A Two-Layer Dynamic Generative Model of Natural Image Sequences

Jarmo Hurri and Aapo Hyvärinen

Neural Networks Research Centre

Helsinki University of Technology

P.O.Box 9800, 02015 HUT, Finland

{jarmo.hurri,aapo.hyvarinen}@hut.fi

September 30, 2002

Abstract

We present a two-layer dynamic generative model of the statistical structure of natural image sequences. The second layer of the model is a linear mapping from simple cell outputs to pixel values, as in most work on natural image statistics. The first layer models the dependencies of the activity levels (amplitudes or variances) of the simple cells, using a multivariate autoregressive model. The second layer shows emergence of basis vectors that are localized, oriented and have different scales, just like previous work. But our new model enables the first layer to learn connections between the simple cells that are similar to complex cell pooling: connections are strong among cells with similar

location, frequency and orientation. In contrast to previous work in which one of the layers needed to be fixed in advance, the dynamic model enables us to estimate both of the layers simultaneously from natural data.

1 Introduction

A central question in the study of sensory neural networks is how stimuli are represented or coded by neurons. Knowledge of the properties of this neural code is needed when one wants to study the computations which take place at different stages of neural processing. One approach to studying the neural code is to examine how its properties are related to the statistics of natural stimuli (Simoncelli and Olshausen, 2001). In this approach it is assumed that the statistics of the natural input have affected the structure of the networks via natural selection or during development.

In the visual system, the primary visual cortex is an area which is relatively well known from the point of view of neurophysiology. There is a large amount of knowledge about what different types of cells exist in this area, the responses of these cells to different visual stimuli, and the connections and physical layout of these cells. Within the past ten years, computational principles relating the properties of cells in this area to the statistics of natural stimuli have been proposed. The most influential of these theories have been sparse coding (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2001), independent component analysis (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998; Hyvärinen et al.,

2001), and temporal coherence (Földiák, 1991; Kayser et al., 2001; Wiskott and Sejnowski, 2002; Hurri and Hyvärinen, 2002). In sparse coding, the fundamental property of the neural code is that only a small proportion of the cells is activated by a given stimulus. In independent component analysis, the outputs of different cells are as independent of each other as possible. In the case of image data, these two principles are closely related (Hyvärinen et al., 2001).

The principle of temporal coherence (Földiák, 1991; Mitchison, 1991; Stone, 1996) is based on the idea that when processing temporal input, the representation changes as little as possible over time. This principle has been traditionally associated with complex cells (Földiák, 1991; Kayser et al., 2001; Wiskott and Sejnowski, 2002; Einhäuser et al., 2002; Berkes and Wiskott, 2002), which are considered to be invariant detectors. However, in a recent paper (Hurri and Hyvärinen, 2002) we have shown that a nonlinear form of temporal coherence is also related to the structure of simple cell receptive fields. According to the results presented in (Hurri and Hyvärinen, 2002), simple cell receptive fields are optimally temporally coherent in the sense that the *activity levels* of simple cells are stable over short time intervals. By activity level we mean the amplitude or energy of the output of a linear filter that models a simple cell. (However, the principle seems to be somewhat applicable even in the case of non-negative cell outputs – see (Hurri and Hyvärinen, 2002) for a discussion.)

The measure of temporal activity coherence introduced in (Hurri and Hyvärinen, 2002) took the sum of the temporal activity coherences of single cells. Therefore, there was no possibility of an interaction between the activ-

ity levels of different cells. In this paper, we introduce a model which includes inter-cell activity dependencies. This is accomplished by a generative model in which the activity levels are generated in an autoregressive manner.

The idea of describing a generative model of natural stimuli, and interpreting the hidden variables of this model as a neural representation may at first seem counterintuitive, because the stimuli are not generated by the neural network. However, when the task of vision is considered as a problem of inverse graphics, the approach makes a lot of sense (Hinton and Ghahramani, 1997; Olshausen, 2003). A generative model can express explicitly information about the regularities in the stimuli as properties of hidden variables. If these regularities can be used to make inferences about the underlying real world, the visual system might utilize such an efficient internal representation of its stimuli.

The generative model presented in this paper is a dynamic two-layer model of natural image sequences. We will show that estimation of the model from natural image sequence data yields simple-cell-like receptive fields, and a completely unsupervised complex-cell-like pooling between the outputs of simple cells. In what follows we will first describe the generative model. This is followed by the description of an estimation algorithm, and simulations with artificially generated data. Then we apply the algorithm to natural image sequences data, and analyze the structure of the resulting model.

2 Definition of the model

The generative model of natural image sequences introduced in this paper has two layers (see Figure 1). The first layer is a multivariate autoregressive model of the activity levels (amplitudes) of simple cell responses at time t and time $t - \Delta t$. The signs of cell responses are generated by a second latent signal between the first and second layer. The second layer is linear, and maps cell responses to image features.

[Figure 1 about here.]

We start the formal description of the model with the second, linear layer. We restrict ourselves to linear spatial models of simple cells. Let vector $\mathbf{x}(t)$ denote an image patch taken from natural image sequences at time t. (Vectorization of image patches can be done by scanning images column-wise into vectors.) Let the vector $\mathbf{y}(t) = [y_1(t) \cdots y_K(t)]^T$ represent the outputs of K simple cells. The linear generative model for $\mathbf{x}(t)$ is similar to the one in (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2001):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{y}(t). \tag{1}$$

Here $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_K]$ denotes a matrix which relates the image patch $\mathbf{x}(t)$ to the outputs of simple cells, so that each column $\mathbf{a}_k$, $k = 1, ..., K$, gives the feature that is coded by the corresponding simple cell. When the parameters of the model are estimated, what we obtain first is the mapping from $\mathbf{x}(t)$ to $\mathbf{y}(t)$, denoted by

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t). \tag{2}$$

Conceptually, the set of filters (vectors) $\mathbf{w}_1, ..., \mathbf{w}_K$ corresponds here to the receptive fields of simple cells, and $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]^T$ denotes a matrix with all the filters as rows. The dimension of $\mathbf{x}(t)$ is typically larger than the dimension of $\mathbf{y}(t)$, so that (2) is generally not invertible but an underdetermined set of linear equations. A one-to-one correspondence between $\mathbf{W}$ and $\mathbf{A}$ can be established by computing the pseudoinverse solution[1] $\mathbf{A} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$.

In contrast to sparse coding (Olshausen and Field, 1996) or independent component analysis (Hyvärinen et al., 2001) we do *not* assume that the components of $\mathbf{y}(t)$ are independent. Instead, we assume that the activity levels (amplitudes) of the components of $\mathbf{y}(t)$ are correlated. We model these dependencies with a multivariate autoregressive model in the first layer of our model. Let $\mathbf{abs}\left(\mathbf{y}(t)\right) = \left[|y_1(t)| \cdots |y_K(t)|\right]^T$, and let $\mathbf{v}(t)$ denote a driving noise signal (the distribution of $\mathbf{v}(t)$ is constrained by the non-negativity of the process, and will be discussed in more detail below). Let $\mathbf{M}$ denote a $K \times K$ matrix, and let Δt denote a time lag. Our model for the activities is a *constrained multidimensional first-order autoregressive process*, defined by

$$\mathbf{abs}\left(\mathbf{y}(t)\right) = \mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t - \Delta t)\right) + \mathbf{v}(t), \tag{3}$$

and unit energy constraints

$$\mathrm{E}_t\left\{y_k^2(t)\right\} = 1 \tag{4}$$

[1]When the solution is computed with the pseudoinverse, the solved $\mathbf{x}(t)$ is orthogonal to the nullspace of $\mathbf{W}$, $\mathcal{N}\left(\mathbf{W}\right) = \left\{\mathbf{b} \,|\, \mathbf{W}\mathbf{b} = \mathbf{0}\right\}$. In other words, that part of $\mathbf{x}(t)$ which would be ignored by the linear mapping in equation (2) is set to $\mathbf{0}$.

for $k = 1, ..., K$ and $k_2 = 1, ..., K$. Actually, the constraint of unit energy is not really a constraint but rather a convention. The scale of the latent variables is not well defined because we can arbitrarily multiply a latent variable by a constant and divide the corresponding column of $\mathbf{A}$ by the same constant without affecting the model (a similar situation is found in ICA). Thus, we can define the scale of the $y_k(t)$ as we like.

There are dependencies between the driving noise $\mathbf{v}(t)$ and filter output activities $\mathbf{abs}(\mathbf{y}(t))$, caused by the non-negativity of $\mathbf{abs}(\mathbf{y}(t))$. To define a generative model for the driving noise $\mathbf{v}(t)$ so that the non-negativity of the absolute values holds, we proceed as follows. Let $\mathbf{u}(t)$ denote a zero-mean random vector with components which are statistically independent of each other. We define

$$\mathbf{v}(t) = \mathbf{max}\left(-\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t - \Delta t)\right), \mathbf{u}(t)\right), \tag{5}$$

where, for vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{max}(\mathbf{a}, \mathbf{b}) = [\max(a_1, b_1) \ \cdots \ \max(a_n, b_n)]^T$. We assume that $\mathbf{u}(t)$ and $\mathbf{abs}(\mathbf{y}(t))$ are uncorrelated.

To make the generative model complete, a mechanism for generating the signs of cell responses $\mathbf{y}(t)$ must be included. We specify that the signs are generated randomly with equal probability for plus or minus after the strengths of the responses have been generated. All the signs are mutually independent, both over time and the cell population, and also independent of the activity levels. Note that one consequence of this random generation of signs is that that filter outputs are uncorrelated, which can be shown as follows. Let $k_1 \neq k_2$, and let $s_{k_1}(t)$ and $s_{k_2}(t)$ denote the generated signs. Then we have $\mathrm{E}_t\{y_{k_1}(t)y_{k_2}(t)\} = \mathrm{E}_t\{s_{k_1}(t)\,|y_{k_1}(t)|\,s_{k_2}(t)\,|y_{k_2}(t)|\} =$

$\underbrace{\mathrm{E}_t\left\{s_{k_1}(t)\right\}}_{=0}\underbrace{\mathrm{E}_t\left\{s_{k_2}(t)\right\}}_{=0}\mathrm{E}_t\left\{|y_{k_1}(t)|\,|y_{k_2}(t)|\right\} = 0.$

Note that the unit energy constraints and the uncorrelatedness of the outputs can be represented by a single matrix equation

$$\mathbf{W}\mathbf{C}_{\mathbf{x}(t)}\mathbf{W}^T = \mathbf{I}, \tag{6}$$

where $\mathbf{C}_{\mathbf{x}(t)} = \mathrm{E}_t\left\{\mathbf{x}(t)\mathbf{x}(t)^T\right\}$, and that they imply $\mathrm{E}_t\left\{\|\mathbf{y}(t)\|^2\right\} = K$. Therefore, because the sign generation mechanism also implies that each $y_k(t)$ has zero mean, the variances of the outputs will also be constant.

In equation (3), a large positive matrix element $\mathbf{M}(i,j)$, or $\mathbf{M}(j,i)$, indicates that there is a strong dependency between the activities of cells i and j. Thinking in terms of grouping cells with large activity dependencies together, matrix $\mathbf{M}$ can be thought of as containing similarities (reciprocals of distances) between different cells. We will use this property in the experimental section to derive a spatial organization of the simple cells from $\mathbf{M}$.

Note that the driving noise $\mathbf{v}(t)$ could be considered as closely related to complex cell outputs, that is, higher-order features. Typically, this innovation process would be very sparse for image data. When the process does take a positive value, this will cause activity in some simple cells, and this activity will spread to other simple cells in the next time steps, though diminished at every step.

3 Estimation of the model

To estimate the model defined above we need to estimate both $\mathbf{M}$ and $\mathbf{W}$ (the pseudoinverse of $\mathbf{A}$). In this section we first show how to estimate $\mathbf{M}$, given $\mathbf{W}$. Then we describe an objective function which can be used to estimate $\mathbf{W}$, given $\mathbf{M}$. Each iteration of the estimation algorithm consists of two steps. During the first step $\mathbf{M}$ is updated, and $\mathbf{W}$ is kept constant; during the second step these roles are reversed.

First, regarding the estimation of $\mathbf{M}$, consider a situation in which $\mathbf{W}$ has been fixed. It is shown in Appendix A that $\mathbf{M}$ can be estimated by using approximative method of moments, and that the estimate is given by

$$\begin{aligned}\widehat{\mathbf{M}} \approx \beta \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\} \\ \times \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\}^{-1},\end{aligned} \tag{7}$$

where $\beta > 1$. We will return to the role of the scalar multiplier β below.

The estimation of $\mathbf{W}$ is more complicated. A rigorous derivation of an objective function based on well-known estimation principles is very difficult because the statistics involved are non-Gaussian, and the processes have difficult interdependencies. Therefore, instead of deriving an objective function from first principles, we derived an objective function heuristically, and verified through simulations that the objective function is capable of estimating models generated according to the two-layer model. The objective function is a weighted sum of the covariances of filter output amplitudes at times $t - \Delta t$

and t, defined by

$$f(\mathbf{W}, \mathbf{M}) = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{M}(i,j) \operatorname{cov} \left\{ |y_i(t)|, |y_j(t - \Delta t)| \right\}, \tag{8}$$

which can also be expressed as

$$\begin{aligned} f(\mathbf{W}, \mathbf{M}) = \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t \left\{ \mathbf{abs}\left(\mathbf{y}(t)\right) \right\} \right)^T \mathbf{M} \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t - \Delta t)\right) - \mathrm{E}_t \left\{ \mathbf{abs}\left(\mathbf{y}(t)\right) \right\} \right) \Big\}. \end{aligned} \tag{9}$$

(The function f depends on $\mathbf{W}$ through the relationship (2).) The estimation of $\mathbf{W}$ is thus accomplished by maximizing this objective function

$$\widehat{\mathbf{W}} = \arg \max_{\mathbf{W}} f(\mathbf{W}, \mathbf{M}). \tag{10}$$

Optimization of the objective function f over $\mathbf{W}$ under constraint (6) uses a gradient projection approach (Hurri and Hyvärinen, 2002). The initial value of $\mathbf{W}$ is selected randomly.

Note that the scalar multiplier β in (7) has a constant linear effect in objective function (9). Because this scaling does not affect the optima of (9), and because we are more interested in the relative magnitudes of elements of $\mathbf{M}$ than their absolute values, we can discard β in the estimation process. Therefore, in the estimation we set $\beta = 1$. However, in the validation of the estimation method the real value of this coefficient must be taken into account. This case will be considered in detail below.

4 Experiments with artificial data

Before applying the estimation method to natural data, we wanted to verify its validity using artificial data. In order to do this we first generated 100

different matrices $\mathbf{M}$ and $\mathbf{A}$, and used these to generate data which followed our model. The dimension of the data was $K = 10$, and both $\mathbf{M}$ and $\mathbf{A}$ were 10×10 matrices. Input noise $\mathbf{u}(t)$ was Gaussian white noise. In generating the data, care must be taken so that the the constraints are fulfilled, and that the resulting autoregressive model is stable. Details on how the data was generated are given in Appendix B.1.

After data generation we ran our estimation algorithm 100 times, once for each of the data sets, to obtain estimates $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{W}}$ (estimate of the pseudoinverse of $\mathbf{A}$) of all the original matrices. Because of the insensitivity of the objective function (9) to a different ordering of the components of $\mathbf{y}(t)$, care had to be taken to compensate for a possible permutation; details on how this was done are described in Appendix B.2. After compensating for the possible permutation, the effect of the unknown scalar multiplier β in equation (7) had to be accounted for. This was done by using equation (7) to estimate β by

$$\hat{\beta} = \frac{\|\mathbf{M}\|_{\mathrm{F}}}{\left\|\widehat{\mathbf{M}}\right\|_{\mathrm{F}}} \tag{11}$$

(remember that estimate $\widehat{\mathbf{M}}$ is obtained by setting $\beta = 1$ in equation (7)). To analyze the convergence of the algorithm, we examined how the relative estimation errors $\frac{\|\mathbf{M}-\widehat{\mathbf{M}}\|_{\mathrm{F}}}{\|\mathbf{M}\|_{\mathrm{F}}}$ and $\frac{\|\mathbf{W}-\widehat{\mathbf{W}}\|_{\mathrm{F}}}{\|\mathbf{W}\|_{\mathrm{F}}}$ change as a function of number of iterations (here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, i.e., the sum of the squares of all the elements of its argument).

Figure 2 shows the resulting plots of the relative errors. The plots show the median and the maximum of the errors of the estimates of $\mathbf{M}$ and $\mathbf{W}$, computed over the whole set of 100 runs. The median and maximum of

the errors are plotted as a function of iteration number. As we can see, the estimate of $\mathbf{W}$ converges very reliably to the true value.

[Figure 2 about here.]

As for the estimation of $\mathbf{M}$, the scalar multiplier β estimated as in equation (11) was consistently greater than 1, as predicted in Appendix A. The relative error of the estimate of $\mathbf{M}$ decreases considerably in the estimation, but the final estimate is not as good as in the case of $\mathbf{W}$. This is probably due to the approximation made in its estimation (see Appendix A). However, a large part of the error is caused by a systematic bias in the estimate that does not seem to be critical in our analysis of the results. The nature of the bias can be seen in Figure 3, which shows a scatter plot of the true elements of the 100 matrices $\mathbf{M}$ vs. their estimates. We can see that the systematic bias is largely a nonlinear *element-wise* relationship between the true value of an element of $\mathbf{M}$ and its estimate. This nonlinear relationship is a monotonic convex function, characterized by large positive deviations from the true value when the absolute value of the element of $\mathbf{M}$ is large. Remembering that the Frobenius norm – which is used to measure the relative error – emphasizes large errors, we can see that a large part of the relative error results from this systematic bias.

[Figure 3 about here.]

In the analysis of results with real data we are mostly interested in the magnitudes of the elements of $\mathbf{M}$ with respect to other elements of the same matrix. These relationships are preserved by a smooth monotonic mapping

of the elements of $\mathbf{M}$, like the systematic bias described above. In Figure 4 we have plotted four first matrices $\mathbf{M}$ from the set of 100 matrices, along with their estimates $\widehat{\mathbf{M}}$. Although there are some differences in some individual elements of the matrices, especially in elements with large absolute values, the structures of the true matrices and their estimates look very much alike. This is because the relative values of the elements with respect to the values of the other elements are similar.

[Figure 4 about here.]

5 Experiments with natural image sequences

5.1 Data collection and preprocessing

The data and preprocessing used in the experiments were very similar to those in (Hurri and Hyvärinen, 2002), so we will describe them only shortly here, and refer the reader to (Hurri and Hyvärinen, 2002) for details.

The natural image sequences used in data collection consisted of 129 image sequences, which were a subset of natural image sequences used in (van Hateren and Ruderman, 1998). The sampling rate in these sequences was 25 Hz. Initially 200,000 image sequences with a duration of 440 ms, and spatial size 16×16 pixels, were sampled from these sequences. The fairly long duration of these initial samples was necessary because of the temporal filtering used in preprocessing,

The preprocessing consisted of four steps: temporal decorrelation, subtraction of local mean, normalization, and dimensionality reduction. Tempo-

ral decorrelation enhances temporal changes in the data, and differentiates our results from those obtained with static images (Hurri and Hyvärinen, 2002). It can also be motivated as a model of temporal processing at the lateral geniculate nucleus (Dong and Atick, 1995). Temporal decorrelation was performed with a temporal filter of length 400 ms. The length of the resulting sequences, which was also the time delay Δt in our experiment, was 40 ms. After temporal decorrelation the spatial local mean (spatial DC component) was subtracted from each of the 400,000 image patches, and the patches were normalized to unit norm. This normalization can be considered as a form of contrast gain control (Carandini et al., 1997; Heeger, 1992). Finally, to reduce the effect of noise and aliasing artifacts, the dimensionality of the data was reduced to 160 using principal component analysis (Hyvärinen et al., 2001).

5.2 Results

The estimation algorithm described in Section 3 was applied to the preprocessed natural image sequence data to obtain estimates for $\mathbf{M}$ and $\mathbf{A}$ (the pseudoinverse of $\mathbf{W}$). Figure 5 shows the resulting basis vectors – that is, columns of $\mathbf{A}$. As can be seen, the resulting basis vectors are localized, oriented, and have multiple scales. These are the most important defining criteria of simple cell receptive fields (Palmer, 1999). These qualitative features are also characteristic of results obtained with independent component analysis or sparse coding (Olshausen and Field, 1996; van Hateren and van der Schaaf, 1998) and purely temporal activity coherence (Hurri and Hyvärinen,

2002). This suggests that, as far as receptive field structure is concerned, these methods are roughly equivalent to each other in that similar receptive fields emerge when the methods are applied to natural stimuli.

[Figure 5 about here.]

The estimated matrix $\mathbf{M}$ captures the spatiotemporal activity dependencies between the basis vectors shown in Figure 5. The diagonal elements of the estimated $\mathbf{M}$ are relatively large, ranging from 0.31 to 0.74 with a mean of 0.44, indicating that for all the basis vectors, activities at time $t - \Delta t$ and time t have considerable correlation. This is in concordance with the results in (Hurri and Hyvärinen, 2002). A histogram of the non-diagonal elements of $\mathbf{M}$, which contain the information about spatiotemporal dependencies between the basis vectors, is shown in Figure 6. In order to examine these dependencies more closely, we first plotted the basis vectors with the highest and lowest activity dependency values for a set of representative reference vectors. The results, shown in Figure 7, suggest that basis vectors with high positive activity dependencies code for similar features at nearby positions, whereas basis vectors with low (negative) dependencies code for features with different scale or orientation.

[Figure 6 about here.]

[Figure 7 about here.]

To visualize the spatiotemporal dependencies of all of the basis vectors, we used the interpretation of $\mathbf{M}$ as a similarity matrix (see Section 2). Matrix

$\mathbf{M}$ was first converted to a non-negative similarity matrix $\mathbf{M}_\mathrm{s}$ by subtracting $\min_{i,j} \mathbf{M}(i,j)$ from the elements of $\mathbf{M}$, and by setting the diagonal elements to value 1. Multidimensional scaling (MDS) was then applied to $\mathbf{M}_\mathrm{s}$ by interpreting the values $1 - \mathbf{M}_\mathrm{s}(i,j)$ and $1 - \mathbf{M}_\mathrm{s}(j,i)$ as distances between cells i and j. The objective of MDS is to map the points in a (high-dimensional) space to a two-dimensional space (a plane) so that the distances between the points in the original space are preserved as well as possible on the plane. A central concept in the application of MDS to a particular problem is the measurement scale (Borg and Groenen, 1997; SAS/STAT, 2000), which is a mathematical description of the type of information contained in the measurements of proximity. We applied MDS to our data so that the interval measurement scale (Borg and Groenen, 1997; SAS/STAT, 2000) was assumed. Informally, the interval measurement scale is characterized so that relative sizes of differences between measurements are meaningful, but there is no absolute zero. This makes sense in our case, because firstly, the differences between the elements of $\mathbf{M}_\mathrm{s}$ should tell us something about the differences of strengths of spatiotemporal dependencies, and secondly, we do not know the maximum possible spatiotemporal dependency in natural image sequence data (the absolute zero).

The resulting spatial layout produced by the MDS procedure is shown in Figure 8. Because some of the points in the planar representation were very close to each other, some small distances were stretched (some of the tightest clusters were magnified) in order to be able to show the basis vectors in a reasonable scale (without overlap between the basis patches). As in Figure 7, we can see that those basis vectors which are very close to each other seem

to be mostly coding for similarly oriented features with the same frequencies at nearby spatial positions. This kind of grouping is characteristic of pooling of simple cell outputs at complex cell level, as well as of the topographic organization of the visual cortex (Palmer, 1999). Note that this grouping effect is not a result of the magnification of the tightest clusters described above; in fact, the magnification reduces the effect. In addition to the local topography described above, some global topography also emerges in the results: those basis vectors which code for horizontal features are on the left in Figure 8, while those that code for vertical features are on the right.

[Figure 8 about here.]

Thus, the estimation of our two-layer model from natural image sequences yields both simple-cell-like receptive fields (Figure 5), and grouping similar to the pooling of simple cell outputs and local topography in the primary visual cortex (Figures 7 and 8). The receptive fields emerge in the second layer (matrix $\mathbf{A}$), and cell output grouping emerges in the first layer (matrix $\mathbf{M}$). Both of these layers emerge simultaneously during the estimation of the model. This is a significant improvement on earlier statistical models of early vision (Hyvärinen and Hoyer, 2000; Hyvärinen and Hoyer, 2001; Wainwright and Simoncelli, 2000), because no a priori fixing of either of these layers is needed.

6 Discussion

There are two main contributions in this paper. First, to our knowledge, the generative model presented here is the first two-layer generative model of natural image sequences presented in literature. A multi-layered description of the stimuli is biologically important because it enables us to capture dependencies within the different layers of sensory processing. In our case, the results suggest that simple cell outputs have temporal activity dependencies, and that cells at the next level of processing (complex cells) pool simple cell outputs so that cells with high activity dependencies are pooled together. This can provide important cues as to how different layers in the visual pathway are connected.

In addition, the results suggest that temporal activity dependencies could also be reflected in the topography of the primary visual cortex – that is, cells with high temporal activity dependency seem to be physically located close to each other within the cortex. Earlier research has shown that *simultaneous* activity dependency is also reflected in the organization of the cortex in a similar manner (Hyvärinen and Hoyer, 2001). Therefore, it seems possible that "activity bubbles" (Hyvärinen et al., 2002), activations of simple cells which are contiguous both in space and time, appear on the cortical surface when a stimulus with appropriate characteristics (orientation, scale) is present in the visual field. This is an intriguing characterization of the neural code at the simple cell level, the implications of which are a subject of future research.

Second, this paper also makes a rather different contribution, describ-

ing a general-purpose two-layer model that is a generalization of the basic generative models used in blind source separation. The generative model described in this paper employs nonlinearities and interdependencies, resulting in a model which is difficult to solve using well-known estimation principles. Therefore, when developing the estimation algorithm, we had to resort to approximation and heuristics. However, as we have shown above, the resulting algorithm can estimate fairly well the unknown parameters from data which follows our model. Matrix $\mathbf{A}$ can be estimated with great accuracy. Matrix $\mathbf{M}$ can also be recovered up to a fairly small relative error, and a systematic bias which is irrelevant for our purposes. This generative model could be applied to many of those applications in which blind source separation algorithms have been successful, such as brain imaging and data analysis (Hyvärinen et al., 2001).

Research related to the results presented here can be found in research concerning natural image statistics, blind source separation, and econometrics. The outputs of related wavelet filters with uncorrelated outputs exhibit a similar dependency in natural images (Wainwright and Simoncelli, 2000; Schwartz and Simoncelli, 2001): the conditional variance of the output of one filter is larger when the output of the other filter has a large amplitude. In a more generative-model setting, dependence of the simultaneous activity levels between simple cells have been used in modeling complex cells and topography (Hyvärinen and Hoyer, 2000; Hyvärinen and Hoyer, 2001). In these models, the second (pooling) layer was fixed and only the first layer was estimated. In blind source separation, Bayesian methods have been used to extract sources with nonlinear dynamics and nonlinear mapping from state

space to observations (Valpola and Karhunen, 2002). In econometrics, autoregressive conditional heteroskedasticity (ARCH) models (e.g., (Bera and Higgins, 1993)) are used to model econometric time series in which variance changes over time, and is highly correlated over time, thereby exhibiting temporal coherence of high activity. Multivariate ARCH models can be used to model cases where the variances of different time series have dependencies.

To conclude, we have described a two-layer dynamic generative model of image sequences, and an algorithm for estimating the model from sample data. Application of the estimation algorithm to natural image sequences yields a set of linear filters, or basis vectors, which are similar to simple cell receptive fields, and connections between the simple cells that are similar to the way in which simple cell outputs are pooled at the complex cell level. The basis vectors are learned in one layer of the model, and the pooling property in the other. Both layers are learned simultaneously and in a completely unsupervised manner.

Acknowledgements

We would like to thank Bruno Olshausen, Patrik Hoyer, Jarkko Venna, and Kai Puolamäki for comments and interesting discussions. Funding was provided by Helsinki Graduate School in Computer Science and Engineering (J.H.) and the Academy of Finland, Academy Fellow position (A.H.).

A Estimation of M

We estimate $\mathbf{M}$ using the method of moments. From (3) we get

$$\mathrm{E}_t\left\{\mathbf{v}(t)\right\} = \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\} - \mathbf{M}\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}.$$

Therefore

$$\begin{aligned}
&\mathrm{E}_t\left\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\right\} \\
&\quad = \mathrm{E}_t\Big\{\left(\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) + \mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\
&\qquad\quad \times \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\} \\
&\quad = \mathrm{E}_t\Big\{\big(\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathbf{M}\mathrm{E}_t\left\{\mathbf{y}(t)\right\} \\
&\qquad\quad + \mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\} + \mathbf{M}\mathrm{E}_t\left\{\mathbf{y}(t)\right\}\big) \\
&\qquad\quad \times \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\} \\
&\quad = \mathrm{E}_t\Big\{\left(\mathbf{M}(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{y}(t)\right\}\right) + \mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{v}(t)\right\}\right) \\
&\qquad\quad \times \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\} \\
&\quad = \mathbf{M}\mathrm{E}_t\Big\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\
&\qquad\quad \times \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\} \\
&\qquad + \mathrm{E}_t\left\{\left(\mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{v}(t)\right\}\right)\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\right\}.
\end{aligned} \tag{12}$$

The second term in equation (12) is non-zero because of the non-negativity of $\mathbf{abs}\left(\mathbf{y}(t)\right)$, which is implemented by equation (5). However, we can approximate this term. Let us say that the non-negativity constraint in (5) is

active for a proportion $\alpha \in (0,1)$ of the whole sample. Then we approximate

$$\begin{aligned}
&\mathrm{E}_t\left\{\left(\mathbf{v}(t)-\mathrm{E}_t\left\{\mathbf{v}(t)\right\}\right)\left(\mathbf{y}(t-\Delta t)-\mathrm{E}_t\left\{\mathbf{y}(t)\right\}\right)^T\right\}\\
&\quad\approx \alpha \mathrm{E}_t\left\{\left(-\mathbf{M}\mathbf{y}(t-\Delta t)-\mathrm{E}_t\left\{\mathbf{M}\mathbf{y}(t-\Delta t)\right\}\right)\left(\mathbf{y}(t-\Delta t)-\mathrm{E}_t\left\{\mathbf{y}(t)\right\}\right)^T\right\}\\
&\qquad(1-\alpha)\underbrace{\mathrm{E}_t\left\{\left(\mathbf{u}(t)-\mathrm{E}_t\left\{\mathbf{u}(t)\right\}\right)^T\left(\mathbf{y}(t-\Delta t)-\mathrm{E}_t\left\{\mathbf{y}(t-\Delta t)\right\}\right)^T\right\}}_{=\mathbf{0}}\\
&\quad=-\alpha\mathbf{M}\mathrm{E}_t\left\{\left(\mathbf{y}(t)-\mathrm{E}_t\left\{\mathbf{y}(t)\right\}\right)\left(\mathbf{y}(t)-\mathrm{E}_t\left\{\mathbf{y}(t)\right\}\right)^T\right\}.
\end{aligned}$$

Using this approximation we get from equation (12)

$$\begin{aligned}
\widehat{\mathbf{M}} \approx \frac{1}{1-\alpha}&\mathrm{E}_t\Big\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\\
&\qquad\times\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\}\\
&\times\mathrm{E}_t\left\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\right\}^{-1}.
\end{aligned}\tag{13}$$

Setting $\beta = 1/(1-\alpha)$ in this equation yields (7).

B Mathematical details of the validation of the estimation algorithm

B.1 Data generation

The generated data must follow equations (1), (3) and (4). In addition, $\mathbf{M}$ must be specified so that the autoregressive model (3) is stable.

The main steps of data generation were as follows (details are given below). First, we chose a random $\mathbf{M}$ that is stable. In order to generate $\mathbf{y}(t)$ we first generated positive magnitude data according to the autoregressive

model (3), and then assigned a random sign for each value. We then modified the data so that the constraints specified in (4) were fulfilled. This latter step also affects the temporal model in (3), so during the latter step the parameters of (3) were updated. After this we chose a random $\mathbf{A}$, and used it to generate observed data $\mathbf{x}(t)$ linearly from $\mathbf{y}(t)$.

To generate data according to the temporal equation (3), a non-negative matrix $\mathbf{M}_0$ was first generated by assigning the absolute value of a random number from a normal distribution with mean zero and variance one, to each of its elements, and then ensuring the stability of the autoregressive model by normalizing $\mathbf{M}_0$ so that its spectral norm[2] was between 0.6 and 0.8 (the actual value of the norm was chosen randomly from this interval during each run). After this a sample of $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ of length 40000 points was generated using equations (3) and (5) with a random (non-negative) starting point $\left|\mathbf{y}_0(0)\right|$.

Signed data $\mathbf{y}_0(t)$ was generated from $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ by assigning a random sign with equal probability for plus or minus. As was shown in Section 2, this step alone guarantees that the components of $\mathbf{y}(t)$ are uncorrelated.

The unit energy constraint on each of the components of $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ was enforced by normalizing the components. This is equivalent to premultiplying $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ with a diagonal matrix $\mathbf{\Lambda}$, where $\mathbf{\Lambda}(k,k) = \frac{1}{\sqrt{\mathrm{E}_t\left\{y_{0,k}^2(t)\right\}}}$, so that $\mathbf{abs}\left(\mathbf{y}(t)\right) = \mathbf{\Lambda}\,\mathbf{abs}\left(\mathbf{y}_0(t)\right)$. Substituting $\mathbf{y}(t)$ with $\mathbf{y}_0(t)$ in equation (3), and

[2]The spectral norm of a matrix $\mathbf{B}$, denoted by $\left\|\mathbf{B}\right\|_2$, is defined to be the square root of the largest eigenvalue of $\mathbf{B}^T\mathbf{B}$. If $\left\|\mathbf{M}\right\|_2 < 1$, then the autoregressive model is stable because $\left\|\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t)\right)\right\| \leq \left\|\mathbf{M}\right\|_2 \left\|\mathbf{abs}\left(\mathbf{y}(t)\right)\right\|$ (Horn and Johnson, 1985).

premultiplying with $\mathbf{\Lambda}$ yields

$$\begin{aligned}
\mathbf{\Lambda}\,\mathbf{abs}\,(\mathbf{y}_0(t)) &= \mathbf{\Lambda}\mathbf{M}_0\,\mathbf{abs}\,(\mathbf{y}_0(t-\Delta t)) + \mathbf{\Lambda}\mathbf{v}_0(t) \\
\mathbf{abs}\,(\mathbf{y}(t)) &= \underbrace{\mathbf{\Lambda}\mathbf{M}_0\mathbf{\Lambda}^{-1}}_{=\mathbf{M}}\mathbf{\Lambda}\,\mathbf{abs}\,(\mathbf{y}_0(t-\Delta t)) + \underbrace{\mathbf{\Lambda}\mathbf{v}_0(t)}_{=\mathbf{v}(t)} \\
\mathbf{abs}\,(\mathbf{y}(t)) &= \mathbf{M}\,\mathbf{abs}\,(\mathbf{y}(t-\Delta t)) + \mathbf{v}(t),
\end{aligned}$$

where $\mathbf{M} = \mathbf{\Lambda}\mathbf{M}_0\mathbf{\Lambda}^{-1}$ is the final parameter matrix of the generated data, and $\mathbf{v}(t) = \mathbf{\Lambda}\mathbf{v}_0(t)$ is the driving noise of the model. This scaling also affects the spectral norm of $\mathbf{M}$ – the values of these norms varied between 0.7 and 1.2. The variances of the components of $\mathbf{v}(t)$ varied between 0.4 and 1.2.

To generate the observed data $\mathbf{x}(t)$ from $\mathbf{y}(t)$, a random number from a normal distribution with mean zero and variance one was first assigned to each of the elements of matrix $\mathbf{A}$, which was then applied to $\mathbf{y}(t)$ according to equation (1).

B.2 Compensating for a possible permutation of components of $\mathbf{y}(t)$

The objective function (9) is insensitive to a reordering of the components of $\mathbf{y}(t)$, and possible sign changes. Let $\mathbf{y}_2(t) = \mathbf{P}\mathbf{y}(t)$, where $\mathbf{P}$ is a signed permutation matrix. This permutation needs to be compensated in both layers of the model (equations (1) and (3)).

First, concerning the linear layer, let $\mathbf{A}_2$ denote the linear basis corresponding to $\mathbf{y}_2(t)$ (see equation (1)). We have $\mathbf{A}\mathbf{y}(t) = \mathbf{x}(t) = \mathbf{A}_2\mathbf{y}_2(t) = \mathbf{A}_2\mathbf{P}\mathbf{y}(t)$, or

$$\mathbf{A} = \mathbf{A}_2\mathbf{P}. \tag{14}$$

Second, concerning the temporal layer, let $\mathbf{P}_{\mathrm{a}}$ denote an unsigned permutation matrix $\mathbf{P}_{\mathrm{a}} = \mathbf{abs}\left(\mathbf{P}\right)$, where $\mathbf{abs}\left(\cdot\right)$ takes an absolute value of each of the elements of its argument, and let $\mathbf{M}_2$ denote the temporal matrix corresponding to $\mathbf{y}_2(t)$ (see equation (3)). For the magnitudes of $\mathbf{y}_2(t)$ we have $\mathbf{abs}\left(\mathbf{y}_2(t)\right) = \mathbf{P}_{\mathrm{a}}\,\mathbf{abs}\left(\mathbf{y}(t)\right)$, so $\mathbf{abs}\left(\mathbf{y}(t)\right) = \mathbf{P}_{\mathrm{a}}^{-1}\,\mathbf{abs}\left(\mathbf{y}_2(t)\right) = \mathbf{P}_{\mathrm{a}}^{T}\,\mathbf{abs}\left(\mathbf{y}_2(t)\right)$ Substituting $\mathbf{y}(t)$ with $\mathbf{y}_2(t)$ in equation (3), and premultiplying with $\mathbf{P}_{\mathrm{a}}^{T}$ yields

$$\begin{aligned}
\mathbf{P}_{\mathrm{a}}^{T}\,\mathbf{abs}\left(\mathbf{y}_2(t)\right) &= \mathbf{P}_{\mathrm{a}}^{T}\mathbf{M}_2\,\mathbf{abs}\left(\mathbf{y}_2(t-\Delta t)\right) + \mathbf{P}_{\mathrm{a}}^{T}\mathbf{v}_2(t) \\
\mathbf{abs}\left(\mathbf{y}(t)\right) &= \mathbf{P}_{\mathrm{a}}^{T}\mathbf{M}_2\mathbf{P}_{\mathrm{a}}\mathbf{P}_{\mathrm{a}}^{T}\,\mathbf{abs}\left(\mathbf{y}_2(t-\Delta t)\right) + \mathbf{P}_{\mathrm{a}}^{T}\mathbf{v}_2(t) \\
\mathbf{abs}\left(\mathbf{y}(t)\right) &= \underbrace{\mathbf{P}_{\mathrm{a}}^{T}\mathbf{M}_2\mathbf{P}_{\mathrm{a}}}_{=\mathbf{M}}\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) + \underbrace{\mathbf{P}_{\mathrm{a}}^{T}\mathbf{v}_2(t)}_{=\mathbf{v}(t)},
\end{aligned}$$

so

$$\mathbf{M} = \mathbf{P}_{\mathrm{a}}^{T}\mathbf{M}_2\mathbf{P}_{\mathrm{a}}. \tag{15}$$

To convert the previous equations into a procedure, let $\widehat{\mathbf{W}}_{\mathrm{p}}$ and $\widehat{\mathbf{M}}_{\mathrm{p}}$ denote the estimates computed with the estimation method (corresponding to possibly permutated outputs), and $\mathbf{A}$ and $\mathbf{M}$ denote the correct parameter matrices corresponding to the generated data. We first use (14) to compute an estimate of the permutation matrix, $\widehat{\mathbf{P}} = \mathbf{round}\left(\widehat{\mathbf{W}}_{\mathrm{p}}\mathbf{A}\right)$, where $\mathbf{round}\left(\cdot\right)$ rounds every element of its argument to the nearest integer. (If $\widehat{\mathbf{P}}$ is singular, as usually happens during the first few rounds of the algorithm when examining its convergence, then it is not sensible to carry on to compute the estimates for $\mathbf{A}$ or $\mathbf{M}$.) An estimate for $\mathbf{A}$ can be computed using (14) again: $\widehat{\mathbf{A}} = \widehat{\mathbf{W}}_{\mathrm{p}}^{-1}\widehat{\mathbf{P}}$. The unsigned permutation matrix $\widehat{\mathbf{P}}_{\mathrm{a}} = \mathbf{abs}\left(\widehat{\mathbf{P}}\right)$ can be used to compute an estimate of $\mathbf{M}$ with equation (15): $\widehat{\mathbf{M}} = \widehat{\mathbf{P}}_{\mathrm{a}}^{T}\widehat{\mathbf{M}}_{\mathrm{p}}\widehat{\mathbf{P}}_{\mathrm{a}}$.

References

Bell, A. and Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338.

Bera, A. K. and Higgins, M. L. (1993). ARCH models: Properties, estimation and testing. *Journal of Economic Surveys*, 7(4):305–366.

Berkes, P. and Wiskott, L. (2002). Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties. In Dorronsoro, J. R., editor, *Artificial Neural Networks – ICANN 2002*, volume 2415 of *Lecture notes in computer science*, pages 81–86. Springer.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer.

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, 17(21):8621–8644.

Dong, D. W. and Atick, J. (1995). Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178.

Einhäuser, W., Kayser, C., König, P., and Körding, K. (2002). Learning the invariance properties of complex cells from their responses to natural stimuli. *European Journal of Neuroscience*, 15(3):475–486.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200.

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

Hinton, G. and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B*, 352:1177–1190.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Hurri, J. and Hyvärinen, A. (2002). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*. In press.

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 Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423.

Hyvärinen, A., Hurri, J., and Väyrynen, J. (2002). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. Submitted.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. (2001). Extracting slow subspaces from natural videos leads to com-

plex cells. In Dorffner, G., Bischof, H., and Hornik, K., editors, *Artificial Neural Networks – ICANN 2001*, volume 2130 of *Lecture notes in computer science*, pages 1075–1080. Springer.

Mitchison, G. (1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, 3(3):312–320.

Olshausen, B. A. (2003). Principles of image representation in visual cortex. In Chalupa, L. and Werner, J., editors, *The Visual Neurosciences*. The MIT Press. In press.

Olshausen, B. A. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Palmer, S. E. (1999). *Vision Science – Photons to Phenomenology*. The MIT Press.

SAS/STAT (2000). *SAS/STAT Users Guide, version 8*. SAS Publishing.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216.

Stone, J. (1996). Learning visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492.

Valpola, H. and Karhunen, J. (2002). An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*. In press.

van Hateren, J. H. and Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1412):2315–2320.

van Hateren, J. H. and van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265(1394):359–366.

Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 855–861. The MIT Press.

Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

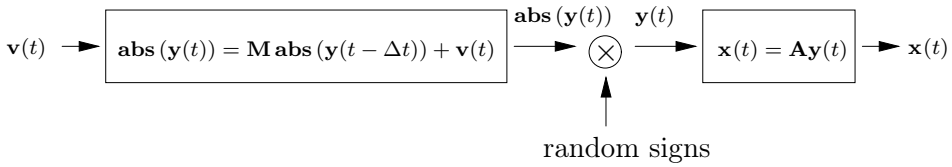

Figure 1: The two layers of the generative model with spatiotemporal activity dependencies. Let $\mathbf{abs}\,(\mathbf{y}(t)) = [|y_1(t)| \cdots |y_K(t)|]^T$ denote the activity levels (amplitudes) of simple cell responses. In the first layer, the driving noise signal $\mathbf{v}(t)$ generates the activities of simple cells via a multivariate autoregressive model. Matrix $\mathbf{M}$ captures the spatiotemporal activity dependencies in the model. The signs of the responses are generated randomly between the first and second layer to yield signed responses $\mathbf{y}(t)$. In the second layer, natural image sequence $\mathbf{x}(t)$ is generated linearly from simple cell responses. In addition to the relations shown here, the generation of $\mathbf{v}(t)$ is affected by $\mathbf{M}\,\mathbf{abs}\,(\mathbf{y}(t-\Delta t))$ to ensure non-negativity of $\mathbf{abs}\,(\mathbf{y}(t))$. See text for details.

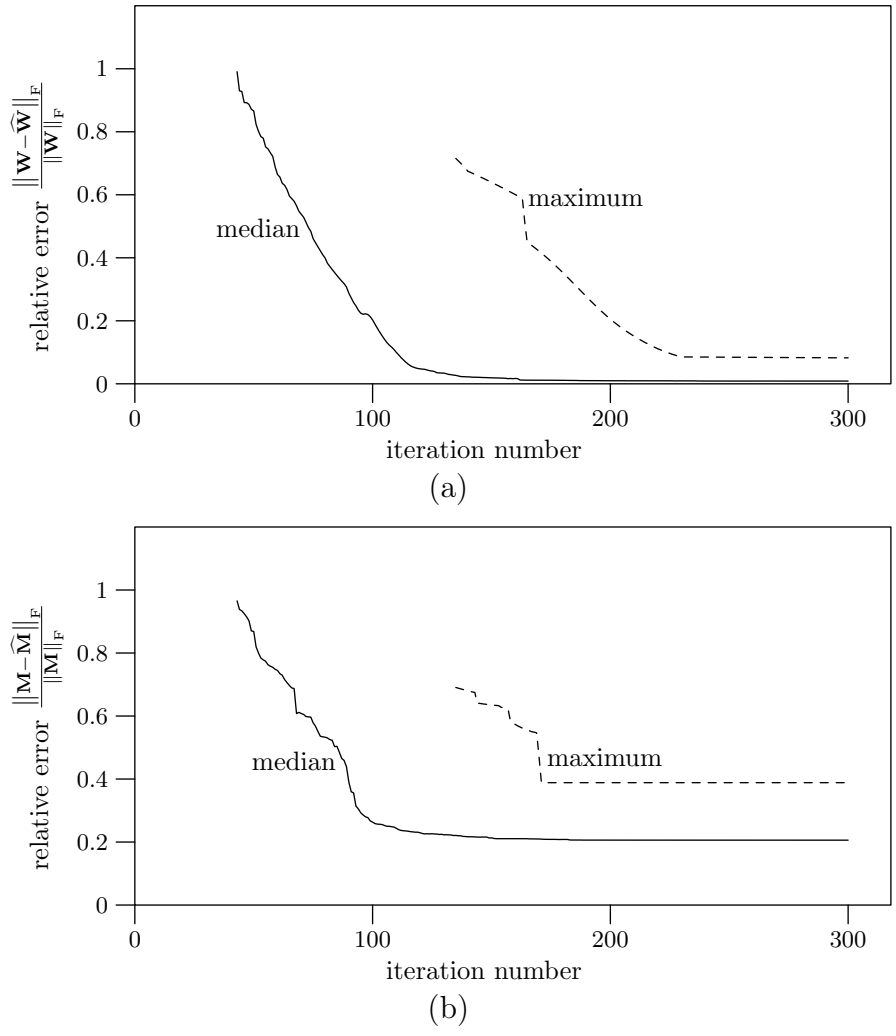

Figure 2: The median and the maximum of the relative errors made in the estimation of $\mathbf{W}$ and $\mathbf{M}$, computed over the estimates of 100 different instances of our two-layer model. Each run of the algorithm used a different data set corresponding to different values of of $\mathbf{M}$ and $\mathbf{A}$ (the pseudoinverse of $\mathbf{W}$), as well as different driving noise $\mathbf{u}(t)$, and different random signs of components of $\mathbf{y}(t)$. (a) The median and the maximum of the relative error made in the estimation of $\mathbf{W}$, plotted as a function of iteration number. (b) The median and the maximum of the relative error made in the estimation of $\mathbf{M}$, plotted as a function of iteration number. Note that none of the plots begin from iteration 1, because the possible permutation of the components of $\mathbf{y}(t)$ can not be determined during the first rounds of the algorithm (see Appendix B.2).

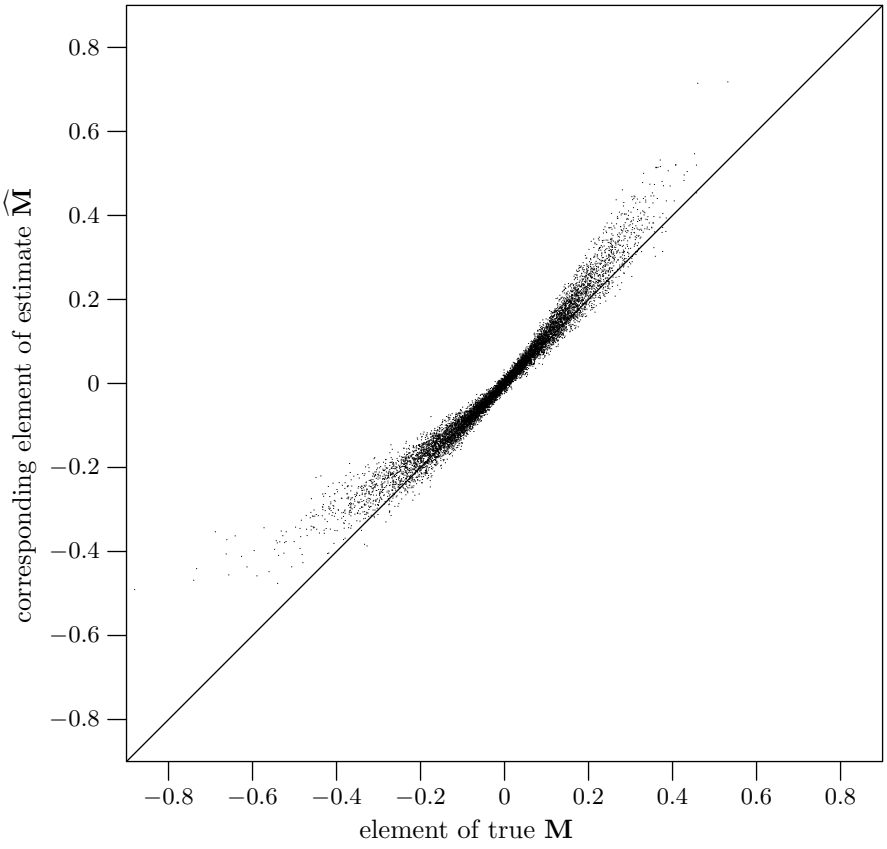

Figure 3: The approximation used in the estimation of $\mathbf{M}$ introduces a systematic bias in the estimate $\widehat{\mathbf{M}}$. The figure shows a scatter plot of the 10000 elements of all 100 matrices $\mathbf{M}$ vs. the corresponding elements of estimates $\widehat{\mathbf{M}}$. Let $\mathbf{M}(i,j)$ denote an element of $\mathbf{M}$. The scatter plot shows that in addition to the variance of the estimates growing as a function of $|\mathbf{M}(i,j)|$, there is also a positive bias in $\widehat{\mathbf{M}}(i,j)$ when $|\mathbf{M}(i,j)|$ is large. This bias is characterized by a convex monotonic mapping from $\mathbf{M}(i,j)$ to $\widehat{\mathbf{M}}(i,j)$. Notice, however, that such a monotonic bias tends to preserve the ordering of the magnitudes of the elements of $\mathbf{M}$ – that is, if an element $\mathbf{M}(i_1,j_1) > \mathbf{M}(i_2,j_2)$, then typically also $\widehat{\mathbf{M}}(i_1,j_1) > \widehat{\mathbf{M}}(i_2,j_2)$. In the analysis of the results we are mostly interested in this ordering, while the convergence analysis presented above employs Frobenius norm which emphasizes large errors. The scatter plot shows that a large part of the relative error is a consequence of this systematic bias.

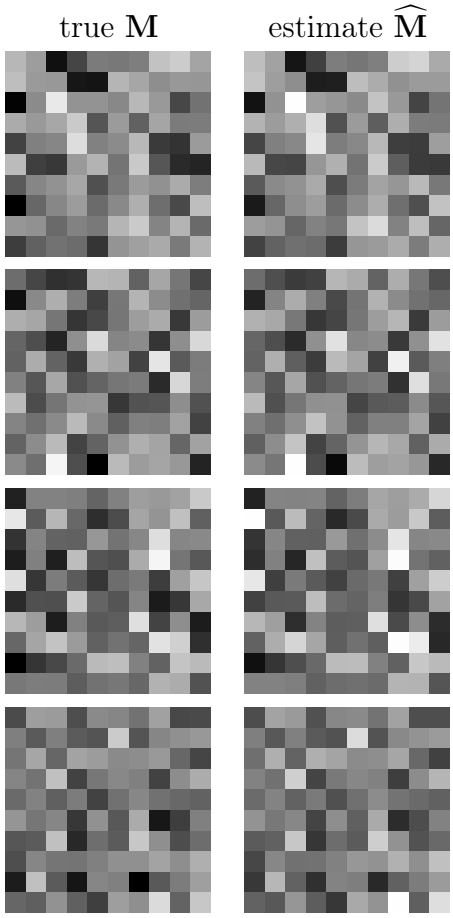

Figure 4: Estimates $\widehat{\mathbf{M}}$ are very similar to the true $\mathbf{M}$, except for positive differences at elements with high absolute values. This is a consequence of the fairly small relative error and the fact that the systematic bias made in the estimation of $\mathbf{M}$ accounts for a large proportion of the remaining error. The plots show the true matrices $\mathbf{M}$ (left column) and their estimates $\widehat{\mathbf{M}}$ (right column) from the first four runs of the 100 runs of the validation experiment. Bright pixels indicates high positive values, dark pixels low negative ones (zero is medium gray). Each ($\mathbf{M}$,$\widehat{\mathbf{M}}$)-pair was plotted using a common colormap, so similar pixel intensities in the true value and the estimate indicate that the elements have similar values. The estimates look very similar to the true matrices. A closer inspection reveals that in the estimates the brightest and the darkest pixels are typically brighter than in the true matrices. This is in accordance with the systematic bias illustrated in Figure 3.

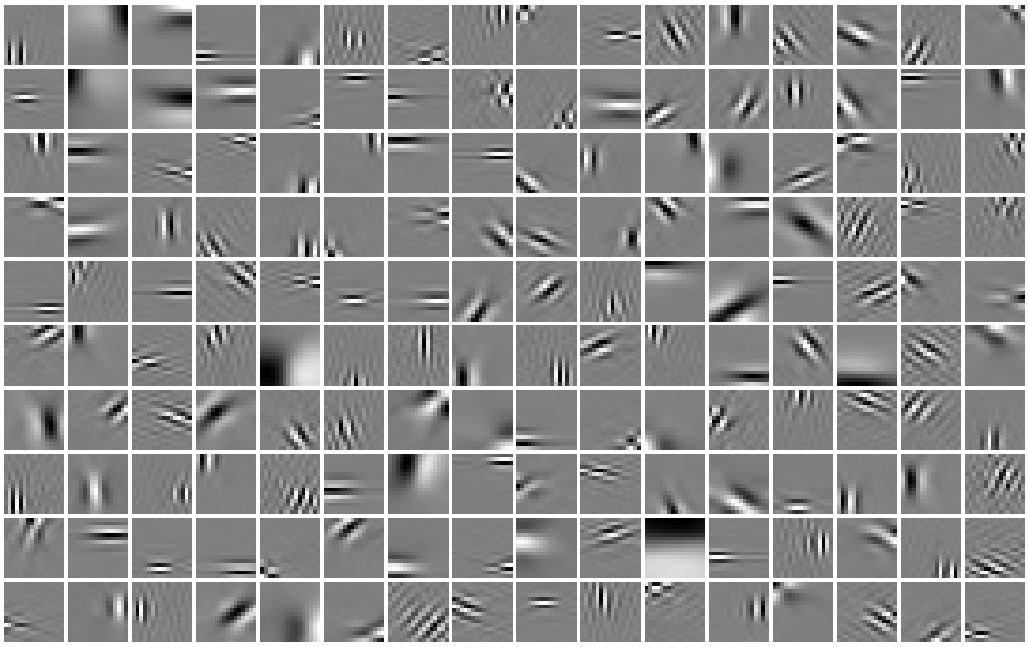

Figure 5: The estimation of the generative model from natural visual stimuli results in the emergence of localized, oriented receptive fields with multiple scales. These basis vectors (columns of $\mathbf{A}$) were obtained by applying the estimation procedure described in Section 3 to a large set of samples from natural image sequences. The basis vectors are in no particular order in this figure.

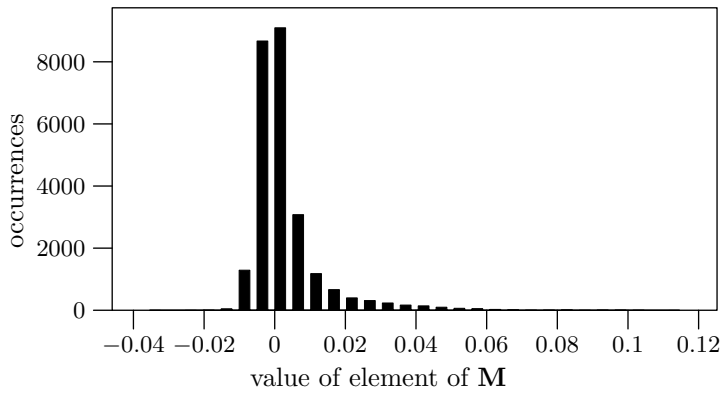

Figure 6: Histogram of the non-diagonal elements of $\mathbf{M}$ estimated from natural image sequence data.

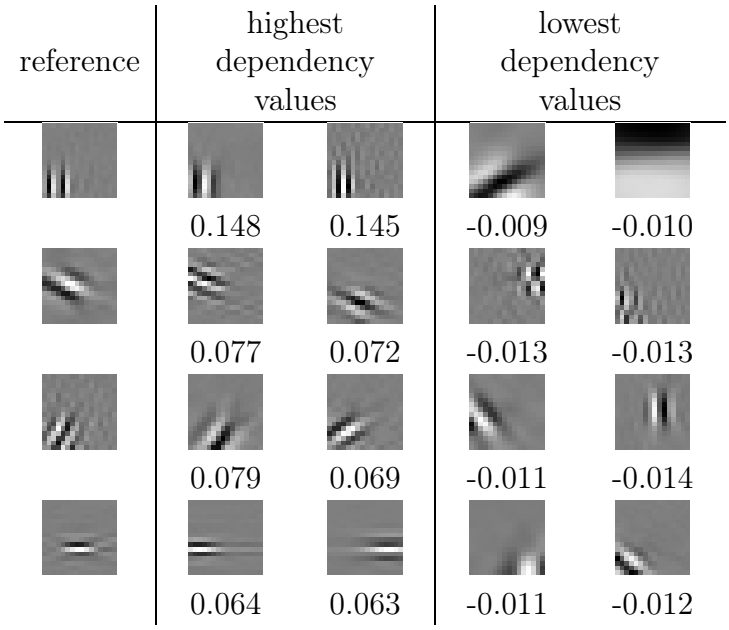

Figure 7: Basis vectors (columns of $\mathbf{A}$) with high activity dependency values code for similar features at nearby positions, whereas basis vectors with low dependency values code for features with different scale and/or orientation. Each row shows the basis vectors with highest and lowest dependency values with respect to the reference vector in the leftmost column. The reference vectors were chosen from the set of vectors in Figure 5 as representatives of four different orientations. The measure of spatiotemporal dependency used was $\frac{\mathbf{M}(i,j)+\mathbf{M}(j,i)}{2}$, where i and j denote the columns of the basis vectors in $\mathbf{A}$. The dependency value of each of the basis vectors with respect to the reference is shown under the vector. As can be seen, basis vectors with high positive activity dependency code for similar features (orientation, frequency) as the reference vectors, whereas those with low (negative) dependency code for different scale and/or orientation.

Figure 8: Grouping similar to complex cell pooling of simple cell outputs emerges from spatiotemporal activity dependencies. Here we have plotted each of the basis vectors (columns of $\mathbf{A}$) at a 2D position obtained by applying multidimensional scaling to the similarity values defined by $\mathbf{M}$. As can be seen, nearby basis vectors seem to be mostly coding for similarly oriented features with similar frequencies at nearby spatial positions. In addition, some global topographic organization also emerges: those basis vectors which code for horizontal features are on the left in the figure, while those that code for vertical features are on the right. Some short distances have been extended in order to be able to show the basis vectors in a reasonable scale.