Topographic Independent Component Analysis

Aapo Hyvärinen, Patrik O. Hoyer, and Mika Inki
Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland
aapo.hyvarinen@hut.fi

Neural Computation 13(7):1527-1558 [July, 2001]

Abstract

In ordinary independent component analysis, the components are assumed to be completely independent, and they do not necessarily have any meaningful order relationships. In practice, however, the estimated "independent" components are often not at all independent. We propose that this residual dependence structure could be used to define a topographic order for the components. In particular, a distance between two components could be defined using their higher-order correlations, and this distance could be used to create a topographic representation. Thus we obtain a linear decomposition into approximately independent components, where the dependence of two components is approximated by the proximity of the components in the topographic representation.

1 Introduction

Indendent component analysis (ICA) (Jutten and Herault, 1991) 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 (Comon, 1994); the most fundamental is that the independent components s_i must be nongaussian (Comon, 1994). A considerable amount of research has been recently conducted on the estimation of this model, see e.g. (Amari et al., 1996; Bell and Sejnowski, 1995; Cardoso and Laheld, 1996; Cichocki and Unbehauen, 1996; Delfosse and Loubaton, 1995; Hyvärinen and Oja, 1997; Hyvärinen, 1999a; Karhunen et al., 1997; Oja, 1997; Pajunen, 1998).

In classic ICA, the independent components s_i have no particular order, or other relationships. It is possible, though, to define an order relation between the independent components by such criteria as nongaussianity or contribution to the observed variance (Hyvärinen, 1999c); the latter is given by the norms of the corresponding columns of the mixing matrix as the independent components are defined to have unit variance. Such trivial order relations may be useful for some purposes, but they are not very informative in general.

The lack of an inherent order of independent components is related to the assumption of complete statistical independence. In practical applications of ICA, however, one can very often observe clear violations of the independence assumption. It is possible to find, for example, couples of estimated independent components such that they are clearly dependent on each other. This dependence structure is often very informative, and it would be useful to somehow estimate it.

Estimation of the "residual" dependency structure of estimates of independent components could be based, for example, on computing the cross-cumulants. Typically these would be higher-order cumulants since second-order cross-cumulants, i.e. the covariances, are typically very small, and are in fact forced to be zero in many ICA estimation methods, e.g. (Comon, 1994; Hyvärinen and Oja, 1997; Hyvärinen, 1999a). A more information-theoretic measure for dependence would be given by mutual information. Whatever measure is used, however, the problem remains as to how such numerical estimates of the dependence structure should be visualized or otherwise utilized. Moreover, there is another serious problem associated with simple estimation of some dependency measures from the estimates of the independent components. This is due to the fact that often the independent components do not form a well-defined set. Especially in image decomposition (Bell and Sejnowski, 1997; Olshausen and Field, 1996; Olshausen and Field, 1997; Hyvärinen, 1999b), the set of potential independent components seems to be larger than what can be estimated at one time, in fact the set might be infinite. A classic ICA method gives an arbitrarily chosen subset of such independent components, corresponding to a local minimum of the objective function. (This can be seen in the fact that the basis vectors are different for different initial conditions.) Thus, it is important in many applications that the dependency information is utilized during the estimation of the independent components, so that the estimated set of independent components is one that can be ordered in a meaningful way.

In this paper, we propose that the residual dependency structure of the "independent" components, i.e. dependencies that cannot be cancelled by ICA, could be used to define a *topographic order* between the components. The topographic order is easy to represent by visualization, and has the usual computational advantages associated with topographic maps that will be discussed below. We propose a modification of the ICA model that explicitly formalizes a topographic order between the components. This gives a topographic map where the distance of the components in the topographic representation is a function of the dependencies of the components. Components that are near to each other in the topographic representation are relatively strongly dependent in the sense of higher-order correlations, or mutual information. This gives a new principle for topographic organization. Furthermore, we derive a learning rule for the estimation of the model. Experiments on image feature extraction and blind separation of magnetoencephalographic data demonstrate the usefulness of the model.

This paper is organized as follows. First, topographic ICA is motivated and formulated as a generative model in Section 2. Since the likelihood of the model is intractable, a tractable approximation is derived in Section 3. A gradient learning rule for performing the estimation of the model is then introduced in Section 4. Discussion on the relation of our model to some other methods, as well as on the utility of topography is given in Section 5. Simulations and experiments are given in Section 6. Finally, some conclusions are drawn in Section 7.

2 Topographic ICA Model

2.1 Dependence and topography

In this section, we define topographic ICA using a generative model that is a hierarchical version of the ordinary ICA model. The idea is to relax the assumption of the independence of the components s_i in (1) so that components that are close to each other in the topography are not assumed to be independent in the model. For example, if the topography is defined by a lattice or grid, the dependency of the components is a function of the distance of the components on that grid. In contrast, components that are not close to each other in the topography *are* independent, at least approximately; thus most pairs of components are independent. Of course, if independence would not hold for most component pairs, any connection to ICA would be lost, and the model would not be very useful in those applications where ICA has proved useful.

2.2 What kind of dependencies should be modelled?

The basic problem is then to choose what kind of dependencies are allowed between near-by components. The most basic dependence relation is linear correlation[1]. However, allowing linear correlation between the components does not seem very useful. In fact, in many ICA estimation methods, the components are constrained to be uncorrelated (Cardoso and Laheld, 1996; Comon, 1994; Hyvärinen and Oja, 1997; Hyvärinen, 1999a), so the requirement of uncorrelatedness seems natural in any extension of ICA as well.

A more interesting kind of dependency is given by a certain kind of higher-order correlation, namely correlation of energies. This means that

$$\text{cov}\,(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0 \tag{2}$$

if s_i and s_j are close in the topography. Here, we assume that this covariance is positive. Intuitively, such a correlation means that the components tend to be active, i.e. non-zero, at the same time, but the actual values of s_i and s_j are not easily predictable from each other. For example, if the variables are defined as products of two zero-mean independent components z_i, z_j and a common "variance" variable σ:

$$s_i = z_i \sigma \tag{3}$$

$$s_j = z_j \sigma \tag{4}$$

then s_i and s_j are uncorrelated, but their energies are not. In fact the covariance of their energies equals $E\{z_i^2\sigma^2 z_j^2\sigma^2\} - E\{z_i^2\sigma^2\}E\{z_j^2\sigma^2\} = E\{\sigma^4\} - E\{\sigma^2\}^2$, which is non-negative because it equals the variance of σ^2 (we assumed here for simplicity that z_i and z_j are of unit variance). Energy correlation is illustrated in Fig. 1.

Using this particular kind of higher-order correlation could be initially motivated by mathematical and conceptual simplicity. Correlation of energies is arguably the simplest and most intuitive kind of higher-order dependency because it can be interpreted as *simultaneous activation*. The variable σ in (3-4) can be considered as a higher-order component controlling the activations of the components s_i and s_j. This kind of higher-order correlation is therefore relatively easy to analyze and understand, and likely to have applications in many different areas.

Moreover, an important empirical motivation for this kind of dependency can be found in image feature extraction. In (Simoncelli and Schwartz, 1999), it was shown that the predominant dependence of wavelet-type filter outputs is exactly the strong correlation of their energies; this property was utilized for improving ordinary shrinkage denoising methods. Similarly, in (Hyvärinen and Hoyer, 2000), a subspace version of ICA was introduced (to be discussed in more detail in Sec. 5.2) in which the components in each subspace have energy correlations. It was shown that meaningful properties, related to complex cells, emerge from natural image data using this model.

2.3 The generative model

Now we define a generative model that implies correlation of energies for components that are close in the topographic grid. In the model, the observed variables $\mathbf{x} = \mathbf{A}\mathbf{s}$ are generated as a linear transformation of the components $\mathbf{s}$, just as in the basic ICA model in (1). The point is to define the joint density of $\mathbf{s}$ so that it expresses the topography. The topography is defined by simultaneous activation as discussed in the previous subsection.

[1] In this paper, we mean by correlation a normalized form of covariance: $\text{corr}(s_1, s_2) = [E\{s_1 s_2\} - E\{s_1\}E\{s_2\}][\text{var}\,(s_1)\text{var}\,(s_2)]^{-1/2}$

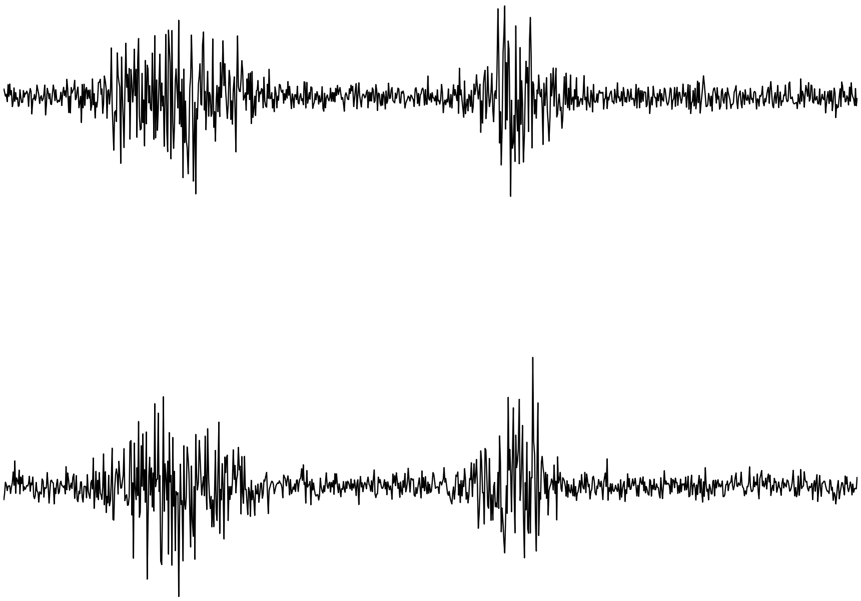

Figure 1: Illustration of higher-order dependencies. The two signals in the figure are uncorrelated but they are not independent. In particular, their energies are correlated. The signals were generated as in (3-4), but for purposes of illustration, the random variable σ was replaced by a time-correlated signal.

We define the joint density of $\mathbf{s}$ as follows. The variances σ_i^2 of the s_i are not constant, instead they are assumed to be random variables, generated according to a model to be specified. After generating the variances, the variables s_i are generated independently from each other, using some conditional distributions to be specified. In other words, the s_i are *independent given their variances*. Dependence among the s_i is implied by the dependence of their variances. According to the principle of topography, the variances corresponding to near-by components should be (positively) correlated, and the variances of components that are not close should be independent, at least approximatively.

To specify the model for the variances σ_i^2, we need to first define the topography. This can be accomplished by a neighborhood function $h(i,j)$, which expresses the proximity between the i-th and j-th components. The neighborhood function can be defined in the same ways as with the self-organizing map (Kohonen, 1995). Neighborhoods can thus be defined as one-dimensional or two-dimensional; 2-D neighborhoods can be square or hexagonal. Usually, the neighborhood function is defined as a monotonically decreasing function of some distance measure, which implies among other things that it is symmetric: $h(i,j) = h(j,i)$, and has constant diagonal: $h(i,i) = const.$ for all i. A simple example is to define a 1-D neighborhood relation by

$$h(i,j) = \begin{cases} 1, & \text{if } |i-j| \leq m \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The constant m defines here the width of the neighborhood: The neighborhood of the component with index i consists of those components whose indices are in the range $i-m, \ldots, i+m$. The neighborhood function $h(i,j)$ is thus a matrix of hyperparameters. In this paper, we consider it to be known and fixed. Future work may provide methods for estimating the neighborhood function from the data.

Using the topographic relation $h(i,j)$, many different models for the variances σ_i^2 could be used. We prefer here to define them by an ICA model followed by a nonlinearity:

$$\sigma_i = \phi\left(\sum_{k=1}^{n} h(i,k) u_k\right) \tag{6}$$

where u_i are the "higher-order" independent components used to generate the variances, and ϕ is some scalar nonlinearity. This particular model can be motivated by two facts. First, taking sparse u_i, we can model sparse local activations, that is, the case where activation is limited to a few regions in the map. This is what seems to happen in image features. Second, the model is mathematically quite simple, and in particular, it enables a simple approximation of likelihood that will be derived in Sec. 3.

In the model, the distributions of the u_i and the actual form of ϕ are additional hyperparameters; some suggestions will be given below. It seems natural to constrain the u_k to be non-negative. The function ϕ can then be constrained to be a monotonic transformation in the set of non-negative real numbers. This ensures that the σ_i's are non-negative.

The resulting topographic ICA model is summarized in Fig. 2. Note that the two stages of the generative model can be expressed as a single equation, analogously to (3-4), as follows:

$$s_i = \phi\left(\sum_{k} h(i,k) u_k\right) z_i \tag{7}$$

where z_i is a random variable that has the same distribution as s_i given that σ_i^2 is fixed to unity. The u_i and the z_i are all mutually independent.

2.4 Basic Properties of the Topographic ICA model

Here we discuss some basic properties of the generative model defined above.

1. All the components s_i are uncorrelated. This is because according to (7) we have

   $$E\{s_is_j\} = E\{z_i\}E\{z_j\}E\{\phi(\sum_k h(i,k)u_k)\phi(\sum_k h(j,k)u_k)\} = 0 \tag{8}$$

   due to the independence of the u_k from z_i and z_j. (Recall that z_i and z_j are zero-mean.) To simplify things, one can define that the marginal variances (i.e. integrated over the distibution of σ_i) of the s_i are equal to unity, as in ordinary ICA. In fact, we have

   $$E\{s_i^2\} = E\{z_i^2\}E\{\phi(\sum_k h(i,k)u_k)^2\}, \tag{9}$$

   so we only need to rescale $h(i,j)$ (the variance of z_i is equal to unity by definition). Thus the vector $\mathbf{s}$ can be considered to be sphered, i.e. white.

2. Components that are far from each other are more or less independent. More precisely, assume that s_i and s_j are such that their neighborhoods have no overlap, i.e. there is no index k such that both $h(i,k)$ and $h(j,k)$ are non-zero. Then the components s_i and s_j are independent. This is because their variances are independent, as can be seen from (6). Note, however, that independence need not be strictly true for the estimated components, just as independence does not need to hold for the components estimated by classic ICA.

3. Components s_i and s_j that are near to each other, i.e. such that $h(i,j)$ is significantly non-zero, tend to be active (non-zero) at the same time. In other words, their energies s_i^2 and s_j^2 are usually positively correlated. This property cannot be strictly proven in general, since it depends on the form of ϕ and the distributions of the u_i. However, the following intuitive argument can be made. Calculating

   $$\begin{aligned} &E\{s_i^2s_j^2\} - E\{s_i^2\}E\{s_j^2\} \\ &\quad = E\{z_i^2\}E\{z_j^2\}[E\{\phi^2(\sum_k h(i,k)u_k)\phi^2(\sum_k h(j,k)u_k)\} - E\{\phi^2(\sum_k h(i,k)u_k)\}E\{\phi^2(\sum_k h(j,k)u_k)\}] \end{aligned} \tag{10}$$

   we see that the covariance of the energies of s_i and s_j is equal to the covariance of σ_i^2 and σ_j^2. The covariance of the sums $\sum_k h(j,k)u_k$ and $\sum_k h(j,k)u_k$ can be easily evaluated as $\sum_k h(i,k)h(j,k)\mathrm{var}\, u_k$. This is clearly positive, if the components s_i and s_j are close to each other. Since we constrained ϕ to be monotonic in the set of nonnegative real numbers, ϕ^2 is monotonic in that set as well, and we therefore conjecture that the covariance is still positive when the function ϕ^2 is applied on these sums, since this amounts to computing the covariance of the nonlinear transforms. This would imply that the covariance of σ_i^2 and σ_j^2 is still positive, and this would imply the result.

4. An interesting special case of topographic ICA is obtained when every component s_i is assumed to have a gaussian distribution when the variance is given. This means that the marginal, unconditional distributions of the components s_i are continuous mixtures of gaussians. In fact these distributions are always supergaussian, i.e. have positive kurtosis. This is because

   $$\mathrm{kurt}\, s_i = E\{s_i^4\} - 3(E\{s_i^2\})^2 = E\{\sigma_i^4z_i^4\} - 3(E\{\sigma_i^2z_i^2\})^2 = 3[E\{\sigma_i^4\} - (E\{\sigma_i^2\})^2] \tag{11}$$

   which is always positive because it is the variance of σ_i^2 multiplied by 3. Since most independent components encountered in real data are supergaussian (Bell and Sejnowski, 1997; Hyvärinen, 1999b; Olshausen and Field, 1996; Vigário, 1997), it seems realistic to use a gaussian conditional distribution for the s_i.

5. Classic ICA is obtained as a special case of the topographic model, by taking a neighborhood function $h(i,j)$ that is equal to the Kronecker delta function, $h(i,j) = \delta_{ij}$.

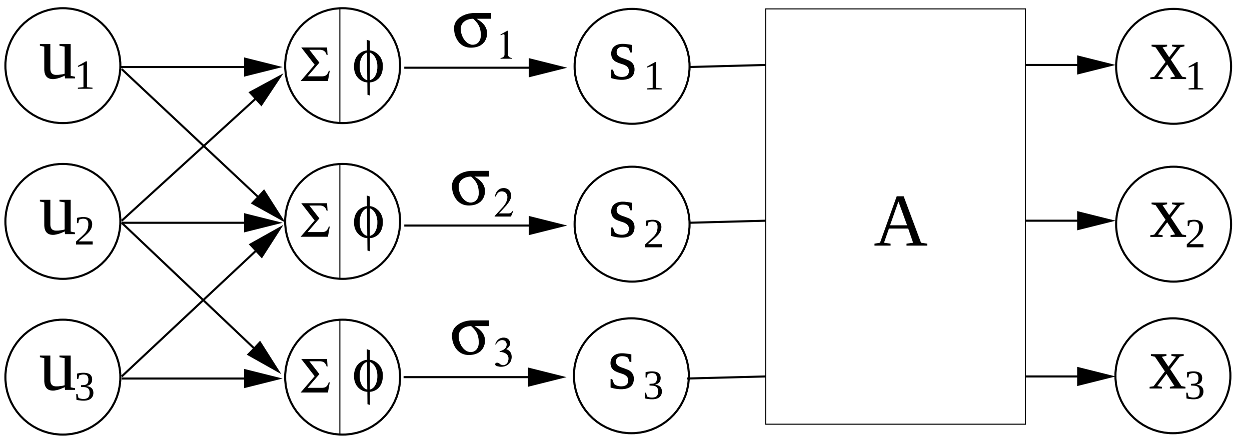

Figure 2: An illustration of the topographic ICA model. First, the “variance-generating” variables u_i are generated randomly. They are then mixed linearly inside their topographic neighborhoods. (The figure shows a one-dimensional topography.) The mixtures are then transformed using a nonlinearity ϕ, thus giving the local variances σ_i^2. Components s_i are then generated with variances σ_i^2. Finally, the components s_i are mixed linearly to give the observed variables x_i.

3 Approximating the likelihood of the model

In this section, we discuss the estimation of the topographic ICA model introduced in the previous section. The model is a missing variables model in which the likelihood cannot be obtained in closed form. However, to simplify estimation, we derive a tractable approximation of the likelihood.

The joint density of $\mathbf{s}$, i.e. the topographic components, and $\mathbf{u}$, i.e. the “higher-order” independent components generating the variances, can be expressed as

$$p(\mathbf{s},\mathbf{u}) = \prod_i p_i^s\Big(\frac{s_i}{\phi(\sum_k h(i,k)u_k)}\Big)\frac{p_i^u(u_i)}{\phi(\sum_k h(i,k)u_k)} \tag{12}$$

where the p_i^u are the marginal densities of the u_i and the p_i^s are the densities of p_i^s for variance fixed to unity. The marginal density of $\mathbf{s}$ could be obtained by integration:

$$p(\mathbf{s}) = \int \prod_i p_i^s\Big(\frac{s_i}{\phi(\sum_k h(i,k)u_k)}\Big)\frac{p_i^u(u_i)}{\phi(\sum_k h(i,k)u_k)}d\mathbf{u} \tag{13}$$

and using the same derivation as in ICA (Pham et al., 1992), this gives the likelihood as

$$L(\mathbf{W}) = \prod_{t=1}^{T}\int \prod_i p_i^s\Big(\frac{\mathbf{w}_i^T\mathbf{x}(t)}{\phi(\sum_k h(i,k)u_k)}\Big)\frac{p_i^u(u_i)}{\phi(\sum_k h(i,k)u_k)}|\det\mathbf{W}|d\mathbf{u} \tag{14}$$

where $\mathbf{W} = (\mathbf{w}_1,...,\mathbf{w}_n)^T = \mathbf{A}^{-1}$, and the $\mathbf{x}(t), t = 1,...,T$ are the observations of $\mathbf{x}$. It is here assumed that the neighborhood function and the nonlinearity ϕ as well as the densities p_i^u and p_i^s are known.

The problem with (14) is that it contains an intractable integral. One way of solving this problem would be to use the EM algorithm (Dempster et al., 1977), but it seems to be intractable as well. Estimation could still be performed by Monte Carlo methods, but such methods would be computationally expensive. Therefore, we prefer to approximate the likelihood by an analytical expression. To simplify the notation, we assume in the following that the densities p_i^u are equal for all i, and likewise for p_i^s.

To obtain the approximation, we first fix the density $p_i^s = p_s$ to be gaussian, as discussed in Section 2.4, and we define the nonlinearity ϕ as

$$\phi(\sum_k h(i,k)u_k) = (\sum_k h(i,k)u_k)^{-1/2} \tag{15}$$

The main motivation for these choices is algebraic simplicity that makes a simple approximation possible. Moreover, the assumption of conditionally gaussian s_i, which implies that the unconditional distribution of s_i supergaussian, is compatible with the preponderance of supergaussian variables in ICA applications.

With these definitions, the marginal density of $\mathbf{s}$ equals:

$$p(\mathbf{s}) = \int \frac{1}{\sqrt{2\pi}^n} \exp(-\frac{1}{2}\sum_i s_i^2[\sum_k h(i,k)u_k]) \prod_i p_u(u_i) \sqrt{\sum_k h(i,k)u_k}\, d\mathbf{u} \tag{16}$$

which can be manipulated to give

$$p(\mathbf{s}) = \int \frac{1}{\sqrt{2\pi}^n} \exp(-\frac{1}{2}\sum_k u_k[\sum_i h(i,k)s_i^2]) \prod_i p_u(u_i) \sqrt{\sum_k h(i,k)u_k}\, d\mathbf{u}. \tag{17}$$

The interesting point in this form of the density is that it is a function of the “local energies” $\sum_i h(i,k)s_i^2$ only. The integral is still intractable, though. Therefore, we use the simple approximation:

$$\sqrt{\sum_k h(i,k)u_k} \approx \sqrt{h(i,i)u_i}. \tag{18}$$

This is actually a lower bound, and thus our approximation will be an lower bound of the likelihood as well. This gives us the following approximation $\tilde{p}(\mathbf{s})$:

$$\tilde{p}(\mathbf{s}) = \prod_k \exp(G(\sum_i h(i,k)s_i^2)) \tag{19}$$

where the scalar function G is obtained from the p_u by:

$$G(y) = \log \int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}uy) p_u(u) \sqrt{h(i,i)u}\, du. \tag{20}$$

Recall that we assumed $h(i,i)$ to be constant.

Thus we obtain finally the following approximation of the log-likelihood:

$$\log \tilde{L}(\mathbf{W}) = \sum_{t=1}^{T} \sum_{j=1}^{n} G(\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{x}(t))^2) + T \log|\det \mathbf{W}|. \tag{21}$$

This is a function of local energies. Every term $\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{x}(t))^2$ could be considered as the energy of a neighborhood, possibly related to the output of a higher-order neuron as in visual complex cell models (Hyvärinen and Hoyer, 2000). The function G has a similar role as the log-density of the independent components in classic ICA.

The formula for G in (20) can be exactly evaluated only in special cases. One such case is obtained if the u_k are obtained as squares of standardized gaussian variables. Straight-forward calculation then gives the following function

$$G_0(y) = -\log(1+y) + \frac{1}{2}\log \pi^2 h(i,i). \tag{22}$$

In ICA, it is well-known that the exact form of the log-density does not affect the consistency of the estimators, as long as the overall shape of the function is correct. This is probably true in topographic ICA as well. The simulations and experiments that we have performed support this conjecture, see Sec. 6. If the data is sparse, i.e. supergaussian, convergence seems to be obtained by almost any $G(y)$ that is *convex* for non-negative y, like the function in (22). Therefore, one could use many other more or less heuristically chosen functions. For example, one could use the function proposed in (Hyvärinen and Hoyer, 2000):

$$G_1(y) = -\alpha_1\sqrt{y} + \beta_1, \tag{23}$$

This function is related to the exponential distibution (Hyvärinen and Hoyer, 2000). The scaling constant α_1 and the normalization constant β_1 are determined so as to give a probability density that is compatible with the constraint of unit variance of the s_i, but they are irrelevant in the following. In practice, a small constant may be added inside the square root for reasons of numerical stability:

$$G_1^*(y) = -\alpha_1\sqrt{\varepsilon + y} + \beta_1, \tag{24}$$

Another possibility would be a simple polynomial that could be considered as a Taylor approximation of the real G_i:

$$G_2(y) = \alpha_2 y^2 + \beta_2, \tag{25}$$

where the first-order term is omitted because it corresponds to second-order statistics that stay constant if the decomposition is constrained to be white. Again, the constants α_2 and β_2 are immaterial.

One point that we did not treat in the preceding was the scaling of the neighborhood function $h(i,j)$. As shown in Sec. 2.4, to obtain unit variance of the s_i, $h(i,j)$ has to be scaled according to (9). However, since the functions in (23) and (25) are homogenic, i.e. any scalar multiplying their arguments is equivalent to a scalar multiplying the functions themselves, any rescaling of $h(i,j)$ only multiplies the log-likelihood by a constant factor. (We ignored here the irrelevant constants β_i.) Therefore, when using (23) and (25), the s_i can be considered to have unit variance without any further complications. This is not the case with (22), however. In practice, however, this complication does not seem very important, and was completely ignored in our simulations and experiments.

4 Learning rule

In this section, we derive a learning rule for performing the maximization of the approximation of likelihood derived in the previous section. The approximation enables us to derive a simple gradient learning rule.

First, we assume here that the data is preprocessed by whitening

$$\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} \tag{26}$$

where the whitening matrix $\mathbf{V}$ can be computed as $\mathbf{V} = (E\{\mathbf{x}\mathbf{x}^T\})^{-1/2}$, for example. The inverse square root is here defined by the eigenvalue decomposition of $E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{E}\mathbf{D}\mathbf{E}^T$ as $\mathbf{V} = (E\{\mathbf{x}\mathbf{x}^T\})^{-1/2} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T$. Alternatively, one can use PCA whitening $\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{E}^T$, which also allows one to reduce the dimension of the data.

Then we can constrain the $\mathbf{w}_i^T$, which here denote the estimates of the rows of the new separating matrix $(\mathbf{VA})^{-1}$, to form an orthonormal system (Comon, 1994; Hyvärinen and Oja, 1997; Hyvärinen, 1999a; Cardoso and Laheld, 1996). This implies that the estimates of the components are uncorrelated. Such a simplification is widely used in ICA, and it is especially useful here since it allows us to concentrate on higher-order correlations.

Thus we can simply derive (see Appendix) a gradient algorithm in which the i-th (weight) vector $\mathbf{w}_i$ is updated as

$$\Delta\mathbf{w}_i \propto E\{\mathbf{z}(\mathbf{w}_i^T\mathbf{z})r_i)\} \tag{27}$$

where

$$r_i = \sum_{k=1}^{n} h(i,k) g(\sum_{j=1}^{n} h(k,j)(\mathbf{w}_j^T \mathbf{z})^2). \tag{28}$$

The function g is the derivative of G, defined, e.g. as in (23) or (25). Note that rigorously speaking, the expectation in (27) should of course be the sample average, but for simplicity, we use this notation. Of course, a stochastic gradient method could be used as well, which means omitting the averaging and taking only one sample point at a time. Due to our constraint, the vectors $\mathbf{w}_i$ must be normalized to unit variance and orthogonalized after every step of (27). The orthogonalization and normalization can be accomplished, e.g., by the classical method involving matrix square roots,

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\mathbf{W} \tag{29}$$

where $\mathbf{W}$ is the matrix $(\mathbf{w}_1, ..., \mathbf{w}_n)^T$ of the vectors. For further methods, see (Hyvärinen and Oja, 1997; Hyvärinen, 1999a).

In a neural interpretation, the learning rule in (27) can be considered as "modulated" Hebbian learning, since the learning term is modulated by the term r_i. This term could be considered as top-down feedback as in (Hyvärinen and Hoyer, 2000), since it is a function of the local energies which could be the outputs of higher-order neurons (complex cells).

After learning the $\mathbf{w}_i$, the original mixing matrix $\mathbf{A}$ can be computed by inverting the whitening process as

$$\mathbf{A} = (\mathbf{W}\mathbf{V})^{-1} = \mathbf{V}^{-1}\mathbf{W}^T \tag{30}$$

On the other hand, the rows of the inverse of $\mathbf{A}$ give the filters (weight vectors) in the original, not whitened space.

5 Discussion

5.1 Comparison with some other topographic mappings

Our method is different from ordinary topographic mappings in several ways.

The first minor difference is that whereas in most topographic mappings a single weight vector represents a single point in the data space, every vector in topographic ICA represents a direction, i.e. a one-dimensional subspace. This difference is not of much consequence, however. For example, there are versions of the Self-Organizing Map (SOM) (Kohonen, 1995) that use a single weight vector in much the same say as topographic ICA.

Second, since topographic ICA is a modification of ICA, it still attempts to find a decomposition into components that are independent. This is because only near-by components are not independent, at least approximately, in the model. In contrast, most topographic mappings choose the representation vectors by principles similar to vector quantization and clustering. This is the case, for example, with the SOM, the Generative Topographic Mapping (GTM, Bishop et al, 1997) and related models, e.g. (Kiviluoto and Oja, 1998).

Most interestingly, the very principle defining topography is different in topographic ICA and most topographic maps. Usually, the similarity of vectors in the data space is defined by Euclidean geometry: either the Euclidean distance, as in the SOM and the GTM, or the dot-product, as in the "dot-product SOM" (Kohonen, 1995). In topographic ICA, the similarity of two vectors in the data space is defined by their higher-order correlations, which cannot be expressed as Euclidean relations. It could be expressed using the general framework developed in (Goodhill and Sejnowski, 1997), though. For another non-Euclidean topographic mapping that uses proximity information, see (Graepel and Obermayer, 1999).

In fact, the topographic similarity defined in topographic ICA could be seen as a higher-order version of the dot-product measure. If the data is prewhitened, the dot-product in the data space is equivalent to correlation in the original space. Thus, topography based on dot-products could be used to define a "second-order" topography, where components near to each other in the topography have larger linear correlations. As explained above, one can constrain the components to be uncorrelated in ICA, and thus also in topographic ICA. Then any statistical dependency that could be used to create the topographic organization must be obtained from higher-order correlations, and this is exactly what happens in topographic ICA.

5.2 Connection to Independent Subspace Analysis

Our approximation of the likelihood shows clearly the connection to another modification of the classic ICA model: Independent Subspace Analysis (Hyvärinen and Hoyer, 2000). In that model, as in topographic ICA, the components s_i were not assumed to be all mutually independent. Instead, it was assumed that the s_i can be divided into couples, triplets or in general *n*-tuples, such that the s_i inside a given *n*-tuple could be dependent on each other, but dependencies between different *n*-tuples were not allowed. Related relaxations of the independence assumption were proposed in (Cardoso, 1998; Lin, 1998).

Inspired by Kohonen's principle of feature subspaces (Kohonen, 1996), the probability densities for the *n*-tuples of s_i were assumed in (Hyvärinen and Hoyer, 2000) to be *spherically symmetric*, i.e. depend only on the norm. In other words, the probability density $p_q(.)$ of the *n*-tuple with index $q \in \{1,...,Q\}$, could be expressed as a function of the sum of the squares of the $s_i, i \in S_q$ only, where we denote by S_q the set of indices of the components s_i that belong to the q-th *n*-tuple. (For simplicity, it was assumed further that the $p_q(.)$ were equal for all q, i.e. for all subspaces.) In this model, the logarithm of the likelihood can thus be expressed as

$$\log L(\mathbf{W}) = \sum_{t=1}^{T} \sum_{q=1}^{Q} G(\sum_{i \in S_q} (\mathbf{w}_i^T \mathbf{x}(t))^2) + T \log |\det \mathbf{W}| \tag{31}$$

where $G(\sum_{i \in S_q} s_i^2) = \log p_q(s_i, i \in S_q)$ gives the logarithm of the probability density inside the q-th *n*-tuple of s_i. Thus the likelihood is a function of the norms of the projections of the data onto the subspaces spanned by the $\mathbf{w}_i$ in each *n*-tuple.

It is to be expected that the norms of the projections on the subspaces represent some higher-order, invariant features. The exact nature of the invariances has not been specified in the model but will emerge from the input data, using only the prior information on their independence.

The independent subspace model introduces a certain dependence structure for the independent components. Consider two variables generated with a common variance as in (3-4). If the original variables z_1, z_2 are gaussian, the joint distribution of s_1 and s_2 is spherically symmetric, which is obvious by symmetry arguments. As was proven in connection with Eq. (3-4), the two variables then have positive correlation of energies. This shows that the higher-order correlation structure in independent subspace analysis is closely connected to that found in topographic ICA.

In fact, topographic ICA can be considered a generalization of the model of independent subspace analysis. The likelihood in (31) could be expressed as a special case of the approximative likelihood (21) with a neighborhood function of the form

$$h(i,j) = \begin{cases} 1, & \text{if } \exists q : i,j \in S_q \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

It is also easy to see that the generative model obtained from topographic ICA using this neighborhood generates spherically symmetric densities, if conditionally gaussian s_i are used.

This connection shows that topographic ICA is closely connected to the principle of invariant-feature subspaces. In topographic ICA, the invariant-feature subspaces, which are actually no longer independent, are completely overlapping. Every component has its own neighborhood, which could be considered to define a subspace. Each of the local energies $\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{x})^2$ could be considered as the square of the (weighted) norm of the projection on a feature subspace. Thus the local energies, possibly after a nonlinear transform, give the values of invariant features. In fact, this connection is one of the motivations for our approximation of the likelihood. In vision science, computation of local energy is a widely used mechanism, and has some biological plausibility (Mel et al., 1998; Emerson et al., 1992; Gray et al., 1998). The wiring diagram for such higher-order feature detectors is shown in Fig. 3.

5.3 Utility of topography

A valid question at this point is: What could be the utility of a topographic representation as opposed to the unordered representation given by ordinary ICA.

The first well-known utility is visualization (Kohonen, 1995). In exploratory data analysis, topography shows the connections between the different components, and this may give a lot of additional information. This is probably the main benefit of topography when our model is applied on blind source separation.

When the model is applied on feature extraction, arguments advanced in computational neuroscience may be used. The utility of a topographic representations in the cortex has been discussed in detail by several authors, for a review see (Kohonen, 1995; Swindale, 1996). The most relevant argument in connection to topographic ICA may be that centered on minimal wiring length (Durbin and Mitchison, 1990). The pooling into complex cell responses, i.e. computation of local energies requires that the squares of the s_i are summed, and this requires some wiring or information channels. To minimize the total length of the wiring, it would be most useful that the units whose responses are to be pooled would be as close to each other as possible. For detailed neuroanatomical arguments linking complex cell pooling and topography, see (Blasdel, 1992; DeAngelis et al., 1999).

The minimal wiring length argument is not restricted to the pooling required for complex cells. It is applicable for any operations performed on the s_i; it is reasonable to assume that such operations require mostly interactions between components that are statistically related to each other. Gain control (Heeger, 1992) is one such operation. Topographic ICA minimizes wiring length in this more general case as well. A related benefit of topography may be minimizing communication load in a parallel processing environment (Nelson and Bower, 1990). It has to be admitted, though, that these arguments are quite speculative.

One may wonder if the redundancy introduced by the topographic ordering is in contradiction to the general principle of reducing the redundancy of a representation, which has been used to justify the application of ICA for feature extraction (Olshausen and Field, 1996; Bell and Sejnowski, 1997). Here we must stress that introduction of topography does not seem to increase the redundancy of the components by any significant amount: Measuring the mutual information between the components of ICA and topographic ICA shows only a small increase in redundancy. [2] Rather, topographic ICA makes *explicit* the redundancy that cannot be cancelled in a linear decomposition.

Making the redundancy explicit by topography may in fact help further processing stages to reduce the redundancy of their components. Whether topography helps in reducing redundancy or not, it is reasonable to assume that there is some utility of the estimated dependency structure in further processing. How exactly the topography could be used is a question for future research; some speculation can be found in (Blasdel and Obermayer, 1994).

[2] We investigated change in the mutual information in the feature extraction experiments reported in Section 6.2. Mutual information cannot be easily computed in itself. However, since the weight vectors are orthogonal in the whitened space, the mutual information is given by the difference of an unknown constant and the sum of the negentropies of the components (Comon, 1994; Hyvärinen, 1999a). Therefore, we can compare the sums of negentropies, which can be much more easily approximated. Approximating the negentropies with the method in (Hyvärinen, 1998), we obtained that the topographic ICA (3×3 neighborhood, see Sec. 6.2) had a sum of negentropies that was approximately 2% larger than in ordinary ICA. This is to be compared with the 50% increase when moving from ICA to PCA. Likewise, we computed by nonparametric histogram estimation the average mutual information of two components in the three component sets (using only 100 pairs of components for computational reasons). The decrease was approximately 30% when comparing PCA either to topographic ICA or ordinary ICA; the difference in the two cases could not be made statistically significant with a reasonable amount of computational resources.

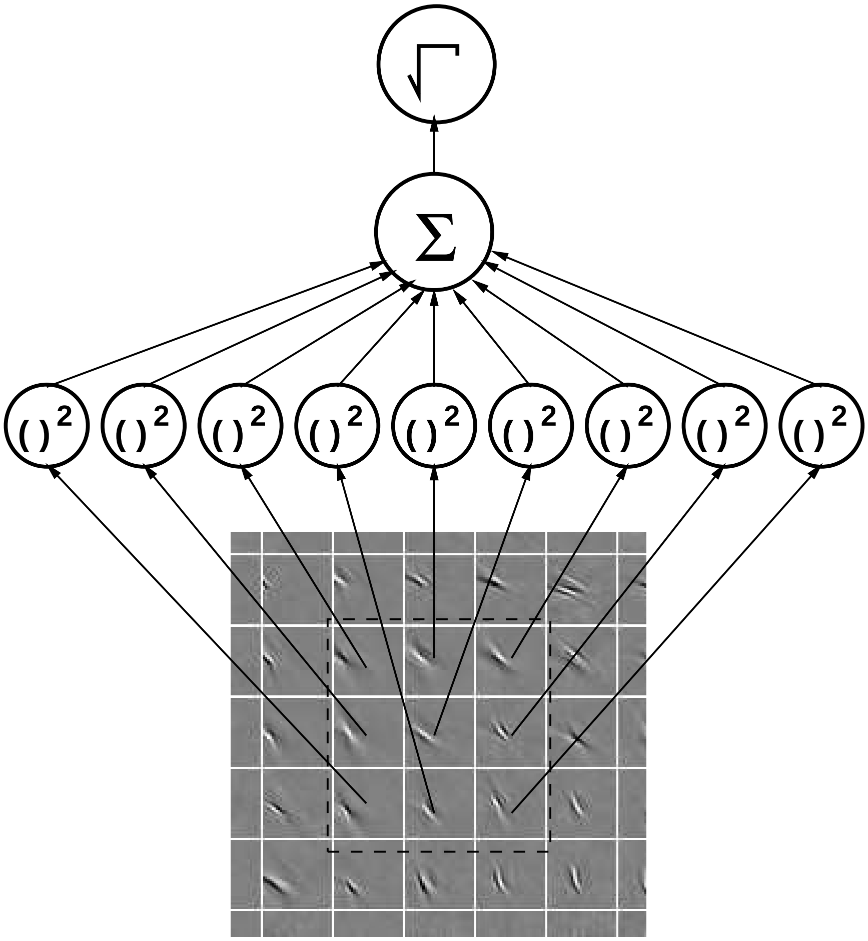

Figure 3: The wiring diagram of the higher-order feature detectors given by the local energies in topographic ICA, for a two-dimensional topography. These feature detectors are a generalization of independent feature subspaces, and could be interpreted as complex cells. The local energies are computed by first taking the squares of the outputs of linear filters, and then summing these squares inside a topographic neighborhood. A square root may be taken for normalization.

6 Simulations and experiments

6.1 Validating the likelihood approximation

Our algorithms are based on the approximation of the likelihood given in Sec. 3. To see whether this approximation is valid in practice, we generated data according to the generative model, and then estimated the model using the approximation.

To generate data according to the model, the variables u_i were first generated by taking absolute values of gaussian variables. A 1-D neighborhood function in a 20-dimensional space was defined by convolving the vector of the form $(...,0,0,0,1,1,1,1,1,0,0,0,....)$ (with 5 ones) three times with itself. The topography was ring-like, i.e. without borders. The s_i were then generated as in Section 3, with a random mixing matrix. The approximation of likelihood in Section 3 was used with G defined as in (24), with $\varepsilon = 0.005$. This was maximized by gradient ascent as in (27), using an adaptive step size. The neighborhood in the algorithm was the same one as was used to generate the data, and was assumed known.

A typical resulting matrix **WVA** is depicted in Fig. 4 a). (For simplicity, the absolute values of the elements of this matrix are shown). This matrix is a permutation matrix (up to irrelevant signs), which shows that the components s_i were estimated correctly. In fact, if the data were generated by ordinary ICA and the estimation were performed by ordinary ICA methods, the fact that we have a permutation matrix would show that the method performs adequately. But in contrast to ordinary ICA, the matrix **WVA** is here such that it completely preserves the topographic structure of the components.

Several other random initial values were used, and they all converged in equivalent results, one is shown in Fig. 4 b). Likewise, using the different nonlinearities given in (22) and (25) did not change the convergence significantly, as shown in Fig. 4 c) and d). The best results were obtained with (24), though. Experimenting with different neighborhood sizes in the algorithm, it was found that if the neighborhood used in the algorithm is smaller that the neighborhood used in generating the data, the convergence deteriorates: the global topography is not found, only a local patchy topography. In contrast, if the algorithm used a neighborhood that was somewhat larger than the one in the generation, the convergence was improved, which was rather unexpected.

Thus, these simulation results supports the validity of our approximation of the likelihood. Moreover, they support our conjecture that the exact form of nonlinearity G is not important.

6.2 Experiments with image data

A very interesting application of topographic ICA can be found with image data. Image patches (windows) can be analyzed as a linear superposition of basis vectors, as in the ICA and topographic ICA models. This gives a useful description on a low level where we can ignore such higher-level nonlinear phenomena as occlusion.

6.2.1 Data and methods

The data was obtained by taking 16×16 pixel image patches at random locations from monochrome photographs depicting wild-life scenes (animals, meadows, forests, etc.). The images were taken directly from PhotoCDs, and are available on the World Wide Web[3]. The image patches were then converted into vectors of length 256. The mean gray-scale value of each image patch (i.e. the DC component) was subtracted. The data was then low-pass filtered by reducing the dimension of the data vector by principal component analysis, retaining the 160 principal components with the largest variances, after which the data was whitened by normalizing the variances of the principal components. These preprocessing steps are essentially similar to those used in (Hyvärinen and Hoyer, 2000;

[3] WWW address: `http://www.cis.hut.fi/projects/ica/data/images/`

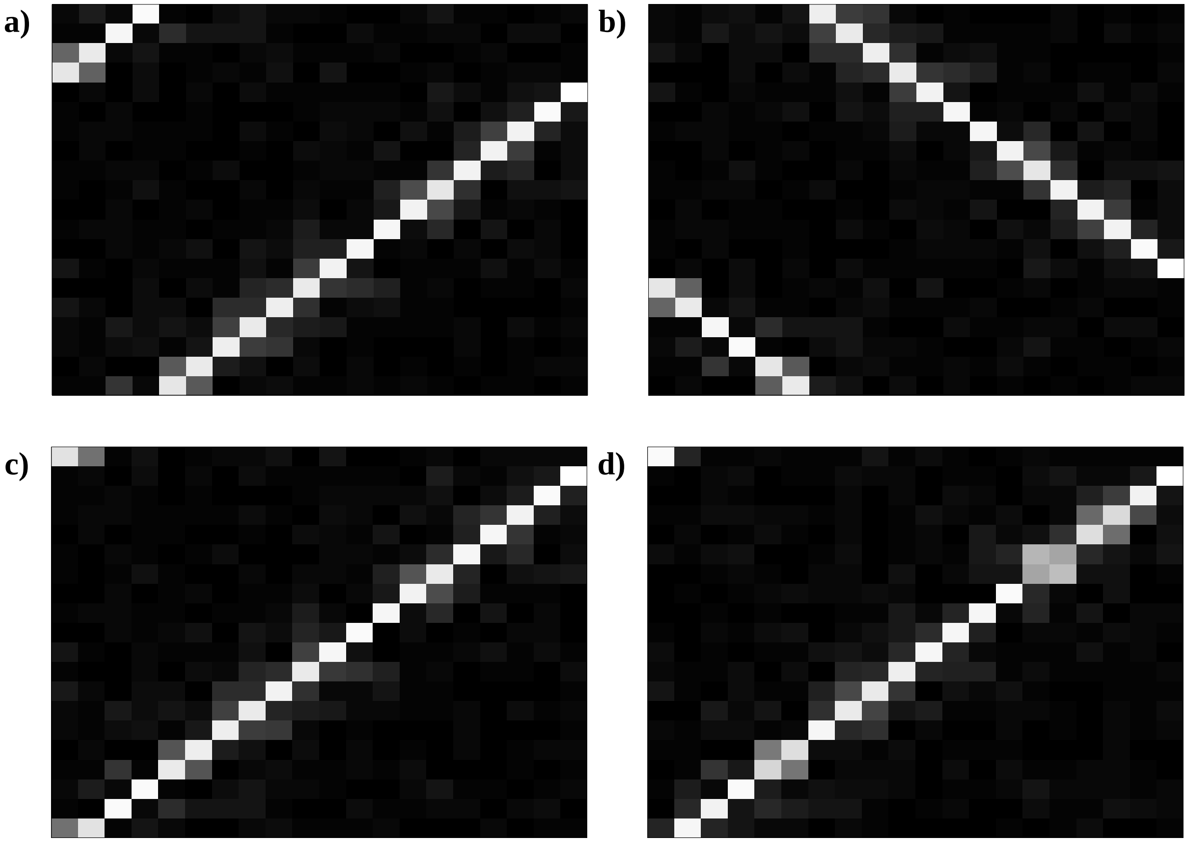

Figure 4: Validation of the approximation of the likelihood. Data was artificially generated according to the model, and the (inverse of the) mixing matrix **A** was estimated by maximizing the approximation of the likelihood by a gradient method as in Eq. (27), giving **W**. The plots a) - d) shows the matrix **WA**. a) and b) were estimated using the nonlinearity in (24) and two different initial values. c) and d) were estimated using nonlinearities (22) and (25), respectively; the initial values were different from the previous ones. The matrices in the plots are all permutation matrices, which shows that the components were estimated correctly. Moreover, any neighboring components remain neighboring after multiplication with this matrix, which shows that the topographic structure was preserved.

Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998). In the results shown below, the inverse of these preprocessing steps was performed. The fact that the data was contained in a 160 dimensional subspace meant that the 160 basis vectors now formed an orthonormal system for that subspace and not for the original space, but this did not necessitate any changes in the learning rule.

As the topography we chose a 2-D torus lattice (Kohonen, 1995). The choice of a two-dimensional grid was here motivated by convenience of visualization only; further research is needed to see what the "intrinsic dimensionality" of natural image data could be. The torus was chosen to avoid border effects. We used three different neighborhood sizes. The first one was defined so that every neighborhood consisted of a 3×3 square of 9 units. In other words, we defined the neighborhood function $h(i, j)$ so that it equals one if the components i and j are adjacent in the 2-D lattice, even obliquely; otherwise, it equals zero. The second one was defined similarly, but this this time with a 5×5 neighborhood. The third neighborhood function was defined by taking only the five non-obliquely adjacent elements in the neighborhood.

We used the three different functions in (22), (24), with parameter ε fixed at 0.001, and (25). The approximation of likelihood in Eq. (21) for 50,000 observations was maximized under the constraint of orthonormality of the filters in the whitened space, using the gradient method in (27).

6.2.2 Results

We show the two basic results, using 3×3 and 5×5 neighborhoods with the nonlinearity in (24). The obtained basis vectors, i.e. columns of the mixing matrix **A**, are shown in Fig. 5 for the smaller neighborhood and in Fig. 6 for the larger one. In both cases, the basis vectors are similar to those obtained by ordinary ICA of image data (Olshausen and Field, 1996; Bell and Sejnowski, 1997). In addition, they have a clear topographic organization. The topographic organization is somewhat different in the two cases. With the smaller neighborhood, the spatial frequency is an important factor, whereas with the larger neighborhood, it has little influence and is overridden by orientation and location. Furthermore, the basis vectors are slightly different: with the larger neighborhood, more elongated features are estimated.

The results for the nonlinearity (25) are not reported since they are not satisfactory. This was to be expected since using (25) is related to using kurtosis, and kurtosis gives poor results in image decomposition due to its adverse statistical properties (Hyvärinen, 1999a). The results for the two nonlinearities (22) and (24) were essentially identical, so we only show the results for (24). On the other hand, results for the third neighborhood are not reported since the resulting topographic ordering was too weak; this neighborhood seemed to be too small, and we obtained results closer to ordinary ICA.

The topographic organization was investigated in more detail by computing the correlations of the energies of the components, for the whole input data set (with 3×3 neighborhood). In Fig. 7, the correlations of the energies are plotted as a function of the distance on the topographic grid. The results are shown for the smaller neighborhood. One can see that the correlations are decreasing as the distance increases. This result was predicted by the model. After a certain distance, however, the correlations no longer decrease, reaching a constant value. According to the model, the correlations should continue decreasing and reach zero, but this does not happen exactly because image data does not exactly follow the model. It is probable, however, that for a much larger window size, the correlations would go to zero.

We also investigated the distributions of the higher-order components u_i in the generative model. It is not possible to definitively estimate these, since they are not observed directly. However, we were able to find distributions that generated distributions for the s_i that were very close to those observed in the data. The family of distribution for u_i that we used is as follows:

$$u_i = t_i^{\rho}, \text{where } p(t_i) = \lambda \exp(-\lambda t_i). \tag{33}$$

In other words, the variables in this family are obtained as the ρ:th power of an exponential variable with parameter λ. Thus we have a two-parameter family of densities, where λ can be considered as a scale parameter, and ρ determines the sparsity of the distribution. We estimated the parameters ρ and λ from the distribution of a typical s_i, and obtained the parameter values $\lambda = 0.09$ and $\rho = -1.8$. This provided a very good fit to the observed distibution, see Fig. 8. The important point is that this distribution has a pdf that is monotonically decreasing and has a very heavy tail; in other words, it is very sparse. This seems to be the essential requirement for the distribution of the u_i.

6.2.3 Complex cell interpretation

The connection to independent subspace analysis (Hyvärinen and Hoyer, 2000), which is basically a complex cell model, can also be found in these results. Two neighboring basis vectors in Fig. 5 tend to be of the same orientation and frequency. Their locations are near to each other as well. In contrast, their phases are very different. This means that a neighborhood of such basis vectors, i.e. simple cells, functions as a complex cell: The local energies that are summed in the approximation of the likelihood in (21) can be considered as the outputs of a complex cell, possibly after a nonlinear transformation like the square root (Hyvärinen and Hoyer, 2000). Likewise, the feedback r_i in the learning rule could be considered as coming from complex cells.

The complex cell interpretation was investigated in more detail using the same methods as in (Hyvärinen and Hoyer, 2000). We compared the responses of topographic neighborhoods with the responses of the underlying linear filters $\mathbf{w}_i$, for different stimulus configurations. The results for the two bases shown were not very satisfactory, probably because the dimension of the data was too small to allow the basis vectors to change smoothly enough as a function of position. Therefore, we computed the basis from much larger windows: 32×32 pixels. The data size was 50,000, and the dimension was reduced by PCA to 625 dimensions. The toroidal topology was thus 25×25 units, and the neighborhood was 5×5. The nonlinearity was as in (24).

First, an optimal stimulus, i.e. the one that elicits maximum response, was computed for each neighborhood and linear filter in the set of Gabor filters. The response of topographic neighborhoods was computed as the local energy. The response of linear filters was computed as the absolute value of the dot-product. One of the stimulus parameters was changed at a time to see how the response changes, while the other parameters were held constant at the optimal values. Some typical simuli are depicted in Fig. 9. The investigated parameters were phase, orientation, and location (shift). The response values were normalized so that the maximum response for each neighborhood or linear filter was equal to 1. Fig. 10 shows the median responses of the whole populations, together with the 10% and 90% percentiles. (Plotting individual response curves as in (Hyvärinen and Hoyer, 2000) gave similar results.) In Fig. 10 a), the responses are given for varying phase. The top row shows the absolute responses of the linear filters, and in the bottom row the corresponding results for the neighborhoods are depicted. The figures show that phase invariance is a rather strong property of the neighborhoods: the minimum response was usually at least half of the maximum response. This was not the case for the linear filters. Fig. 10 b) shows the results for location shift. The “receptive field” of a typical neighborhood is larger and more invariant than that of a typical linear filter. As for orientation, Fig. 10 c) and depicts the corresponding results, showing that the orientation selectivity was approximately equivalent in linear filters and neighborhoods.

Thus we see that invariances with respect to translation and especially phase, as well as orientation selectivity, are general properties of the neighborhoods. This shows that topographic ICA shows emergence of properties similar to those of complex cells. These results are very similar, qualitatively as well as quantitatively, to those obtained by independent subspace analysis in (Hyvärinen and Hoyer, 2000). Thus we have a connection in both of these models between natural image statistics and the most basic properties of complex cells. Complex cells have many other properties as well, and it remains to be seen which of them can be explained by natural image statistics.

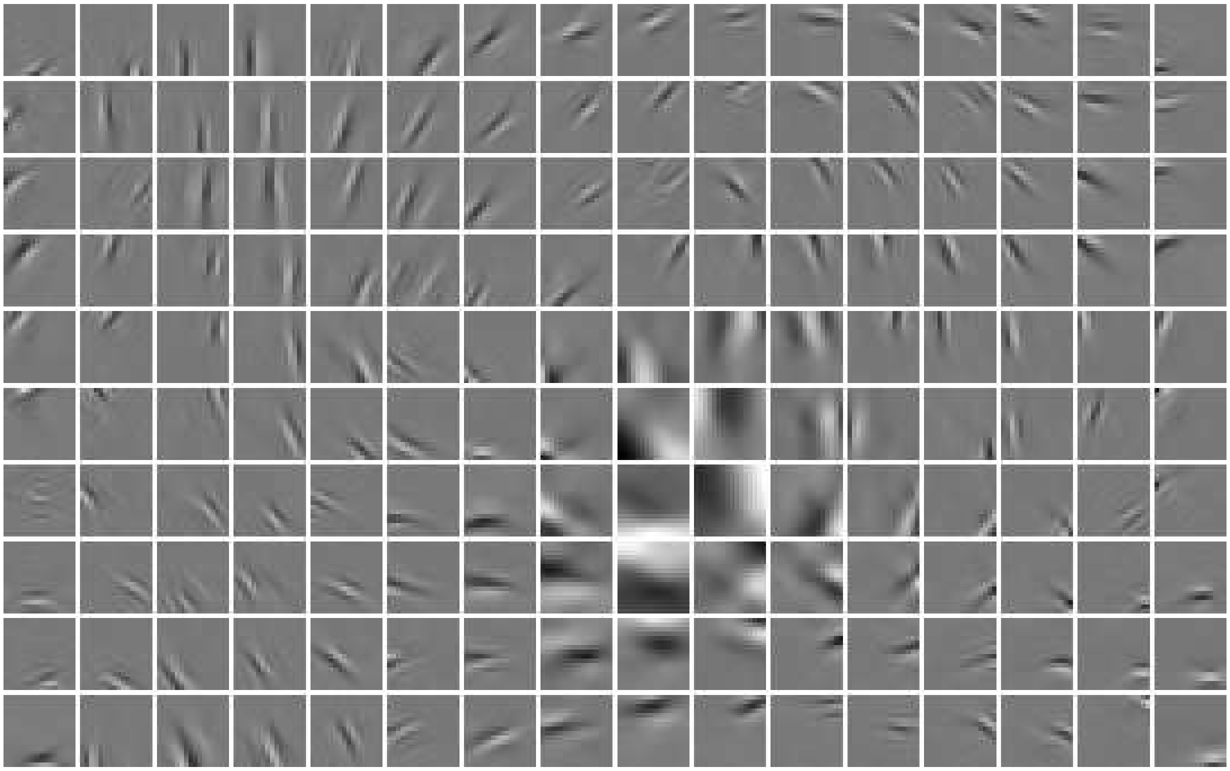

Figure 5: Topographic ICA of natural image data. Neighborhood size 3×3. The model gives Gabor-like basis vectors for image windows. Basis vectors that are similar in location, orientation and/or frequency are close to each other. The phases of nearby basis vectors are very different, giving each neighborhood properties similar to those of complex cells (see Fig. 10).

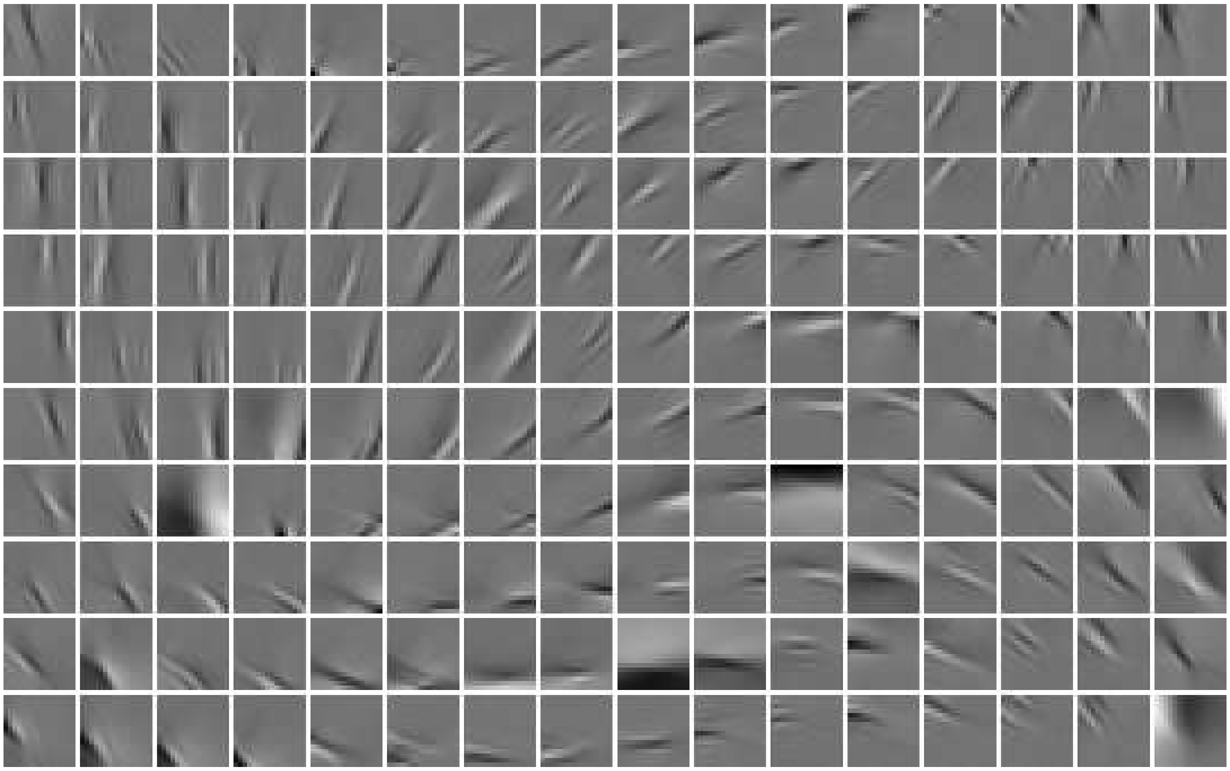

Figure 6: Topographic ICA of natural image data, this time with neighborhood size 5 × 5. With this bigger neighborhood, the topographic order is more strongly influenced by orientation.

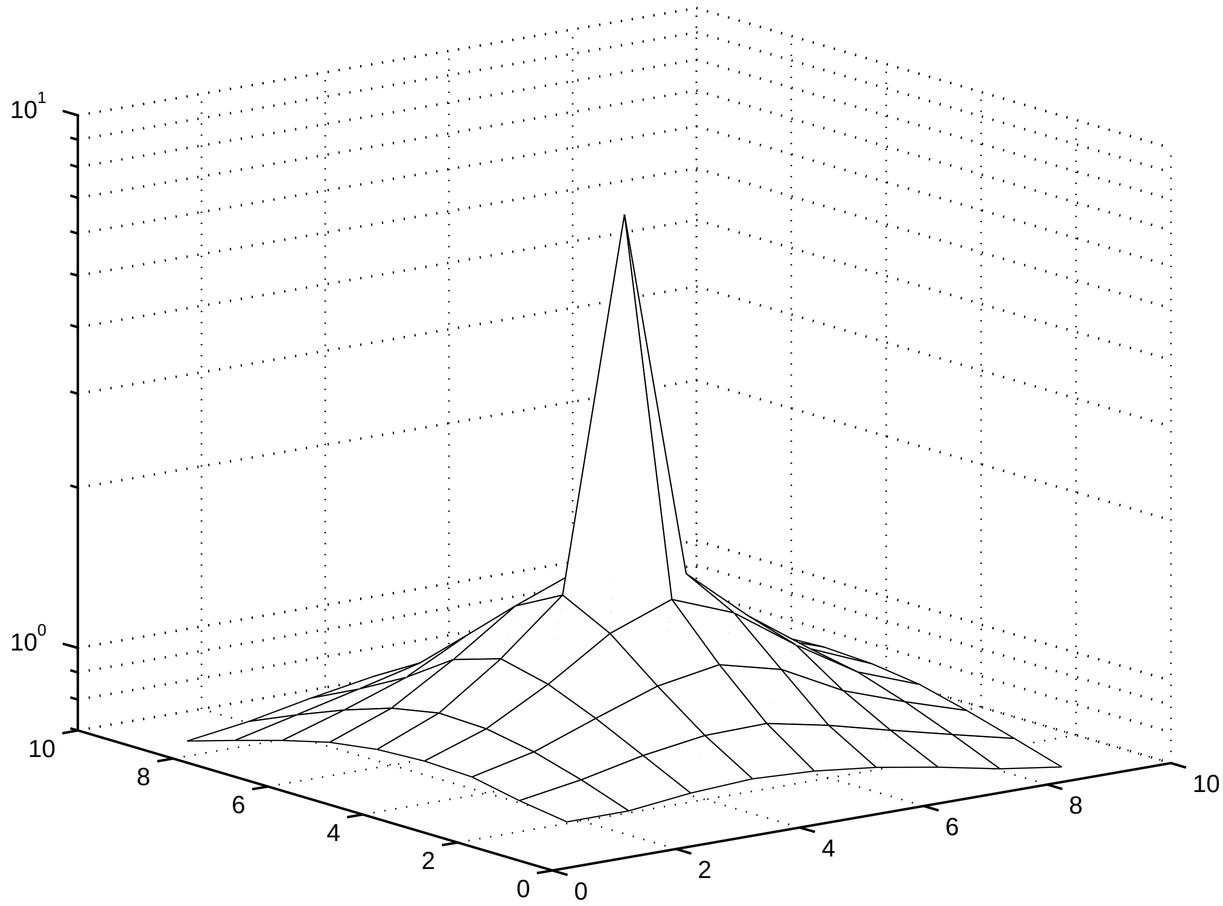

Figure 7: Analysis of the higher-order correlations of the components estimated from image data. The plot shows the covariances of energies (in log-scale) of the components as a function of the relative position on the topographic grid. The covariances were averaged over all components. The plot shows that the covariances are a decreasing function of distance.

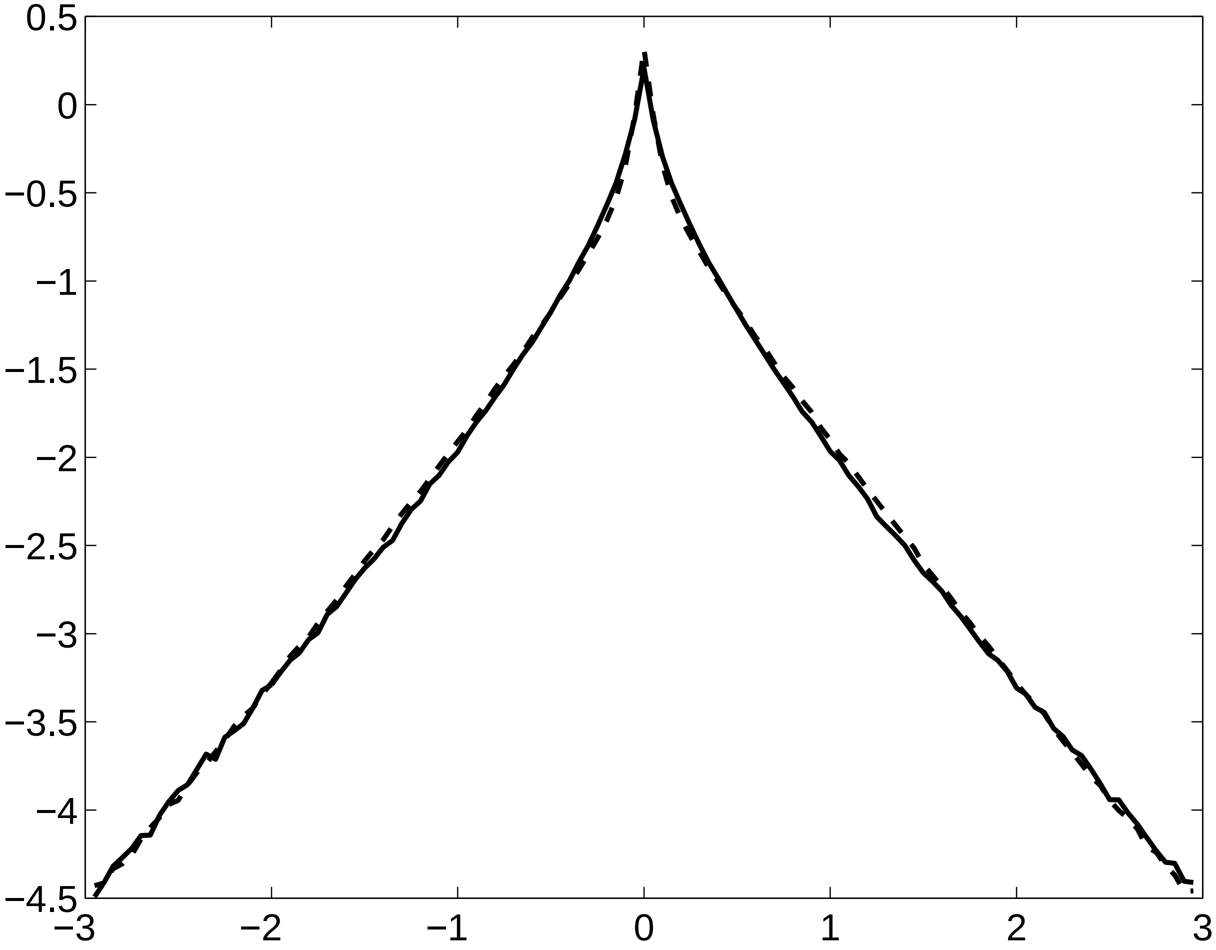

Figure 8: Distribution of the independent components obtained by the distribution in (33), compared to the one observed in image data. The fit is very good.

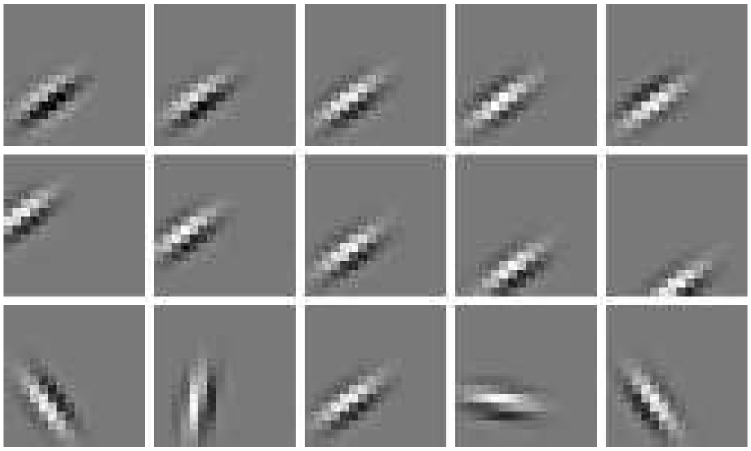

Figure 9: Typical stimuli used in the experiments in Fig. 10 below. The middle column shows an original Gabor stimulus. One of the parameters was varied at a time. Top row: varying phase. Middle row: varying location (shift). Bottom row: varying orientation.

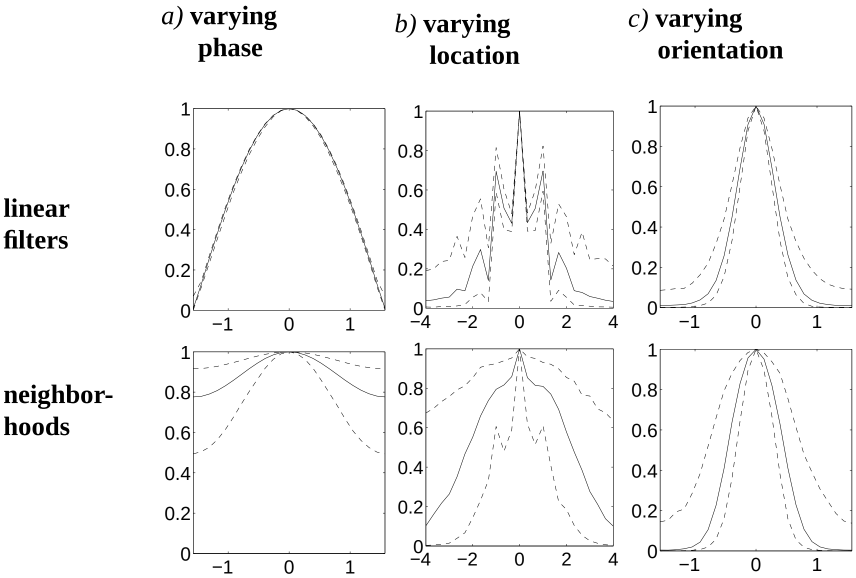

Figure 10: Statistical analysis of the properties of the neighborhoods, with the corresponding results for linear filters given for comparison. In all plots, the solid line gives the median response in the population of all neighborhoods (or linear filters), and the dotted lines give the 90% and 10% percentiles of the responses. Stimuli were as in Fig. 9. Top row: responses (in absolute values) of linear filters (simple cells). Bottom row: responses of neighborhoods (complex cells), given by local energies. a) Effect of varying phase. b) Effect of varying location (shift). c) Effect of varying orientation.

6.3 Experiments with magnetoencephalographic recordings

An application of topographic ICA that is very different from feature extraction can be found in blind separation of artifacts in magnetoencephalographic recordings.

6.3.1 Data and methods

Two minutes of magnetoencephalographic (MEG) data was collected using a 122-channel whole-scalp neuromagnetometer device. The sensors measured the gradient of the magnetic field in two orthogonal directions at 61 distinct locations. The measurement device and the data are described in detail in (Vigário et al., 1998). The test subject was asked to blink and make horizontal eye saccades in order to produce typical ocular artifacts and bite the teeth for 20 seconds in order to create myographic artifacts. This 122 dimensional input data was first reduced to 20 dimensions by PCA, in order to eliminate noise and "bumbs", which appear in the data if the dimensionality is not sufficiently reduced (Hyvärinen et al., 1999). Then a band-pass filtering was performed with the pass band between 0.5 and about 45 Hz. This eliminated most of the powerful low-frequency noise and the effect of the power grid at 50 Hz.

The topographic ICA algorithm was then run on the data using a one dimensional ring-shaped topography. The neighborhood was formed by convolving a vector of three ones with itself four times. The nonlinearity G was as in (24).

6.3.2 Results

The resulting separated signals are shown in Fig. 11. The signals themselves are very similar to those found by ordinary ICA in (Vigário et al., 1998). As for the topographic organization, we can see that

1. The signals corresponding to bites (#9-#15) are now adjacent. When computing the field patterns corresponding to these signals, one can also see that the signals are ordered according to whether they come from the left or the right side of the head.

2. Two signals corresponding to eye artifacts are adjacent as well (#18 and #19). The signal #18 corresponds to horizontal eye saccades and the signal #19 to eye blinks. A signal which seems to relate to eye activity has been separated into #17.

We can also see signals that do not seem to have any meaningful topographic relations, probably because they are quite independent form the rest of the signals. These include the heart beat (signal #7), and a signal corresponding to a digital watch which was at a distance of 1 m from the magnetometer (signal #6). Their adjacency in the results shown seems to be a pure coincidence, and was not consistently found when repeating the estimation with different initial values. In contrast, the cluster related to muscle artifacts emerged practically always, and the cluster related to eye activity emerged quite often.

Thus we see that topographic ICA finds largely the same components as those found by ICA in (Vigário et al., 1998). Using topographic ICA has the advantage, however, that signals are grouped together according to their dependencies. Here we see two clusters, one created by the signals coming from the muscle artifact, and the other by eye muscle activity.

7 Conclusion

We introduced topographic ICA, which is a generative model that combines topographic mapping with ICA. As in all topographic mappings, the distance in the representation space (on the topographic “grid”) is related to the distance

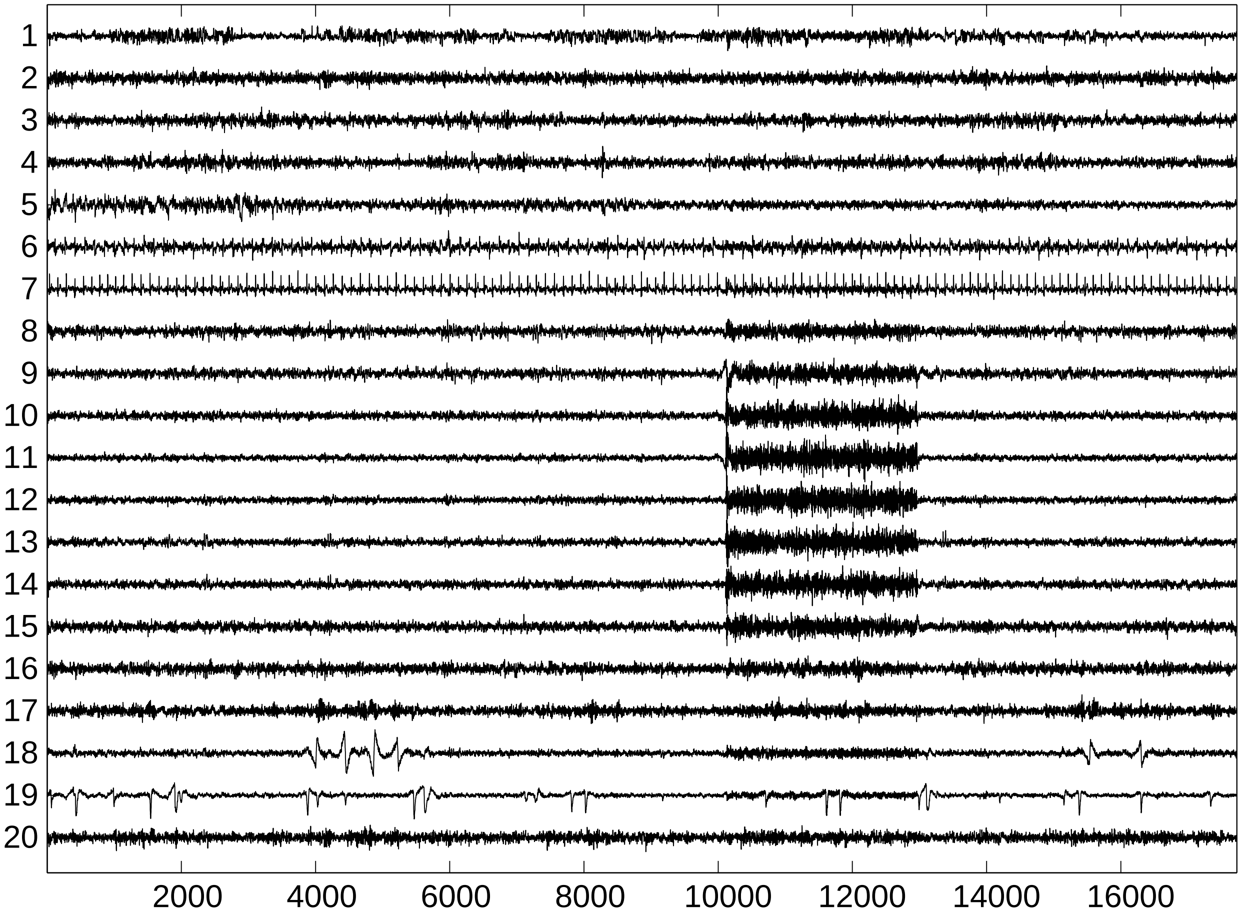

Figure 11: The source signals found by topographic ICA from MEG data.

of the represented components. In topographic ICA, the distance between represented components is defined by the mutual information implied by the higher-order correlations, which gives the natural distance measure in the context of ICA. This is in contrast to most existing topographic mapping methods, where the distance is defined by basic geometrical relations like Euclidean distance or correlation, as in e.g. (Kohonen, 1995; Bishop et al., 1998; Goodhill and Sejnowski, 1997). In fact, our principle makes it possible to define a topography even among a set of orthogonal vectors, whose Euclidean distances are all equal.

To estimate the model in practice, we considered maximum likelihood estimation. Since the likelihood of the model is intractable in general, we derived an approximation (actually, a lower bound) of the likelihood. The approximation makes it possible to derive a simple gradient learning rule for estimation of the model. This leads to an interesting form of Hebbian learning, where the Hebbian term is modulated by top-down feedback.

An interesting application of this novel model of topographic organization is found with natural image data, where topographic ICA gives a linear decomposition into Gabor-like linear features. In contrast to ordinary ICA, the higher-order dependencies that linear ICA could not remove define a topographic order such that near-by cells tend to be active at the same time. Also, the topographic neighborhoods resemble complex cells in their responses. Our model thus shows simultaneous emergence of topographic organization and properties similar to those of complex cells. On the other hand, when applying the model to blind separation of MEG artifacts, we can separate more or less the same artifacts as with ICA, with an additional topographic ordering that can be used for visualizing the results.

References

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

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

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

Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10:215–234.

Blasdel, G. and Obermayer, K. (1994). Putative strategies of scene segmentation in monkey visual cortex. *Neural Networks*, 7:865–881.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, 12(8):3139–3161.

Cardoso, J.-F. (1998). Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98),* Seattle, WA.

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

Cichocki, A. and Unbehauen, R. (1996). 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.

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

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., and Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience*, 19(10):4046–4064.

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

Dempster, A. P., Laird, N., and Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. of the Royal Statistical Society, ser. B*, 39:1–38.

Durbin, R. and Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, 343:644–647.

Emerson, R. C., Bergen, J. R., and Adelson, E. H. (1992). Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research*, 32:203–218.

Goodhill, G. J. and Sejnowski, T. J. (1997). A unifying objective function for topographic mappings. *Neural Computation*, 9(6):1291–1303.

Graepel, T. and Obermayer, K. (1999). A stochastic self-organizing map for proximity data. *Neural Computation*, 11:139–155.

Gray, M. S., Pouget, A., Zemel, R. S., Nowlan, S. J., and Sejnowski, T. J. (1998). Reliable disparity estimation through selective integration. *Visual Neuroscience*, 15(3):511–528.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198.

Hyvärinen, A. (1998). New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279. MIT Press.

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

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

Hyvärinen, A. (1999c). Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128.

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

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

Hyvärinen, A., Särelä, J., and Vigário, R. (1999). 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.

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

Karhunen, J., Oja, E., Wang, L., Vigário, R., and Joutsensalo, J. (1997). A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504.

Kiviluoto, K. and Oja, E. (1998). S-Map: A network with a simple self-organization algorithm for generative topographic mappings. In *Advances in Neural Information Processing Systems*, volume 10. The MIT Press.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, New York.

Kohonen, T. (1996). Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75:281–291.

Lin, J. K. (1998). Factorizing multivariate function classes. In *Advances in Neural Information Processing Systems*, volume 10, pages 563–569. The MIT Press.

Mel, B. W., Ruderman, D. L., and Archie, K. A. (1998). Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *Journal of Neuroscience*, 18:4325–4334.

Nelson, M. E. and Bower, J. M. (1990). Brain maps and parallel computers. *Trends in Neurosciences*, 13(10):403–408.

Oja, E. (1997). The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46.

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

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

Pajunen, P. (1998). Blind source separation using algorithmic information theory. *Neurocomputing*, 22:35–48.

Pham, D.-T., Garrat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774.

Simoncelli, E. P. and Schwartz, O. (1999). 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.

Swindale, N. V. (1996). The development of topography in the visual cortex: a review of models. *Network*, 7(2):161–247.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society ser. B*, 265:359–366.

Vigário, R. (1997). Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph. clin. Neurophysiol.*, 103(3):395–404.

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

A Derivation of learning rule

Here we derive (27). Consider our approximation of the likelihood:

$$\log \tilde{L}(\mathbf{W}) = \sum_{t=1}^{T} \sum_{j=1}^{n} G\left(\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2\right) + T \log|\det \mathbf{W}|, \tag{34}$$

where the data $\mathbf{z}$ is whitened. A first point to note is that when we constrain the weight matrix $\mathbf{W}$ to be orthogonal, the term $T \log|\det \mathbf{W}|$ is zero, since the absolute value of the determinant of an orthogonal matrix is always equal to one. Thus this term can be omitted.

The computation of the gradient is essentially reduced to computing the gradient of

$$F_j(\mathbf{W}) = G\left(\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2\right) \tag{35}$$

Denote by $\mathbf{w}_k^l$ the l-th component of $\mathbf{w}_k$. By the chain rule, we obtain

$$\frac{\partial F_j}{\partial \mathbf{w}_k^l} = g(\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2)[\sum_{i=1}^n 2h(i,j)\delta_{ik}(\mathbf{w}_i^T \mathbf{z}(t))z_l(t)]$$

$$= g(\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2)2h(k,j)(\mathbf{w}_k^T \mathbf{z}(t))z_l(t). \quad (36)$$

This can be written in matrix form:

$$\nabla_{\mathbf{w}_k} F_j = 2g(\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2)h(k,j)(\mathbf{w}_k^T \mathbf{z}(t))\mathbf{z}(t). \quad (37)$$

Thus, we obtain

$$\nabla_{\mathbf{w}_k}(\log \tilde{L}(\mathbf{W})) = \sum_{t=1}^T \sum_{j=1}^n \nabla F_j = \sum_{t=1}^T \sum_{j=1}^n 2g(\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2)h(k,j)(\mathbf{w}_k^T \mathbf{z}(t))\mathbf{z}(t)$$

$$= 2\sum_{t=1}^T \mathbf{z}(t)(\mathbf{w}_k^T \mathbf{z}(t)) \sum_{j=1}^n h(k,j)g(\sum_{i=1}^n h(i,j)(\mathbf{w}_i^T \mathbf{z}(t))^2) \quad (38)$$

Now, we can replace the sum over t by the expectation, and switch the indices i, j, and k, which is merely a notational change. Furthermore, we can omit the constant 2 which does not change the direction of the gradient, and denote the latter sum as in (28). Thus we obtain (27).