Chapter 11
Energy Correlations and Topographic Organization

The energy detection model in the preceding chapter can easily be modified to incorporate topography, i.e. an arrangement of the features on a two-dimensional grid. This is very interesting because such organization is one of the most prominent phenomena found in the primary visual cortex. In this chapter, we shall investigate such a topographic version of the ICA model. It is, mathematically, a rather simple modification of the independent subspace analysis model.

11.1 Topography in the Cortex

Topography means that the cells in the visual cortex are not in any random order; instead, they have a very specific spatial organization. When moving on the cortical surface, the response properties of the neurons change in systematic ways. The phenomenon can also be called topological organization, and sometimes the term "columnar organization" is used in almost the same sense.

Fundamentally the cortex is, of course, three-dimensional. In addition to the surface coordinates, which we denote by x_c and y_c, there is the depth dimension z_c. The depth "axis" goes from the very surface of the cortex through different layers of the grey matter to the white matter.

However, the depth dimension is usually assumed to be different from the other two dimensions. In the most simplistic interpretations, the cells that are on the same surface location (x_c, y_c) are similar irrespective of how deep they are on the cortex. This is most clearly expressed in the classic "ice cube" model of V1. Such a simplistic view has been challenged, and it is now well known that at least some properties of the cells are clearly different in different (depth) layers. In particular, input to V1 is received in some of the layers and others are specialized in outputting the results. Still, it seems that the response properties which we consider in this book, such as location, frequency, and orientation selectivities, depend mainly on the coordinates (x_c, y_c) of the cell with respect to the surface.

Looking at the spatial organization of response properties as a function of the surface coordinates x_c and y_c, the most striking aspect of topographic organization is *retinotopy*, which means that the location of the receptive field in the retinal space is closely correlated with the x_c and y_c coordinates. The global correspondence of the retinal coordinates and the cortical coordinates is somewhat complicated due to such phenomena as the magnification factor (the area in the center of the visual field has a relatively larger representation on the cortex), the division into two hemispheres, some unexpected discontinuities, and so on. The correlation is, therefore, more of a local nature.

A. Hyvärinen, J. Hurri, P.O. Hoyer, *Natural Image Statistics,*
Computational Imaging and Vision 39,

The second important topographic property is the gradual change of orientation tuning. The preferred orientation of simple and complex cells mostly changes smoothly. This phenomenon is often referred to as *orientation columns*. They can be seen most clearly in optical imaging experiments where one takes a "photograph" of the cortex that shows which regions are active when the input consists of a grating of a given orientation. Such activity patterns take the form of stripes (columns).

The third important property of spatial organization is that frequency selectivity seems to be arranged topographically into low-frequency blobs so that the blobs (or at least their centers) contain predominantly cells that prefer low-frequency cells and the inter-blob cells prefer higher frequencies. These low-frequency blobs seem to coincide with the well-known cytochrome oxidase blobs.

A final point to note is that phase is not arranged topographically. In fact, phase seems to be completely random: there is no correlation between the phase parameters in two neighboring cells.

11.2 Modeling Topography by Statistical Dependence

Now, we show how to extend the models of natural image statistics to include topography. The key is to consider the dependencies of the components. The model is thus closely related to the model of independent subspace analysis in Chap. 10. In fact, ISA can be seen as a special case of this model.

11.2.1 Topographic Grid

To model topographic organization, we have to first define which features are "close to each other" on the cortical surface. This is done by arranging the features s_i on a two-dimensional grid or lattice. The restriction to 2D is motivated by cortical anatomy, but higher dimensions are equally possible. The spatial organization on the grid models the organization on the cortical surface. The arrangement on the lattice is illustrated in Fig. 11.1.

The topography is formally expressed by a neighborhood function $\pi(i, j)$ that gives the proximity of the features (components) with indices i and j. Typically, one defines that $\pi(i, j)$ is 1 if the features are sufficiently close to each other (they are "neighbors"), and 0 otherwise. Typically, the neighborhood function is chosen by defining the neighborhood of a feature to be square. For example, $\pi(i, j)$ is 1 if the feature j is in a 5×5 square centered on feature i; otherwise, $\pi(i, j)$ is zero.

11.2.2 Defining Topography by Statistical Dependencies

Consider a number of features s_i, $i = 1, \ldots, n$. How can we order the features on the topographic grid in a meaningful way? The starting point is to define a measure of

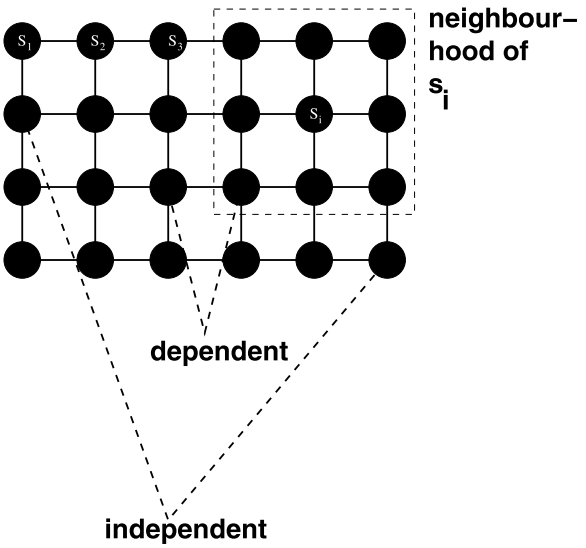

Fig. 11.1 Illustration of topography and its statistical interpretation. The neurons (feature detectors) are arranged on a two-dimensional grid that defines which neurons are near to each other and which are far from each other. It also defines the neighborhood of a cell as the set of cells which are closer than a certain radius. In the statistical model, neurons that are near to each other have statistically dependent outputs, neurons that are far from each other have independent outputs

similarity between two features, and then to order the features so that features that are similar are close to each other on the grid. This is a general principle that seems fair enough. But then, what is a meaningful way of defining similarity between two features? There are actually a couple of different possibilities.

In many models, the similarity of features is defined by similarity of the features weights or receptive fields W_i. Typically, this means the dot-product (also called, somewhat confusingly, the correlation of the receptive fields). This is the case in Kohonen's self-organizing map and related models. However, this seems rather inadequate in the case of the visual cortex. For example, two features of the same frequency need not exhibit large dot-products of weight vectors; in fact, the dot-product can be zero if the features are of orthogonal orientations with otherwise similar parameters. Yet, since the V1 exhibits low-frequency blobs, low-frequency features should be considered similar to each other even if they are quite different with respect to other parameters. What's even worse is that since the phases change randomly when moving a bit on the cortical surface, the dot-products between neighboring components also change rather randomly since the phase has a large influence on the shape of the receptive fields and on the dot-products.

Another candidate for a similarity measure would be correlation of the feature detector outputs s_i when the input consists of natural images. However, this is no good either, since the outputs (components) are typically constrained to be exactly uncorrelated in ICA and related models. Thus, they would all be maximally dissimilar if similarity is based on correlations.

Yet, using correlations seems to be a step in the right direction. The central hypothesis used in this book—visual processing in the cortex is strongly influenced by the statistical structure of the natural input—would suggest that we have to look at the statistics of feature detector outputs in order to find a meaningful measure of

similarity to be used in a model of topography. We just need more information than the ordinary linear correlations.

Our statistical approach to topography thus concentrates on the pattern of statistical dependencies between the s_i, assuming that the joint distribution of the s_i is dictated by the natural image input. The basic idea is that *similarity is defined by the statistical dependency of the outputs*. Thus, features that have strong statistical dependencies are defined to be similar, and features that are independent or weakly dependent are defined to be dissimilar.

The application of this principle is illustrated in Fig. 11.1. The linear feature detectors (simple cells) have been arranged on a grid (cortex) so that any two feature detectors that are close to each other have dependent outputs, whereas feature detectors that are far from each other have independent outputs.

Actually, from Chaps. 9 and 10, we know what are the most prominent statistical dependencies that remains after ordinary ICA: the correlations of squares (or absolute values, which seems to be closely related). Thus, we do not need to model the whole dependency structure of the s_i, which would be most complicated. We can just concentrate on the dependencies of the squares s_i^2.

11.3 Definition of Topographic ICA

As in the ICA and ISA models, we model the image as a linear superposition of features A_i with random coefficients s_i:

$$I(x, y) = \sum_{i=1}^{m} A_i(x, y) s_i. \tag{11.1}$$

As in ICA and ISA, the s_i are obtained as the outputs of linear feature detectors as

$$s_i = \sum_{x,y} W_i(x, y) I(x, y) = \sum_{j=1}^{n} v_{ij} z_j = \mathbf{v}_i^{\mathrm{T}} \mathbf{z} \tag{11.2}$$

where the z_j denotes the jth variable obtained from the image patch by canonical preprocessing.

The point is now to define the joint pdf of the s_i so that it expresses the topographic ordering. First, we define the "local energies" as

$$c_i = \sum_{j=1}^{n} \pi(i, j) s_j^2. \tag{11.3}$$

This is basically the general activity level in the neighborhood of the linear feature s_i. The weighting by $\pi(i, j)$ means that we only sum over s_j which are close to s_i in the topography.

Next, we define the likelihood of the topographic ICA model by a simple modification of the log-likelihood in the ISA model, given in (10.11) on page 219. We replace the subspace energies e_k by these local energies. (The connection between the two models is discussed in more detail later.) Thus, define the pdf of the s_i as

$$\log p(s_1, \ldots, s_n) = \sum_{i=1}^{n} h\left(\sum_{j=1}^{n} \pi(i, j) s_j^2 \right) \tag{11.4}$$

where h is a convex function as in the preceding chapters, e.g. Sect. 6.2.1. Assuming we have observed a set of image patches, represented by z_t, $t = 1, \ldots, T$ after canonical preprocessing, we obtain the likelihood

$$\log L(\mathbf{v}_1, \ldots, \mathbf{v}_n) = T \log \left| \det(\mathbf{V}) \right| + \sum_{i=1}^{n} \sum_{t=1}^{T} h\left(\sum_{j=1}^{n} \pi(i, j) \left(\mathbf{v}_j^{\mathrm{T}} \mathbf{z}_t \right)^2 \right). \tag{11.5}$$

The topography given by $\pi(i, j)$ is considered fixed, and only the linear feature weights $\mathbf{v}_j$ are estimated, so this likelihood is a function of the $\mathbf{v}_j$ only. As in earlier models, the vectors $\mathbf{v}_j$ are constrained to form an orthogonal matrix, so the determinant is constant (one) and the term $T \log |\det(\mathbf{V})|$ can be ignored.

The central feature of this model is that the responses s_i of near-by simple cells are *not* statistically independent in this model. The responses are still linearly uncorrelated, but they have non-linear dependencies. In fact, the energies s_i^2 are strongly positively correlated for neighboring cells. This property is directly inherited from the ISA model; that connection will be discussed next.

11.4 Connection to Independent Subspaces and Invariant Features

Topographic ICA can be considered a generalization of the model of independent subspace analysis. The likelihood of ISA (see (10.11)) can be expressed as a special case of the likelihood in (11.5) with a neighborhood function which is one if the components are in the same subspace and zero otherwise, or more formally:

$$\pi(i, j) = \begin{cases} 1, & \text{if there is some subspace with index } q \text{ so that } i, j \in S(q), \\ 0, & \text{otherwise.} \end{cases}$$

This shows that topographic ICA is closely connected to the principle of invariant feature subspaces in Chap. 10. In topographic ICA, every component has its own neighborhood, which corresponds to a subspace in ISA. Each of the local energies c_i could be considered as the counterpart of the energies e_k in ISA. Thus, the local energies, possibly after a non-linear transform, can be interpreted as the values of invariant features. The pooling process is controlled by the neighborhood function

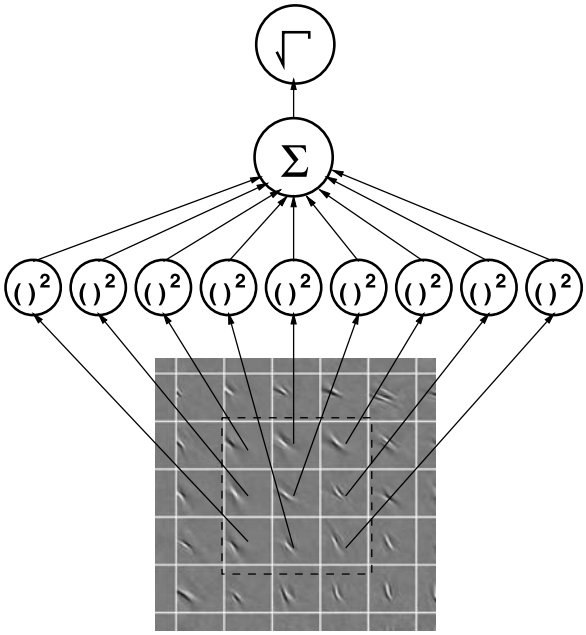

Fig. 11.2 Computation of invariant features in the topographic ICA model. Invariant features (complex cell outputs) are obtained by summing the squares of linear features (simple cell outputs) in a neighborhood of the topographic grid. From Hyvärinen et al. (2001a), Copyright ©2001 MIT Press, used with permission

$\pi(i, j)$. This function directly gives the pooling weights, i.e. the connections between the linear features with index i and the invariant feature cell with index j. Note that the number of invariant features is here equal to the number of underlying linear features.

The dependencies of the components can also be deduced from this analogy with ISA. In ISA, components which are in the same subspace have correlations of energies. In topographic ICA, components which are close to each other in the topographic grid have correlations of squares. Thus, all the features in the same neighborhood tend to be active (non-zero) at the same time.

In a biological interpretation, our definition of the pooling weights from simple cells to complex cells in topographic ICA is equivalent to the assumption that complex cells only pool outputs of simple cells that are near-by on the topographic grid. Neuroanatomic measurements indicate that the wiring of complex cells may indeed be so constrained; see References below. Such a two-layer network is illustrated in Fig. 11.2.

11.5 Utility of Topography

What is the computational utility of a topographic arrangement? A widely used argument is that such a spatial arrangement is useful to *minimize wiring length*. Wiring

length means here the length of the physical connections (axons) needed to send signals from one neuron to another. Consider, for example, the problem of designing the connections from simple cells to complex cells so that the "wires" are as short as possible. It is rather obvious that topographic ICA is related to minimizing that wiring length because in topographic ICA all such connections are very local in the sense that they are not longer that the radius of the neighborhoods. A more general task may be to pool of responses to reduce noise: if a cell in a higher area wants to "read", say, the orientation of the stimulus, it could reduce noise in V1 cell responses by looking at the average of the responses of many cells which have the same orientation selectivity.

In general, if we assume that two cells need to communicate with each other if (and only if) their outputs are statistically dependent, topographic ICA provides optimal wiring. The same applies if the responses of two cells are combined by a third cell only if the outputs of the two cells are statistically dependent. Such assumptions are reasonable because if the cells represent pieces of information which are related (in some intuitive sense), it is likely that their outputs are statistically dependent, and vice versa; so, statistical dependence tells which cells contain related information which has to be combined in higher levels.

Minimization of wiring length may be important for keeping the total brain volume minimal: a considerable proportion of the brain volume is used up in interconnecting axons. It would also speed up processing because the signal travels along the axons with limited speed.

11.6 Estimation of Topographic ICA

A fundamental similarity to ISA is that we do *not* specify what parameters should be considered as defining the topographic order. That is, the model does not specify, for example, that near-by neurons should have receptive fields that have similar locations, or similar orientations. Rather, we let the natural images decide what the topography should be like, based on their statistical structure.

Since we have already defined the likelihood in (11.5), estimation needs hardly any further comment. We use whitened (canonically preprocessed) data, so we constrain $\mathbf{V}$ to be orthogonal just like in ICA and ISA. We maximize the likelihood under this constraint. The computational implementation of such maximization is discussed in detail in Chap. 18, in particular, Sect. 18.5.

The intuitive interpretation of such estimation is that we are maximizing the sparsenesses of the local energies. This is completely analogue to ISA, where we maximize sparsenesses of complex cell outputs. The learning process is illustrated in Fig. 11.3.

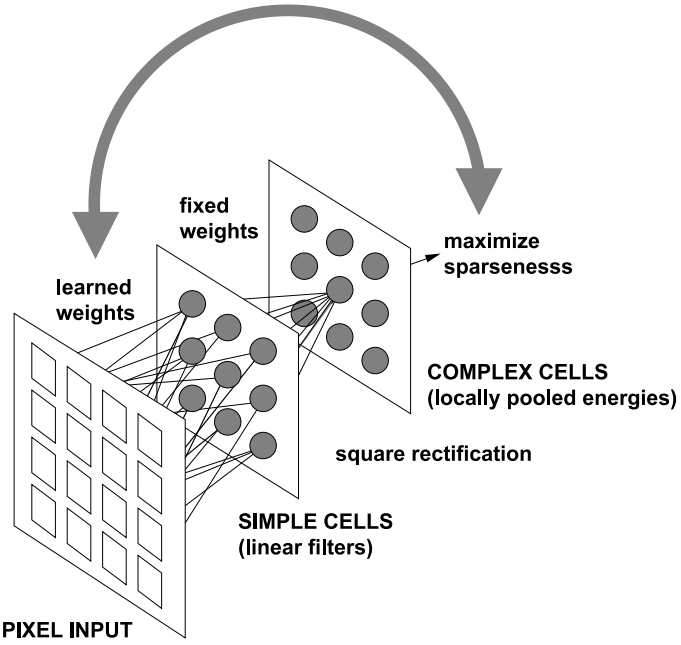

Fig. 11.3 Illustration of learning in the topographic ICA model. From Hyvärinen and Hoyer (2001), Copyright ©2001 Elsevier, used with permission

11.7 Topographic ICA of Natural Images

11.7.1 Emergence of V1-like Topography

11.7.1.1 Data and Preprocessing

We performed topographic ICA on the same data as in previous chapters. We took the same 50 000 natural image patches of size 32×32 as in the preceding chapters. We preprocessed the data in the same way as in the ISA case: This means divisive normalization using (9.11), and reducing the dimension to 256 by PCA. The nonlinearity h was chosen to be a smoothed version of the square root as in (6.14), just like in the ISA experiments.

The topography was chosen so that $\pi(i, j)$ is 1 if the cell j is in a 5×5 square centered on cell i; otherwise $\pi(i, j)$ is zero. Moreover, it was chosen to be cyclic (toroidal) so that the left edge of the grid is connected to the right edge, and the upper edge is connected to the lower edge. This was done to reduce border artifacts due to the limited size to the topographic grid.

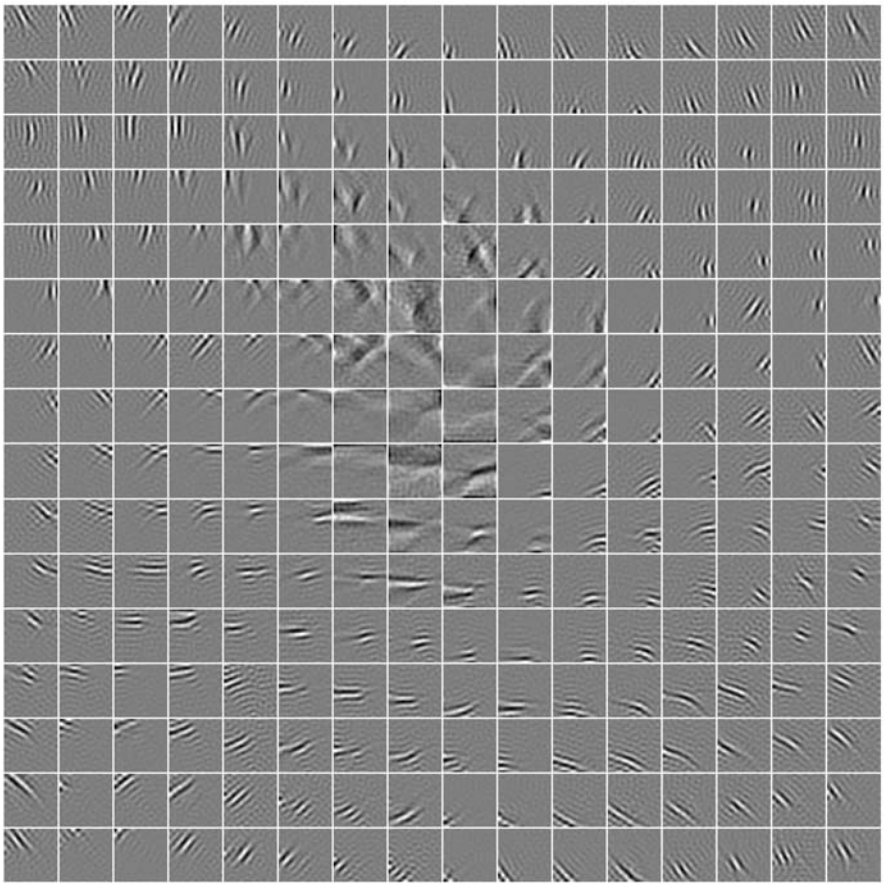

Fig. 11.4 The whole set of vectors W_i obtained by topographic independent component analysis, in the topographic order

11.7.1.2 Results and Analysis

The linear detector weights W_i obtained by topographic ICA from natural images are shown in Fig. 11.4, and the corresponding feature vectors A_i are in Fig. 11.5. The topographic ordering is visually obvious. The underlying linear features are tuned for the three principal parameters: orientation, frequency, and location. Visual inspection of the map shows that orientation and location mostly change smoothly as a function of position on the topographic grid. A striking feature of the map is a "blob" grouping low-frequency features. Thus, the topography is determined by the same set of parameters for which the features are selectively tuned; these are just the same as in ICA and ISA. These are also the three parameters with respect to which a clear spatial organization has been observed in V1.

The topography can be analyzed in more detail by either a global or a local analysis. A local analysis is done by visualizing the correlations of the optimal Gabor

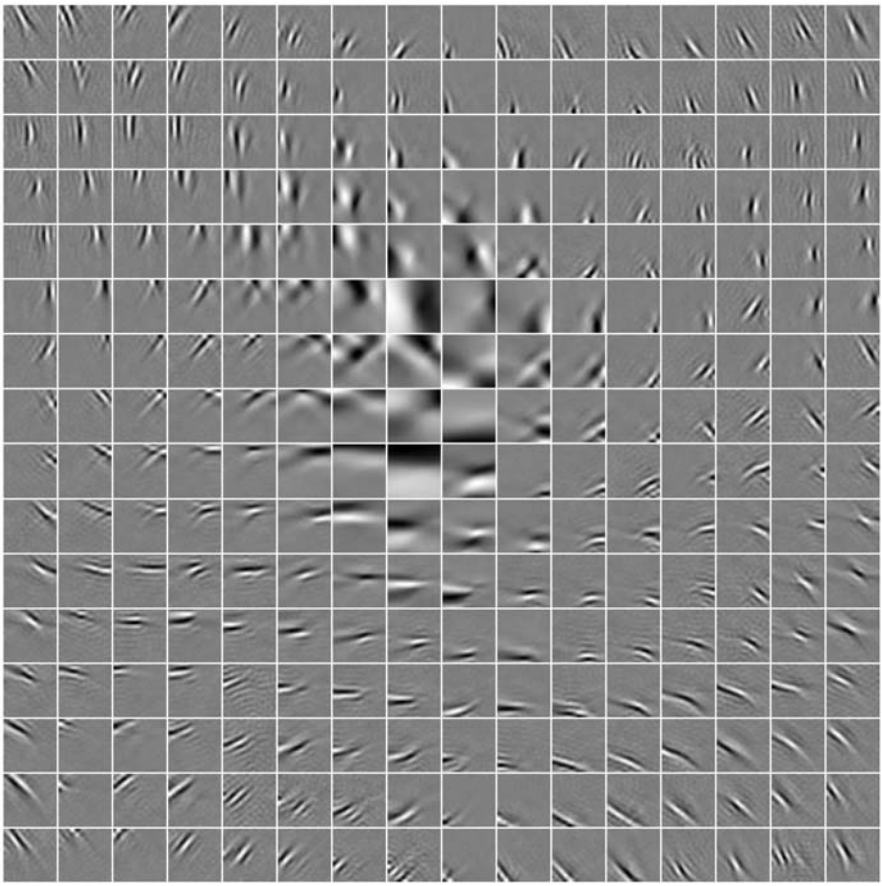

Fig. 11.5 The whole set of vectors A_i obtained by topographic independent component analysis

parameters for two linear features that are immediate neighbors. In Fig. 11.6, we see that the locations (a, b) and orientations (c) are strongly correlated. In the case of frequencies (d), the correlation is more difficult to see because of the overall concentration to high frequencies. As for phases (e), no correlation (or any kind of statistical dependency) can be seen, which is again similar to what has been observed in V1. Furthermore, all these correlations are similar to the correlations inside independent subspaces in Fig. 10.8 on page 229. This is not surprising because of the intimate connection between the two models, explained above in Sect. 11.4.

A global analysis is possible by color-coding the Gabor parameters of linear features. This gives "maps" whose smoothness shows the smoothness of the underlying parameter. The maps are shown in Fig. 11.7. The locations (a and b) can be seen to change smoothly, which is not obvious from just looking at the features in Fig. 11.4. The orientation and frequency maps (c and d) mainly change smoothly, which was rather obvious from Fig. 11.4 anyway. In some points, the orientation

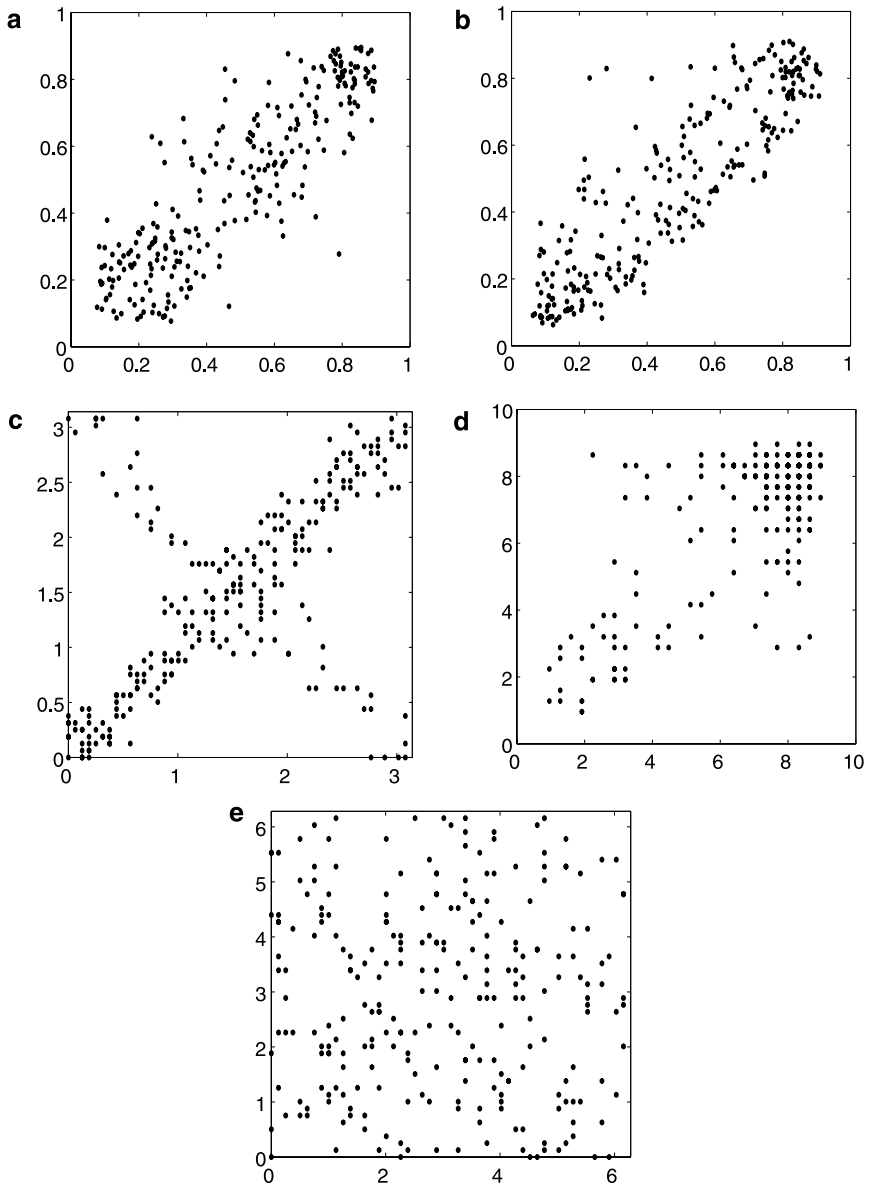

Fig. 11.6 Correlation of parameters characterizing the linear features of two neighboring features in Fig. 11.4. An immediate neighbor for each cell chosen as the one immediately to the right. Each point in the scatter plots is based on one such couple. **a** scatter plot of locations along x-axis, **b** locations along y-axis, **c** orientations, **d** frequencies, and **e** phases. The plots are very similar to corresponding plots for ISA in Fig. 10.8 on page 229; the main visual difference is simply due to the fact that here we have twice the number of dots in each plot

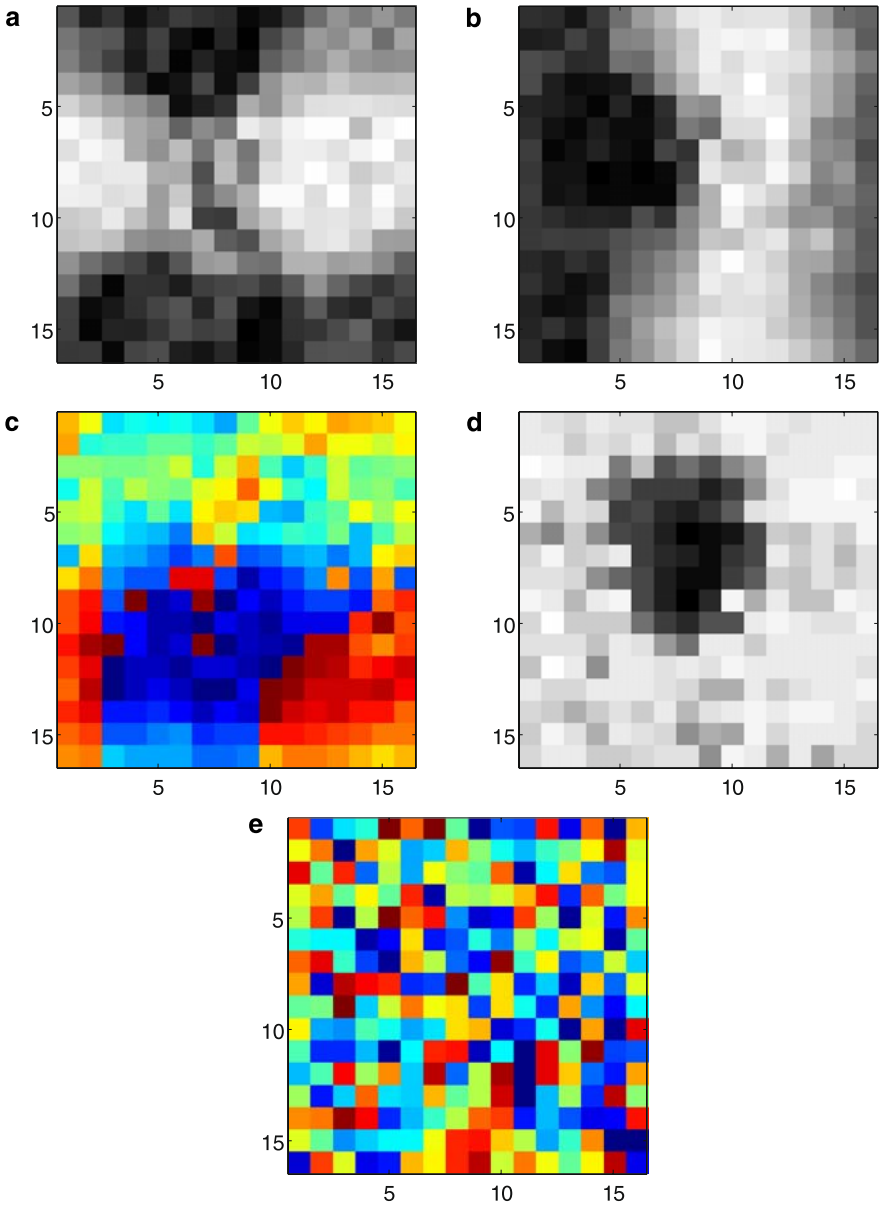

Fig. 11.7 Global structure of the topography estimated from natural images in Fig. 11.4. Each parameter of the Gabor functions describing the features is plotted grey-scale or color-coded. Color-coding is used for parameters which are cyclic: orientation and phase, since the color spectrum is also cyclic. The actual values of the parameters are not given because they have little importance. **a** Locations along x-axis, **b** locations along y-axis, **c** orientations, **d** frequencies, and **e** phases

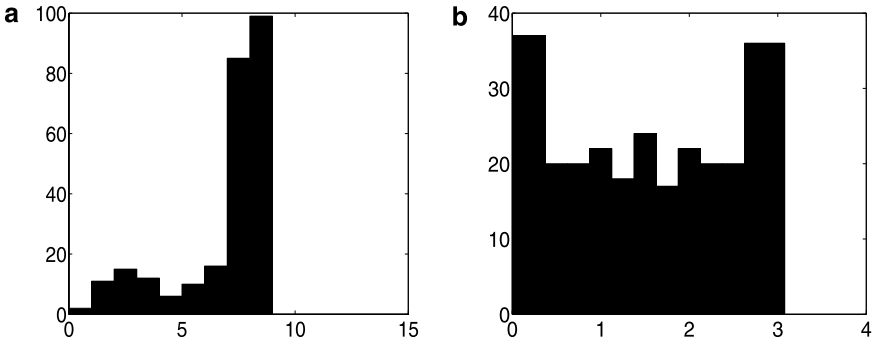

Fig. 11.8 Histograms of the optimal **a** frequencies and **b** orientations of the linear features in topographic ICA

seems to change abruptly, which may correspond to so-called "pinwheels", which are points in which many different orientations can be found next to each other, and have been observed on the cortex. As for phases, the map in (e) shows that they really change randomly.

We can also analyze the distribution of the frequencies and orientations of the features. The plot in Fig. 11.8 shows the histograms preferred orientations and frequencies for the linear features. We see that all orientations are almost equally present, but the horizontal orientation is slightly overrepresented. This is the same anisotropy we have seen all preceding models. In contrast, the frequency distribution is very strongly skewed: most linear features are tuned to high frequencies. However, the distribution of frequencies is a bit closer to uniform than in the cases of ICA (Fig. 6.9) or ISA (Fig. 10.9).

The connection of the model to ISA suggests that the local energies can be interpreted as invariant features. What kind of invariances do we see emerging from natural images? Not surprisingly, the invariances are similar to what we obtained with ISA, because the neighborhoods have the same kinds of parameters correlations (Fig. 11.6) as in ICA; we will not analyze them in more detail here. The main point is that *local energies are like complex cells*. That is, the topographic ICA model automatically incorporates a complex cell model.

Basically, the conclusion to draw from these results is that the topographic ICA model produces a spatial topographic organization of linear features that is quite similar to the one observed in V1.

11.7.1.3 Image Synthesis Results and Sketch of Generative Model

Next, we will synthesize images from the topographic ICA model. This is a bit tricky because, in fact, we did not yet introduce a proper generative model for topographic ICA. Such a model can be obtained as a special case of the framework introduced later in Sect. 11.8.2. We will here briefly describe how such a generative model can be obtained.

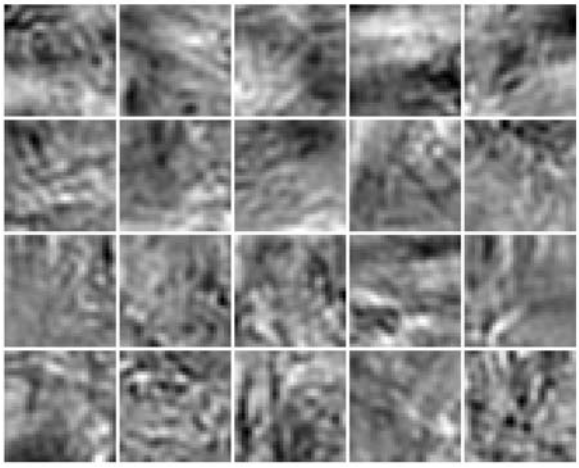

Fig. 11.9 Image synthesis using topographic ICA. Compare with the ICA results in Fig. 7.4 on page 162 and ISA results in Fig. 10.10 on page 231

Basically, the idea is a simple generalization of the framework using variance variables as in Sects. 10.4 and 9.3. Here, we have a separate variance variable d_i for each component s_i:

$$s_i = \tilde{s}_i d_i \tag{11.6}$$

where the $\tilde{s}_i$ are Gaussian and independent from each other (and from the d_i). The point is to generate the d_i so that their dependencies incorporate the topography. This can be accomplished by generating them using a higher-order ICA model, where the mixing matrix is given by the neighborhood function. Denoting the higher-order components by u_i, we simply define

$$d_i = \sum_j \pi(i, j) u_i . \tag{11.7}$$

This produces approximately the same distribution as the pdf which we used to define the topographic ICA model earlier in this chapter. See Sect. 11.8.2 for details. A problem we encounter here is that it is not obvious how to estimate the distributions of the u_i. So, we have to fix them rather arbitrarily, which means the results are not quite directly comparable with those obtained by ISA and ICA where we could use the observed histograms of the features.

Results of synthesizing images with this generative model are shown in Fig. 11.9. The u_i were generated as the fourth powers of Gaussian variables. The synthesized images seem to have more global structure than those obtained by ICA or ISA, but as we just pointed out, this may be related to the way we fixed the distributions of the u_i.

11.7.2 Comparison with Other Models

When compared with other models on V1 topography, we see three important properties in the topographic ICA model:

1. The topographic ICA model shows emergence of a topographic organization using the above-mentioned three principal parameters: location, frequency, and orientation. The use of these particular three parameters is not predetermined by the model, but determined by the statistics of the input. This is in contrast to most models that only model topography with respect to one or two parameters (usually orientation possibly combined with binocularity) that are chosen in advance.
2. No other model has shown the emergence of a low-frequency blob.
3. Topographic ICA may be the first one to explicitly show a connection between topography and complex cells. The topographic, columnar organization of the simple cells is such that complex cell properties are automatically created when considering local activations. This is related to the randomness of phases, which means that in each neighborhood, there are linear features with very different phases, like in the subspaces in ISA.

It is likely that the two latter properties (blobs and complex cells) can only emerge in a model that is based on simultaneous activation (energy correlation) instead of similarity of receptive fields as measured by Euclidean distances or receptive field correlations. This is because Euclidean distances or correlations between feature vectors of different frequencies, or of different phases, are quite arbitrary: they can obtain either large or small values depending on the other parameters. Thus, they do not offer enough information to qualitatively distinguish the effects of phase vs. frequency, so that phase can be random and frequency can produce a blob.

11.8 Learning Both Layers in a Two-Layer Model *

In this section, we discuss estimation of a two-layer model which is a generalization of the topographic ICA. The section is quite sophisticated mathematically, and presents ongoing work with a lot of open problems, so it can be skipped by readers not interested in mathematical details.

11.8.1 Generative vs. Energy-Based Approach

Many of the results in the preceding chapters are related to a two-layer generative model. In the model, the observed variables $\mathbf{z}$ are generated as a linear transformation of components $\mathbf{s}$, just as in the basic ICA model: $\mathbf{z} = \mathbf{A}\mathbf{s}$. The point is to define the joint density of $\mathbf{s}$ so that it expresses the correlations of squares that seem to be dominant in image data.

There are two approaches we can use. These parallel very much the sparse coding and ICA approaches in Chaps. 6 and 7. In the first approach, typically called "energy-based" for historical reasons,[1] we just define an objective function which expresses sparseness or some related statistical criterion, and maximize it. In the second approach, we formulate a generative model which describes how the data is generated starting from some elementary components. We shall consider here first the generative-model approach; the energy-based model is considered in Sect. 11.8.5.

11.8.2 Definition of the Generative Model

In the generative-model approach, we define the joint density of $\mathbf{s}$ as follows. The variances d_i^2 of the s_i are not constant, instead they are assumed to be random variables. These random variables d_i are, in their turn, generated according to a model to be specified. After generating the variances d_i^2, the variables s_i are generated independently from each other, using some conditional distributions to be specified. In other words, the s_i are *independent given their variances*. Dependence among the s_i is implied by the dependence of their variances.

This is a generalization of the idea of a common variance variable presented in Sect. 7.8.3. Here, there is no single common variance variable, since there is a separate variance variable d_i^2 corresponding to each s_i. However, these variance variables are correlated, which implies that the squares of the s_i are correlated. Consider the extreme case where the d_i are completely correlated. Then the d_i^2 are actually the same variable, possibly multiplied by some constants. Thus, in this extreme case, we actually have just a single variance variable as in the divisive normalization model in Chap. 9.

Many different models for the variances d_i^2 could be used. We prefer here to use an ICA model followed by a non-linearity:

$$d_i = r\left(\sum_{k=1}^{n} \pi(i,k) u_k\right). \tag{11.8}$$

Here, the u_k are the "higher-order" independent components used to generate the variances, and r is some scalar non-linearity (possibly just the identity $r(z) = z$). The coefficients $\pi(i,k)$ are the entries of a higher-order feature matrix. It is closely related to the matrix defining the topography in topographic ICA, which is why we use the same notation.

This particular model can be motivated by two facts. First, taking sparse u_i, we can model a two-layer generalization of sparse coding, where the activations (i.e. the variances) of the components s_i are sparse, and constrained to some groups of "related" components. Related components means here components whose variances are strongly influenced by the same higher-order components u_i.

[1] Note that the word "energy" here has nothing to do with Fourier energy; it comes from a completely different physical analogy.

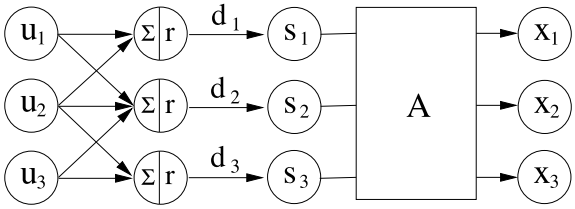

Fig. 11.10 An illustration of the two-layer generative model. First, the "variance-generating" variables u_i are generated randomly. They are then mixed linearly. The resulting variables are then transformed using a non-linearity r, thus giving the local variances d_i^2. Components s_i are then generated with variances d_i^2. Finally, the components s_i are mixed linearly to give the observed variables x_i (which are subsequently whitened to give the z_i)

In the model, the distributions of the u_i and the actual form of r are additional parameters; some suggestions will be given below. It seems natural to constrain the u_k to be non-negative. The function r can then be constrained to be a monotonic transformation in the set of non-negative real numbers. This ensures that the d_i's are non-negative, so is a natural constraint since they give the standard deviation of the components.

The resulting two-layer model is summarized in Fig. 11.10. Note that the two stages of the generative model can be expressed as a single equation, analogously to (9.4), as follows:

$$s_i = r\left(\sum_k \pi(i,k)u_k\right)\tilde{s}_i \tag{11.9}$$

where $\tilde{s}_i$ is a random variable that has the same distribution as s_i given that d_i is fixed to unity. The u_k and the $\tilde{s}_i$ are all mutually independent.

11.8.3 Basic Properties of the Generative Model

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

11.8.3.1 The Components s_i Are Uncorrelated

This is because according to (11.9) we have

$$E\{s_i s_j\} = E\{\tilde{s}_i\}E\{\tilde{s}_j\}E\left\{r\left(\sum_k \pi(i,k)u_k\right)r\left(\sum_k \pi(j,k)u_k\right)\right\} = 0 \tag{11.10}$$

due to the independence of the u_k from $\tilde{s}_i$ and $\tilde{s}_j$. (Recall that $\tilde{s}_i$ and $\tilde{s}_j$ are zero-mean.) To simplify things, one can define that the marginal variances (i.e. integrated

over the distribution of d_i) of the s_i are equal to unity, as in ordinary ICA. In fact, we have

$$E\{s_i^2\} = E\{\tilde{s}_i^2\} E\left\{ r\left(\sum_k \pi(i,k) u_k \right)^2 \right\}, \tag{11.11}$$

so we only need to rescale $\pi(i, j)$ (the variance of $\tilde{s}_i$ is equal to unity by definition).

11.8.3.2 The Components s_i Are Sparse

This is true in the case where component s_i is assumed to have a Gaussian distribution when the variance is given. This follows from the proof given in Sect. 7.8.3: the logic developed there still applies in this two-layer model, when the marginal distribution of each component s_i is consider separately. Then the marginal, unconditional distributions of the components s_i are called Gaussian scale mixtures.

11.8.3.3 Topographic Organization Can Be Modeled

This is possible simply by constraining the higher-order matrix $\pi(i, j)$ to equal a topographic neighborhood matrix as in Sect. 11. We can easily prove that components which are far from each other on the topography are then independent. Assume that s_i and s_j are such that their neighborhoods have no overlap, i.e. there is no index k such that both $\pi(i, k)$ and $\pi(j, k)$ are non-zero. Then their variances d_i and d_j are independent because no higher-order component influences both of these variances. Thus, the components s_i and s_j are independent as well.

11.8.3.4 Independent Subspaces Are a Special Case

This is more or less implied by the discussion in Sect. 11.4 where independent subspace analysis was shown to be a special case of topographic ICA. A more direct connection is seen by noting that each variance variable could determine the variance inside a single subspace, with no interactions between the variance variables. Then we get the ISA model as explained in Sect. 10.4.

11.8.4 Estimation of the Generative Model

11.8.4.1 Integrating Out

In this section, we discuss the estimation of the two-layer model introduced in the previous section. In principle, this can be done by "integrating out" the latent variables. Integrating out is an intuitive appealing method: since the likelihood depends

on the values of the variance variables u_i which we don't know, why not just compute the likelihood averaged over all possible values of u_i? Basically, if we have the joint density of the s_i and the u_i, we could just compute the integral over the u_i to get the density over s_i alone:

$$p(\mathbf{s}) = \int p(\mathbf{s}, \mathbf{u})\, d\mathbf{u} \tag{11.12}$$

The problem is, as always with integration, that we may not be able to express this integral with a simple formula, and numerical integration may be computationally impossible.

In our case, the joint density of **s**, i.e. the topographic components, and **u**, i.e. the higher-order independent components generating the variances, can be expressed as

$$p(\mathbf{s}, \mathbf{u}) = p(\mathbf{s}|\mathbf{u})\, p(\mathbf{u}) = \prod_i p_i^s \left(\frac{s_i}{r(\sum_k \pi(i,k) u_k)} \right) \frac{1}{r(\sum_k \pi(i,k) u_k)} \prod_j p_j^u(u_j) \tag{11.13}$$

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

$$p(\mathbf{s}) = \int \prod_i p_i^s \left(\frac{s_i}{r(\sum_k \pi(i,k) u_k)} \right) \frac{\prod_j p_j^u(u_j)}{r(\sum_k \pi(i,k) u_k)}\, d\mathbf{u} \tag{11.14}$$

Possibly, for some choices of the non-linearity r and the distributions p_i^u, this integral could be computed easily, but no such choices are known to us.

11.8.4.2 Approximating the Likelihood

One thing which we can do is to *approximate* the likelihood by an analytical expression. This approximation actually turns out to be rather useless for the purpose of estimating the two-layer model, but it shows an interesting connection to the likelihood of the topographic ICA model.

To simplify the notation, we assume in the following that the densities p_i^u are equal for all i, and likewise for p_i^s. To obtain the approximation, we first fix the density $p_i^s = p_s$ to be Gaussian, as discussed in Sect. 11.8.3, and we define the non-linearity r as

$$r\left(\sum_k \pi(i,k) u_k \right) = \left(\sum_k \pi(i,k) u_k \right)^{-1/2}. \tag{11.15}$$

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

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

$$p(\mathbf{s}) = \int \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2}\sum_i s_i^2\left[\sum_k \pi(i,k)u_k\right]\right)\prod_i p_u(u_i)\sqrt{\sum_k \pi(i,k)u_k}\, d\mathbf{u} \tag{11.16}$$

which can be manipulated to give

$$p(\mathbf{s}) = \int \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{1}{2}\sum_k u_k\left[\sum_i \pi(i,k)s_i^2\right]\right)\prod_i p_u(u_i)\sqrt{\sum_k \pi(i,k)u_k}\, d\mathbf{u}. \tag{11.17}$$

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

$$\sqrt{\sum_k \pi(i,k)u_k} \approx \sqrt{\pi(i,i)u_i}. \tag{11.18}$$

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

$$\tilde{p}(\mathbf{s}) = \prod_k \exp\left(G\left(\sum_i \pi(i,k)s_i^2\right)\right) \tag{11.19}$$

where the scalar function G is obtained from the p_u by

$$G(y) = \log \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}uy\right) p_u(u)\sqrt{\pi(i,i)u}\, du. \tag{11.20}$$

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

Next, using the same derivation as in ICA, we obtain the likelihood of the data as

$$\log \tilde{L}(\mathbf{V}) = \sum_{t=1}^{\mathrm{T}}\sum_{j=1}^{n} G\left(\sum_{i=1}^{n}\pi(i,j)\left(\mathbf{v}_i^{\mathrm{T}}\mathbf{z}(t)\right)^2\right) + T\log|\det \mathbf{V}|. \tag{11.21}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)^{\mathrm{T}} = \mathbf{A}^{-1}$, and the $\mathbf{z}(t)$, $t = 1, \ldots, T$ are the observations of $\mathbf{z}$. It is here assumed that the neighborhood function and the non-linearity r as well as the densities p_i^u and p_i^s are known. This approximation is a function of local energies. Every term $\sum_{i=1}^{n}\pi(i,j)(\mathbf{v}_i^{\mathrm{T}}\mathbf{z}(t))^2$ could be considered as the energy of a neighborhood, related to the output of a higher-order neuron as in complex cell models. The function G has a similar role as the log-density of the independent components in ICA; the corresponding function h is basically obtained as $h(u) = G(\sqrt{|u|})$.

The formula for G in (11.20) can be analytically evaluated only in special cases. One such case is obtained if the u_k are obtained as squares of standardized Gaussian

variables. Straightforward calculation then gives the following function

$$G_0(y) = -\log(1+y) + \text{const.} \tag{11.22}$$

However, in ICA, it is well known that the exact form of the log-density does not affect the consistency of the estimators, as long as the overall shape of the function is correct. This is probably true in topographic ICA as well.

11.8.4.3 Difficulty of Estimating the Model

What we have really shown in deriving the approximation of likelihood in (11.21) is that the heuristically justified objective function in (11.5) can be obtained from the two-layer generative model as an approximation. But we have not really got any closer to the goal of estimating both layers of weights. This is because the approximation used here approximates the dependence of the likelihood from π quite badly. To see why, consider maximization of the approximative likelihood in (11.21) with respect to the $\pi(i, j)$. Take G as in (11.22). Now, $\sum_{i=1}^{n} \pi(i, j)(\mathbf{v}_i^{\mathrm{T}}\mathbf{z}(t))^2$ is always non-negative. On the other hand, G attains its maximum at zero. So, if we simply take $\pi(i, j) = 0$ for all i, j, G is actually always evaluated at zero and the approximative likelihood is maximized. So, taking all zeros in π is the maximum, which is absurd!

One approach would be to find the values of the latent variables u_i which maximize the likelihood, treating the u_i like the parameters. Thus, we would not try to integrate out the u_i, but rather just formulate the joint likelihood of $\mathbf{V}$, $\pi(i, j)$, $u_i(t)$ for all i, j and all $t = 1, \ldots, T$. This is computationally very difficult because the latent variables d_i are different for each image patch, so there is a very large number of them. The situation could be simplified by first estimating the first layer by ordinary ICA, and then fixing $\mathbf{V}$ once and for all (Karklin and Lewicki 2005). However, this does not reduce the number of dimensions.

So, we see that the estimation of both layers in a generative two-layer model is quite difficult. However, abandoning the generative-model approach simplifies the situation, and provides a promising approach, which will be treated next.

11.8.5 Energy-Based Two-Layer Models

A computationally simpler alternative to estimation of the two layers is provided by an "energy-based" approach. The idea is to take the likelihood in (11.5) as the starting point. As pointed out above, it does not make sense to try to maximize this with respect to the π, because the maximum is obtained by taking all zeros as the second layer weights.

There is a deep mathematical reason why we cannot maximize the likelihood in (11.5) with respect to the π. The reason is that the likelihood is *not normalized*. That is, when we interpret the likelihood as a pdf, its integral over the data variables

is not equal to one: the integral depends on the values of the π. This means it is not a properly defined pdf because a pdf must always integrate to one, so the likelihood is not a properly defined likelihood either. To alleviate this, we have to introduce what is called a *normalization constant* or a *partition function* in the likelihood. The normalization constant, which is actually not a constant but a function of the model parameters, is chosen so that it makes the integral equal to one. Denoting the normalization constant by $Z(\pi)$, we write

$$\log L(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \sum_{t=1}^{\mathrm{T}} \sum_i h\left(\sum_{j=1}^{n} \pi(i,j) \left(\mathbf{v}_i^{\mathrm{T}} \mathbf{z}_t\right)^2 \right) - \log |\det \mathbf{V}| - \log Z(\pi). \tag{11.23}$$

See Sect. 13.1.5 and Chap. 21 for more discussion on the normalization constant.

In principle, the normalization constant can be computed by computing the integral of the underlying pdf over the space of the $\mathbf{v}$, but this is extremely complicated numerically. Fortunately, there is a way around this problem, which is to use special estimation methods which do not require the normalization constant. Thus, we abandon maximization of likelihood because it requires that we compute the normalization constant. See Chap. 21 for information on such methods.

Attempts to estimate both layers in a two-layer model, using an energy-based approach, and estimation methods which circumvent the need for a normalization constant, can be found in Osindero et al. (2006), Köster and Hyvärinen (2007, 2008). This is a very active area of research (Karklin and Lewicki 2008). Some more remotely related work is in Köster et al. (2009a).

11.9 Concluding Remarks and References

A simple modification of the model of independent subspace analysis leads to emergence of topography, i.e. the spatial arrangement of the features. This is in contrast to ICA and ISA, in which the features are in random order. (In ISA, it is the subspaces which are in random order, but the linear features have some organization because of their partition to subspaces.) The basic idea in modeling topography is to consider subspaces which are overlapping, so that the neighborhood of each cell is one subspace. It is also possible to formulate a proper generative model which incorporates the same kind of statistical dependencies using variance variables which are generated by a higher-order ICA model, but that approach is mathematically difficult and still under construction.

Basic though old papers on topography are Hubel and Wiesel (1968), DeValois et al. (1982). Optical imaging results are shown in Blasdel (1992), and a recent high-resolution imaging study is in Ohki et al. (2005). Topography with respect to spatial frequency is investigated in Tootell et al. (1988), Silverman et al. (1989), Edwards et al. (1996). Seminal papers on pinwheels are Bonhoeffer and Grinvald (1991), Maldonado et al. (1997). A most interesting recent paper is DeAngelis et al. (1999) that also shows that the phases are not correlated in neighboring cells. The

relationships of the topographic representation for different parameters are considered in Hübener et al. (1997). An important point is made in Yen et al. (2007), who show that the topography of responses is not so clear when the stimuli are complex, presumable due to non-linear interactions. The connection between topography and complex cell pooling is discussed in Blasdel (1992), DeAngelis et al. (1999).

The idea of minimum wiring length, or wiring economy, goes back to Ramón y Cajal, cited in Chen et al. (2006). The metabolic advantages of topography are further considered in Durbin and Mitchison (1990), Mitchison (1992), Koulakov and Chklovskii (2001), Attwell and Laughlin (2001). Comparisons between white and grey matter volume also point out how brain (skull) size limits the connectivity (Zhang and Sejnowski 2000).

Original papers describing the topographic ICA models are Hyvärinen and Hoyer (2001), Hyvärinen et al. (2001a). Kohonen's famous self-organizing map is also closely related (Kohonen 1982, 2001), but it has not been shown to produce a realistic V1-like topography; reasons for this were discussed in Sect. 11.7.2. A model which produces more a realistic topography (but still no low-frequency blobs) is Kohonen's ASSOM model (Kohonen 1996; Kohonen et al. 1997). However, in that model the nature of the topography is strongly influenced by an artificial manipulation of the input (a sampling window that moves smoothly in time), and it does not really emerge from the structure of images alone.

A related idea on minimization of wiring length has been proposed in Vincent and Baddeley (2003), Vincent et al. (2005), in which it is proposed that the retinal coding minimizes wiring, whereas cortical coding maximizes sparseness of activities.