrix of the whitened data, and the value of this matrix depends on the whitening matrix. There is an infinity of whitening matrices, so imposing sparseness on the whitened separating matrix may have different meanings.

On the other hand, it is not necessary to whiten the data. The above framework can be used for non-white data as well. If the data is not whitened, the meaning of the sparse prior is somewhat different, though. This is because every row of $\mathbf{b}_i$ is not constrained to have unit norm for general data. Thus our measure of sparsity does not anymore measure the sparsities of each $\mathbf{b}_i$. On the other hand, the developments of the preceding section show that the sum of squares of the whole matrix $\sum_{ij} b_{ij}$ does stay constant. This means that the sparsity measure is now rather measuring the global sparsity of $\mathbf{B}$, instead of the sparsities of individual rows.

In practice, one usually wants to whiten the data for technical reasons. Then the problems arises: How to impose the sparseness on the original separating matrix even when the data used in the estimation algorithm needs to be whitened? The above framework can be easily modified so that the sparseness is imposed on the original separating matrix. Denote by $\mathbf{V}$ the whitening matrix and by $\mathbf{B}$ the separating matrix for original data. Thus, we have $\mathbf{WV} = \mathbf{B}$ and $\mathbf{z} = \mathbf{Vx}$ by definition. Now, we can express the prior in (9) as

$$\log p(\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{n} G(\mathbf{b}_i^{\mathrm{T}} \mathbf{e}_j) + const. = \sum_{i=1}^{n} \sum_{j=1}^{n} G(\mathbf{w}_i^{\mathrm{T}}(\mathbf{V}\mathbf{e}_j)) + const. \tag{16}$$

Thus, we see that the virtual sample added to $\mathbf{z}(t)$ now consists of the columns of the whitening matrix, instead of the identity matrix.

Incidentally, a similar manipulation of (9) shows how to put the prior on the original mixing matrix instead of the separating matrix. We always have $\mathbf{VA} = \mathbf{W}^{-1} = \mathbf{W}^{\mathrm{T}}$. Thus, we obtain $\mathbf{a}_i^{\mathrm{T}}\mathbf{e}_j = \mathbf{a}_i^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}(\mathbf{V}^{-1})^{\mathrm{T}}\mathbf{e}_j = \mathbf{w}_i^{\mathrm{T}}(\mathbf{V}^{-1})^{\mathrm{T}}\mathbf{e}_j$. This shows that imposing a sparse prior on $\mathbf{A}$ is done by using the virtual sample given by the rows of the inverse of the whitening matrix. (Note that for whitened data, the mixing matrix is the transpose of the separating matrix, so the fourth logical possibility of formulating prior for the whitened mixing matrix is not different from using a prior on the whitened separating matrix.)

In practice, the problems implied by whitening can often be solved by using a whitening matrix that is sparse in itself. Then imposing sparseness on the whitened separating matrix is meaningful. In the context of image feature extraction, a sparse whitening matrix is obtained by the zero-phase whitening matrix (see [3] for discussion), for example. Then it is natural to impose the sparseness for the whitened separating matrix, and the complications discussed in this subsection can be ignored.

4. Connection to spatiotemporal ICA

When using sparse priors, we actually make rather similar assumptions on both the independent components and the mixing matrix. Both are assumed to be generated so that the values are taken from independent, typically sparse, distributions. In the limit, we might develop a model where the very same assumptions are made on the mixing matrix and the independent components. Such a model was introduced in [21], independently from us, and called "spatiotemporal" ICA since in a way, it does ICA both in the temporal domain (if the independent components are considered time signals), and in the spatial domain (which corresponds to the spatial mixing defined by the mixing matrix).

In spatiotemporal ICA, the distinction between independent components and the mixing matrix is completely abolished. To see why this is possible, consider the data as a single matrix of the observed vectors as its columns: $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$, and likewise for the independent components. Then the ICA model can be expressed as

$$\mathbf{X} = \mathbf{AS}. \tag{17}$$

Now, taking a transpose of this equation, we obtain

$$\mathbf{X}^{\mathrm{T}} = \mathbf{S}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}. \tag{18}$$

Now we see that the matrix $\mathbf{S}$ is like a mixing matrix, with $\mathbf{A}$ giving the realizations of the "independent components". Thus, by taking the transpose, we flip the roles of the mixing matrix and the independent components.

In the basic ICA model, the difference between $\mathbf{s}$ and $\mathbf{A}$ is due to the statistical assumptions made on $\mathbf{s}$, which are the independent random variables, and on $\mathbf{A}$, which is a constant matrix of parameters. But with sparse priors, we made assumptions on $\mathbf{A}$ that are very similar to those usually made on $\mathbf{s}$. So, we can simply consider both $\mathbf{A}$ and $\mathbf{S}$ as being generated by independent random variables, in which case either one of the mixing equations (with or without transpose) are equally valid. This is the basic idea in spatiotemporal ICA.

There is an important difference between **S** and **A**, though. The dimensions of **A** and **S** are typically very different: **A** is square whereas **S** has many more columns than rows. This difference can be abolished in spatiotemporal ICA by considering that there **A** has much less columns than rows, that is, there is some redundancy in the signal. In this paper, however, we consider **A** to be square, which makes our model different from spatiotemporal ICA.

The estimation of the spatiotemporal ICA model can be performed in a manner somewhat similar to using sparse priors. In [21], it was proposed to form something similar to a virtual sample where the data consists of two parts, the original data and the data obtained by transposing the data matrix. The dimensions of these data sets must be strongly reduced and made equal to each other by principal component analysis or related methods. This is possible because it is assumed that both **A** and $\mathbf{S}^{\mathrm{T}}$ have the same kind of redundancy: much more rows than columns. In [21] the infomax criterion [2,18] was applied on this estimation task. The exact connection between estimation of spatiotemporal ICA and our sparse priors is an important problem for future work.

5. Experiments

We performed experiments in image feature extraction to explore the applicability of sparse priors.

The basic idea is as in [3,19,9]. The data was obtained by taking 20 × 20 pixel image patches at random locations from monochrome photographs depicting wild-life scenes (animals, meadows, forests, etc.). The patches were normalized to unit norm. The data was whitened by the zero-phase whitening filter, which means multiplying the data by $\mathbf{C}^{-1/2}$, where **C** is the covariance of the data (see e.g. [3]). In the results shown above, the inverse of these preprocessing steps was performed.

The sample size was fixed at 20 000. This is insufficient for such a large window size. The estimated basis vectors are shown in Fig. 2 (For reasons of space, only 200 of the 400 basis vectors are shown; these were randomly selected). Using prior information with the parameter α fixed at 25, we obtained a much better basis. This basis is shown in Fig. 3. Visually, one sees that the features are much better. The features are oriented, as well as localized in space and in frequency. The aspect of multiresolution seems to be less developed than in results with sufficient data [3,19], however.

To validate the prior quantitatively, we computed the sparsities of the bases corresponding to different values of the parameter α, which correspond to different strengths given to the prior information. The sparsity is here measured as the (negative) expectation of the absolute value of the estimated independent components: this is essentially an approximation of the likelihood. The sparsity was measured using a test set that was separate from the training set used in learning the basis vectors. This is plotted in Fig. 1. The values of sparsity can be seen to increase with increasing α, i.e. increasing strength placed on prior information. At a certain value, the sparsity has a maximum and starts decreasing. This is natural because too large a value for α means that only prior information is used, and the data is neglected.

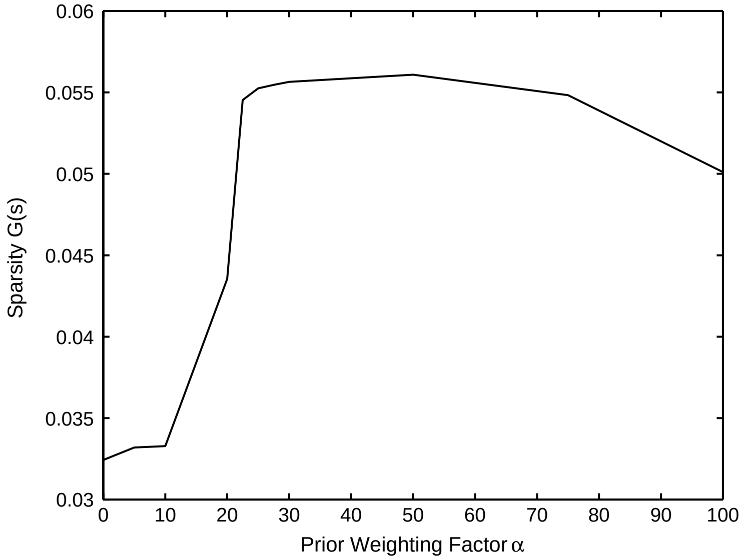

Fig. 1. Sparsities as function of prior information strength α. A suitable value for α gives sparser components than ordinary ICA.

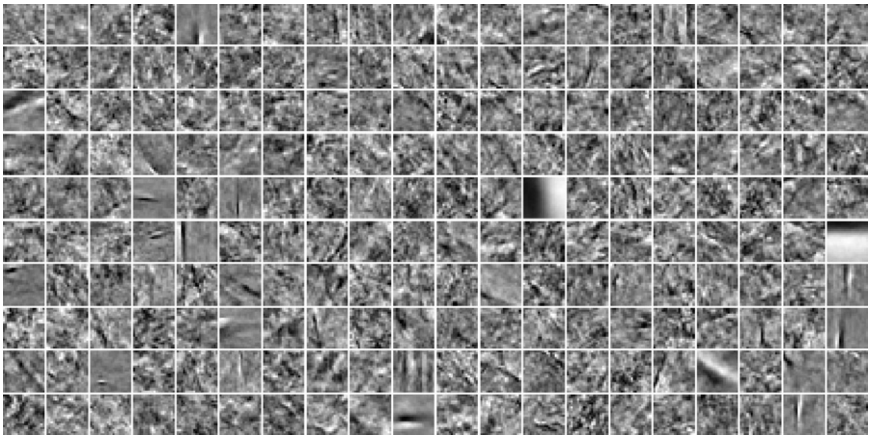

Fig. 2. Estimation of the image features with no prior information. The sample size was insufficient to give useful estimates.

6. Conclusion

We introduced sparse priors on the mixing matrix. We argued that such priors may be useful in a wide area of applications. Computationally they are very convenient because they are conjugate priors, which means that many existing ICA algorithms can

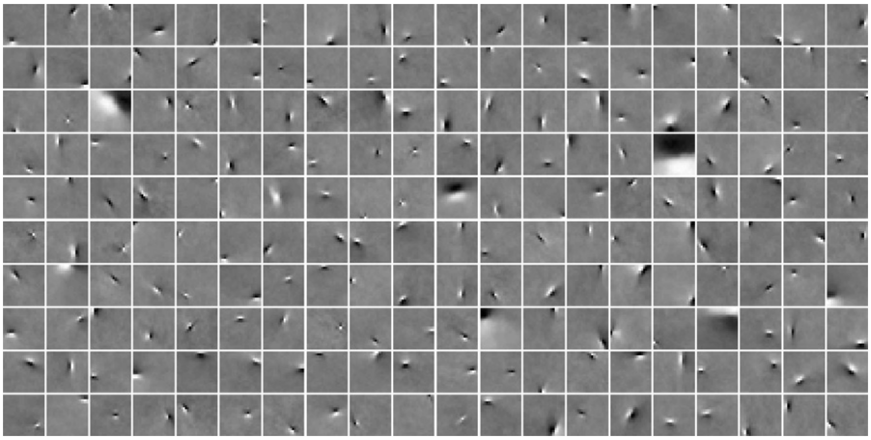

Fig. 3. Estimation of the image features with suitable prior information. The estimation was successful even with this small sample size.

be directly used by simply introducing a virtual sample. Experiments show that sparse priors can be successfully used in image feature extraction.

References

[1] S.-I. Amari, A. Cichocki, H.H. Yang, A new learning algorithm for blind source separation, Advances in Neural Information Processing Systems, Vol. 8, MIT Press, Cambridge, MA, 1996, pp. 757–763.

[2] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] A.J. Bell, T.J. Sejnowski, The 'independent components' of natural scenes are edge filters, Vision Res. 37 (1997) 3327–3338.

[4] J.-F. Cardoso, B. Hvam Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. 44 (12) (1996) 3017–3030.

[5] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian signals, IEE Proc.-F 140 (6) (1993) 362–370.

[6] A. Cichocki, R. Unbehauen, Robust neural networks with on-line learning for blind identification and blind separation of sources, IEEE Trans. Circuits and Systems 43 (11) (1996) 894–906.

[7] P. Comon, Independent component analysis—a new concept? Signal Process. 36 (1994) 287–314.

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

[9] A. Hyvärinen, Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation, Neural Comput. 11 (7) (1999) 1739–1768.

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

[11] A. Hyvärinen, R. Karthikesh, Sparse priors on the mixing matrix in independent component analysis, Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000), Helsinki, Finland, 2000, pp. 477–482.

[12] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (7) (1997) 1483–1492.

[13] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.

[14] A. Hyvärinen, J. Särelä, R. Vigário, Spikes and bumps: artefacts generated by independent component analysis with insufficient sample size, Proceedings of the International Workshop on Independent Component Analysis and Signal Separation (ICA'99), Aussois, France, 1999, pp. 425–429.

[15] J. Igual, L. Vergara, Prior information about mixing matrix in BSS-ICA formulation, Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000), Helsinki, Finland, 2000, pp. 123–128.

[16] C. Jutten, J. Hérault, Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture, Signal Process. 24 (1991) 1–10.

[17] H. Knuth, A bayesian approach to source separation, Proceedings of the International Workshop on Independent Component Analysis and Signal Separation (ICA'99), Aussois, France, 1999, pp. 283–288.

[18] J.-P. Nadal, N. Parga, Non-linear neurons in the low noise limit: a factorial code maximizes information transfer, Network 5 (1994) 565–581.

[19] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.

[20] P. Pajunen, Blind source separation using algorithmic information theory, Neurocomputing 22 (1998) 35–48.

[21] J.V. Stone, J. Porrill, C. Buchel, K. Friston, Spatial, temporal, and spatiotemporal independent component analysis of fMRI data, in: R.G. Aykroyd, K.V. Mardia, I.L. Dryden (Eds.), Proceedings of the 18th Leeds Statistical Research Workshop on Spatial-Temporal Modelling and its Applications, Leeds University Press, Leeds, 1999, pp. 23–28.

[22] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, E. Oja, Independent component analysis for identification of artifacts in magnetoencephalographic recordings, Advances in Neural Information Processing Systems, Vol. 10, MIT Press, Cambridge, MA, 1998, pp. 229–235.

Aapo Hyvärinen studied mathematics and statistics at the universities of Helsinki (Finland), Vienna (Austria), and Paris (France), obtaining his Master's degree at the University of Paris-Dauphine in 1994. Subsequently, he obtained a Ph.D. degree at the Laboratory of computer and information science of the Helsinki University of Technology in 1997. After post-doc work at the Neural Network Research Center of the Helsinki University of Technology and a civilian service at the Department of Psychology of the University of Helsinki, he was appointed an Academy Research Fellow in 2001, positioned at the Neural Networks Research Centre. His research interests include computational neuroscience, the visual system, and independent component analysis.

Karthikesh Raju obtained a Bachelor of Engineering in Electronics and Communication from Bharathiar University, India, in 1997 and his M.Sc. in Communication Technology from the University of Ulm, Ulm, Germany in 2000. He is now a graduate student at the Laboratory of Computer and Information Science of the Helsinki University of Technology. His research interests are in the application of ICA to Detector Structures for CDMA.