Cogn Neurodyn (2007) 1:261–272
DOI 10.1007/s11571-007-9021-1

RESEARCH ARTICLE

Robustness of neural codes and its implication on natural image processing

Sheng Li · Si Wu

Received: 10 February 2007 / Accepted: 15 May 2007 / Published online: 12 July 2007

Abstract In this study, based on the view of statistical inference, we investigate the robustness of neural codes, i.e., the sensitivity of neural responses to noise, and its implication on the construction of neural coding. We first identify the key factors that influence the sensitivity of neural responses, and find that the overlap between neural receptive fields plays a critical role. We then construct a robust coding scheme, which enforces the neural responses not only to encode external inputs well, but also to have small variability. Based on this scheme, we find that the optimal basis functions for encoding natural images resemble the receptive fields of simple cells in the striate cortex. We also apply this scheme to identify the important features in the representation of face images and Chinese characters.

Keywords Robust coding · Neural codes · Natural image processing · Neuronal variability · V1

Introduction

In natural environments, variations on the view angle, distance and background, and deformations of images, mean that the external inputs from the same object to neural systems are highly fluctuated. In order to recognize objects reliably, it is important for neural systems to extract the important features of external inputs. Mathematically this is expressed as statistical inference.

We write down the external inputs to a neural system as $I(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$, with $\boldsymbol{x}$ denoting the spatial location of the data points in the image and $f(\boldsymbol{x})$ the important features of external inputs which are necessary to define the stimulus and $\epsilon(\boldsymbol{x})$ representing those un-important components which are regarded as noise. We consider that $f(\boldsymbol{x})$ is encoded as an activity pattern $\boldsymbol{a}$ in the neural system. The goal of a neural estimator is to infer the value of $\boldsymbol{a}$ based on the noisy input $I(\boldsymbol{x})$. Because of noise, the inferred result $\hat{\boldsymbol{a}}$ is in general different from the true value $\boldsymbol{a}$. Robustness, or sensitivity, of neural coding refers to the discrepancy between $\hat{\boldsymbol{a}}$ and $\boldsymbol{a}$ due to noise. Low robustness, or large sensitivity, implies that neural responses to the same object corrupted with different kinds of noise are dramatically different. Clearly, in order to recognize objects reliably, it is desired that neural coding is robust against noise.

Two issues concerning the robustness of neural codes are explored in the present study. Firstly, we identify what are the key factors influencing the robustness of neural coding, and find that the overlap between neural receptive fields plays a critical role, i.e., the larger the overlap, the more susceptive to noise the neural responses are. This property implies that for achieving robust coding, neural systems should use receptive fields having as small as possible overlap (under the proviso that external objects are adequately encoded).

Secondly, we investigate, under the requirement of robustness, how neuronal receptive fields are shaped in order to encode natural images accurately. It turns out that the obtained results resemble the localized and oriented receptive fields of simple cells in the striate cortex (Hubel and Wiesel 1968; Palmer 1999). This finding is very

S. Li · S. Wu
Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QH, UK

Present Address:
S. Li (✉)
School of Psychology, University of Birmingham, Edgbaston, Birmingham B15-2TT, UK
e-mail: s.li.1@bham.ac.uk

interesting, which provides a novel justification for the receptive field properties of simple cells, different from those in the literature (see, e.g., Olshausen and Field 1996; Bell and Sejnowski 1997; van Hateren and van der Schaaf 1998; Lewicki and Olshausen 1999; Simoncelli and Olshausen 2001; Hurri and Hyvärinen 2003; Vincent and Baddeley 2003). The robust coding algorithm we formulate is quite close to the well-known sparse coding approach (Olshausen and Field 1996). The main difference is that rather than restricting the sparseness of neural activities, we minimize the total variability of neural responses. It turns out that apart from predicting the similar basis functions, our method makes the neural system being more robust to external noise.

The organization of the paper is as follows. In Section ''Statistical inferential sensitivity'', we first study a simple statistical inference model and elucidate the factors which influence the sensitivity of neural responses. In Section ''Robust coding for natural images'', a robust coding scheme which minimizes the sensitivity of neural responses to noise is proposed. This scheme enforces the neural responses not only to encode external inputs well, but also to have small variability. Based on this scheme, we optimize the basis functions for encoding natural images, and compare the results with the receptive fields of simple cells. In Section ''Discussions'', robust coding is compared with other efficient coding schemes, such as sparse coding (Olshausen and Field 1996, 1997; Simoncelli and Olshausen 2001) and temporal coherence (Földiák, 1991; Becker 1993; Stone 1996; Wiskott and Sejnowski 2002; Hurri and Hyvärinen 2003). Its implications on our understanding of neural information processing are discussed. Finally, in Section ''Conclusion'', the overall conclusion of this work is given.

Statistical inferential sensitivity

Our study on the robustness of neural codes is based on a simple model of neural encoding. It assumes that external stimuli $f(\boldsymbol{x})$ is represented as a linear superposition of a set of basis functions (Olshausen and Field 1996; Bell and Sejnowski, 1997)

$$f(\boldsymbol{x}) = \sum_{l=1}^{M} a_l \phi_l(\boldsymbol{x}), \tag{1}$$

where $\phi(\boldsymbol{x}) = \{\phi_l(\boldsymbol{x})\}$, for $l = 1, \ldots, M$, represent the basis functions, and the coefficients $\boldsymbol{a} = \{a_l\}$, for $l = 1, \ldots, M$, the representation of the stimulus $f(\boldsymbol{x})$ in the basis set $\phi(\boldsymbol{x})$. It is known that the basis functions can be associated with neuronal receptive fields, and the variable $\boldsymbol{a}$ to neural activities (Olshausen and Field 1996; Bell and Sejnowski, 1997).

We first study a simple toy model, which allows us to analytically quantify the sensitivity of estimating $\boldsymbol{a}$ due to noise.

A toy model study

Consider an extremely simple case, in which there are only two coding units and the basis functions are fixed to be Gaussian. The noises in external inputs are also assumed to be independent Gaussian, which are written as

$$I(x^i) = f(x^i) + \epsilon^i, \quad \text{for} \quad i = 1, \ldots, N, \tag{2}$$

where x^i represents the ith sampling point of the stimulus, and ϵ^i a Gaussian random number of zero mean and variance σ^2. N is the number of data points sampled from the stimulus.

We consider a simple inference method, called Least Square Error (LSE), whose estimate for $\hat{\boldsymbol{a}}$ is given by

$$\hat{\boldsymbol{a}} = \min \sum_{i=1}^{N} \left[I(x^i) - a_1 \phi_1(x^i) - a_2 \phi_2(x^i) \right]^2, \tag{3}$$

that is, the inferred result is the one that has the minimum reconstruction error for the external inputs.

It can be proved that for the above simple model, when N is sufficiently large, the LSE satisfies a normal distribution (see Appendix A),

$$P(\hat{a}_1, \hat{a}_2) = \frac{1}{Z} \exp\{-\frac{1}{2} \boldsymbol{a}'^T \Omega^{-1} \boldsymbol{a}'\}, \tag{4}$$

where the vector, $\boldsymbol{a}' = (\hat{a}_1 - a_1, \hat{a}_2 - a_2)$, denotes the decoding errors. The matrix Ω is the covariance matrix, whose inverse is given by

$$\Omega^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_i \phi_1(x^i)^2 & \sum_i \phi_1(x^i)\phi_2(x^i) \\ \sum_i \phi_1(x^i)\phi_2(x^i) & \sum_i \phi_2(x^i)^2. \end{pmatrix} \tag{5}$$

It can be checked that Ω is the inverse of the Fisher information (see Appendix A). This implies that the inferential sensitivity of LSE with respect to noise has reached the minimum possible value. This is because, according to the Cramér-Rao bound, the inverse of the Fisher information is the lower bound for decoding errors of un-biased estimators.[1] Thus, our discussion below on the statistical inferential sensitivity is independent of the decoding method used.

[1] Here we only consider un-biased estimators. A biased estimator may achieve lower inferential sensitivity, but it is at the expense of biased estimation.

The sensitivity of decoding can be described by the marginal distribution of $\hat{\boldsymbol{a}}$, which is calculated as (see Appendix A),

$$P(\hat{a}_l) = \frac{1}{\sqrt{2\pi\tau_l^2}} \exp\left[-\frac{(\hat{a}_l - a_l)^2}{2\tau_l^2}\right], \quad \text{for} \quad l = 1, 2, \tag{6}$$

where the variance τ_l^2 is given by (only τ_1 is shown, the result for τ_2^2 is similar)

$$\tau_1^2 = \frac{\sigma^2 \sum_i \phi_2(x^i)^2}{\sum_i \phi_1(x^i)^2 \sum_i \phi_2(x^i)^2 - \left(\sum_i \phi_1(x^i)\phi_2(x^i)\right)^2}. \tag{7}$$

The magnitude of τ_l^2, i.e., the broadness of the marginal distribution, quantifies the sensitivity of decoding. Intuitively, the broader the distribution, the more the estimate can be expected to deviate from the true value. Large sensitivity implies that inferred results for the same stimulus corrupted with different instantiations of noise can be dramatically different.

From Eq. (7), we see that the sensitivity of decoding is determined by several factors, namely, the noise strength (σ^2), the number of data points (N) and the overlap between basis functions.

The overlap between basis functions is measured by (see the denominator of Eq. (7))

$$OP = \frac{\sum_i \phi_1(x^i)\phi_2(x^i)}{\sqrt{\sum_i \phi_1(x^i)^2 \sum_i \phi_2(x^i)^2}}. \tag{8}$$

Figure 1 illustrates that inferential sensitivity increases with the amount of overlap. This effect is intuitively understandable. In the overlapping region, the contributions from the two basis functions (i.e., the two statistical components) on generating the stimulus are mixed. As a result, data points in this region are less informative for 'distinguishing' the activities of the two encoding units. More overlap leads to more ambiguity in inference. Consider the extreme case when the two basis functions are completely overlapping, there exists an infinite number of pairs of $\hat{a}_1$ and $\hat{a}_2$, with $\hat{a}_1 + \hat{a}_2$ a constant, that can equally interpret the stimulus, and the inferential sensitivity is infinitely large.

The moral of the toy model study

The model studied above is quite simple, but the moral it reveals to us is profound, which has at least two important implications:

- First, it tells us that if neural receptive fields are overlapped, then neural responses are inevitably sensitive to noise.

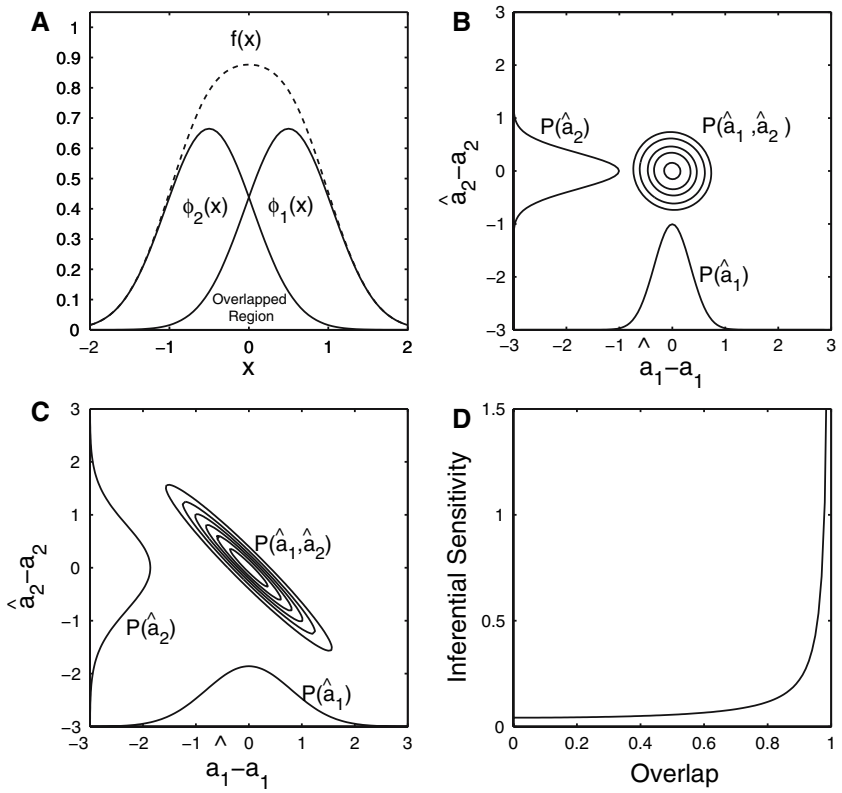

Fig. 1 The results of the toy model study. (**A**) An illustration of the stimulus and two overlapped basis functions; (**B**) In the case of small overlap, the joint and marginal distributions of the estimation are narrow; (**C**) In the case of large overlap, the joint and marginal distributions are broad; (**D**) The inferential sensitivity (measured by τ_1^2) vs. the amount of overlap

- Second, it tells us that if robustness is desired for neural responses, then neural systems should choose those basis functions having as small as possible overlap (under the proviso that objects can be adequately encoded).

These two issues will be further explored in the study below.

Robust coding for natural images

Our hypothesis is that to recognize objects reliably, neural coding should be constructed to be as robust as possible against noise. As shown above, this requires neural receptive fields to have small overlap. To test this idea, we carry out the following experiment. Firstly, we construct a coding scheme, which not only encodes external inputs well, but also suppress the sensitivity of neural responses to noise. Based on this scheme, we optimize the basis functions to encode natural images, and compare the result with the receptive fields of simple cells. This is motivated by that the optimal basis functions for encoding natural images under the linear framework of Eq. (1) and proper constraints tend to resemble the receptive fields of simple cells (Olshausen and Field 1996; Bell and Sejnowski 1997; van Hateren and van der Schaaf 1998; Lewicki and Olshausen 1999; Simoncelli and Olshausen 2001; Hurri and Hyvärinen 2003; Vincent and Baddeley 2003).

The coding scheme

The coding scheme we consider minimizes the following cost function,

$$E = \left\langle \frac{1}{2}[I(\boldsymbol{x}) - \boldsymbol{a}\cdot\phi(\boldsymbol{x})]^2 \right\rangle + \lambda H(\boldsymbol{a}|\phi), \tag{9}$$

where $I(\boldsymbol{x})$ represents a natural image. The term $H(\boldsymbol{a}|\phi)$ is a measure about the variability of neural responses given the basis functions φ. The bracket $\langle\cdot\rangle$ denotes averaging over the set of natural images. The parameter λ controls the balance between the reconstruction error and the variability of neural responses.

The choice of $H(\boldsymbol{a}|\phi)$

To quantify the inferential sensitivity of neural code, ideally, we should set $H(\boldsymbol{a}|\phi)$ to be the summation of the marginal entropy of neural responses when a fixed stimulus is presented, for instance,

$$H(\boldsymbol{a}|\phi) = \sum_{l=1}^{M}\sum_{t} H(a_l|\phi,\boldsymbol{S}_t), \tag{10}$$

where $H(a_l|\phi,\boldsymbol{S}_t)$ is the marginal entropy of the lth neuron's responses given a particular stimulus $\boldsymbol{S}_t$, and the summations are over all neurons and the stimuli to be encoded. But unlike the above toy model (where the stimulus is fixed), here for encoding natural images, we do not know in advance what are the important features and what should be regarded as noise.[2] Thus, the measure Eq. (10) cannot be used.

To overcome this difficulty, our strategy is to choose $H(\boldsymbol{a}|\phi)$ to be the summation of the marginal entropy of neural responses given the set of natural images, rather than a particular important feature, i.e.,

$$H(\boldsymbol{a}|\phi) = \sum_{l=1}^{M} H(a_l|\phi,\boldsymbol{I}), \tag{11}$$

where $H(a_l|\phi,\boldsymbol{I})$ is the marginal entropy of the lth neuron's responses given the set of natural images $\boldsymbol{I}$. $H(\boldsymbol{a}|\phi)$ therefore measures the total variability of neural responses for representing a set of external inputs. Our expectation is that through properly combining the neural response variability and the reconstruction error, the neural system will, on one hand, learn to encode the important features of external inputs, as this is most effective for decreasing the reconstruction error, and on the other hand, become insensitive to those un-important components, as a consequence of restricting the total variability of neural responses (since in this case the variability of neural responses is mainly used for the encoding of important features). Therefore, minimizing the cost function Eq. (9) can indirectly achieve our goal of implementing robust coding. This idea is confirmed by the simulation result in Section ''Sensitivity of robust coding''.

Furthermore, we choose the Renyi's quadratic entropy to quantify $H(a_l|\phi,\boldsymbol{I})$ (Renyi 1976), which is defined as

$$H(a_l|\phi,\boldsymbol{I}) = -\ln\int p(a_l|\phi,\boldsymbol{I})^2 da_l. \tag{12}$$

The good point of this measure is that although we do not know the analytic form of $p(a_l|\phi,\boldsymbol{I})$, the Renyi's quadratic entropy allows us to approximate $H(a_l|\phi,\boldsymbol{I})$ in a data-dependent way, i.e., by using sampled values of a_l when a number of images are presented.

Let us denote $\{a_l^k\}$, for $k = 1,\ldots,K$, as the sampled values of a_l after K natural images are presented and the basis functions are φ. According to the Parzen window

[2] Indeed, this knowledge can be only obtained after the basis functions are optimized. For instance, if the basis functions turn out to be localized and oriented, it tells us that the bar or edge like features are important in the representation of natural images, and other elements are relatively un-important and can be regarded as noise.

approximation (Parzen 1962), the distribution $p(a_l|\phi,\boldsymbol{I})$ can be approximated as

$$p(a_l|\phi,\boldsymbol{I}) \approx \frac{1}{\sqrt{2\pi}dK}\sum_{k=1}^{K}\exp[-(a_l-a_l^k)^2/(2d^2)], \qquad (13)$$

where the Gaussian kernel has been used in the Parzen window approximation with d the width of the kernel.

Substituting Eq. (13) in (12), we obtain (see Appendix B)

$$H(a_l|\phi,\boldsymbol{I}) \approx -\ln\frac{1}{\sqrt{2\pi}dK^2}\sum_{k=1}^{K}\sum_{m=1}^{K}\exp[-(a_l^k-a_l^m)^2/(4d^2)], \qquad (14)$$

which now fully depends on sampled values of a_l.

In practice, sampled values of $\{a_l\}$, for $l = 1,\ldots,M$, are easily obtained. We use the gradient descent method to minimize the function E (for detail, see Appendix B). At each step of updating $\phi(\boldsymbol{x})$, we present $K \gg 1$ training examples, which automatically generates K sampled values for each a_l.

Finally, combining Eqs. (9) and (14), we get the cost function

$$E = \left\langle \frac{1}{2K}\sum_{k=1}^{K}[I^k(\boldsymbol{x}) - \boldsymbol{a}^k\cdot\phi(\boldsymbol{x})]^2 \right\rangle - \lambda\sum_{l=1}^{M}\ln\frac{1}{\sqrt{2\pi}dK^2}\sum_{k=1}^{K}\sum_{m=1}^{K}\exp[-(a_l^k-a_l^m)^2/(4d^2)]. \qquad (15)$$

Here, for clarity, we have written down explicitly the batch of images, denoted as $\{I^k(\boldsymbol{x})\}$, for $k = 1,\ldots,K$, presented at each step of training. The detail of the training procedure is presented in Appendix B.

Simulation results

For natural scenes

We first choose natural scenes, such as leaves, trees, rocks and river etc., as external inputs. They are constructed as follows. First, we select ten, 480 × 480 pixels, pre-whitened natural scenes as the repository of training examples (for examples of natural scenes and their whiten images, see Fig. 2A, B). Then, at each step of training, we randomly choose 100 (i.e., the sampling size $K = 100$) 20 × 20 image patches from the repository as inputs. The neural system consists of 400 neurons. Initially, all basis functions are randomly initialized in the 20 × 20 image space. After training, these basis functions are optimized, as shown in Fig. 2C.

We see that the optimal basis functions are oriented and localized in the space, resembling the properties of receptive fields of simple cells in the striate cortex (Hubel and Wiesel 1968; Palmer 1999). How should we interpret this result in the view of robustness? From Eq. (9), we see that optimal basis functions need to satisfy two conditions, namely, to encode natural scenes well and to maintain small variability in neural responses. First, to encode external inputs well, from the theory of function approximation (Bishop 1996; Schölkopf and Smola 2001), this requires basis functions to 'match' the structures of the external inputs. Second, to maintain small variability in neural responses, this requires basis functions to have a small overlap. It is straightforward to see that the above localized and oriented basis functions best meet these two requirements. Firstly, they have the same shapes as the oriented lines and edges, the salient features in natural scenes. Secondly, they are dispersed in the space and have a relatively small extent of overlap.

For human faces

To validate further our analysis, we use parts of human faces as external inputs. The advantage of these inputs is that they have some new localized features, such as the shapes of eye, nose or mouth, which provides us an opportunity to check whether our coding scheme works consistently.

The face images are taken from the ORL face database (Olivetti Research Laboratory in Cambridge, UK), and they are properly re-scale into a size of 70 × 70 pixels (for examples, see Fig. 3A). Again, we first select 117 face images as the repository, and then, at each step of training, we randomly choose 100, 20 × 20 image patches as training examples. The training results are shown in Fig. 3B. We see that, apart from the shapes of localized lines and edges, some basis functions display curved or round shapes.

For Chinese characters

Essentially, the coding scheme Eq. (9) can also be used as a general method for feature extraction: It selects those important features so that the coding system non only encodes external inputs but also has small entropy. To test this idea, we apply this method to the representation of Chinese characters.

Chinese characters, whose history can be tracked back to thousands of years ago, appear to be very complicated in terms of recognition (see, e.g., examples in Fig. 4). In fact, however, they are composed of some simple pictographs and ideographs in a logical way that can be easily remembered. The most elementary component of the Chinese character is strokes, which further compose radicals or other structured components. A radical may be

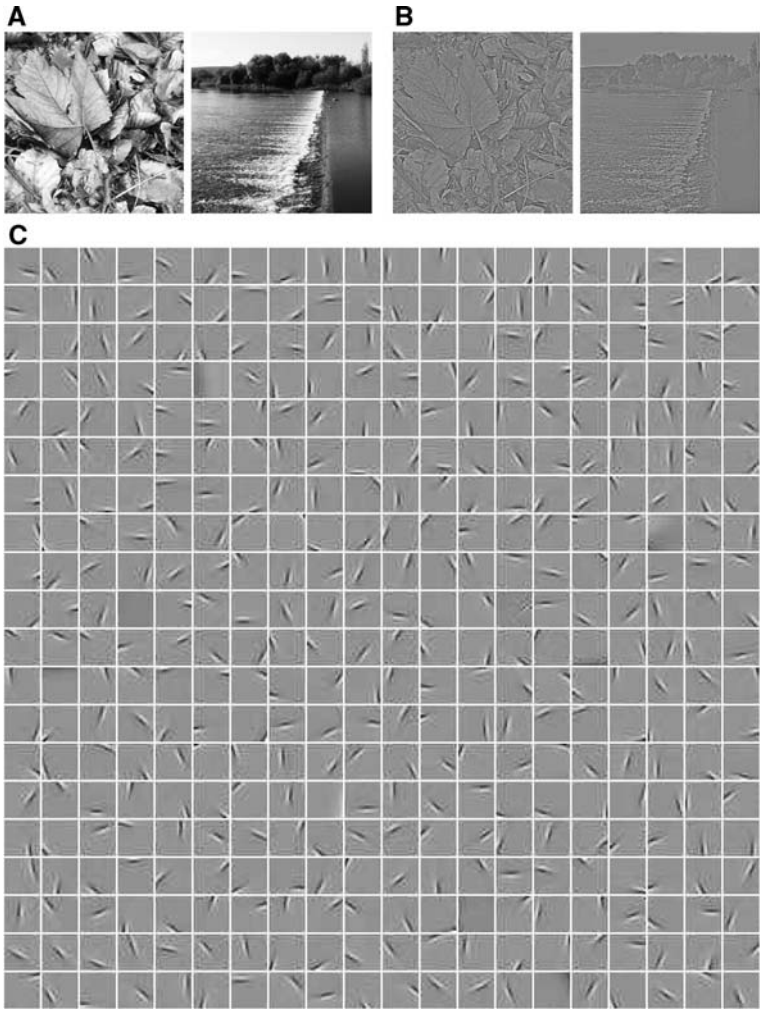

Fig. 2 Simulation on natural images. (**A**) Two examples of natural scenes. (**B**) Whitened images of (**A**). (**C**) The obtained optimal basis functions. The parameters used are $\lambda = 0.13$, $d = 0.3$

shared by characters having the same category meaning. An example of such a Chinese character, which means ''chess'' in Chinese, is illustrated in Fig. 4B. This character is left and right structured. The left part (indicated as part 1 in the figure) is a radical, which is shared by many characters having a meaning related to wood. The right part is actually another character that points out the pronunciation of this character but is meaningless in the structure, being just composed of a few strokes. Many research work including psychophysical and functional Magnetic Resonance Imaging (fMRI) studies have revealed that strokes and radicals play important roles in Chinese character recognition (Hildebrandt and Liu 1993; Peng et al. 2004). They are the salient features for Chinese characters, like lines and edges for natural scenes.

We choose some of the most frequently used Chinese characters as the training dataset, which consists 2,500 words.[3] Each character is presented as a 25×25 pixels image, see Fig. 4A for example. During the training, the number of basis functions was set to 100, and at each step of updating, 200 character images were randomly chosen as inputs.

The training results are shown in Fig. 5, which demonstrate that the optimal basis functions resemble the shapes of strokes, the salient features of Chinese characters, supporting that our method can extract the important features of external inputs.

Discussions

Relationship with other efficient coding schemes

The idea of assuming neural codes are designed to satisfy certain efficient criteria is a valuable top–down method for us to explore the properties of neural information processing (see, e.g., Attneave 1954; Barlow 1961; Laughlin 1981; Atick 1992; Field 1994; Li and Atick 1994; Lewicki and Olshausen 1999). This kind of method is of

[3] We follow the recommendation of *Modern Chinese Language Frequently Used Characters*.

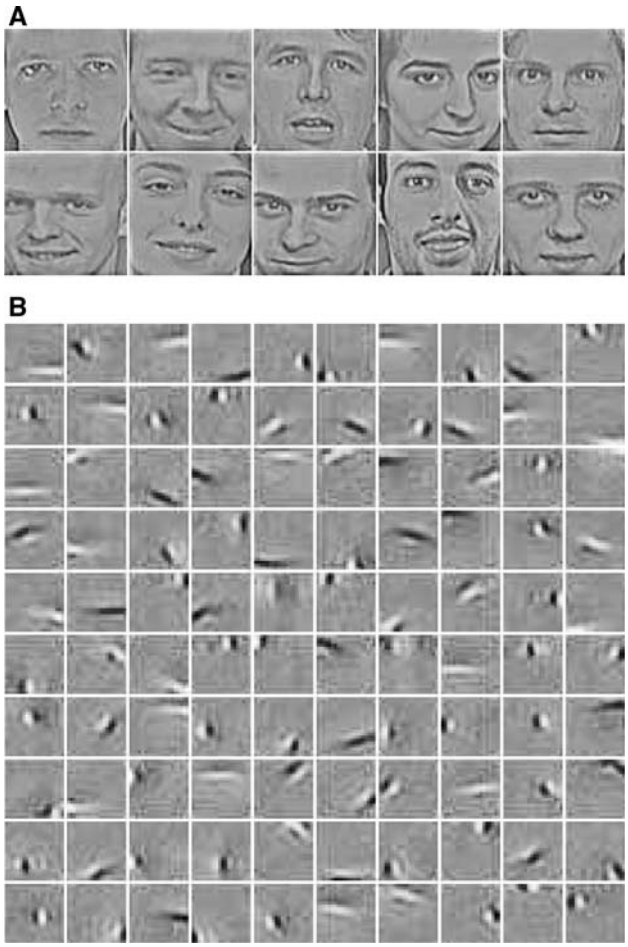

Fig. 3 Simulation on face images. (**A**) Examples of face images from ORL database, the size of the images is 70×70 pixels. (B) The obtained optimal basis functions. The parameters used are $\lambda = 0.13$, $d = 0.3$

particular importance at the current stage since we do not have much knowledge about the detail of neural systems. In this study, we assume that neural codes are constructed to be as robust as possible against noise. Based on this hypothesis, we show that the receptive fields of simple cells may be justified as the consequence of encoding natural images optimally. We should point out that our work is neither the only one nor the first that predicts these simple-cell like basis functions. Indeed, quite a few methods based on different efficiency criteria have successfully achieved this, which particularly include sparse coding (Olshausen and Field 1996, 1997), independent component analysis (ICA) (Bell and Sejnowski 1997; van Hateren and van der Schaaf 1998), temporal coherence (Hurri and Hyvärinen 2003) and energy-efficient coding (Vincent and Baddeley 2003).

Sparse coding was the first approach to make this breakthrough, whose fundamental assumption is that the neural system interprets external stimuli by using very few statistically independent components, and hence neural responses should be sparse. For natural images, these statistically independent components happen to be the oriented lines and edges. ICA, when applied to analyze natural images, has a similar spirit to sparse coding.

Temporal coherence, on the other hand, has the same origin as robust coding, in the sense of that it also assumes the robustness of neural codes. But, unlike the formulation here which restricts the variability of neural responses, temporal coherence imposes the robustness constraint in a different way, which takes into account that in natural environments, temporal sequences of images, such as videos, vary slowly. It is therefore reasonable to assume that the variations between two temporally consecutive

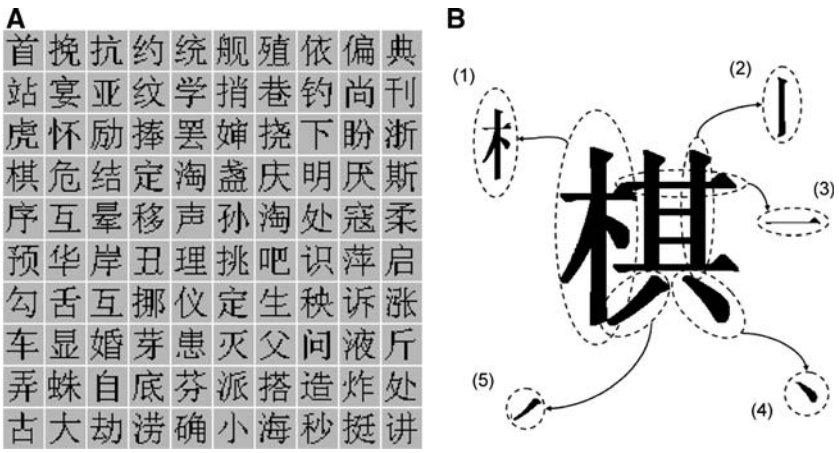

Fig. 4 Chinese characters. (**A**) Examples of Chinese character images, the dataset is the most frequently used 2,500 Chinese characters, the size of the images is 25×25 pixels. (**B**) Decomposition of a typical Chinese character, which means ''chess''. This left and right structured character is composed of a radical (part 1) at the left side and another character at the right side. For the right component, it is composed of four kinds of strokes (parts 2–5)

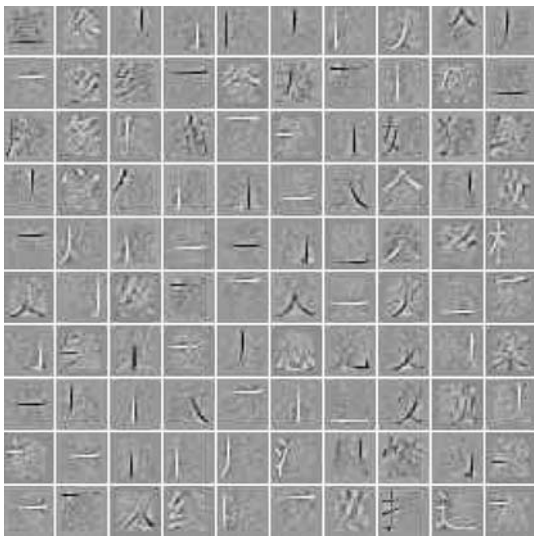

Fig. 5 Simulation results on Chinese character images. The training set is chosen as the most frequently used 2,500 Chinese characters, the size of the images is 25×25 pixels. The obtained optimal basis functions are shown here with the parameter setting $\lambda = 0.13$ and $d = 0.3$

inputs are mainly due to noise, and enforce neural systems to respond least to such changes. Temporal coherence and robust coding can be seen as two complementary approaches demonstrating that robustness is important in the construction of neural coding.

It is worth further clarifying the difference between robust and sparse coding. Although both schemes predict the same, or similar, basis functions for encoding natural images, their natures are different. In the formulation of sparse coding, the cost function is set to be

$$E = \left\langle \frac{1}{2}[I(\boldsymbol{x}) - \boldsymbol{a} \cdot \phi(\boldsymbol{x})]^2 \right\rangle + \lambda \sum_l S(a_l), \tag{16}$$

where $S(a_l)$ is the sparseness constraint, which can be chosen to be, e.g., $S(a_l) = \log(1 + a_l^2)$. $S(a_l)$ takes small value when the neural activity a_l are small. Thus, sparse coding tends to restrict the total activity of neurons (under the proper forms). This is different to robust coding, where the total variability of neurons is restricted (see Eq. (14), robust coding tends to minimize the difference between the activities of individual neurons when different inputs are presented, i.e., it is to minimize $(a_l^k - a_l^m)$ rather than the absolute value of a_l.). This difference will be confirmed by the simulation in Section ''‘‘Sensitivity of robust coding''.

In some sense, the coding scheme Eq. (9) is to implement the minimum entropy code under the constraint that external inputs are adequately encoded (Barlow 1989). Sparse coding can be also interpreted equally, but it evaluates the entropy based on a specific assumption about the distribution of neural responses, that is, this distribution is peaked at zero and has heavy tails (Lewicki and Olshausen 1999). Here, for robust coding Eq. (9), the entropy is evaluated based on the sampled data when a set of images are presented. We should point out that the idea of using the Renyi's formula to estimate entropy in a data-dependent way has been used in other places, for instance, Principle et al. use this strategy to solve the blind source separation problem (Principe et al. 2000). But here, it is the first time this strategy is used to impose a robustness constraint for investigating the encoding of natural image.

All the above coding schemes we discussed focus on exploring how the input statistics determine the coding properties of neurons. In reality, however, the neural coding strategies are also strongly influenced by the tasks the neural systems aim to perform. In a recent work (Salinas 2006), Salinas has shown how the output statistics may shape the neuronal tuning curves. It will be interesting in our future to combine the robustness constraint and the specific task requirements.

We should also note that the coding model we consider is different from the de-noising method in image processing (e.g., Hyvärinen 1999). Here, we do not care about how to remove noise in external inputs (although the scheme can have this effect), but instead how to construct a code so that the responses of coding units are robust against input noise.

Sensitivity of robust coding

To confirm that the coding scheme Eq. (9) indeed suppresses the sensitivity of neural responses to noise, we carry out the following experiment.

The noisy inputs to the neural system are generated as follows (Li and Wu 2005). First, we choose a large-size natural image as a seed, and randomly down-sample this image 100 times with a rate of 25% (this can be done, for instance, by randomly and uniformly choosing 25% data points from the seed image). These down-sampled images are almost visually indistinguishable, and can be regarded as noisy inputs from the same object (for examples of down-sampled images, see Fig. 6A, B). Then, we calculate the neural responses to these noisy inputs, which are obtained by minimizing E with the basis functions fixed to be optimal. The sensitivity of coding is measured by the variance of the neural responses.

For comparison, we also calculate the sensitivity of sparse coding and a scheme using Gaussian functions as the basis. The Gaussian basis functions are chosen to be $G_l(\boldsymbol{x}) = A\exp\left[-||\boldsymbol{x} - \mu_l||^2/2d^2\right]$, for $l = 1, \ldots, M$, with the center μ_l being at each pixel in the image space. To ensure that the comparison is fair, we normalize all basis functions, including the optimal and the Gaussian ones, i.e.,

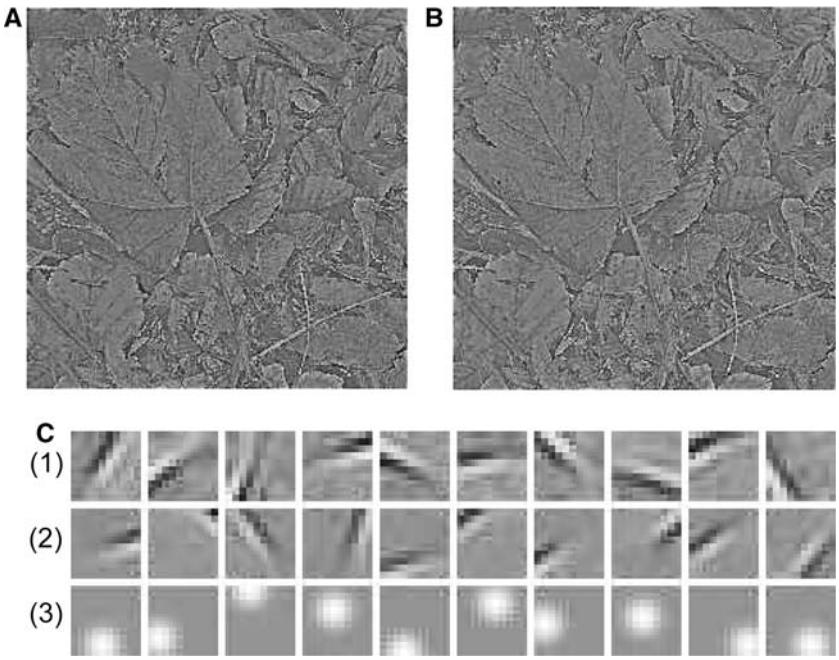

Fig. 6 (**A**) & (**B**) Two down-sampled images at the rate of 25% from the same seed. (**C**) Examples of basis functions in three coding schemes: (1) robust code, (2) sparse code, and (3) Gaussian

$\int \phi_l(\boldsymbol{x})^2 d\boldsymbol{x} = C$, for $l = 1, \ldots, M$ and C a constant. This is to avoid the problem that under the transformations of $\tilde{\phi}(\boldsymbol{x}) = k\phi(\boldsymbol{x})$ and $\tilde{\boldsymbol{a}} = \boldsymbol{a}/k$, the reconstruction error is invariant, but the variance of $\boldsymbol{a}$ decreases $1/k^2$ times. Also, for the coding scheme using Gaussian basis functions, we set the width d of Gaussian functions to be the value, such that it has the same reconstruction error as robust coding. The basis functions used in the three coding schemes are illustrated in Fig. 6C.

Figure 7 shows the distribution of the variation of neuronal responses to their mean values, i.e., $(a_l^k - \langle a_l \rangle)$, for $l = 1, \ldots, M$ and $k = 1, \ldots, K$, where $\langle a_l \rangle$ is the mean activity of the lth neuron. We observe that: (1) The scheme using Gaussian basis functions has the broadest distribution, indicating that in this scheme neural response is most sensitive to noise. This is understandable, since its basis functions are not optimized and have the largest overlap. (2) The neural response variation in robust coding is much more sharply distributed than that in sparse coding, although they have the similar basis functions. This indicates that in robust coding neural responses are much more robust to noise, confirming our analysis. The sensitivities of the three schemes, measured by $\langle (a_l^k - \langle a_l \rangle)^2 \rangle$, are calculated to be 0.0021, 0.025 and 0.076 for robust coding, sparse coding and Gaussian basis functions, respectively.

How robust can neural coding be?

We argue that neural codes are constructed to be robust against noise and also show that the overlap between receptive fields critically determines the sensitivity of neural responses. Then, an interesting question is: how robust can neural codes be? This turns out that the robustness of neural codes depends on the statistical property of external inputs to be encoded. Let us consider first an extreme case where a coding system only needs to encode a single object, then we can easily construct a set of non-overlapping basis functions with the minimum sensitivity to represent the object well, e.g., by simply dividing the image space into N non-overlapping pieces. In practice, however, the task the neural system faces is to use a single set of basis functions to encode a large number of objects with varied structures. In such a case, in order to achieve high encoding accuracy, it is important for the basis functions to match the coherent structures of external inputs, and the overlap between basis functions becomes inevitable.[4] That is, there is in general a trade-off between encoding accuracy and inferential sensitivity. The situation for natural images seems to be a little bit special, where the external inputs happen to have some salient features which are oriented and localized in the space, and this property enables the neural system to encode natural images well at a cost of relatively low statistical inferential sensitivity.

[4] This is similar to the situation of using radial basis function networks for function approximation, in which the overlap between basis functions reflects the smooth structure of inputs (Bishop 1996; Schölkopf and Smola 2001).

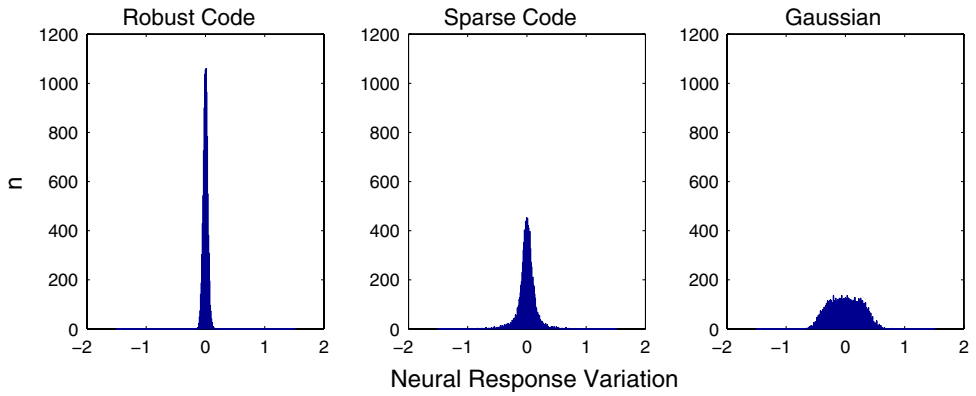

Fig. 7 The distribution of the variation of neuronal responses for three coding schemes. The variation of the lth neuron is calculated to be $\langle (a_l - \langle a_l \rangle)^2 \rangle$, where the symbol $\langle\rangle$ denotes the average over all stimuli. The figure shows the distribution of the variances of all neurons

Conclusion

The present study investigates the robustness of neural codes, an issue of critical importance for our understanding of neural information processing. Based on the view of statistical inference, we show that the overlap between neural receptive fields critically determines the sensitivity of neural responses. We then investigate the optimal basis functions for encoding natural images under the requirement of robust coding, and find that they resemble the receptive fields of simple cells. This provides a new justification for the receptive fields of simple cells. Interestingly, we find that that although both robust and sparse coding predict the same, or similar, optimal basis functions for encoding natural images, the neural response's variabilities the two scheme are different. This may provide a clue for future research to clarify which one is more biologically plausible. Through investigating the encoding of Chinese characters, we also show that the coding scheme we propose may serve as a general method for feature extraction. This issue will be studied in the future work.

Acknowledgements We are very grateful to Peter Dayan. Without his instructive and inspirational discussions, the paper would exist in a rather different form. We also acknowledge valuable comments from Kingsley Sage and Jim Stone.

Appendix A: The Fisher information and the performance of LSE

The Fisher information

Since noise is independent Gaussian, the conditional probability of observing data $I(\boldsymbol{x})$ given $\boldsymbol{a}$ is written as

$$p(I|\boldsymbol{a}) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\{-\frac{1}{2\sigma^2}\sum_i^N [I(x^i) - a_1\phi_1(x^i) - a_2\phi_2(x^i)]^2\}, \quad (17)$$

where N is the number of data points and σ^2 the noise strength.

The Fisher information matrix $\boldsymbol{F}$ is calculated to be

$$F_{mn} = -\int \frac{\partial^2 \ln P(I|\boldsymbol{a})}{\partial a_m \partial a_n} P(I|\boldsymbol{a}) dI, \quad \text{for} \quad m, n = 1, 2, \quad (18)$$

where F_{mn} is the element in the mth row and nth column of $\boldsymbol{F}$.

It is straightforward to check that

$$\boldsymbol{F} = \frac{1}{\sigma^2}\begin{pmatrix} \sum_i \phi_1(x^i)^2 & \sum_i \phi_1(x^i)\phi_2(x^i) \\ \sum_i \phi_1(x^i)\phi_2(x^i) & \sum_i \phi_2(x^i)^2 \end{pmatrix} \quad (19)$$

According to the Cramér-Rao bound, the inverse of the Fisher information defines the lower bound for decoding errors of un-biased estimators. Consider the covariance matrix for estimation errors of an un-biased estimator is given by $\Omega_{mn} = \langle (\hat{a}_m - a_m)(\hat{a}_n - a_n) \rangle$. The Cramér-Rao bound states that $\Omega \geq \boldsymbol{F}^{-1}$, or more exactly, the matrix $(\Omega - \boldsymbol{F}^{-1})$ is semi-positive definite. Intuitively, this means that the inverse of the Fisher information quantifies the minimum inferential sensitivity of un-biased estimators.

The asymptotical performance of LSE

It is straightforward to check that for independent Gaussian noise, LSE is equivalent to maximum likelihood inference, i.e., its solution is obtained through maximizing the log likelihood, $\ln p(I|\boldsymbol{a})$ (comparing Eq. (3) with Eq. (17)). This implies the solution of LSE satisfies the condition,

$$\frac{\partial \ln p(I|\hat{\boldsymbol{a}})}{\partial a_l} = 0, \quad \text{for} \quad l = 1, 2. \quad (20)$$

Consider $\hat{\boldsymbol{a}}$ is sufficiently close to the true value $\boldsymbol{a}$, the above equations can be approximated as (the first-order Taylor expansion at the point $\boldsymbol{a}$)

$$\frac{\partial \ln p(I|\boldsymbol{a})}{\partial a_l}+\frac{\partial^2 \ln p(I|\boldsymbol{a})}{\partial a_l^2}(\hat{a}_l-a_l)+\frac{\partial^2 \ln p(I|\boldsymbol{a})}{\partial a_l \partial a_{m\neq l}}(\hat{a}_m-a_m) \approx 0, \quad \text{for} \quad l,m=1,2. \tag{21}$$

By using Eq. (17), the above equations can be simplified as

$$\frac{\partial \ln p(I|\boldsymbol{a})}{\partial a_l}+\sum_i \frac{\phi_l(x^i)^2}{\sigma^2}(\hat{a}_l-a_l)+\sum_i \frac{\phi_l(x^i)\phi_{m\neq l}(x^i)}{\sigma^2} (\hat{a}_m-a_m) \approx 0, \quad \text{for} \quad l,m=1,2. \tag{22}$$

Since noise are independent Gaussian, we have

$$\frac{\partial \ln p(I|\boldsymbol{a})}{\partial a_l}=\sum_i \frac{\epsilon_i}{\sigma^2}\phi_l(x^i), \quad \text{for} \quad l=1,2, \tag{23}$$

where ϵ_i, for i = 1,...,N, are independent Gaussian random numbers of zero mean and variance σ^2.

Combining Eqs. (22) and (23), we obtain the estimation error of LSE. Here, we only show the result for $\hat{a}_1$ (the case for $\hat{a}_2$ is similar), which is given by,

$$\hat{a}_1-a_1=-\frac{\sum_i \varepsilon_i \phi_1(x^i)\sum_j \phi_2(x^j)^2-\sum_i \varepsilon_i \phi_2(x^i)\sum_j \phi_1(x^j)\phi_2(x^j)}{\sum_i \phi_1(x^i)^2 \sum_j \phi_2(x^j)^2-\left(\sum_i \phi_1(x^i)\phi_2(x^i)\right)^2}. \tag{24}$$

It is easy to check LSE is un-biased, i.e.,

$$\langle(\hat{a}_1-a_1)\rangle=0. \tag{25}$$

The variance of $\hat{a}_1$ is calculated to be

$$\langle(\hat{a}_1-a_1)^2\rangle=\frac{\sigma^2\sum_i \phi_2(x^i)^2}{\sum_i \phi_1(x^i)^2 \sum_i \phi_2(x^i)^2-\left(\sum_i \phi_1(x^i)\phi_2(x^i)\right)^2}. \tag{26}$$

According to the Central Limiting Theorem, when the number of data points N is sufficiently large, the random variable $(\hat{a}_1-a_1)$ will satisfy a normal distribution with the variance given by Eq. (26).

The covariance between the estimation errors of the two components can also be calculated, which is given by

$$\langle(\hat{a}_1-a_1)(\hat{a}_2-a_2)\rangle = \frac{-\sigma^2\sum_i \phi_1(x^i)\phi_2(x^i)}{\sum_i \phi_1(x^i)^2 \sum_i \phi_2(x^i)^2-\left(\sum_i \phi_1(x^i)\phi_2(x^i)\right)^2}. \tag{27}$$

It is straightforward to check that the covariance matrix of estimation errors of LSE, given by $\Omega_{mn}=\langle(\hat{a}_m-a_m)(\hat{a}_n-a_n)\rangle$, for m,n = 1,2, is the inverse of the Fisher information matrix $\boldsymbol{F}$, i.e., $\Omega\boldsymbol{F}=1$. This implies LSE is asymptotically efficient.

Appendix B: Optimizing the basis functions of robust coding

The sensitivity measure $H(\boldsymbol{a})$

We choose the Renyi's quadratic entropy to measure the variability of neural responses when natural images are presented, which is given by

$$H(a|\boldsymbol{I})=-\ln\int p(a|\boldsymbol{I})^2 da. \tag{28}$$

Here for simplicity, we use a to replace a_l, for $l=1,\ldots,M$.

Suppose we have K sampled values of a which are obtained when K natural images are presented, then according to the Parzen window approximation (with the Gaussian kernel), $p(a|\boldsymbol{I})$ can be approximated as

$$p(a|\boldsymbol{I})=\frac{1}{\sqrt{2\pi}dK}\sum_{k=1}^{K} e^{-\frac{(a-a^k)^2}{2d^2}}, \tag{29}$$

where $\{a^k\}$, for k = 1,...,K, represents the sampled values, and d is the width of Gaussian kernel.

Note that

$$\begin{aligned}\int p(a|\boldsymbol{I})^2 da &= \int \tfrac{1}{\sqrt{2\pi}dK}\sum_{k=1}^{K} e^{-\left(a-a^k\right)^2/2d^2} \times \tfrac{1}{\sqrt{2\pi}dK}\sum_{m=1}^{K} e^{-(a-a^m)^2/2d^2} da, \\ &= \tfrac{1}{\sqrt{2\pi}dK^2}\sum_{k=1}^{K}\sum_{m=1}^{K} e^{-\frac{(a^k-a^m)^2}{4d^2}}.\end{aligned} \tag{30}$$

Thus, we have

$$H(a|\boldsymbol{I})=-\ln\left[\frac{1}{\sqrt{2\pi}dK^2}\sum_{k=1}^{K}\sum_{m=1}^{K} e^{-\frac{(a^k-a^m)^2}{4d^2}}\right], \tag{31}$$

which fully depends on the sampled values.

The training procedure

Minimizing Eq. (15) is carried out by using the gradient descent method in two alternative steps, namely, (1) updating $\boldsymbol{a}$ while fixing φ and (2) updating φ while fixing $\boldsymbol{a}$.

(1) Updating $\boldsymbol{a}$

To apply the gradient descent method, the key is to calculate the gradient of E with respect to a_l^k, for $l = 1, \ldots, M$ and k = 1,...,K.

For the first term in Eq. (15), we have

$$\frac{\partial}{\partial a_l^k}\frac{1}{2K}\sum_{k=1}^{K}\left|I^k(\boldsymbol{x}) - \boldsymbol{a}^k\phi(\boldsymbol{x}))\right|^2 = -\frac{1}{K}\left[I^k(\boldsymbol{x}) - \boldsymbol{a}^k\phi(\boldsymbol{x})\right]\phi_l(\boldsymbol{x}). \tag{32}$$

For the second term, we have

$$\frac{\partial}{\partial a_l^k}\lambda H(\boldsymbol{a}) = \lambda\sum_{m=1}^{K} e^{-(a_l^k - a_l^m)^2/4d^2}(a_l^k - a_l^m)\Big/ \left[d^2\sum_{j=1}^{K}\sum_{m=1}^{K} e^{-(a_l^j - a_l^m)^2/4d^2}\right]. \tag{33}$$

Combining Eqs. (32) and (33), we obtain the update rule for $\boldsymbol{a}$:

$$a_l^k(new) = a_l^k(old) + \eta\Delta a_l^k, \quad \text{for} \quad l = 1, \ldots, M, \text{ and } k = 1, \ldots, K, \tag{34}$$

where η is the learning rate, and Δa_l^k is given by

$$\Delta a_l^k = \tfrac{1}{K}\left[I^k(\boldsymbol{x}) - \boldsymbol{a}^k\phi(\boldsymbol{x})\right]\phi_l(\boldsymbol{x}) - \lambda\sum_{m=1}^{K} e^{-(a_l^k - a_l^m)^2/4d^2}(a_l^k - a_l^m)\Big/\left[d^2\sum_{j=1}^{K}\sum_{m=1}^{K} e^{-(a_l^j - a_l^m)^2/4d^2}\right]. \tag{35}$$

(2) Updating φ

Similarly, the update rule for φ is given by

$$\phi_l(\boldsymbol{x})(new) = \phi_l(\boldsymbol{x})(old) + \frac{\eta}{K}\sum_{k=1}^{K} a_l^k\left[I(\boldsymbol{x}) - \boldsymbol{a}^k\phi(\boldsymbol{x})\right], \quad \text{for} \quad l = 1, \ldots, K. \tag{36}$$

References

Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network-Comp Neural 3:213–251

Attneave F (1954) Some informational aspects of visual perception. Psychol Rev 61:183–193

Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA (ed) Sensory communication. MIT Press, Cambridge, MA

Barlow HB (1989) Unsupervised learning. Neural Comput 1:295–311

Becker S (1993) Learning to categorize objects using temporal coherence. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems 5. Morgan Kaufmann, San Mateo, CA

Bell AJ, Sejnowski TJ (1997) The independent components of natural scenes are edge filters. Vision Res 37:3327–3338

Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press

Field DJ (1994) What is the goal of sensory coding? Neural Comput 6:559–601

Földiák P (1991) Learning invariance from transformation sequences. Neural Comput 3:194–200

Hildebrandt TH, Liu WT (1993) Optical recognition of handwritten Chinese characters: Advances since 1980. Pattern Recognit 26:205–225

Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195:215–243

Hurri J, Hyvärinen A (2003) Simple-cell-like receptive fields maximize temporal coherence in natural video. Neural Comput 15:663–691

Hyvärinen A (1999) Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. Neural Comput 11:1739–1768

Laughlin SB (1981) A simple coding procedure enhances a neuron's information capacity. Z Naturforsch C 36:910–912

Lewicki MS, Olshausen BA (1999) Probabilistic framework for the adaptation and comparison of image codes. J Opt Soc Am A 16:1587–1601

Li S, Wu S (2005) On the variability of cortical neural responses: a statistical interpretation. Neurocomputing 65–66:409–414

Li Z, Atick JJ (1994) Toward a theory of the striate cortex. Neural Comput 6:127–146

Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607–609

Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? Vision Res 37:3311–3325

Palmer SE (1999) Vision science: photons to phenomenology. MIT Press, Cambridge, MA

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076

Peng D, Ding G, Perry C, Xu D, Jin Z, Luo Q, Zhang L, Deng Y (2004) fMRI evidence for the automatic phonological activation of briefly presented words. Cognitive Brain Res 20:156–164

Principe JC, Xu D, Fisher JW (2000) Information-theoretic learning. In: Haykin S (ed) Unsupervised adaptive filtering, vol 1: Blind Source Separation. Wiley

Renyi A (1976) Some fundamental questions of information theory. In: Turan P (ed) Selected papers of Alfred Renyi, vol 2. Akademiai Kiado, Budapest

Salinas E (2006) How behavioral constraints may determine optimal sensory representations. PLoS Biol 4(12):e387

Schölkopf B, Smola AJ (2001) Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA

Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24:1193–1216

Stone JV (1996) Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. Neural Comput 8:1463–1492

van Hateren JH, van der Schaaf A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. Proc R Soc Lond B 265:359–366

Vincent BT, Baddeley RJ (2003) Synaptic energy efficiency in retinal processing. Vision Res 43:1283–1290

Wiskott L, Sejnowski TJ (2002) Slow feature analysis: unsupervised learning of invariances. Neural Comput 14:715–770