Temporal and spatiotemporal coherence in simple-cell responses: A 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

May 14, 2003

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 in our new model, the first layer learns connections between the simple cells that are similar to complex cell pooling: connections are strong among cells with similar preferred 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. 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 data on 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 (see, e.g., (Palmer, 1999)). Within the past ten years, researchers have proposed computational principles that relate the properties of cells in this area to the statistics of natural stimuli. The most influential of these theories have been sparse coding (Olshausen and Field, 1996; Hyvärinen and Hoyer, 2001), independent component analysis (ICA) (Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998; van Hateren and Ruderman, 1998; Hyvärinen et al., 2001b), and temporal coherence (Földiák, 1991; Kayser et al., 2001; Wiskott and Sejnowski, 2002; Hurri and Hyvärinen, 2003). 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., 2001b).

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, 2003) we showed 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, 2003), 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. (The principle seems to be somewhat applicable even in the case of non-negative, half-wave rectified cell outputs – see (Hurri and Hyvärinen, 2003) for a discussion.)

The measure of temporal activity coherence introduced in (Hurri and Hyvärinen, 2003) took the sum of the temporal activity coherences of single cells. Therefore, there was no possibility of interaction between the activity 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 depend on each other in an autoregressive manner.

The idea of describing natural stimuli by a generative model, and inter-

preting 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, if vision is considered as inverse graphics (Hinton and Ghahramani, 1997; Olshausen, 2003), the approach makes a lot of sense. 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 will benefit from such an internal representation of its stimuli.

The organization of this paper is as follows. In Section 2 we first give an intuitive interpretation of activity level dependencies in simple cell responses in the case of natural stimuli. A dynamic two-layer generative model of natural image sequences which captures these dependencies is then introduced in Section 3. In Section 4 we describe an algorithm for estimating the model. The validity of the algorithm is assessed using artificial (generated) data in Section 5. In Section 6, estimation of the model from natural image sequence data is shown to yield, in one of the layers, receptive fields that have the principal properties of simple-cell receptive fields. The other layer gives connections between simple-cell outputs that seem to be related to both the topographic properties of the primary visual cortex, and to the way in which complex cells pool the outputs of simple cells. We conclude the paper in Section 7 by comparing our model against independent component analysis, addressing some biological considerations of the model, and discussing the merits of this work.

2 Activity-level dependencies of simple-cell-like filters

In independent component analysis, simple cells are modeled as linear filters whose outputs are statistically independent of each other. However, previous research has already shown that the independence assumption does not hold, not even for static image input (Zetzsche and Krieger, 1999; Hyvärinen and Hoyer, 2000; Wainwright and Simoncelli, 2000; Hyvärinen et al., 2001a; Schwartz and Simoncelli, 2001). In the case of dynamic input (image sequences), modeling dependencies in the resulting neural code yields an intriguing interpretation of both the structure of simple-cell receptive fields and the connectivity (pooling and topographic properties) of the primary visual cortex. In this section, we motivate such models intuitively, before the formal treatment of Section 3.

In particular, it seems that the key to modeling these dependencies is to

model dependencies between the *activity levels* – that is, amplitudes, energies or variances – of the filters. We have shown in an earlier paper that maximization of time-correlation of output energy is an alternative to sparse coding and independent component analysis as a computational principle underlying simple-cell receptive field structure (Hurri and Hyvärinen, 2003). A simplified intuitive illustration of why simple-cell outputs have such strong energy correlation over time is shown in Figure 1. Most transformations of objects in the 3D world result in something similar to local translations of lines and edges in image sequences. This is obvious in the case of 3D translations, and is illustrated in Figure 1A for two other types of transformations: rotation and bending. In the case of a local translation, a suitably oriented simple-cell-like filter responds strongly at consecutive time points, but the sign of the response may change (see (Hurri and Hyvärinen, 2003) for additional analysis of why the optimal filters are localized and oriented). We call these kinds of dependencies – dependencies over time in the outputs of individual filters – *temporal* activity level dependencies. Note that when the output of a filter is considered as a continuous signal, the change of sign implies that the signal reaches zero at some intermediate time point, which can lead to a weak measured correlation. Thus, a better model of the dependencies would be to consider dependencies of variances (Pham and Cardoso, 2000; Valpola et al., 2003). However, for simplicity, we consider here the magnitude that is a crude approximation of the underlying variance.

[Figure 1 about here.]

Temporal activity level dependencies, described above, are not the only type of activity level dependencies in a set of simple-cell-like filters. Figure 2 illustrates how two *different* cells with similar receptive field profiles – having the same orientation but slightly different positions – respond at consecutive time instances when the input is a translating line. The receptive fields are otherwise identical, except that one is a slightly translated version of the other. It can be seen that *both cells* are highly active at *both time instances,* but again, the signs of the outputs vary. This means that in addition to temporal activity dependencies (the activity of a cell is large at time $t - \Delta t$ and time t), there are two other kinds of activity level dependencies.

spatial (static) dependencies Both cells are highly active at a single time instance. This kind of dependency is an example of the energy dependencies modeled in previous research on static images (Zetzsche and Krieger, 1999; Hyvärinen and Hoyer, 2000; Wainwright and Simoncelli, 2000; Hyvärinen et al., 2001a; Schwartz and Simoncelli, 2001).

spatiotemporal dependencies The activity levels of different cells are also related over time. For example, the activity of cell 1 at time $t - \Delta t$ is related to the activity of cell 2 at time t.

[Figure 2 about here.]

What makes these dependencies important is that they seem to be reflected in the structure of the primary visual cortex. As was already mentioned above, our earlier results showed that simple-cell-like receptive fields emerge when temporal activity level dependencies are maximized for natural image sequence data (Hurri and Hyvärinen, 2003). To be more precise, in the class of linear filters, the outputs of simple-cell-like receptive fields have maximal correlation of energies over short time for natural image sequence input. In this paper we show that combining temporal activity level dependencies with *spatiotemporal* dependencies yields both simple-cell-like receptive fields and a set of connections between these receptive fields. These connections can be related to both the way in which complex cells seem to pool simple-cell outputs, and to the topographic organization observed in the primary visual cortex. Therefore, according to the results presented in this paper, the principle of activity level dependencies seems to underlie *both receptive field structure and their organization.*

3 Definition of the model

The generative model of natural image sequences introduced in this paper has two layers, as illustrated in Figure 3. The first layer, which captures the activity level dependencies discussed above in Section 2, is a multivariate autoregressive model between 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 latent signal between the first and second layer. The second layer is linear, and maps cell responses to the image space.

[Figure 3 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 the pixel grayscale values in a natural image sequence at time t. (Vectorization of a frame of an image sequence can be done by scanning the two-dimensional frame column-wise into a vector.) 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 sequence grayscale values $\mathbf{x}(t)$ to the outputs of simple cells $\mathbf{y}(t)$, 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

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

The dimension of $\mathbf{x}(t)$ is typically larger than the dimension of $\mathbf{y}(t)$, so that equation (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] $\mathbf{A} = \mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-1}$.

As was discussed above in Section 2, in contrast to sparse coding (Olshausen and Field, 1996) or independent component analysis (Hyvärinen et al., 2001b) 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 us define the activity levels by $\mathbf{abs}\left(\mathbf{y}(t)\right) = \left[|y_1(t)| \cdots |y_K(t)|\right]^T$, and let $\mathbf{v}(t)$ denote a driving noise signal (the distribution of $\mathbf{v}(t)$ 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

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

for $k = 1, ..., K$. Actually, the constraint of unit energy is not 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

[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 $\mathbf{0}$.

the model (a similar situation is found in independent component analysis). Thus, we can define the scale of the $y_k(t)$'s as we like.

There are dependencies between the driving noise $\mathbf{v}(t)$ and cell activity levels $\mathbf{abs}\,(\mathbf{y}(t))$ because of 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 whose components are statistically independent of each other. We define

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

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 probability that a latent signal $y_k(t)$ retains its sign is P_{ret}, that is,

$$\mathrm{P}\,(\,y_k(t) > 0\,|\,y_k(t-\Delta t) > 0\,) = \mathrm{P}\,(\,y_k(t) < 0\,|\,y_k(t-\Delta t) < 0\,) = P_{\text{ret}}. \tag{6}$$

For simplicity, we assume that the sign of a latent signal at time t is independent of the magnitude of the signal at time $t - \Delta t$, and the signs of different latent signals are independent of each other. Note that one consequence of this random generation of signs is 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

$$\begin{aligned}
\mathrm{E}_t\,\{y_{k_1}(t)y_{k_2}(t)\} &= \mathrm{E}_t\,\{s_{k_1}(t)\,|y_{k_1}(t)|\,s_{k_2}(t)\,|y_{k_2}(t)|\} \\
&= \underbrace{\mathrm{E}_t\,\{s_{k_1}(t)\}}_{=0}\underbrace{\mathrm{E}_t\,\{s_{k_2}(t)\}}_{=0}\mathrm{E}_t\,\{|y_{k_1}(t)|\,|y_{k_2}(t)|\} \\
&= 0.
\end{aligned} \tag{7}$$

Similarly, the means of the $y_k(t)$'s are all zero.

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

$$\mathbf{W}\mathbf{C}_{\mathbf{x}(t)}\mathbf{W}^T = \mathbf{I}, \tag{8}$$

where $\mathbf{C}_{\mathbf{x}(t)} = \mathrm{E}_t\left\{\mathbf{x}(t)\mathbf{x}(t)^T\right\}$.

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 level 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 the estimated $\mathbf{M}$.

One interpretation of the driving noise $\mathbf{v}(t)$ is closely related to this idea of grouping cells with strong interdependencies together: $\mathbf{v}(t)$ can be considered as coding for higher-order features in the dynamic data. Consider the case in which a component of $\mathbf{v}(t)$, say $v_k(t)$, takes a high positive value at time t. Then the corresponding component of $\mathbf{y}(t)$ – that is, $y_k(t)$ – would also become highly active at time t. Because of the dynamic properties of the autoregressive model, during consecutive time instances the activity of $y_k(t)$ would spread to those components of $\mathbf{y}(t)$ for which the corresponding elements of $\mathbf{M}$ are large. Therefore, a large value in $v_k(t)$ codes for the activation of such a group of units with strong dependencies. In other words, $v_k(t)$ signals the occurrence of a higher-order feature that is common for this group. We will see below in Section 6.2 how this interpretation relates $\mathbf{v}(t)$ to complex cells.

4 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 an approximative method of moments. The estimate is given by

$$\begin{aligned}\widehat{\mathbf{M}} \approx \beta \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\} \\ \times \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right) \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \Big\}^{-1},\end{aligned} \tag{9}$$

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 starting from the least squares estimate (see Appendix B), and verified through simulations that the objective function is capable of estimating 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

$$f(\mathbf{W}, \mathbf{M}) = \sum_{i=1}^{K} \sum_{j=1}^{K} \mathbf{M}(i,j) \operatorname{cov} \left\{ |y_i(t)|, |y_j(t - \Delta t)| \right\}, \tag{10}$$

which can also be expressed as

$$\begin{aligned} f(\mathbf{W}, \mathbf{M}) = \mathrm{E}_t \Big\{ & \left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t \left\{ \mathbf{abs}\left(\mathbf{y}(t)\right) \right\} \right)^T \mathbf{M} \\ & \times \left(\mathbf{abs}\left(\mathbf{y}(t - \Delta t)\right) - \mathrm{E}_t \left\{ \mathbf{abs}\left(\mathbf{y}(t)\right) \right\} \right) \Big\}. \end{aligned} \tag{11}$$

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

$$\widehat{\mathbf{W}} = \arg \max_{\mathbf{W}} f(\mathbf{W}, \widehat{\mathbf{M}}). \tag{12}$$

Optimization of the objective function f over $\mathbf{W}$ under constraint (8) uses a gradient projection approach (Hurri and Hyvärinen, 2003). The initial value of $\mathbf{W}$ is selected randomly.

When the model is estimated from natural image sequence data, the value of the scalar multiplier β in (9) can not be estimated. However, first note that this multiplier has a constant linear effect in objective function (10). This means that the value of β does not affect the optima of (10), so the correct value of β is not needed to estimate $\mathbf{W}$. Second, multiplier β only scales the elements of $\mathbf{M}$ with a constant value. This rescaling does not affect the ratios of the elements, or their ordering. In addition, as was discussed above, matrix $\mathbf{M}$ can be thought of as containing similarity measurements between different cells. The multiplication of $\mathbf{M}$ with a positive scalar does not modify the information contained in the measurements when an interval measurement scale (Borg and Groenen, 1997) is used. We will see below in Section 6.2 that in our case the interval scale is a natural measurement scale for the measurement distances in $\mathbf{M}$. Therefore, in the estimation we just set $\beta = 1$.

Note, however, that in the validation of the estimation method this possible rescaling of $\mathbf{M}$ must be taken into account, because we want to measure the convergence of the algorithm quantitatively. This will be considered in detail below.

5 Experiments with artificial data

Before applying the estimation method to natural data, we verified its validity using artificial data. 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 $K = 10$, so 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 this can be done are given in Appendix C.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 (10) to a different ordering of the components of $\mathbf{y}(t)$, which is similar to the case of independent component analysis (Hyvärinen et al., 2001b), care had to be taken to compensate for a possible permutation; details on how this was done are described in Appendix C.2.

After compensating for the possible permutation, the effect of the unknown scalar multiplier β in equation (9) had to be accounted for. In the estimation process above we just set $\beta = 1$, because in the case of natural image sequence data this coefficient can be discarded as was discussed in Section 4 above. Here, however, the convergence of the algorithm is examined quantitatively, so this multiplier has to be accounted for to get an exact performance measure. This was done by using equation (9) to estimate β by

$$\hat{\beta} = \frac{\|\mathbf{M}\|_{\mathrm{F}}}{\left\|\widehat{\mathbf{M}}\right\|_{\mathrm{F}}} \tag{13}$$

(remember that estimate $\widehat{\mathbf{M}}$ is obtained by setting $\beta = 1$ in equation (9)). Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, that is, $\|M\|_{\mathrm{F}}^2 = \sum_i \sum_j (\mathbf{M}(i,j))^2$.

To analyze the convergence of the algorithm, we examined how the relative estimation errors $\left(\left\|\mathbf{M}-\widehat{\mathbf{M}}\right\|_{\mathrm{F}}\right)/\left\|\mathbf{M}\right\|_{\mathrm{F}}$ and $\left(\left\|\mathbf{W}-\widehat{\mathbf{W}}\right\|_{\mathrm{F}}\right)/\left\|\mathbf{W}\right\|_{\mathrm{F}}$ change as a function of number of iterations. Figure 4 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 are plotted as a function of iteration number. Two different values of P_{ret} (see Figure 3 and equation (6)) were used in two different sets of experiments. The results for $P_{\mathrm{ret}} = 0.5$, corresponding to perfectly temporally decorrelated data, are shown in Figures 4A and C. We also estimated the value of P_{ret} from results obtained from natural image sequences, where a softer form of temporal decorrelation was used (see Section 6). Figures 4B and D show the results obtained when the data is

generated using the estimated value $P_{\text{ret}} = 0.7$. These results show that the estimation method is not sensitive to small variations in P_{ret}.

[Figure 4 about here.]

As we can see, the estimate of $\mathbf{W}$ converges fairly reliably to the true value. As for the estimation of $\mathbf{M}$, the scalar multiplier β estimated as in equation (13) 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 simple bias in the estimate, and this bias does not seem to be critical in our analysis of the results. The nature of the bias can be seen in Figure 5, which shows a scatter plot of the true elements of the 100 matrices $\mathbf{M}$ vs. their estimates. We can see that the 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 larger 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 bias.

[Figure 5 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 simple bias described above. In Figure 6 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 6 about here.]

6 Experiments with natural image sequences

6.1 Data collection and preprocessing

The data and preprocessing used in the experiments were very similar to those in (Hurri and Hyvärinen, 2003), so we will describe them only shortly here, and refer the reader to (Hurri and Hyvärinen, 2003) 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 (see Section 7.1 for an experiment in which neither temporal decorrelation nor normalization was performed). Temporal decorrelation enhances temporal changes in the data, and differentiates our results from those obtained with static images (Hurri and Hyvärinen, 2003). 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. That is, each preprocessed sequence consisted of two 16×16 frames separated by $\Delta t = 40$ ms. After temporal decorrelation, the spatial local mean (spatial DC component) was subtracted from each of the 400,000 frames, and the frames 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 (van Hateren and van der Schaaf, 1998), the dimensionality of the data was reduced to 160 using principal component analysis (Hyvärinen et al., 2001b).

6.2 Results

The estimation algorithm described in Section 4 was applied to the preprocessed natural image sequence data to obtain estimates for $\mathbf{M}$ and $\mathbf{A}$. Figure 7 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, 2003). This suggests that, as far as receptive field structure is concerned, these methods are rather similar to each other in that receptive fields with similar qualitative properties emerge when the methods are applied to natural visual stimuli.

[Figure 7 about here.]

The estimated matrix $\mathbf{M}$ captures the temporal and spatiotemporal activity level dependencies between the basis vectors shown in Figure 7. The diagonal elements of the estimated $\mathbf{M}$ were relatively large, ranging from 0.31 to 0.74 with a mean of 0.44, indicating that for all the basis vectors, activity levels at time $t - \Delta t$ and time t have considerable correlation. This is in concordance with the results in (Hurri and Hyvärinen, 2003). 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 8. In order to examine these dependencies more closely, we first plotted the basis vectors with the highest and lowest activity level dependency values for a set of representative reference vectors. The results, shown in Figure 9, suggest that basis vectors with high positive activity level dependencies code for similar features at nearby positions, whereas basis vectors with low (negative) dependencies code for features with different scale and/or orientation and/or location.

[Figure 8 about here.]

[Figure 9 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 3). 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 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 (reciprocals of similarities) between cells i and j. The objective of multidimensional scaling 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 multidimensional scaling 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 multidimensional scaling to our data so that the interval measurement scale (Borg and Groenen, 1997; SAS/STAT, 2000) was assumed. Informally, use of the interval measurement scale means 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 $\mathbf{M}_s$ 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 multidimensional scaling procedure is shown in Figure 10. 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 9, 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 10, while those that code for vertical features are on the right.

[Figure 10 about here.]

When examining the preferred orientations of the basis vectors in Figure 10, we can see that there are more vectors that prefer horizontal or vertical orientations than those that prefer oblique orientations. A similar imbalance has been observed in the visual cortex, in the number of cells preferring oblique orientations vs. the number of cells preferring horizontal/vertical orientations (see, e.g., (Li et al., in press)). This imbalance is thought to underlie the *oblique effect*, the fact that in psychophysical tests vertical and horizontal orientations are discriminated better than oblique ones. Our results suggest that there may be a connection between the oblique effect and the statistics of natural stimuli. We suspect that these results emerge from natural image sequences because horizontal and vertical lines and edges are prevalent when natural scenes are examined in an upright position, but further research is needed to verify this.

It should also be noted that in Figure 10, in some cases oblique basis vectors whose preferred directions are orthogonal to each other are located close to each other in the spatial layout. Note that this is *not* the case in Figure 9, where we can see that also in case of an oblique preferred direction, the filters with highest dependencies have a similar orientation. Therefore, this effect does not seem to be a property of the extracted dependencies (matrix $\mathbf{M}$); it is more likely due to distortions caused by the multidimensional scaling procedure that forces the points to lie in a plane.

To summarize the results presented in this section, the estimation of our two-layer model from natural image sequences yields, firstly, simple-cell-like receptive fields (Figure 7), and secondly, grouping similar to the pooling of simple cell outputs and local topographic organization in the primary visual cortex (Figures 9 and 10). 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.

We mentioned in Section 3 that the driving noise process $\mathbf{v}(t)$ can code for the occurrence of higher-order features by signalling the activation of a group of units with strong dependencies. Statistically, a large positive value of an element of $\mathbf{v}(t)$ tends to indicate the activation of such a group for a certain period of time. From Figures 9 and 10 we can now see what the higher-order features coded by $\mathbf{v}(t)$ would be: short contours with a certain orientation and scale, differing in their phase and spatial position. This kind of learning of invariant features – invariant to the phase and position of an edge or line in this case – was originally associated with temporal coherence by Földiák in his theoretical work and simulations (Földiák, 1991). In the mammalian visual system, complex cells are traditionally considered to be invariant to phase. The driving noise signal $\mathbf{v}(t)$ thus gives the values of higher-order features that could be related to complex cells.

7 Discussion

7.1 Multivariate AR model estimation vs. independent component analysis

What happens if we try to estimate the second (linear) layer of the model with standard independent component analysis? Under what conditions are

the results different, and how? Intuitively it would seem that independent component analysis would not be applicable if the activity level dependencies between different components of $\mathbf{y}(t)$ are sufficiently strong. The strength of these dependencies is governed by matrix $\mathbf{M}$. To examine this closer we made two experiments, one with simulated data and the other with natural image sequence data.

In the first experiment, we used the matrix $\mathbf{M}$ estimated from natural image sequence data (see Section 6.2) to define the strength of the dependencies. Using the whole matrix $\mathbf{M}$ for a repeated experiment was not computationally feasible, so a 15×15 submatrix $\mathbf{M}_{\text{sub}}$ was used instead. Taking a submatrix meant that we examined the dependencies between *some* components of $\mathbf{y}(t)$, instead of all of them. Technically this was done by selecting a set of indices $\mathcal{I} \subset \{1, ..., 160\}$ whose size was 15 ($|\mathcal{I}| = 15$). For example, a block of those elements of $\mathbf{M}$ whose column and row index is smaller than or equal to 15 forms one such submatrix. However, with this experiment we wanted to examine whether there are dependencies in $\mathbf{M}$ which are *strong enough* so that our estimation method works better than independent component analysis, so we wanted to select a set of components of $\mathbf{y}(t)$ with relatively strong interdependencies. By default the ordering of the components of $\mathbf{y}(t)$ is arbitrary (see Appendix C.2), so components of $\mathbf{y}(t)$ were first ordered according to strengths of their dependencies as follows. The first component was the one that corresponded to the largest diagonal element of $\mathbf{M}$. Once the kth component had been selected to be the one with original index j, the index of the $(k+1)$th component was chosen to be $\arg\max_i \left(\mathbf{M}(j,i) + \mathbf{M}(i,j)\right)$. The ordering of the components of $\mathbf{y}(t)$ was accompanied by a corresponding rearrangement of the rows and columns of $\mathbf{M}$ (see Appendix C.2). After this rearrangement, the upper left block of $\mathbf{M}$ should correspond to components with relatively large dependencies, so it was selected to be $\mathbf{M}_{\text{sub}}$. The selected dependency matrix $\mathbf{M}_{\text{sub}}$ was used to generate 100 different data sets, each having their own random initial starting point, random $\mathbf{A}$, and random signs of the components of $\mathbf{y}(t)$ (see Appendix C.1). Our algorithm and the FastICA algorithm were then used to estimate matrix $\mathbf{W}$ from the data, and the relative error was used as performance criterion. The resulting scatter plot of the relative errors is shown in Figure 11. It is clear that our estimation method succeeds better in this estimation task. This shows that in natural image sequence data, the activity level dependencies *are* so strong that standard independent component analysis is not able to estimate the corresponding basis.

[Figure 11 about here.]

The purpose of the second experiment was to examine how the results ob-

tained from natural image sequence data differ for the two methods. To make the comparison possible, we modified the preprocessing used above (see Section 6.1) so that no temporal preprocessing or normalization was performed – that is, only the local mean was subtracted. No temporal decorrelation was done because it is not meaningful in the case of static independent component analysis, and in order for the results of the two methods to be comparable the same data has to be used. It seems that normalization is only necessary as a preprocessing step when temporal decorrelation is done (probably because of the very large temporal changes that temporal decorrelation introduces into the data), so it was also left out. This experiment also served as a control experiment to show that the qualitative properties of our original results (see Figure 7) were not a consequence of temporal preprocessing or normalization. The results of both methods – our algorithm and independent component analysis – are shown in Figure 12. The results are qualitatively similar in that in both cases the resulting filters are oriented, localized and bandpass. However, some differences can also be seen. First, the number of subregions (dark or light regions in the receptive fields) seems to be smaller in the ICA results. Second, most of the ICA basis vectors include a small global step-like grayscale change. Thus, the difference between ICA and our present algorithm can also be seen in the case of experiments on real data.

[Figure 12 about here.]

7.2 Biological considerations: underlying mechanisms, and nonnegative simple-cell models

The theory and simulations presented in this paper model the relationship between the properties of the primary visual cortex and statistics of natural image sequences. The underlying assumption is that the visual system has specialized to account for the properties of typical stimuli. In terms of biological mechanisms, such specialization could be genetic, or could take place during development. Our model is not intended to specify at all how or when such specialization would take place. It only models the resulting relationships between an organism and its environment, not the dynamic interaction of these two, nor the role of development vs. genetic instructions.

In computational neuroscience, linear filters are typically applied as models of rate coding in simple cells (Olshausen and Field, 1996; Bell and Sejnowski, 1997; van Hateren and van der Schaaf, 1998). In reality, however, the firing rate of a cell can only be positive. This can be modeled, for example, with half-wave rectification (Heeger, 1992), in which the negative output values of a linear filter are squashed to zero in the actual cell output, and

each output unit consists of two cells with otherwise identical receptive fields except for reversed polarities (see, e.g., (Hurri and Hyvärinen, 2003)). The interpretation of our model in this case is that two *cell pairs*, each pair consisting of two cells with reversed polarities, have activity level dependencies. Concerning the results obtained from natural image sequences, consider replacing each receptive field in Figure 10 with a corresponding two-cell unit. The most important qualitative observations made from the results in Figure 10 were that the receptive fields are oriented, localized and multiscale, and that connections implied by elements of matrix $\mathbf{M}$ are strongest between cells with similar orientation and scale and nearby location. Replacing each receptive field by the corresponding two-cell unit does not change these observations. The previous discussion suggests that use of a basic nonnegative simple-cell model (half-wave rectification) should not change the qualitative nature of our results obtained from natural image sequences. However, it must be noted that in the case of purely temporal activity level dependencies (horizontal direction in Figure 2), our earlier results suggest that such dependencies are present even for *individual* half-wave rectified simple-cell models, not only for cell pairs (Hurri and Hyvärinen, 2003). Therefore, additional model development and experimentation is needed before the results of this paper can be generalized to nonnegative cell models.

7.3 Conclusions and related work

There are two main contributions in this paper. First, to our knowledge, the generative model presented here is the first attempt to model the visual system using a two-layer generative model of natural image sequences. A multi-layered description of the stimuli is important because it enables us to capture dependencies within the different layers of sensory processing. In a multi-layered model, the processing in higher layers has an influence on the optimality of features on the lower level and vice versa; thus joint modeling of lower and higher layer features is the only way to find out what the optimal features and processing methods are. In our case, the results suggest that simple-cell outputs have temporal and spatiotemporal activity level dependencies, and that cells at the next level of processing (complex cells) pool simple-cell outputs so that cells with high activity level dependencies are pooled together. This can provide important cues as to how different layers in the visual pathway are connected. (For on-going work on application of another multi-layer model of image sequences for segmentation see (Felderhof et al., 2002).)

The results obtained from natural image sequence data also suggest that spatiotemporal activity level dependencies could also be reflected in the to-

pography of the primary visual cortex – that is, cells with high spatiotemporal activity level dependencies seem to be physically located close to each other within the cortex. This complements earlier research on how *simultaneous* activity dependency (see Figure 2) is reflected in the organization of the cortex in a similar manner. The outputs of related wavelet filters with uncorrelated outputs exhibit a similar dependency in natural images (Zetzsche and Krieger, 1999; 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, dependencies between simultaneous activity levels of 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. When these earlier results on simultaneous activity dependencies are combined with our results on temporal dependencies, it seems possible that "activity bubbles" (Hyvärinen et al., in press), 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, describing 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. On the average, matrix $\mathbf{A}$ can be estimated with very good accuracy. Matrix $\mathbf{M}$ can also be recovered up to a fairly small relative error, and a systematic bias which is irrelevant for most 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 data analysis (Hyvärinen et al., 2001b). Further work on this problem can be found in (Hyvärinen and Hurri, submitted).

Research related to the results presented here can also be found in previous research on unsupervised learning and econometrics. 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; Valpola et al., 2003). A related study can also be found in (Charles et al., 2002), where it was shown that in the case of

simulated data (vertically or horizontally moving lines), temporally related features can be forced to be coded in the same areas in a noisy nonlinear principal component analysis network. This was done by introducing spatially separate noisy areas in a network having time-delayed lateral connections between neighboring cells – both the separate noisy areas and the lateral connections introduce the pooling property a priori into the network. 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, as well as a matrix of connections between the simple cells. These connections seem to be related both to the topography of simple cells in the primary visual cortex, and 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

$$\mathrm{E}_t \{\mathbf{v}(t)\} = \mathrm{E}_t \{\mathbf{abs}\,(\mathbf{y}(t))\} - \mathbf{M}\mathrm{E}_t \{\mathbf{abs}\,(\mathbf{y}(t))\}\,. \tag{14}$$

Therefore we have

$$
\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{abs}\left(\mathbf{y}(t)\right)\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{abs}\left(\mathbf{y}(t)\right)\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{abs}\left(\mathbf{y}(t)\right)\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+\underline{\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{15}
$$

The underlined term in equation (15) is non-zero because of the dependencies between $\mathbf{v}(t)$ and $\mathbf{abs}\left(\mathbf{y}(t)\right)$ that are introduced through equation (5). We will approximate this term: we make the approximation that the non-negativity constraint in equation (5) is active for a random proportion $\alpha \in (0,1)$ of the whole sample (in reality the constraint is not active for a random proportion, but tends to be activated more frequently when $\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)$ has a small value, so this approximation introduces a systematic bias). Then

we approximate

$$\begin{aligned}
&\mathrm{E}_t\left\{\left(\mathbf{v}(t)-\mathrm{E}_t\left\{\mathbf{v}(t)\right\}\right)\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\right\}\\
&\quad\approx \alpha\mathrm{E}_t\Big\{\left(-\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)+\mathrm{E}_t\left\{\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)\right\}\right)\\
&\qquad\qquad\times\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\}\\
&\qquad+(1-\alpha)\mathrm{E}_t\Big\{\underline{\left(\mathbf{u}(t)-\mathrm{E}_t\left\{\mathbf{u}(t)\right\}\right)}\\
&\qquad\qquad\underline{\times\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)\right\}\right)^T\Big\}}\\
&\quad=-\alpha\mathbf{M}\mathrm{E}_t\Big\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\\
&\qquad\qquad\times\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\},
\end{aligned}\tag{16}$$

where, in the last step, we have used the fact that the underlined term is equal to $\mathbf{0}$. Using the approximation (16) we get from equation (15)

$$\begin{aligned}
\widehat{\mathbf{M}}\approx\frac{1}{1-\alpha}&\mathrm{E}_t\Big\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\\
&\qquad\times\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\Big\}\\
\times\,&\mathrm{E}_t\left\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\left(\mathbf{abs}\left(\mathbf{y}(t)\right)-\mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T\right\}^{-1}.
\end{aligned}\tag{17}$$

Setting $\beta=1/(1-\alpha)$ in this equation yields (9).

B Heuristic derivation of the objective function for estimating W

We start from equation (3):

$$\mathbf{abs}\left(\mathbf{y}(t)\right)=\mathbf{M}\,\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)+\mathbf{v}(t).\tag{18}$$

Removing the mean from both sides of this equation yields

$$\begin{aligned}\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\} = \mathbf{M}\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)\right\}\right) \\ + \underbrace{\mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{v}(t)\right\}}_{=\mathbf{v}_0(t)}.\end{aligned} \tag{19}$$

Now, defining $\mathbf{v}_0(t) = \mathbf{v}(t) - \mathrm{E}_t\left\{\mathbf{v}(t)\right\}$, the least squares criterion is given by

$$\begin{aligned}&\mathrm{E}_t\left\{\left\|\mathbf{v}_0(t)\right\|^2\right\} \\ &\quad= \mathrm{E}_t\Big\{\Big\|\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\} \\ &\qquad\qquad - \mathbf{M}\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right)\right\}\right)\Big\|^2\Big\} \\ &\quad= \underline{\mathrm{E}_t\left\{\left\|\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right\|^2\right\}} \\ &\qquad+ \underline{\mathrm{E}_t\left\{\left\|\mathbf{M}\left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\right\|^2\right\}} \\ &\qquad- 2\mathrm{E}_t\Bigg\{\left(\mathbf{abs}\left(\mathbf{y}(t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)^T \\ &\qquad\qquad\times \mathbf{M}\left(\mathbf{abs}\left(\mathbf{y}(t-\Delta t)\right) - \mathrm{E}_t\left\{\mathbf{abs}\left(\mathbf{y}(t)\right)\right\}\right)\Bigg\}.\end{aligned} \tag{20}$$

Consider the two underlined terms in (20). These measure the variance of the overall activity level, and the variance of the autoregressive part of the activity levels. However, because the variance of the overall signal is already fixed in the model in equation (8) (since the means of the $y_k(t)$'s are zero), we discard the first term as redundant. The second term is also discarded as redundant because of the same reason and the fact that $\mathbf{M}$ is fixed during the update of $\mathbf{W}$. This leaves us only with the third term, which equals $-2f\left(\mathbf{W}, \mathbf{M}\right)$, so minimization of $\mathrm{E}_t\left\{\left\|\mathbf{v}_0(t)\right\|^2\right\}$ leads to maximization of the objective function (11).

C Mathematical details of the validation of the estimation algorithm

C.1 Data generation

The generated data must follow equations (1) and (3)–(6). 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 was 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 matrix $\mathbf{M}_0$ was first generated by assigning 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). Then, a sample of $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ of length 60000 points was generated using equations (3) and (5) with a random (non-negative) starting point $\left|\mathbf{y}_0(0)\right|$, and Gaussian white $\mathbf{u}(t)$.

Signed data $\mathbf{y}_0(t)$ was generated from $\mathbf{abs}\left(\mathbf{y}_0(t)\right)$ according to equation (6). As was shown in Section 3, this step 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 premultiplying with $\mathbf{\Lambda}$ yields

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

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 1.1. 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).

[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).

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

The objective function (10) 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

$$\mathbf{A} = \mathbf{A}_2\mathbf{P}. \tag{21}$$

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

$$\mathbf{M} = \mathbf{P}_\mathrm{a}^T\mathbf{M}_2\mathbf{P}_\mathrm{a}. \tag{22}$$

To convert the previous equations into a procedure, let $\widehat{\mathbf{W}}_\mathrm{p}$, $\widehat{\mathbf{A}}_\mathrm{p}$ (the inverse of $\widehat{\mathbf{W}}_\mathrm{p}$) 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 (21) to compute a "permutation matrix" $\mathbf{B} = \widehat{\mathbf{A}}_\mathrm{p}^{-1}\mathbf{A} = \widehat{\mathbf{W}}_\mathrm{p}\mathbf{A}$. Matrix $\mathbf{B}$ is not an exact permutation matrix because during the first rounds of the algorithm we may be far from the solution (even in the last rounds the estimate is not perfect). Therefore, we compute an estimate $\widehat{\mathbf{P}}$ of an exact permutation matrix which is close to $\mathbf{B}$ by iteratively choosing the element $\mathbf{B}(i,j)$ of $\mathbf{B}$ with largest absolute value, setting it to 1, and setting all other elements in the same row and column of $\mathbf{B}(i,j)$ to zero. Using the obtained $\widehat{\mathbf{P}}$ an estimate for $\mathbf{A}$ can be computed using (21) again: $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_\mathrm{p}\widehat{\mathbf{P}}$. 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 (22): $\widehat{\mathbf{M}} = \widehat{\mathbf{P}}_\mathrm{a}^T\widehat{\mathbf{M}}_\mathrm{p}\widehat{\mathbf{P}}_\mathrm{a}$.

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.

Charles, D., Fyfe, C., McDonald, D., and Koetsier, J. (2002). Unsupervised neural networks for the identification of minimum overcomplete basis in visual data. *Neurocomputing*, 47(1–4):119–143.

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.

Felderhof, S. N., Storkey, A. J., and Williams, C. K. I. (2002). Position encoding dynamic trees for image sequence analysis. Technical report, University of Edinburgh.

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. (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691.

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., Hoyer, P. O., and Inki, M. (2001a). Topographic independent component analysis. *Neural Computation*, 13(7):1525–1558.

Hyvärinen, A. and Hurri, J. (submitted). Blind separation of sources that have spatiotemporal variance dependencies.

Hyvärinen, A., Hurri, J., and Väyrynen, J. (in press). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001b). *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 complex 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.

Li, B. W., Peterson, M. R., and Freeman, R. D. (in press). The oblique effect: a neural basis in the visual cortex. *Journal of Neurophysiology*.

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.

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.

Pham, D.-T. and Cardoso, J.-F. (2000). Blind separation of instantaneous mixtures of non stationary sources. In Pajunen, P. and Karhunen, J.,

editors, *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, pages 187–192.

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., Harva, M., and Karhunen, J. (2003). Hierarchical models of variance sources. In Amari, S.-i., Cichocki, A., Makino, S., and Murata, N., editors, *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 83–88.

Valpola, H. and Karhunen, J. (2002). An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692.

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.

Zetzsche, C. and Krieger, G. (1999). Nonlinear neurons and high-order statistics: New approaches to human vision and electronic image processing. In Rogowitz, B. E. and Pappas, T. N., editors, *Human Vision and Electronic Imaging IV*, volume 3644 of *Proceedings of SPIE*, pages 2–33. SPIE—The International Society for Optical Engineering.

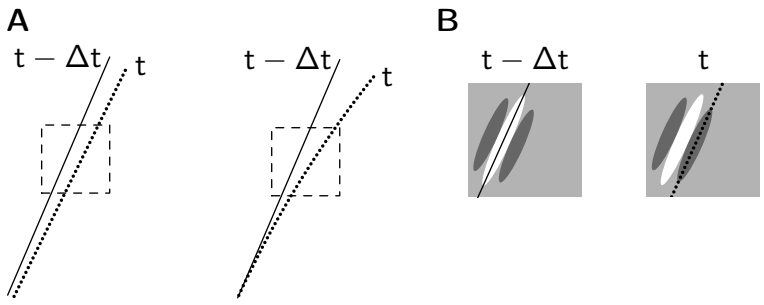

Figure 1: A simplified illustration of temporal activity level dependencies of simple-cell-like filters when the input consists of image sequences. (A) Transformations of objects in the 3D world induce local translations of edges and lines in local regions in image sequences: rotation (left) and bending (right). The solid line shows the position/shape of a line in the image sequence at time $t - \Delta t$, and the dotted line shows its new position/shape at time t. The dashed square indicates the sampling window. (B) Temporal activity level dependencies: in the case of a local translation of an edge or a line, the response of a simple-cell-like filter with a suitable position and orientation is strong at consecutive time points, but the sign may change. The figure shows a translating line superimposed on an oriented and localized receptive field at two different time instances (time $t - \Delta t$, solid line, left; time t, dotted line, right).

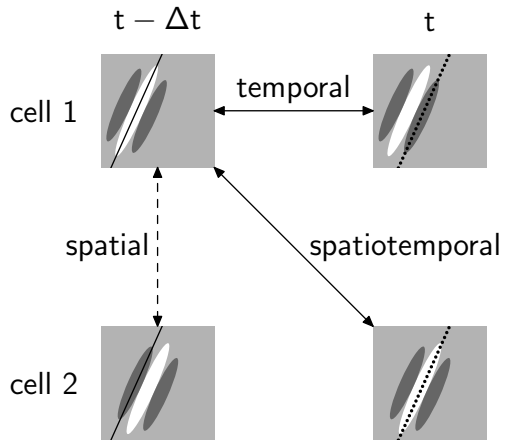

Figure 2: A simplified illustration of static and short-time activity level dependencies of simple-cell-like receptive fields. For a translating edge or line, the responses of two similar receptive fields with slightly different positions (cell 1, top row; cell 2, bottom row) are large at nearby time instances (time $t - \Delta t$, solid line, left column; time t, dotted line, right column). Each subfigure shows the translating line superimposed on a receptive field. The magnitudes of the responses of *both* cells are large at *both* time instances. This introduces three types of activity level dependencies: temporal (in the output of a single cell at nearby time instances), spatial (between two different cells at a single time instance) and spatiotemporal (between two different cells at nearby time instances). The model introduced in this paper includes temporal and spatiotemporal activity level dependencies (marked with solid lines). Spatial activity level dependency (dashed line) is an example of the dependencies modeled in previous work on static images, and is not included in our model.

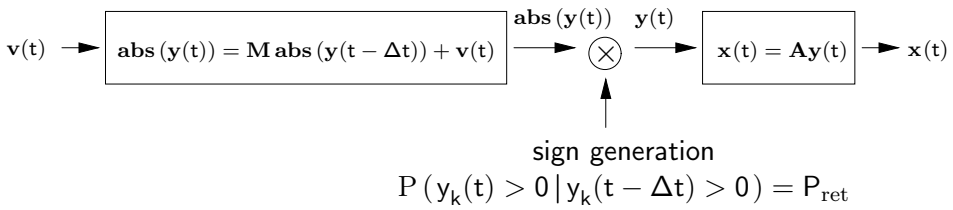

Figure 3: The two layers of the generative model with temporal and spatiotemporal activity level dependencies. Let $\mathbf{abs}\,(\mathbf{y}(t)) = [|y_1(t)| \cdots |y_K(t)|]^T$ denote the activity levels (amplitudes) of simple cell responses. In the first layer, the driving noise signal $\mathbf{v}(t)$ generates the activities of simple cells via a multivariate autoregressive model. Matrix $\mathbf{M}$ captures the spatiotemporal activity level dependencies in the model. The signs of the responses are generated between the first and second layer to yield signed responses $\mathbf{y}(t)$. The probability that a latent signal $y_k(t)$ retains its sign is P_{ret}. 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.

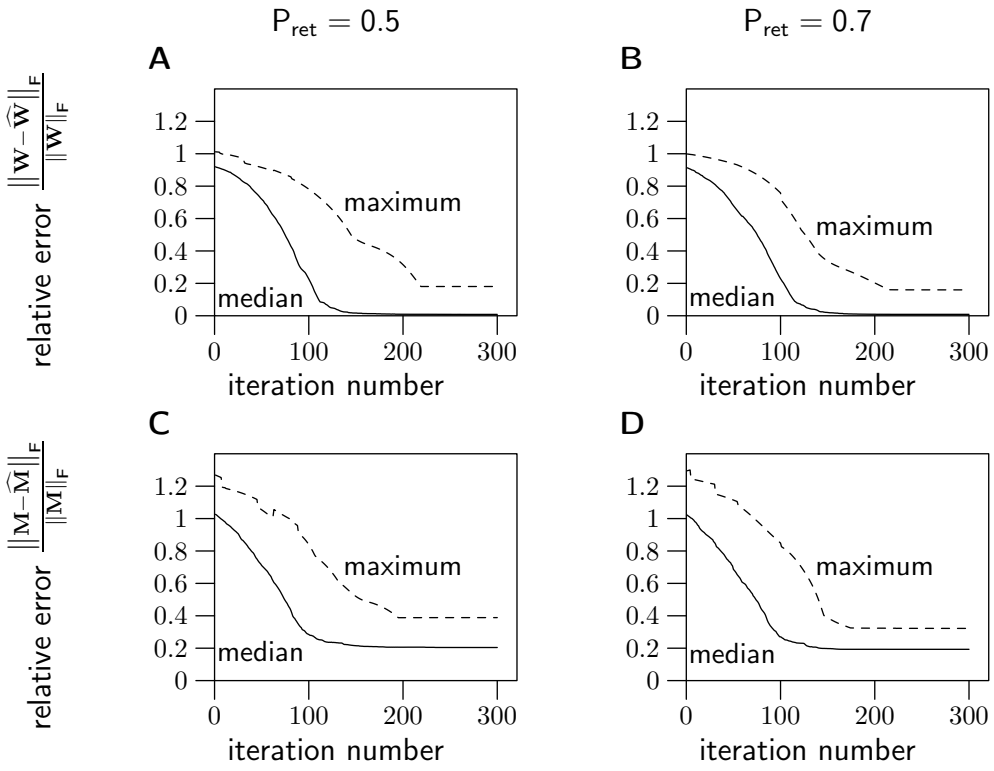

Figure 4: The median and the maximum of the relative errors made in the estimation of $\mathbf{W}$ and $\mathbf{M}$, computed over the estimates of 100 different instances of our two-layer model. Each run of the algorithm used a different data set corresponding to different values of $\mathbf{M}$ and $\mathbf{A}$ (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,C) The median and the maximum of the relative error made in the estimation of $\mathbf{W}$ and $\mathbf{M}$, plotted as a function of iteration number, when the probability that $y_k(t)$ retains the sign of $y_k(t - \Delta t)$ is 0.5 ($P_{\rm ret} = 0.5$). (B,D) The median and the maximum of the relative error made in the estimation of $\mathbf{W}$ and $\mathbf{M}$ when $P_{\rm ret} = 0.7$.

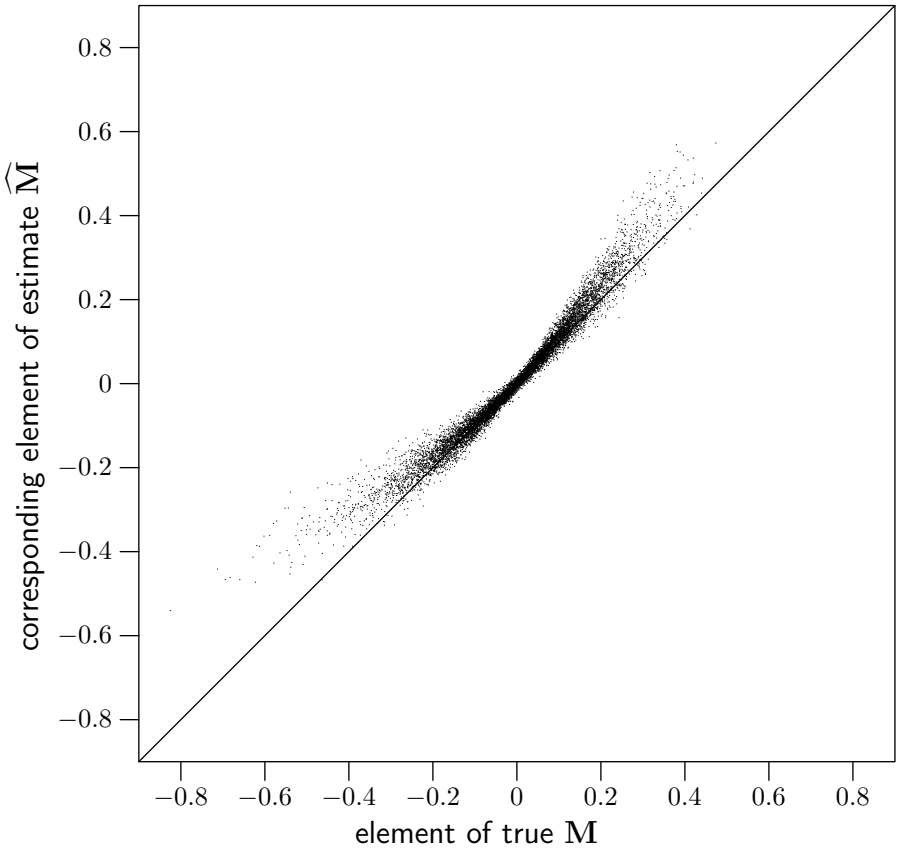

Figure 5: The approximation used in the estimation of $\mathbf{M}$ introduces a systematic bias in the estimate $\widehat{\mathbf{M}}$. The figure shows a scatter plot of the 10000 elements of all 100 matrices $\mathbf{M}$ vs. the corresponding elements of estimates $\widehat{\mathbf{M}}$. Let $\mathbf{M}(i,j)$ 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.

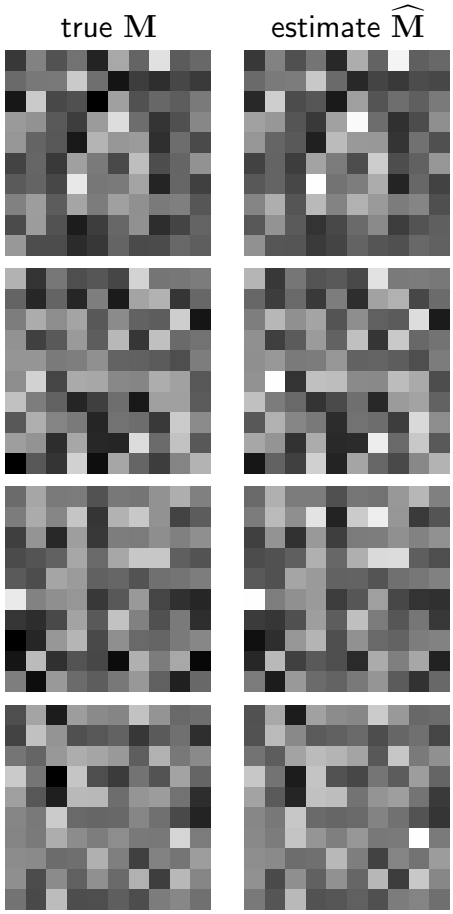

Figure 6: Estimates $\widehat{\mathbf{M}}$ are very similar to the true $\mathbf{M}$, except for positive differences at elements with high absolute values. This is a consequence of the fairly small relative error and the fact that the systematic bias made in the estimation of $\mathbf{M}$ accounts for 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 $\mathbf{M}$ and $\widehat{\mathbf{M}}$ 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 5.

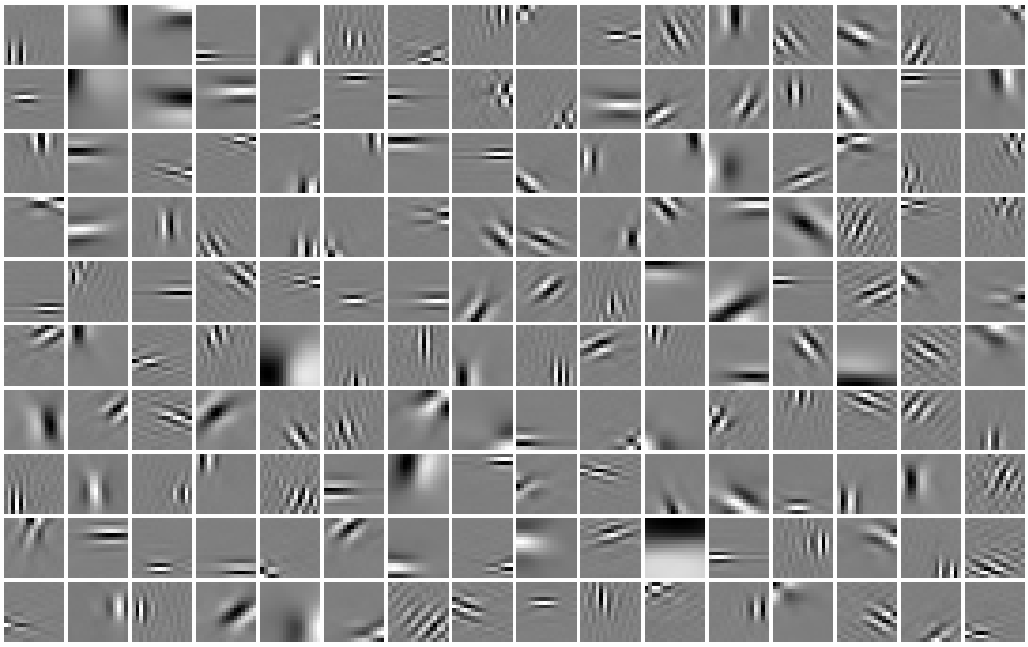

Figure 7: The estimation of the generative model from natural visual stimuli results in the emergence of localized, oriented receptive fields with multiple scales. These basis vectors (columns of $\mathbf{A}$) were obtained by applying the estimation procedure described in Section 4 to a large set of samples from natural image sequences. The basis vectors are in no particular order in this figure.

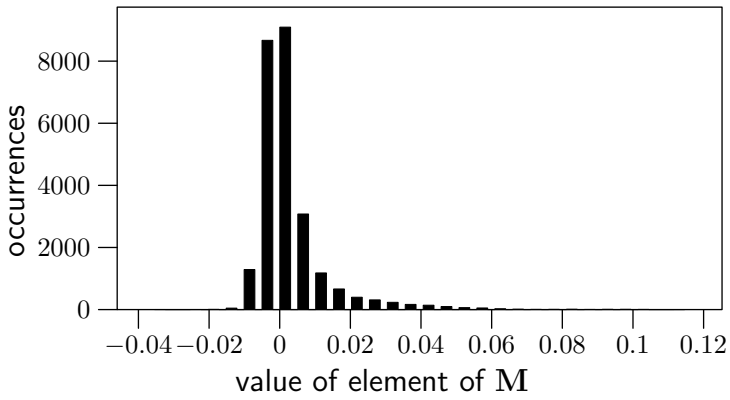

Figure 8: Histogram of the non-diagonal elements of $\mathbf{M}$ estimated from natural image sequence data.

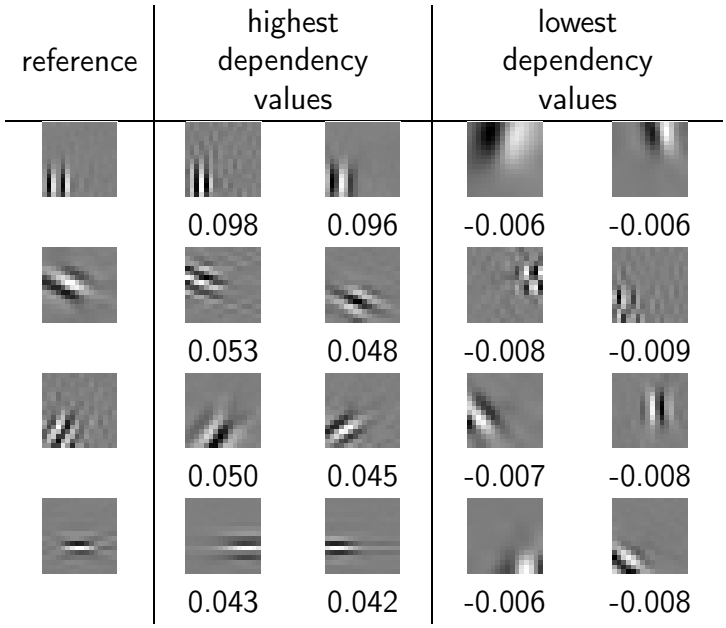

Figure 9: Basis vectors (columns of $\mathbf{A}$) with high activity dependency values code for similar features at nearby positions, whereas basis vectors with low dependency values code for features with different scale and/or orientation and/or location. Each row shows the basis vectors with highest and lowest dependency values with respect to the reference basis vector in the leftmost column. The reference vectors were chosen from the set of vectors in Figure 7 as representatives of four different orientations. The measure of spatiotemporal dependency used was $\big(\mathbf{M}(i,j) + \mathbf{M}(j,i)\big)/2$, 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 level dependency code for similar features (orientation, frequency) as the reference vector, whereas those with low (negative) dependency code for different scale and/or orientation and/or location.

Figure 10: Grouping similar to complex cell pooling of simple cell outputs emerges from spatiotemporal activity level dependencies. Here we have plotted the basis vectors (columns of $\mathbf{A}$) at 2D positions obtained by applying multidimensional scaling to the similarity values defined by $\mathbf{M}$ (see text for details). 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 were magnified in order to be able to show the basis vectors in a reasonable scale.

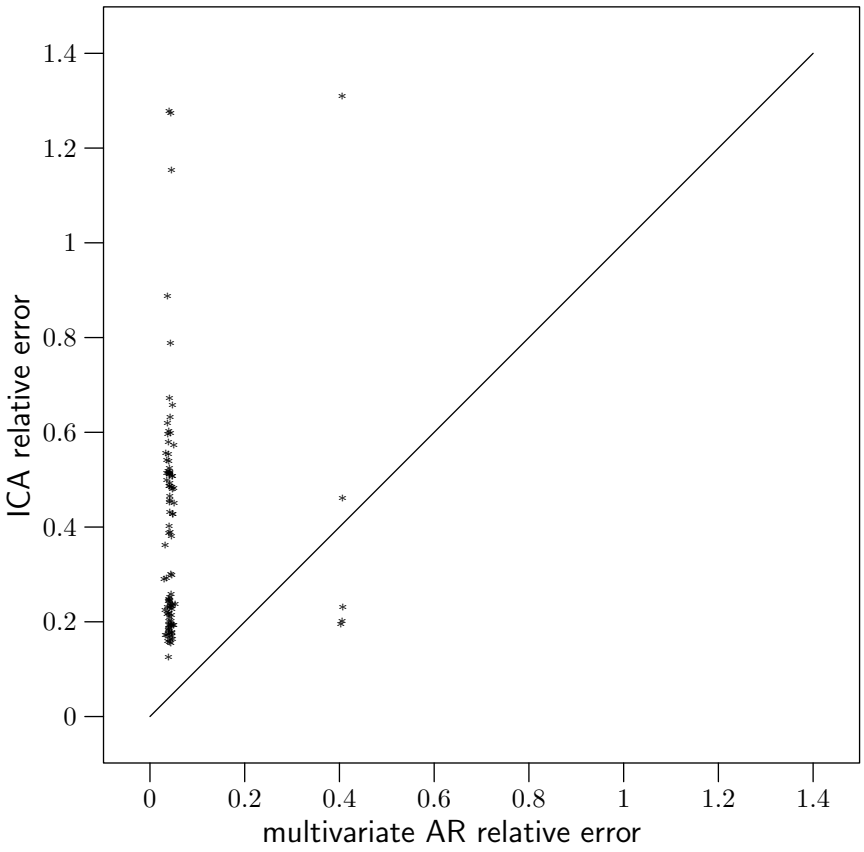

Figure 11: Our estimation method outperforms standard independent component analysis when spatiotemporal activity level dependencies are sufficiently large, as they are in the case of image sequence data. The figure shows a scatter plot of the relative errors made in 100 runs of our algorithm (horizontal axis) against an independent component analysis algorithm (vertical axis) for a matrix $\mathbf{M}$, which was a submatrix of the matrix $\mathbf{M}$ estimated from image sequence data. Each asterisk ($*$) corresponds to one run of both algorithms using the same simulated input data set. For each data set, a different $\mathbf{A}$ and different initial starting points for the algorithms were used. The solid diagonal line represents those points at which the two relative errors are equal; in the area above this line the error made in independent component analysis is larger.

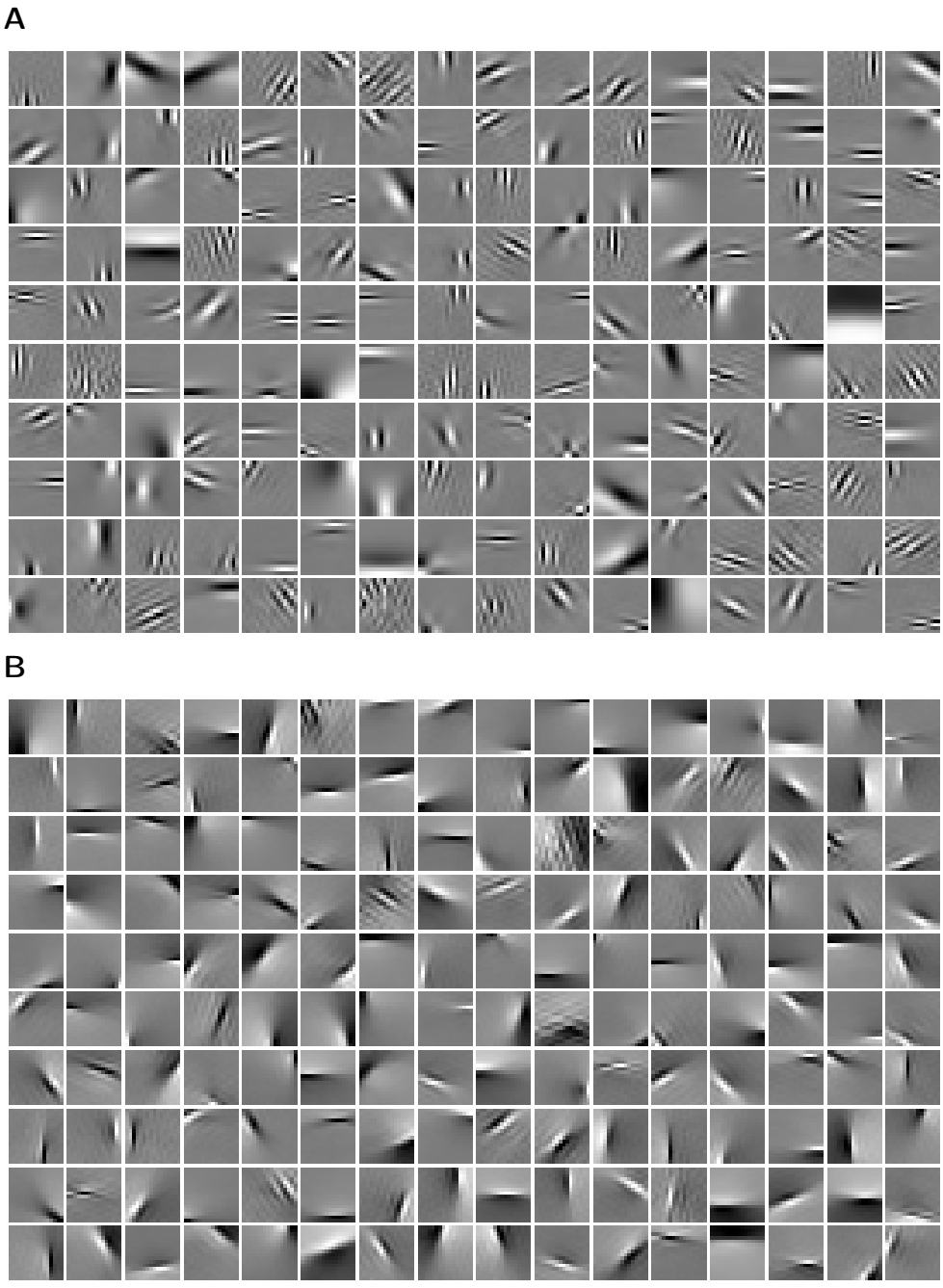

Figure 12: A comparison of results obtained from estimating a basis from natural image sequences with our method and independent component analysis. (A) Basis obtained by our method. (B) Basis obtained using independent component analysis. See text for details.