LOCAL INDEPENDENT COMPONENT ANALYSIS USING CLUSTERING

Juha Karhunen and Simona Malaroiu

Helsinki University of Technology, Laboratory of Computer and Information Science
P.O.Box 2200, FIN-02015 HUT, Espoo, FINLAND
email: Juha.Karhunen@hut.fi

ABSTRACT

In standard ICA, a linear data model is used for a global description of the data. Even though linear ICA yields meaningful results in many cases, it can provide a crude approximation only for nonlinear data distributions. We propose a new structure, where local ICA models are used in connection with a suitable clustering algorithm grouping the data. The clustering part is responsible for an overall coarse nonlinear representation of the underlying data, while linear ICA models of each cluster are used for describing local features of the data. The goal is to represent the data better than in linear ICA while avoiding computational difficulties related with nonlinear ICA. We discuss connections to existing methods, and present experimental results for natural image data.

1. INTRODUCTION

Independent Component Analysis (ICA) [1] is a statistical technique which often characterizes the data in a natural way. Recently, ICA and the related blind source separation (BSS) problem have become important research and application topics both in unsupervised neural learning and statistical signal processing.

The basic data model used in ICA is as follows. Denote by $\mathbf{x}(t) = [x_1(t), \ldots, x_n(t)]^T$ the n-dimensional t:th data vector. The vectors $\mathbf{x}(t)$ have a common unknown zero-mean non-Gaussian statistical distribution. In ICA, we try to fit to the data $\mathbf{x}(t)$ the expansion

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{m} s_i(t)\mathbf{a}_i. \qquad (1)$$

Here the vector $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ contains the m independent components (or source signals) $s_i(t)$ for the data vector $\mathbf{x}(t)$. $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ is a constant full-rank $n \times m$ matrix, called the mixing matrix. The vectors $\mathbf{a}_i$, $i, \ldots, m$, are the basis vectors of ICA; see [1, 2]. In (1), the number of independent components m is usually assumed to be at most equal to the number of mixtures n, and often $m = n$.

The independent components are found by determining an $m \times n$ inverse mapping (separating matrix) $\mathbf{B}$ so that the m-vector

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) \qquad (2)$$

becomes an estimate $\mathbf{y}(t) = \hat{\mathbf{s}}(t)$ of the independent component vector $\mathbf{s}(t)$. To this end, many algorithms have been proposed; see for example [1, 2, 3] for further information and reviews. The estimate of the mixing matrix $\mathbf{A}$ is then obtained as a pseudoinverse of $\mathbf{B}$.

In ICA, the basis vectors $\mathbf{a}_i$ are generally not mutually orthogonal. This can be compared to standard Principal Component Analysis (PCA), where the form of the expansion (1) is the same but the basis vectors $\mathbf{a}_i$ must be mutually orthogonal and the coefficients $s_i(t)$ have maximal variances. The orthogonality requirement of PCA is often an unnatural technical constraint, whereas the independence assumption in ICA is roughly valid. On the other hand, PCA is easier to compute because it requires second-order statistics only.

In spite of its usefulness, the basic ICA model (1) is often too simple for describing well real-world data. Aiming at a general method for describing and analysing the input data rather than BSS applications, we point out two specific weaknesses of the model (1). First, the standard ICA model (1) is *linear*, while generally we need a suitable *nonlinear* model for representing the data adequately. Second, the ICA model (1) tries to describe all the data using the same *global* features (ICA basis vectors) $\mathbf{a}_i$, $i, \ldots, m$. Generally, a "rich" natural data set may have statistically different properties in various parts of its overall distribution. Therefore, *local* features are needed for efficient representation of qualitatively different parts of the data.

In this paper we develop a fairly simple data model that tries to overcome these drawbacks, and is computationally not too demanding.

2. NONLINEAR ICA

A natural extension of the linear ICA (or BSS) model (1) is to assume that the components of the data vectors $\mathbf{x}(t)$ depend nonlinearly on some statistically independent components (source signals) $\mathbf{s}(t)$. Thus the instantaneous mixtures

$$\mathbf{x}(t) = F(\mathbf{s}(t)) \tag{3}$$

of the sources $\mathbf{s}(t) = [s_1(t), \ldots, s_m(t)]^T$ are observed, where $F : R^m \rightarrow R^m$ is an unknown nonlinear mixing function. For simplicity, we assume that the number of independent components m equals to the number of mixtures n. The nonlinear ICA problem now consists of finding a inverse mapping $G : R^m \rightarrow R^m$ which gives estimates of the independent components as

$$\mathbf{y}(t) = G(\mathbf{x}(t)). \tag{4}$$

This important problem has been addressed for example in [4, 5, 6, 8, 9, 10, 11]. However, nonlinear ICA is often difficult to apply for reasons discussed below.

First, solution of the nonlinear ICA problem is usually highly non-unique [7]. For getting more unique solutions, various constraints have been introduced, but it is not yet clear which constraints are most meaningful in different situations. If we find some inverse mapping (4) such that the components of the vector $\mathbf{y}(t)$ are mutually independent, these components may have nothing to do with the original source signals.

Quite recently, Lin [12] has derived some interesting theoretical results on ICA. Let the matrices $\mathbf{H}_s$ and $\mathbf{H}_x$ denote the Hessians of the logarithmic probability densities $\log p_s(\mathbf{s})$ and $\log p_x(\mathbf{x})$ of the source vector $\mathbf{s}$ and data (mixture) vector $\mathbf{x}$, respectively. Then for the the basic linear ICA model (1) it holds that

$$\mathbf{H}_s = \mathbf{A}^T \mathbf{H}_x \mathbf{A}. \tag{5}$$

where $\mathbf{A}$ is the mixing matrix. If the components of $\mathbf{s}$ are truly independent, $\mathbf{H}_s$ should be a diagonal matrix. Due to the symmetry of the Hessian matrices $\mathbf{H}_s$ and $\mathbf{H}_x$, (5) imposes $n(n-1)/2$ constraints for the elements of $\mathbf{A}$. Thus a constant mixing matrix $\mathbf{A}$ can be solved by estimating $\mathbf{H}_x$ at two different points, and assuming some values for the diagonal elements of $\mathbf{H}_s$ [12].

If the nonlinear mapping (3) is twice differentiable, we can locally at any point approximate it by the linear mixing model (1). There $\mathbf{A}$ is defined by the first order term $\partial F(\mathbf{s})/\partial \mathbf{s}$ of the Taylor series expansion of $F(\mathbf{s})$ at the desired point. But now $\mathbf{A}$ generally changes from point to point, so that the constraint conditions (5) still leave $n(n-1)/2$ degrees of freedom for determining $\mathbf{A}$ (omitting the diagonal elements).

Another difficulty in the proposed general nonlinear ICA methods is that they tend to be computationally rather demanding. Moreover, the computational load usually increases very rapidly with the dimensionality of the problem, preventing in practice the application of nonlinear ICA methods to high-dimensional data sets.

3. A LOCAL APPROACH TO ICA

We propose the use of local linear ICA models in order to represent nonlinearly distributed data better than using a global linear ICA model. Our local ICA method consists of the following stages.

1. Cluster the data into K clusters. Denote the set of data vectors $\mathbf{x}^{(i)}$ belonging to the i-th cluster by S_i.

2. Compute the mean vector $\mathbf{m}_i$ of each cluster S_i, and subtract it from the vectors of S_i.

3. Estimate a linear ICA basis $\mathbf{a}_j^{(i)}$, $j = 1, \ldots, N_i$, of dimension N_i for each zero-mean data set S_i using a suitable ICA algorithm.

Thus the total representation of the data is defined by the K clusters, their mean vectors $\mathbf{m}_i$, $i = 1, \ldots, K$, and the associated K ICA bases $\mathbf{a}_j^{(i)}$, $j = 1, \ldots, N_i$, $i = 1, \ldots, K$. Each of these local ICA bases defines local features of the data. For getting reliable estimates, the number of data vectors falling into each of the K clusters should be large enough.

In principle, one could use any reasonable clustering method of the many possible alternatives. We have used the well-known K-means algorithm, which has the following form [13]:

1. Begin with arbitrary assignment of data vectors to K clusters or begin with arbitrary set of K cluster centers and assign data vectors to nearest centers.

2. Compute the sample mean of each cluster S_i.

3. Reassign each data vector to the cluster with the nearest mean.

4. If the clustering of all data vectors has not changed, stop; else go to step 2.

The K-means method is simple and intuitively reasonable, and it has the important feature that it tries to minimize the criterion

$$J = \sum_{k=1}^{K} \sum_{\mathbf{x}^{(i)} \in S_i} \| \mathbf{x}^{(i)} - \mathbf{m}_i \| . \tag{6}$$

The minimum is achieved when the data vectors are divided into K clusters so that the overall mean-square error estimated from the data vectors is the smallest possible. Instead of the data vectors $\mathbf{x}$ themselves, one could use as a basis of clustering vectors $\mathbf{f}(\mathbf{x})$ describing interesting features in the application at hand. In addition to the clustering algorithm there are other choices which affect the results. In this preliminary paper, we have chosen the number K of clusters on heuristic grounds. A rigorous method for selecting K in context with the K-means algorithm is introduced in [14]. We have usually chosen the number N_i of ICA basis vectors in each cluster so that the mean-square error is reduced by a certain amount (typically by at least 80%). Still another choice is the algorithm for estimating the independent components. If the ICA model (1) does not hold for the data sets S_i exactly, various ICA algorithms may yield different results. We have used the symmetric fixed-point algorithm [15]. It uses prewhitening, converges fast, and yields accurate results both for super-Gaussian and sub-Gaussian independent components simultaneously.

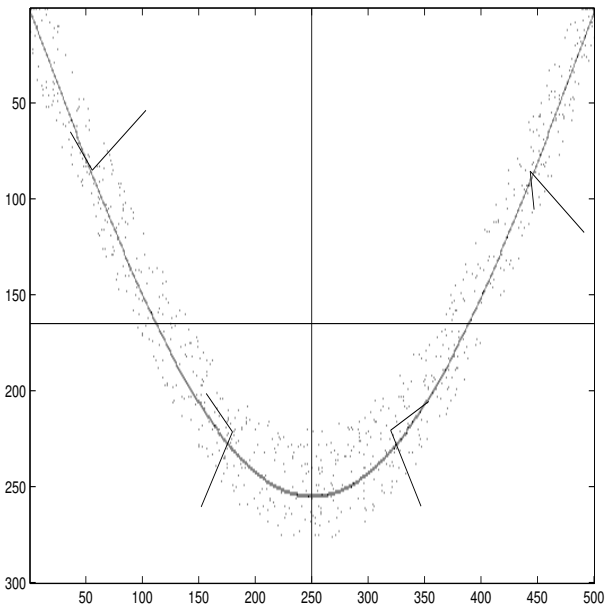

Figure 1: Four clusters and local ICA bases in a simple test example.

4. RELATIONS TO OTHER METHODS

The local ICA method described above has connections with several other approaches. Lin's results [12] described earlier show that local linear ICA expansions are not unique without extra contraints, and that the degree of non-uniqueness is essentially similar as for nonlinear ICA. In our case, the local ICA expansions are defined by the sets S_i of data vectors belonging to each cluster. If the clusters change, the local ICA expansions generally change, too.

The K-means clustering algorithm is closely connected with self-organizing map (SOM), especially its batch version [16]. SOM is a well-known method for nonlinear data mapping, visualization, and classification with numerous applications [16]. Originally we intended to use SOM for clustering, but in making the first experiments described in this paper we have used K-means clustering because of its simplicity. These two methods differ mainly in that in SOM a neighborhood is used in forming the clusters. Usually the neighborhood is decreased during learning. If learning in SOM is continued sufficiently long so that the neighborhood finally approaches to zero, SOM provides quite similar results as K-means clustering.

Usually SOM is used in so that the number of nodes (here clusters) is fairly large (typically tens or hundreds), but each node is represented by one vector only. Ritter [17] has justified that SOM then provides a discrete approximation to the principal manifold, which is a lower dimensional surface fitting the data best in the mean-square error sense. The cluster mean vectors obtained using K-means clustering also form at least a crude approximation to the principal manifold due to the minimization of the estimated mean-square error (6). Our method differs from the standard self-organizing map mainly in that the number of clusters is typically fairly small, and that each cluster is represented by its mean vector and the associated local ICA subspace.

Kambhatla and Leen [18] developed a local PCA method where they used vector quantization (instead of SOM or K-means clustering) for dividing the data into disjoint regions. Then for each region a local PCA expansion is computed. In this method, vector quantization tries to capture the large-scale nonlinear structure in the data, while the local PCA's within each cell describe local variations. This method and a related one introduced by Joutsensalo [19] are useful mainly in data compression.

Kohonen [16] introduced so-called ASSOM method for unsupervised extraction of invariant local features from the input data. This is done by associating a subspace instead of a single weight vector to each node of the self-organizing map. The subspaces are constructed using so-called learning subspace method [16] from the data vectors belonging to each node (cluster). The basis vectors of the subspace are then used as the invariant features for the data vectors classified to that

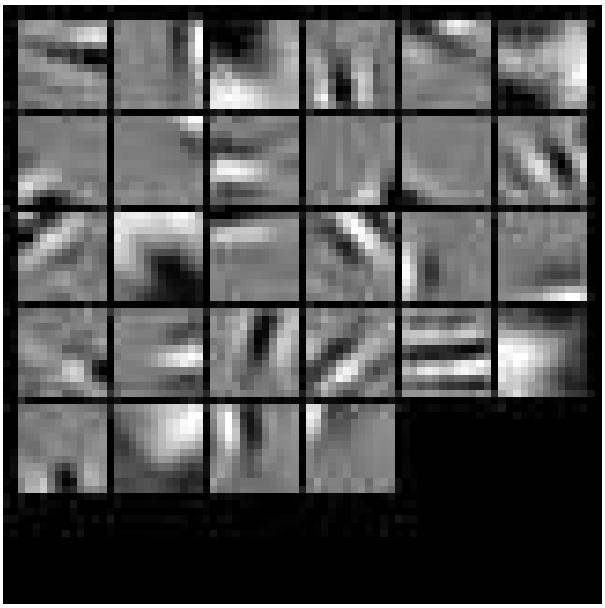

Figure 2: Global ICA basis vectors for image data.

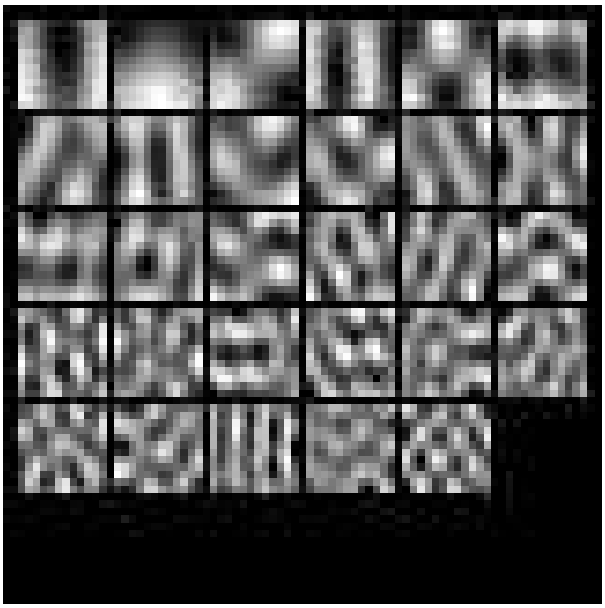

Figure 3: Local PCA basis vectors for the first cluster of the image data.

subspace.

The results are interesting and meaningful for real-world speech and image data [16]. The ASSOM methods have, however, some drawbacks. For obtaining good features sometimes rather heuristic preprocessing is needed. It is theoretically not clear what the iterative learning subspace method actually learns. The subspace model does not apply well to all types of data sets. Furthermore, the basis vectors in ASSOM subspaces are constrained to be mutually orthonormal like in standard PCA.

Recently, McGlinchey and Fyfe [20] used nonlinear PCA in their variant of ASSOM instead of the learning subspace method. Oja and Valkealahti [21] applied ICA to form the subspaces in ASSOM. Their method is the only local ICA method presented thus far to our knowledge, but it differs considerably from ours. In particular, they use a classification based criterion function for clustering, and apply it for extracting ICA type features from textures.

5. EXPERIMENTAL RESULTS

We first present a simple experiment illustrating how the proposed local ICA method works. Here the data is 2-dimensional only for visualization purposes, and it consists of points which are uniformly distributed around a parabola as shown in Figure 1. The number of clusters was somewhat arbitrarily chosen to be $K = 4$, and the 4 clusters found by the K-means algorithms are separated by the horizontal and vertical lines going across Figure 1. The 2 local ICA basis vectors found by the symmetric fixed point algorithm are shown for each cluster attached to the mean vector of the respective cluster. These basis vectors are at least for some clusters clearly different from the respective local PCA basis vectors which would be mutually orthogonal. One can also note that the data vectors belonging to the two lower clusters are still mildly nonlinearly distributed.

We have made more extensive experiments with image data, consisting of 15 natural images describing different scenes, plants and animals. In these experiments a set of typically 100000 possibly overlapping 12×12 subimages was randomly extracted from the original images. The subimages were then represented as 144-component data (mixture) vectors $\mathbf{x}_k$. They were preprocessed by subtracting from each $\mathbf{x}_k$ the mean of its components, and then dividing $\mathbf{x}_k$ by its norm for reducing the variance.

In the first experiments, we used the data vectors $\mathbf{x}_k$ as such in clustering. Later on we found that the proposed local ICA method yields more meaningful results if the data vectors are clustered according to their mean frequency in the K-means algorithm. This was accomplished by applying discrete cosine transform to the data vectors before clustering them. However, discrete cosine transform was used only in finding the clusters. After the clusters were found, the local ICA bases were computed for the data vectors belonging to each cluster

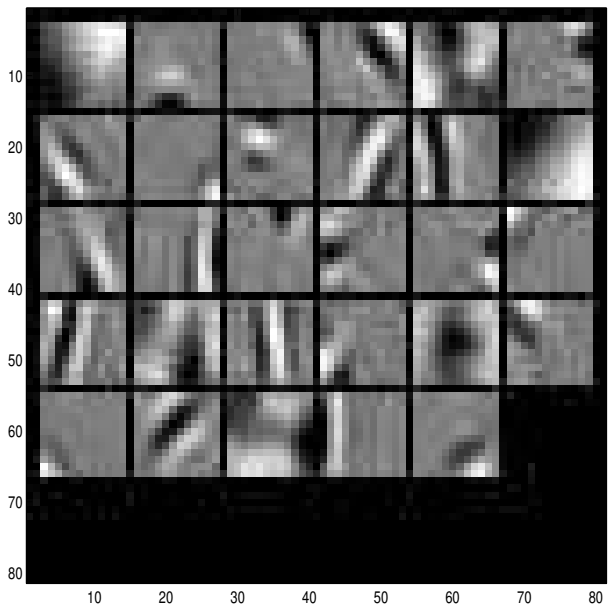

Figure 4: Local ICA basis vectors for the first cluster of the image data.

without application of discrete cosine transform.

Figure 2 shows the 28 first basis vectors, represented again as 12×12 subimages, found by fitting the global linear ICA model (1) to the data without any clustering. These subimages correspond to various types of features, such as smooth changes in subimages, edges in different directions, and local information in the corners of the images.

Figure 3 depicts the 29 basis vectors (subimages) obtained by computing the standard PCA expansion for the data vectors belonging to the first cluster of the image data. These basis images show very clearly typical features extracted by standard PCA. They are sensitive to spatial frequencies, starting from low frequencies and proceeding to higher ones for which the basis images resemble more or less a checkerboard. The local PCA expansions for the other 3 clusters formed using the K-means algorithm were almost similar and are therefore not shown here. Also global PCA expansion provided qualitatively very similar features sensitive to spatial frequencies as in Figure 3.

Finally, Figures 4–7 show the basis images of the local ICA expansions computed for the data vectors belonging to each of the four clusters found by the K-means algorithm. The number of ICA basis vectors is different for each cluster because we determined it by requiring that 80% of the total energy is included. In all the experiments, the symmetric fixed-point algorithm described in [15] was used for computing the needed

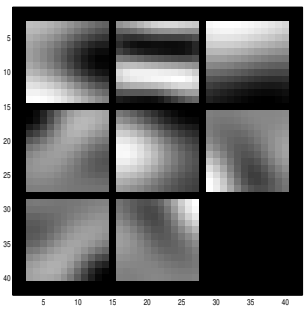

Figure 5: Local ICA basis for the second cluster.

ICA expansions.

Clearly, local ICA basis vectors yield filters that are sensitive to edges and lines of varying orientations, as well as wavelet type filters which are sensitive to local variations in the images. More specifically, one can observe that the basis images in Figure 4 have much more variation in horizontal than in vertical direction. Many of these local ICA basis images correspond to filters sensitive to horizontal edges and lines. The basis images of Figure 6 are largely qualitatively similar as in Fig. 4 but this time sensitive to vertical variations. Most of the 9 subimages in Figure 5 represent smooth variations, while the basis images in Figure 7 are sensitive to local wavelet type features with a few exceptions. These results can be understood by recalling that clustering was based on frequency information provided by the discrete cosine transformation. Further interpretations of ICA in image feature extraction can be found in [3, 22].

Acknowledgments The authors thank Petteri Pajunen, Harri Lappalainen, and Kimmo Valkealahti for helpful comments and discussions.

6. REFERENCES

[1] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287-314, 1994.

[2] J. Karhunen, "Neural approaches to independent component analysis and source separation," in *Proc. 4th European Symp. on Artificial Neural Networks*, Bruges, Belgium, April 1996, pp. 249-266.

[3] E. Oja et al., "Neural independent component analysis - approaches and applications," in S.-I. Amari and N. Kasabov (Eds.), *Brain-Like Computing and Intelligent Information Systems*, Springer, Singapore, 1997, pp. 167-188.

[4] L. Parra, G. Deco, and S. Miesbach, "Statistical independence and novelty detection with information-preserving nonlinear maps," *Neural Computation*, vol. 8, pp. 260-269, 1996.

[5] P. Pajunen and J. Karhunen, "A maximum likelihood approach to nonlinear blind separation," in *Proc. 1997*

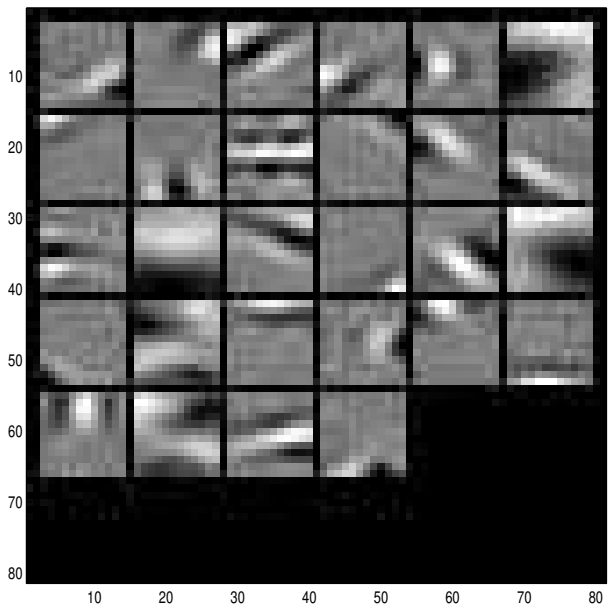

Figure 6: Local ICA basis for third cluster.

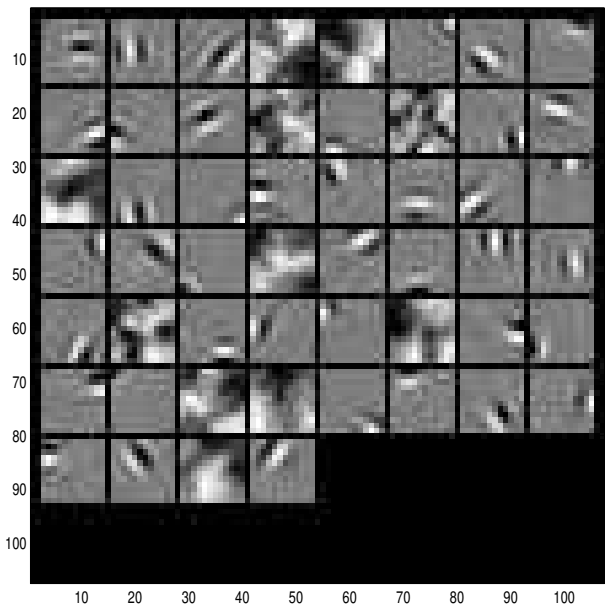

Figure 7: Local ICA basis for the fourth cluster.

Int. Conf. on Artificial Neural Networks, Lausanne, Switzerland, October 1997, pp. 541-546.

[6] J. Lin, D. Grier, and J. Cowan, "Faithful representation of separable distributions," *Neural Computation*, vol. 9, 1997, pp. 1305-1320.

[7] A. Hyvärinen and P. Pajunen, "On existence and uniqueness of solutions in non-linear independent component analysis," in *Proc. of the 1998 Int. Joint Conf. on Neural Networks*, Anchorage, Alaska, May 1998, pp. 1350-1355.

[8] H. Yang, S. Amari, and A. Cichocki, "Information-theoretic approach to blind separation of sources in non-linear mixture," *Signal Processing*, vol. 64, 1998, pp. 291-300.

[9] A. Taleb and C. Jutten, "Nonlinear source separation: the post-nonlinear mixtures," in *Proc. 1997 European Symp. on Artificial Neural Networks*, Bruges, Belgium, April 1997, pp. 279-284.

[10] J. Fisher and J. Principe, "Entropy manipulation of arbitrary nonlinear mappings," in J. Principe et al. (Eds.), *Neural Networks for Signal Processing VII*, IEEE Press, New York 1997, pp. 14-23.

[11] T.-W. Lee, B.-U. Koehler, and R. Orglmeister, 'Blind source separation of nonlinear mixing models" in J. Principe et al. (Eds.), *Neural Networks for Signal Processing VII*, IEEE Press, New York 1997, pp. 406-415.

[12] J. Lin, "Factorizing multivariate function classes," to appear in M. Mozer, M. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems 9*, MIT Press, Cambridge, MA, 1998.

[13] C. Therrien, *Decision, Estimation and Classification.* New York: J. Wiley, 1989.

[14] L. Xu, "Bayesian Ying-Yang machine, clustering and number of clusters," *Pattern Recognition Letters*, vol. 18, 1997, pp. 1167-1178.

[15] A. Hyvärinen, "A family of fixed-point algorithms for independent component analysis," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Munich, Germany, April 1997, pp. 3917-3920.

[16] T. Kohonen, *Self-Organizing Maps.* New York, Berlin: Springer-Verlag, 1995.

[17] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps.* Reading, MA: Addison-Wesley, 1992.

[18] N. Kambhatla and T. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, 1997, pp. 1493-1516.

[19] J. Joutsensalo, "Nonlinear dimension reduction by combining competitive and distributed learning," in *Proc. 1995 Int. Conf. on Artificial Neural Networks*, Paris, France, October 1995, pp. 395-400.

[20] S. McGlinchey and C. Fyfe, "Invariant feature maps for analysis of orientations in image data," in *Proc. 6th European Symp. on Artificial Neural Networks*, Bruges, Belgium, April 1996, pp. 215-220.

[21] E. Oja and K. Valkealahti, "Local independent component analysis by the self-organizing map," in *Proc. 1997 Int. Conf. on Artificial Neural Networks*, Lausanne, Switzerland, October 1997, pp. 553-558.

[22] J. Hurri, A. Hyvärinen, and E. Oja, "Wavelets and natural image statistics," in *Proc. 10th Scan. Conf. on Image Analysis*, Lappeenranta, Finland, June 1997.