SPARSE PRIORS ON THE MIXING MATRIX IN INDEPENDENT COMPONENT ANALYSIS

Aapo Hyvärinen and Raju Karthikesh

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland
http://www.cis.hut.fi/projects/ica/

ABSTRACT

In independent component analysis, prior information on the distributions of the independent components is often used; some weak information is in fact necessary for succesful 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. Moreover, 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.

1. INTRODUCTION

Indendent component analysis (ICA) [13] is a statistical model where the observed data is expressed as a linear transformation of latent variables that are nongaussian and mutually independent. The classic version of the model can be expressed as

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

where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ 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 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 nongaussian [7]. A considerable amount of research has been recently conducted on the estimation of this model, see e.g. [1, 2, 4, 5, 6, 8, 11].

We thus have some prior knowledge on the distribution of the independent components: they are assumed to be nongaussian. Nongaussian variables can be roughly divided into two groups: supergaussian and subgaussian variables, although slightly different definitions exist. In many cases, it is further assumed that we know if the independent components are sub- or supergaussian. This is the case, for example, in image feature extraction [15, 3, 9], in which the components are assumed to be supergaussian, or sparse. This is not an arbitrary assumption, but a simple consequence of the fact that the independent components estimated from image data are supergaussian with very few exceptions.

On the other hand, no prior knowledge on the mixing matrix 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 as 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 [12].

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 paramters. In investigations on application of ICA to magnetoencephalogaphy [17], 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 [14]. 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 larger 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.

2. BACKGROUND: JEFFREYS' AND QUADRATIC PRIORS

In the following, we assume that the estimator $\mathbf{W}$ of the inverse of the mixing matrix $\mathbf{A}$ is constrained so that the estimates of the independent components $\mathbf{y} = \mathbf{W}\mathbf{x}$ are *white*, i.e. decorrelated and of unit variance: $\mathbf{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}$. This restriction facilitates greatly the analysis. For its justification, see e.g. [7, 11]. We concentrate here on formulating priors for $\mathbf{W} = \mathbf{A}^{-1}$. Completely analogue results hold for prior 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 [16] that Jeffreys' prior has the form:

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

Now, the constraint of whiteness of the $\mathbf{y} = \mathbf{W}\mathbf{x}$ means that $\mathbf{W}$ can be expressed as $\mathbf{W} = \mathbf{U}\mathbf{B}$, where $\mathbf{B}$ is a constant matrix, and $\mathbf{U}$ is restricted to be orthogonal. But we have $\det \mathbf{W} = \det \mathbf{U} \det \mathbf{B} = \det \mathbf{B}$, which implies that Jeffreys's prior is constant in the space of allowed estimators (i.e. decorrelating $\mathbf{W}$). 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 $\mathbf{W}$, 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

$$\log p(\mathbf{W}) = \sum_{i=1}^{n} \mathbf{w}_i^T \mathbf{M} \mathbf{w}_i + const. \tag{3}$$

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

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

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

$$E\{\mathbf{y}\mathbf{y}^T\} = E\{\mathbf{W}\mathbf{x}\mathbf{x}^T\mathbf{W}^T\} = \mathbf{W}\mathbf{C}\mathbf{W}^T = \mathbf{I} \quad (5)$$

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

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

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

3. SPARSE PRIORS

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 $\mathbf{W}$ 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 w_{ij} of the filter $\mathbf{w}_i$ 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 [15, 3, 10].

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

$$G(s) = -|s|. \quad (7)$$

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

In feature extraction and probably several other applications as well, the distribution of the elements of $\mathbf{W}$ is zero-mean due to symmetry. Furthermore, let us assume that the data $\mathbf{x}$ is whitened as a preprocessing step. Denote by $\mathbf{z}$ the whitened data vector whose components are thus uncorrelated and have unit variance. 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

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

where G is the logarithm of some supergaussian density function, and again $\mathbf{w}_i^T = (w_{i1}, ..., w_{in})$ are the rows of $\mathbf{A}^{-1}$. The function G in (7) is such 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 supergaussian, 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

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

where we denote by $\mathbf{e}_i$ the canonical basis vectors, i.e. the i-th element of $\mathbf{e}_i$ is equal to one, and all the others are zero. Thus the posterior distribution has the form:

$$\log p(\mathbf{W}|\mathbf{x}) = \sum_{i=1}^{n}[\sum_{t=1}^{T} G(\mathbf{w}_i^T\mathbf{x}(t)) + \sum_{j=1}^{n} G(\mathbf{w}_i^T\mathbf{e}_j)] + const. \quad (10)$$

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 (10) 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 (10) is the likelihood of the augmented (whitened) data sample

$$\mathbf{z}^*(t) = \begin{cases} \mathbf{z}(t), & \text{if } 1 \le t \le T \\ \mathbf{e}_{t-T}, & \text{if } T < t \le T+n. \end{cases} \quad (11)$$

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.

3.3. Modifying prior strength

The conjugate priors given above can be generalized by considering a family of supergaussian priors given by

$$\log p(\mathbf{A}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha G(\mathbf{w}_i^T \mathbf{e}_j) + const. \quad (12)$$

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:

$$\log p(\mathbf{W}|\mathbf{x}) = \sum_{i=1}^{n} [\sum_{t=1}^{T} G(\mathbf{w}_i^T \mathbf{x}(t)) + \sum_{j=1}^{n} \alpha G(\mathbf{w}_i^T \mathbf{e}_j)] + const. \quad (13)$$

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:

$$\log p(\mathbf{W}|\mathbf{x}) = \sum_{i=1}^{n} [\sum_{t=1}^{T} |\mathbf{w}_i^T \mathbf{x}(t)| - \sum_{j=1}^{n} |\mathbf{w}_i^T (\alpha \mathbf{e}_j)|] + const. \quad (14)$$

which is simpler than (13) 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 (13), 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.

3.4. Whitening and priors

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 whitening matrix may have different meanings.

In practice, this problem can 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.

On the other hand, it is not necessary to whiten the data. If the data is not whitened, the meaning of the sparse prior is somewhat different, though. This is because every row of $\mathbf{w}_i$ is not constrained to have unit norm for general data. Thus our measure of sparsity does not correctly measure the sparsities of each $\mathbf{w}_i$. On the other hand, the developments of the preceding section show that the sum of squares of the whole matrix $\sum_{ij} w_{ij}$ does stay constant. This means that the sparsity measure is now measuring the *global* sparsity of $\mathbf{W}$, instead of the sparsities of individual rows.

4. EXPERIMENTS

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

The basic idea is as in [3, 15, 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 $\mathbf{C}$ 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

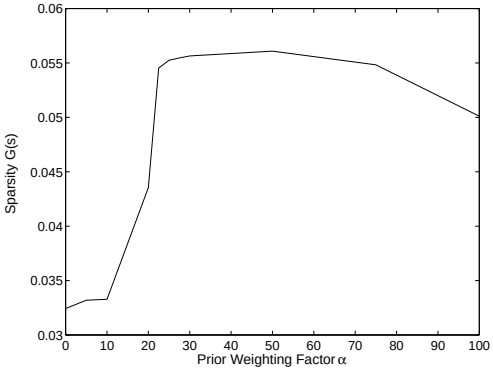

Figure 1: Sparsities as function of prior information strength α. A suitable value for α gives sparser components than ordinary ICA.

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.

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.

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

6. REFERENCES

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

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

[3] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

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

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

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

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

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

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

[10] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 2000. (in press).

[11] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[12] A. Hyvärinen, J. Särelä, and R. Vigário. Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 425–429, Aussois, France, 1999.

[13] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[14] H. Knuth. A bayesian approach to source separation. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 283–288, Aussois, France, 1999.

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

[16] P. Pajunen. Blind source separation using algorithmic information theory. *Neurocomputing*, 22:35–48, 1998.

[17] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems 10*, pages 229–235. MIT Press, 1998.

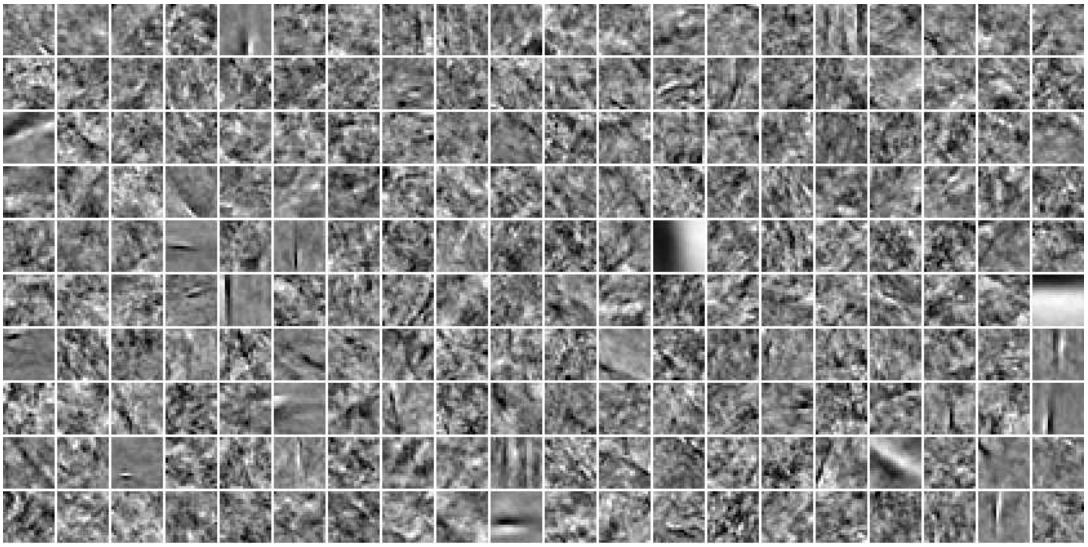

Figure 2: Estimation of the image features with no prior information. The sample size was insufficient to give useful estimates.

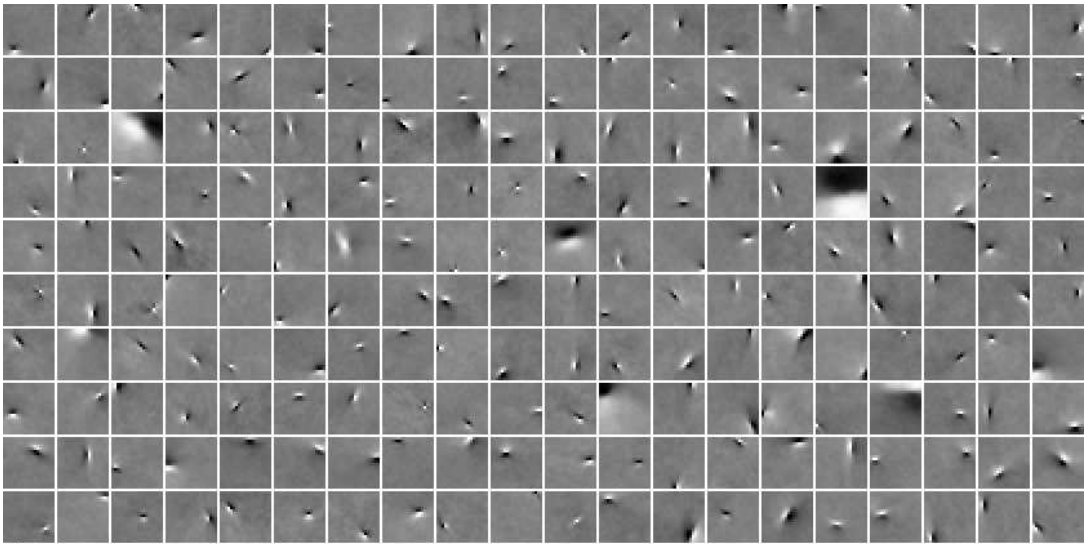

Figure 3: Estimation of the image features with suitable prior information. The estimation was succesful even with this small sample size.