LETTER Communicated by Odelia Schwartz

Learning Nonlinear Statistical Regularities in Natural Images by Modeling the Outer Product of Image Intensities

Peng Qi
pengrobertqi@163.com
Xiaolin Hu
xlhu@tsinghua.edu.cn
State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

It is well known that there exist nonlinear statistical regularities in natural images. Existing approaches for capturing such regularities always model the image intensities by assuming a parameterized distribution for the intensities and learn the parameters. In the letter, we propose to model the outer product of image intensities by assuming a gaussian distribution for it. A two-layer structure is presented, where the first layer is nonlinear and the second layer is linear. Trained on natural images, the first-layer bases resemble the receptive fields of simple cells in the primary visual cortex (V1), while the second-layer units exhibit some properties of the complex cells in V1, including phase invariance and masking effect. The model can be seen as an approximation of the covariance model proposed in Karklin and Lewicki (2009) but has more robust and efficient learning algorithms.

1 Introduction

There have been many computational models for the simple cells and complex cells in the primary visual cortex (V1). Simple cells are usually characterized by Gabor functions for their sensitivity to edges, and complex cells are known to perform a pooling operation among simple cells to achieve phase invariance while remaining selective to orientations.

Simple cell properties have long been studied by modeling pixel intensities of natural images. The output of these models, which are endorsed by information theory and biological economic considerations, are usually constrained to be sparse or independent to each other. Typical models include independent component analysis (ICA) (van Hateren & van der Schaaf, 1998), sparse coding (Olshausen & Field, 1996), sparse restricted Boltzmann machines (RBMs) (Lee, Ekanadham, & Ng, 2008), and K-means clustering (Coates, Lee, & Ng, 2011). The main idea of these models is to reconstruct image intensities with sparsity or independence regularization.

Neural Computation **26, 693–711** (2014)
doi:10.1162/NECO_a_00567

All of these models can produce Gabor-like filters when trained on natural images, which localize in both spatial and frequency domains.

These models cannot capture nonlinear statistical regularities in natural images. In fact, there exists residual dependence among the learned basis vectors (Schwartz & Simoncelli, 2001). This observation has motivated extensive studies for modeling complex cells. The past decade has witnessed great endeavors in this direction. Hyvärinen, Hoyer, and Inki (2001) introduced topographic nonlinear correlation of latent variables to account for residual dependence among the ICA bases. The model is known as topographic ICA (TICA). Other variants of ICA, such as independent subspace analysis (Hyvärinen & Hoyer, 2000) and the two-layer models in Karklin and Lewicki (2005) and Köster and Hyvärinen (2010) have produced complex cell properties like orientation sensitivity and phase invariance. In these models, the first layer is linear, and the second layer is nonlinear.

Instead of modeling image intensities directly like sparse coding (Olshausen & Field, 1996), Karklin and Lewicki (2009) proposed modeling the covariance of image intensities. The model accounted for not only phase invariance but also complicated properties observed in physiological data, such as surround suppression and masking effect. Ranzato and Hinton (2010) proposed modeling image covariance with factorized Boltzmann machines. They also made an effort to unify mean and covariance models in a single energy function. More recently, Coates et al. (2011) incorporated spike and slab prior into Boltzmann machines, and the resulting model unified the heterogeneous hidden units in covariance RBMs (cRBM) (Ranzato & Hinton, 2010). A hallmark of these covariance models is that nonlinearity is present in the first layer.

Our work is largely inspired by Karklin and Lewicki (2009). We propose modeling the outer product of image intensities, a quantity closely related to image covariance. The main idea is to assume a gaussian distribution with parameterized mean for the outer product of image intensities and then learn the parameters on natural images. The proposed model has two layers, with nonlinearity embedded in the first layer. The overall formulation enjoys a simple formulation similar to sparse coding (Olshausen and Field, 1996; Lee, Battle, Raina, & Ng, 2007) and can be seen as an approximation of the model proposed by Karklin and Lewicki (2009). We will show that our model, similar to Karklin and Lewicki's model, is also able to capture some properties of V1 simple cells and complex cells, but in a more efficient, stable, and biologically plausible way.

The rest of the letter is organized as follows. In section 2, the new model is presented, which follows the discussion of connections to existing models. The learning and inference algorithms are presented in section 3. Experimental results are presented in section 4, and discussions are presented in section 5.

2 The Model

Throughout the letter, it is assumed that the mean of image patches $\boldsymbol{x}$ is zero, which can be easily achieved by subtracting the mean of samples. The goal is to model the outer product of image patches $\boldsymbol{x}\boldsymbol{x}^T$. This amounts to model the high-order interactions among the nonlinear relationship between all pairs of components, that is, $x_i x_j$. In contrast, a sparse coding approach (Olshausen & Field, 1996) models the high-order interactions among all components x_i, which is a linear quantity.

Before presenting the models, we introduce some notations. Let $\|\cdot\|_{\mathcal{F}}$ denote the Frobenius matrix norm and $\|\cdot\|_2$ denote the L_2 vector norm. vec(A) stands for a vector by stacking the columns of the matrix A from left to right. "diag" denotes an operator for transforming a vector to the diagonal matrix with the vector elements on its diagonal, or extracting the diagonal vector of a square matrix, which can be determined from the context.

It is assumed that $\boldsymbol{x}\boldsymbol{x}^T$ follows a nonzero-mean gaussian distribution with a constant covariance matrix. Learning and inference operate on the parameterized mean, similar to sparse coding (Olshausen & Field, 1996). First, the mean is parameterized as the weighted sum of the outer products of a set of bases $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K\}$, so

$$p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)|\boldsymbol{u}, \boldsymbol{B}) = \mathcal{N}\left(\mathrm{vec}\left(\sum_k u_k \boldsymbol{b}_k \boldsymbol{b}_k^T\right), \frac{1}{\gamma}\boldsymbol{I}\right),$$

where $\boldsymbol{B}$ is the collective matrix formed by $\boldsymbol{b}_k$'s, $\boldsymbol{u}$ is the latent variable, $\boldsymbol{I}$ is the identity matrix, and γ is a constant. Without loss of generality, we have assumed that the covariance matrix of the gaussian distribution is isotropic.

This model is extended to a two-layer structure by introducing a second layer of hidden units $\boldsymbol{y}$ and connection weights $\boldsymbol{W}$ between $\boldsymbol{u}$ and $\boldsymbol{y}$ as

$$u_k = \sum_j y_j w_{kj}.$$

We have

$$p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)|\boldsymbol{y}, \boldsymbol{B}, \boldsymbol{W}) = \mathcal{N}\left(\mathrm{vec}\left(\sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T\right), \frac{1}{\gamma}\boldsymbol{I}\right).$$

The latent variable (or model response) $\boldsymbol{y}$ is regularized by the Laplacian distribution,

$$p(\boldsymbol{y}) = \mathcal{L}(0, 1).$$

Then the joint distribution is

$$p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{y}|\boldsymbol{B}, \boldsymbol{W}) \propto \exp\left(-\frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^T - \sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T\right\|_{\mathcal{F}}^2 - \gamma \sum_j |y_j|\right).$$

The goal is to maximize the likelihood of the data, that is, $p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)|\boldsymbol{B}, \boldsymbol{W})$. Note that

$$p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)|\boldsymbol{B}, \boldsymbol{W}) = \sum_{\boldsymbol{y}} p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{y}|\boldsymbol{B}, \boldsymbol{W}),$$

but this quantity is intractable. We approximate it with

$$\max_{\boldsymbol{y}} p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{y}|\boldsymbol{B}, \boldsymbol{W}).$$

Then the model becomes

$$\max_{\boldsymbol{B},\boldsymbol{W}} \max_{\boldsymbol{y}} \log p(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{y}|\boldsymbol{B}, \boldsymbol{W})$$

or, equivalently,

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{W},\boldsymbol{y}} \quad & \frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\mathrm{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T\right\|_{\mathcal{F}}^2 + \gamma \sum_j y_j \\ \text{subject to} \quad & \left\|\boldsymbol{b}_k\right\|_2^2 \le 1, \left\|\boldsymbol{w}_j\right\|_2^2 \le 1, \ w_{kj} \ge 0, \ y_j \ge 0 \\ & \forall j \in \{1, 2, \ldots, J\}, k \in \{1, 2, \ldots, K\}. \end{aligned} \tag{2.1}$$

The first two constraints prevent the bases from growing unboundedly by noticing that $\boldsymbol{B}, \boldsymbol{W}$, and $\boldsymbol{y}$ are coupled. The last two constraints ensure that $\boldsymbol{B}\,\mathrm{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T$ is positive semidefinite because $\boldsymbol{x}\boldsymbol{x}^T$ is so.

The proposed model has a hierarchical structure with the first-layer units u_k and the second-layer units y_j. However, u_k does not appear in formulation 2.1 explicitly, and y_j connects the first-layer bases $\boldsymbol{b}_k$ through weight w_{kj} directly.

2.1 Connections to Existing Models. Model 2.1 enjoys a similar form with the sparse coding model (Lee et al., 2007)

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{y}} \quad & \left\|\boldsymbol{x} - \boldsymbol{B}\boldsymbol{y}\right\|_2^2 + \gamma \sum_k |y_k|, \\ \text{subject to} \quad & \left\|\boldsymbol{b}_k\right\|_2^2 \le 1, \ y_k \ge 0, \forall k \in \{1, 2, \ldots, K\}. \end{aligned} \tag{2.2}$$

The essential difference is that equation 2.2 models a linear quantity $\boldsymbol{x}$, while equation 2.1 models a nonlinear quantity $\boldsymbol{x}\boldsymbol{x}^T$.

Unlike the sparse coding model, which assumes that $p(\boldsymbol{x}|\boldsymbol{y})$ is a gaussian distribution with parameterized mean $\boldsymbol{B}\boldsymbol{y}$, Karklin and Lewicki (2009) introduce a framework for learning the covariance statistics of natural images, which assumes that $p(\boldsymbol{x}|\boldsymbol{y})$ is a zero-mean gaussian distribution with parameterized covariance matrix,

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{B}) = \mathcal{N}(0, \boldsymbol{C}), \tag{2.3}$$

where

$$\boldsymbol{C} = \exp\left(\sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T\right)$$

and $\exp(\cdot)$ stands for matrix exponential ($\exp(\boldsymbol{A}) = \sum_{i=0}^{\infty} \boldsymbol{A}^i$). With a sparse prior distribution on $\boldsymbol{y}$, the first-layer bases $\boldsymbol{B}$ and second-layer bases (or pooling matrix) $\boldsymbol{W}$ can be learned.

With the distribution in equation 2.3, the sample covariance matrix of $\boldsymbol{x}$ given $\boldsymbol{y}$ is

$$\boldsymbol{Q} = \frac{1}{N-1} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T,$$

which approximates $\boldsymbol{C}$. Note that the matrix exponential in the expression of $\boldsymbol{C}$ is used to ensure the positive definiteness of the covariance matrix. At this moment, we assume that this operator is absent and the positive definiteness is ensured by some other techniques; then $\boldsymbol{Q}$ approximates $\sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T$. If we neglect the denominator in $\boldsymbol{Q}$ and let $N = 1$, then

$$\boldsymbol{x}\boldsymbol{x}^T \approx \sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T.$$

This is actually what we want to achieve with the proposed model in equation 2.1 (see the first term of the objective function of that equation). In this sense, our model can be regarded as an approximation of the covariance model in Karklin and Lewicki (2009).

Karklin and Lewicki's model suffers numerical difficulties introduced by the matrix exponential operation in the covariance matrix $\boldsymbol{C}$. Our model does not involve matrix exponentials, which makes it possible to devise efficient training algorithms.

3 Inference and Learning

In model 2.1, inference of the hidden variable $\boldsymbol{y}$ given the bases $\boldsymbol{B}$ and $\boldsymbol{W}$ amounts to solving a (convex) quadratic programming problem:

$$\begin{aligned} &\min_{\boldsymbol{y}} \ f_1 = \frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\,\mathrm{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^{\mathrm{T}}\right\|_{\mathcal{F}}^2 + \gamma \sum_{\mathrm{j}} \mathrm{y_j} \\ &\text{subject to} \quad y_j \geq 0, \forall j \in \{1, 2, \ldots, J\}, \end{aligned} \tag{3.1}$$

It can be shown (see appendix A) that this problem is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{y}} \ f_1' = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y} \\ &\text{subject to} \quad y_j \geq 0, \forall j \in \{1, 2, \ldots, J\}, \end{aligned} \tag{3.2}$$

where

$$\boldsymbol{A} = \boldsymbol{W}^T\big((\boldsymbol{B}^T\boldsymbol{B}) \circ (\boldsymbol{B}^T\boldsymbol{B})\big)\boldsymbol{W}, \quad \boldsymbol{b} = -\boldsymbol{W}^T\big((\boldsymbol{B}^T\boldsymbol{x}) \circ (\boldsymbol{B}^T\boldsymbol{x})\big) + \gamma 1. \tag{3.3}$$

In these equations, $\circ$ denotes the Hadamard (element-wise) matrix product and 1 denotes a vector with all 1s. This is a standard quadratic programming problem, and many algorithms, such as the conjugate gradient descent algorithm, are available. We propose modifying the feature-sign algorithm (Lee et al., 2007) for solving it, which was proved to be faster than the conjugate gradient descent algorithm in our experiments (data not shown). (See algorithm 1.) The convergence of the algorithm is stated in theorem 1. The proof is sketched in appendix B.

Theorem 1. *The modified feature-sign algorithm converges to the solution of equation 3.2 within a finite number of iterations.*

This algorithm is suitable for parallelization with minibatches of data. The cost function for a minibatch of M data samples is

$$f = \frac{1}{M}\sum_{i=1}^{M}\left(\left\|\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)^T} - \boldsymbol{B}\mathrm{diag}(\boldsymbol{W}\boldsymbol{y}^{(\mathrm{i})})\boldsymbol{B}^{\mathrm{T}}\right\|_{\mathcal{F}}^2 + \gamma \sum_{\mathrm{j}} \mathrm{y}_{\mathrm{j}}^{(\mathrm{i})}\right),$$

where the inference for different samples can be parallelized on multiple cores.

Algorithm 1: Modified Feature-Sign Algorithm.

1 Initialize $\boldsymbol{y} := \mathbf{0}$, $\boldsymbol{\theta} := \mathbf{0}$, and $active\ set := \{\}$, where $\theta_i \in \{0, 1\}$ denotes sign(y_i).

2 From zero coefficients of $\boldsymbol{y}$, select $i = \arg\max_i -\frac{\partial(1/2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y})}{\partial y_i}$.
If $-\frac{\partial(1/2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y})}{\partial y_i} > 0$, then set $\theta_i := 1$, $active\ set := \{i\} \cup active\ set$.

3 Feature-sign step:

Let $\hat{\boldsymbol{A}}$ denote a submatrix of $\boldsymbol{A}$ that contains only the rows and columns corresponding to the $active\ set$.
Let $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{\theta}}$ be subvectors of $\boldsymbol{b}$, $\boldsymbol{y}$, and $\boldsymbol{\theta}$ corresponding to the $active\ set$.
Compute the analytical solution to the resulting unconstrained QP ($\min\ \hat{\boldsymbol{y}}\frac{1}{2}\hat{\boldsymbol{y}}^T\hat{\boldsymbol{A}}\hat{\boldsymbol{y}} + \hat{\boldsymbol{b}}^T\hat{\boldsymbol{y}}$):

$$\hat{\boldsymbol{y}}_{new} := -\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{b}}.$$

.
Perform a line search on the closed line segment from $\hat{\boldsymbol{y}}$ to $\hat{\boldsymbol{y}}_{new}$:
Check the objective value at $\hat{\boldsymbol{y}}_{new}$ and all points where any coefficient changes sign (while the rest remain nonnegative).
Update $\hat{\boldsymbol{y}}$ (and the corresponding entries in $\boldsymbol{y}$) to the point with the lowest objective value.
Remove zero coefficients of $\hat{\boldsymbol{y}}$ from the $active\ set$ and update $\theta := \text{sign}(\boldsymbol{y})$.

4 Check optimality:

(a) Optimality condition for nonzero coefficients:
$\frac{\partial(1/2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y})}{\partial y_i} = 0, \forall y_j \neq 0.$
If condition (a) is not satisfied, goto step 3 (without any new activation); else check condition b.

(b) Optimality condition for zero coefficients:
$-\frac{\partial(1/2\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y})}{\partial y_i} \leq 0, \forall y_j = 0.$
If condition b is not satisfied, goto step 2; else return $\boldsymbol{y}$ as the solution.

Learning the parameters $\boldsymbol{B}$ and $\boldsymbol{W}$ in model 2.1 given $\boldsymbol{y}$ amounts to solving

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{W}}\ & f_2 = \frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\,\text{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^{\mathrm{T}}\right\|_{\mathcal{F}}^2 \\ \text{subject to}\ & \left\|\boldsymbol{b}_k\right\|_2^2 \leq 1,\ \left\|\boldsymbol{w}_j\right\|_2^2 \leq 1,\ w_{kj} \geq 0 \\ & \forall j \in \{1, 2, \ldots, J\}, k \in \{1, 2, \ldots, K\}, \end{aligned}$$

which is a nonconvex optimization problem. The projected gradient descent method can be used for solving it. The parameters are first updated along

the negative gradient direction with a fixed step size, then projected to the constrained space, that is, all negative components of $\boldsymbol{W}$ are set to zero, and each column of $\boldsymbol{W}$ and $\boldsymbol{B}$ is normalized to unit length.

Since $\boldsymbol{W}$ and $\boldsymbol{B}$ are coupled, a layer-wise scheme can be adopted. First, set the second-layer bases $\boldsymbol{W}$ to the identity matrix and learn the first-layer bases $\boldsymbol{B}$. Then the model becomes a single-layer model in essence (see equation 2.1). Repeat the following two steps alternately until some stopping criterion is met. First, update $\boldsymbol{B}$ with

$$\frac{\partial f_2}{\partial \boldsymbol{B}} = (\boldsymbol{B}\,\mathrm{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\boldsymbol{B}\,\mathrm{diag}(\boldsymbol{W}\boldsymbol{y}), \tag{3.4}$$

and infer $\boldsymbol{y}$ by using algorithm 1. Second, fix $\boldsymbol{B}$ and learn $\boldsymbol{W}$. Similarly, update $\boldsymbol{W}$ with

$$\frac{\partial f_2}{\partial \boldsymbol{W}} = \boldsymbol{B}^{T}(\boldsymbol{B}\,\mathrm{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^{\mathrm{T}} - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\boldsymbol{B}\boldsymbol{y}, \tag{3.5}$$

and infer $\boldsymbol{y}$ by using algorithm 1 until some stopping criterion is met.

Two alternative approaches for learning the bases would be updating $\boldsymbol{B}$ and $\boldsymbol{W}$ simultaneously from the values learned in the first phase or from random initialization of both. In our experiments, these approaches led to qualitatively similar results but took longer.

4 Experiments

The proposed model was trained on the Kyoto Natural Images data set (Doi, Inui, Lee, Wachtler, & Sejnowski, 2003). All images were transformed to gray scale. A large number of random 20×20 image patches were sampled from these images. After removing the mean pixel intensity, we whitened the patches using principal component analysis (PCA) such that

$$\left(\sum_m \left\|\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(m)}_{\text{whitened}}\right\|_{\mathcal{F}}^2\right) \Big/ \left(\sum_m \left\|\boldsymbol{x}^{(m)}\right\|_{\mathcal{F}}^2\right) \leq 0.01,$$

that is, the lost variance after whitening was at most 1%. After whitening, 265 principal components were retained. All results reported in this letter correspond to data transformed back to the original image space. One thousand first-layer units and 100 second-layer units were used in the experiments.

The parameters of the layer-wise approach for learning the bases are as follows. First, $\boldsymbol{W}$ was fixed to the identity matrix $\boldsymbol{I}$, and $\boldsymbol{B}$ was updated starting from random values for 5000 iterations. During the first 4000 iterations, the learning rate was fixed to 0.1, and at every iteration, 200 bases were randomly selected for updating. During the last 1000 iterations, all

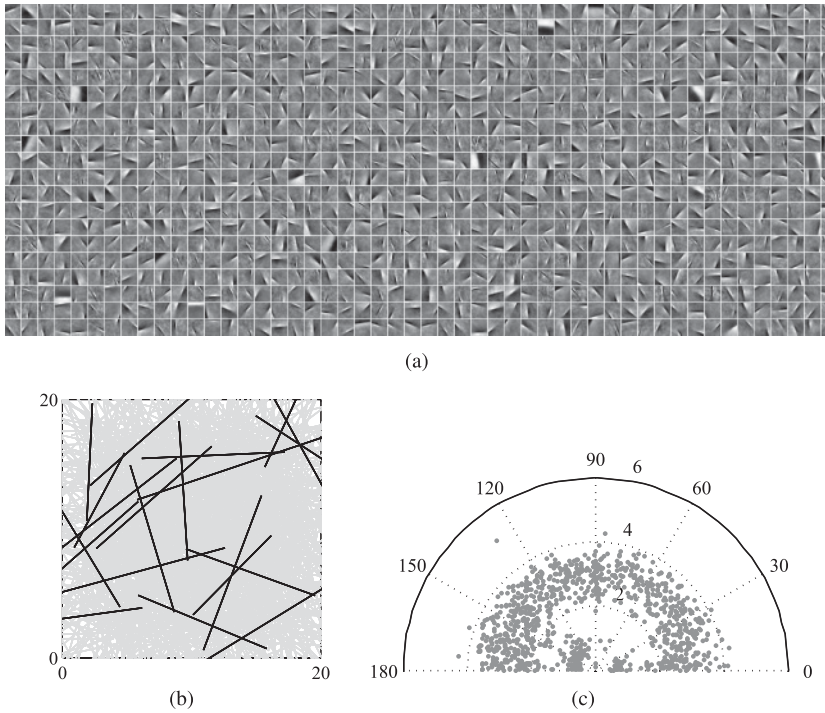

Figure 1: First-layer bases learned on the Kyoto Natural Images data set. (a) Visualization of the 1000 first-layer bases. (b) Fitted Gabor functions for the bases visualized as line segments in a 20×20 patch. Twenty random bases are highlighted in black. (c) Polar visualization of fitted Gabor functions as dots. The angles correspond to the peak spatial frequency orientation, and the radii correspond to log-transformed peak frequency.

bases were updated together with a learning rate of 0.03. We found that this trick made the convergence faster than updating all bases from the very beginning. Then $\boldsymbol{B}$ was fixed and $\boldsymbol{W}$ was initialized to random values and updated for 1000 iterations with a fixed learning rate 0.03. The model was trained with minibatches of 1000 samples. With parallelism, processing each minibatch (including both inference and bases updating) in both phases took an average of 0.7 seconds on an 8-core work station running Linux.

4.1 First-Layer Results. The learned first-layer bases were edge detectors with different scales, positions, and orientations, as plotted in Figure 1a. We fitted parametric Gabor functions to the bases and visualized the functions with line segments in Figure 1b. Each line segment depicts the length

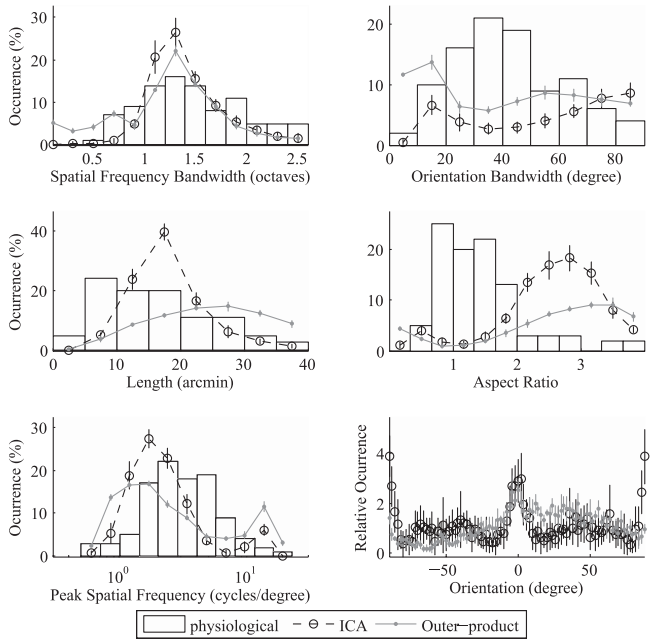

Figure 2: Quantitative comparison between the first-layer bases, ICA filters, and experimental data from macaque's primary visual cortex (De Valois, Albrecht, & Thorell, 1982).

of the gaussian envelope of the Gabor function by its length, the orientation of the Gabor function along the low-pass direction by its direction, and the position of the Gabor function by its position. Figure 1c provides an alternative view of the profile of the bases. For each basis, its peak spatial frequency value (in negative logarithmic scale) was plotted against its orientation. The bases densely covered the spatial-frequency plane.

We compared the spatial-frequency properties of the bases learned by the model with the receptive fields of the V1 simple cells (see Figure 2). As a reference, we also plotted the results of ICA, which were obtained by using the companion codes of Hyvärinen, Hurri, and Hoyer (2009). The following quantities were compared:

1. Spatial frequency bandwidth—the full width at half maximum (FWHM) of each filter along the orientation of the peak in the amplitude spectrum.
2. Orientation bandwidth—the FWHM along a circle through the peak in the amplitude spectrum, centered at zero spatial frequency.
3. Peak spatial frequency and peak orientation—spatial frequency and orientation of the peak in the amplitude spectrum.

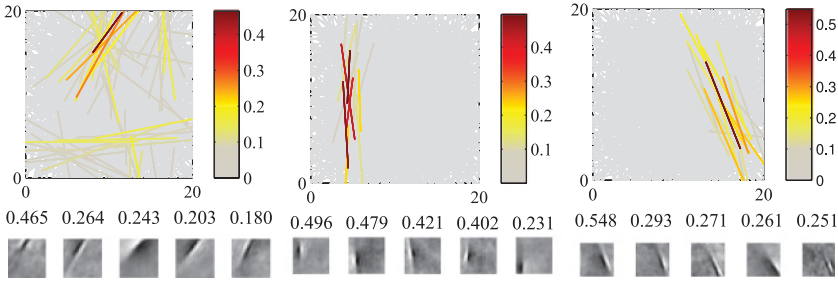

Figure 3: Visualization of three second-layer units with their connection weights to the 1000 first-layer bases. Each column illustrates a second-layer unit. The top insets show the weights to all of the first-layer bases, where each basis is represented by a line segment shaded by the weights. The bottom insets show the first-layer bases corresponding to the five strongest weights connecting to these units. The numbers above indicate the weight values.

4. Length and aspect ratio of the bases or receptive fields—the length was the FWHM of the frequency envelope along the orientation into which the filter was low pass, and the width was the FWHM along the orientation into which the filter was bandpass. The aspect ratio was the ratio of length and width.

The learned bases had a similar distribution to the receptive fields of V1 simple cells in the spatial frequency bandwidth (see the top left of Figure 2) and peak spatial frequency (bottom left). However, the bases tended to be tuned to a narrower orientation bandwidth, which made it more sensitive to orientation changes (see the top right of Figure 2). In addition, they had a longer shape and thus a higher aspect ratio (see the middle right of Figure 2). Such deviations from physiological data can also be observed in the results of ICA (see Figure 2 and van Hateren & van der Schaaf, 1998). The data source, remaining degrees of freedom (number of retained dimensions of PCA whitening), patch size, and number of hidden units could all contribute to these deviations. In addition, the distributions of orientations of the bases obtained by the proposed model and ICA were similar (see the bottom right of Figure 2). But in this case, the physiological data were not available.

4.2 Second-Layer Results. The 1000 fitted Gabor filters were plotted as line segments in a single 20 $\times$ 20 image patch (see Figure 1b). Each second-layer unit was visualized by shading the line segments according to the connection weights between this unit and the first-layer bases. Figure 3 provides three examples. The second-layer units were selective to a range of bars within the patch, which shared similar orientations but located at

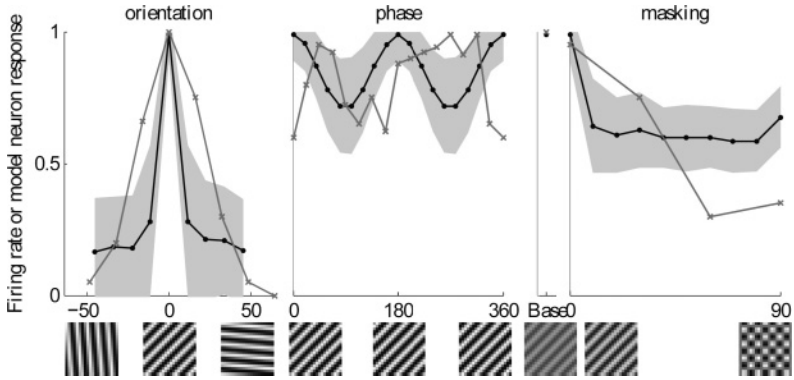

Figure 4: Response properties of the second-layer units are similar to physiological data from V1 complex cells. For each unit, the most responsive input grating was first determined; then its angle (left) and phase (middle) were varied. In addition, another grating with different orientations was imposed on the most responsive grating (right). The curves with filled circles depict the mean response of the 100 second-layer units, and the gray areas illustrate the standard deviation. The curves with "x" marks from left to right depict the firing rates of V1 neurons extracted from Jones, Wang, and Sillito (2002), Movshon, Thompson, and Tolhurst (1978), and Bonds (1989), respectively, which were normalized by dividing the maximum values over the original curves.

different locations, in accordance with the phase invariance property of complex cells in V1.

We then quantitatively compared the properties of the second-layer units with some physiological results of the V1 complex cells. Given the most responsive grating input to a model neuron, we varied its orientation or phase, or superimposed another grating with different orientations, and recorded the responses, which were then normalized to have a maximum value 1. Figure 4 shows the mean response of all second-layer units (filled circles), which is in agreement with physiological data ("x" mark).

It was found that the generalization ability of the model was better than the sparse coding model for describing the nonstationary statistics in natural images. As did Karklin and Lewicki (2009), we projected the responses of the model induced by natural image patches into lower-dimensional space to visualize the ability of the model to distinguish different regions. One thousand random patches of size 20×20 were extracted from each of the four regions on a natural image: waterfall, tree leaves, water ripples, and horizontal wave (see Figures 5a and 5b). We projected the raw pixels (see Figure 5c), the responses of the sparse coding model (see Figure 5d), and the responses of the proposed model (see Figure 5e) into 2D space with linear discriminative analysis (Cai, He, & Han, 2008). The responses of the proposed

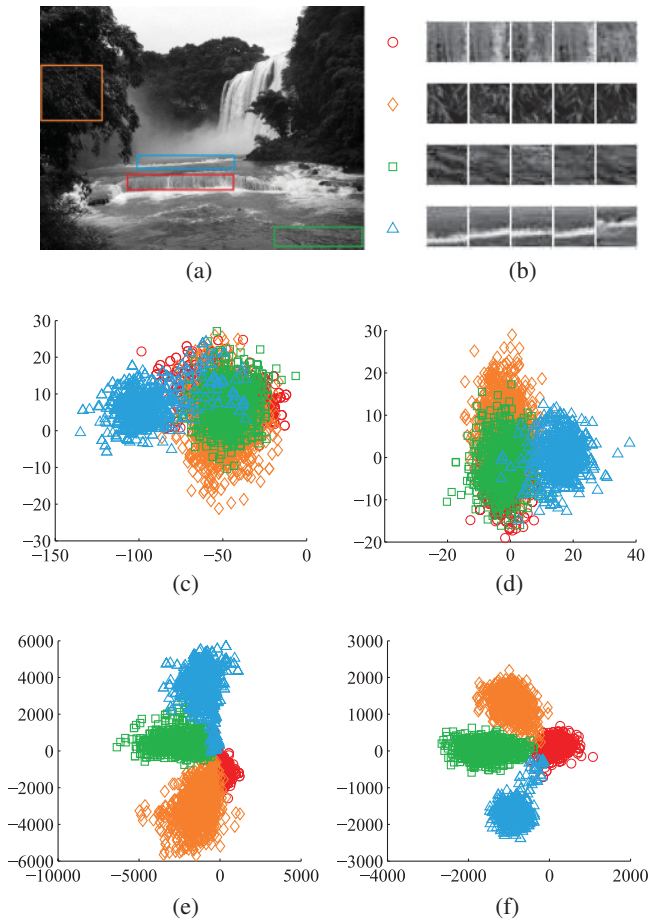

Figure 5: Two-dimensional projection of image patches and model responses. (a) A picture of a natural scene with four selected characteristic regions. (b) Five random image patches from each selected region. (c) Two-dimensional projection of raw pixel intensities. (d) Two-dimensional projection of sparse coding responses. (e) Two-dimensional projection of the two-layer model responses. (f) Two-dimensional projection of the reduced model responses.

model exhibited a clustering effect with respect to the four regions. Sparse coding, though sharing a similar model structure, failed to capture such statistical regularities. The sparse coding results were obtained by using the companion codes of Lee et al. (2007).

Further investigations revealed that this property was introduced not by the hierarchical structure but by the nonlinearity in the first layer. When we set $W = I$, the model degenerated to a single-layer model. The 2D projection

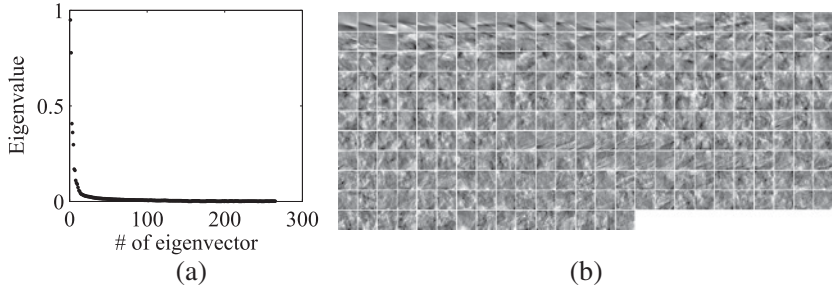

Figure 6: Eigenvalues and eigenvectors of $\tilde{C}_j$ for a second-layer unit. See the text for the definition of $\tilde{C}_j$. (a) The 265 eigenvalues in descending order. (b) The corresponding eigenvectors arranged from left to right and top to bottom.

of responses of this reduced model also exhibited a clustering effect (see Figure 5f).

As discussed in section 2.1, the matrix $\sum_{j,k} y_j w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T$ approximates the covariance matrix defined in equation 2.3. The contribution of each second-layer unit j can be dissociated from the summation as $\tilde{\boldsymbol{C}}_j = \sum_k w_{kj} \boldsymbol{b}_k \boldsymbol{b}_k^T$. Similar to Karklin and Lewicki (2009), we calculated the eigenvalues and eigenvectors of $\tilde{\boldsymbol{C}}_j$ for all j. Figure 6 shows the results of the first unit shown in Figure 3. For this unit, only a few eigenvalues were clearly greater than zero, and others were close to zero. The eigenvectors with the largest eigenvalues corresponded to the directions in image space that were most expanded, which were image features that maximally excited the unit. These features were in agreement with the orientations of the first-layer bases that had the strongest connections to the unit (see Figure 3). But the rest of the eigenvectors did not show any meaningful patterns. This was the case for most other second-layer units. In contrast to Karklin and Lewicki (2009), no eigenvectors for any second-layer unit were found to represent inhibitory image features or other types of complex features.

5 Discussion

We proposed modeling the outer product of image intensities. It was assumed that the distribution of this quantity was gaussian, where the mean was parameterized with latent variables and the covariance matrix was a constant. The latent variables were regularized with sparsity. A two-layer model was presented. Quantitative comparisons with physiological data showed that the first-layer bases resembled receptive fields of V1 simple cells, and the second-layer units exhibited orientation selectivity and phase-invariance properties, similar to V1 complex cells.

Our model differs from the hierarchical models (Hyvärinen et al., 2001; Hyvärinen & Hoyer, 2000; Karklin & Lewicki, 2005; Schwartz, Sejnowski, &

Dayan, 2006; Köster & Hyvärinen, 2010) by the location of nonlinearity. In these models, nonlinearity is always in the second layer, while in our model, it is in the first layer. By capturing some nonlinear statistical regularities in images, our model produced similar results to these models (e.g., orientation tuning and phase invariance properties, see Figures 3 and 4).

In some other hierarchical models (Karklin & Lewicki, 2009; Ranzato & Hinton, 2010; Coates et al., 2011), nonlinearity is present in the first layer, and the covariance of image intensities is modeled. In fact, the outer product of image intensities is closely related to the covariance of the gaussian distribution of image intensities assumed in Karklin and Lewicki (2009). Therefore, the proposed model, though it does not explicitly address the problem of generalization, describes the statistics of input images well.

An advantage of the proposed model over the model proposed by Karklin and Lewicki (2009) is that it does not require the matrix exponential operation. This operation in their model is to ensure the positive semidefiniteness of the covariance matrix. However, it is intuitively not straightforward to incorporate such a complicated function in a biological system. In addition, this operation is computationally expensive. We have experienced difficulties with this model because the learning algorithm was prone to instability. Careful tuning of learning rates seems to be necessary, but it was time-consuming from our experience. The experiments showed that our model was much more robust during learning, suggesting that it may play a more important role in many applications.

The outer product model, though it reproduced some properties of V1 neurons including orientation tuning, phase invariance, and a masking effect (see Figure 4), failed to reproduce the surround suppression effect of some V1 neurons (Bonds, 1989). This last effect was successfully reproduced by the covariance model (Karklin & Lewicki, 2009). In addition, the eigenvalue-eigenvector decomposition analysis revealed that the learned second-layer units of the outer product model were not as diverse of those presented in Karklin and Lewicki (2009). This might be due to the lack of matrix exponential operation and inhibitory units in the second layer of the proposed model.

Appendix A: Standardizing the Inference Problem

We show that problems 3.1 and 3.2 are equivalent. Notice that $||\boldsymbol{A}||_{\mathcal{F}}^2 = \text{trace}(\boldsymbol{A}^T\boldsymbol{A})$ and $\text{trace}(\boldsymbol{ABC}) = \text{trace}(\boldsymbol{BCA})$; the objective function in problem 3.1 can be expanded as follows:

$$\begin{aligned} f_1 &= \frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\,\text{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T\|_{\mathcal{F}}^2 + \gamma\sum_j y_j \\ &= \frac{1}{2}\text{trace}((\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\,\text{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T)^T(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{B}\,\text{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T)) + \gamma\sum_j y_j \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\operatorname{trace}(\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T) - \operatorname{trace}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T) \\
&\quad + \gamma \sum_j y_j + \operatorname{const}(\boldsymbol{y}) \\
&= \frac{1}{2}\operatorname{trace}(\boldsymbol{B}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})\boldsymbol{B}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})) - \operatorname{trace}(\boldsymbol{B}^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})) \\
&\quad + \gamma 1^T\boldsymbol{y} + \operatorname{const}(\boldsymbol{y}),
\end{aligned}
$$

where $\operatorname{const}(\boldsymbol{y})$ denotes a term independent of $\boldsymbol{y}$. Let $\boldsymbol{T} = \boldsymbol{B}^T\boldsymbol{B}\operatorname{diag}(\boldsymbol{W}\boldsymbol{y})$, which follows

$$
t_{ik} = \sum_j w_{kj} y_j \boldsymbol{b}_i^T \boldsymbol{b}_k,
$$

where t_{ik} denotes an element of $\boldsymbol{T}$. We then have

$$
\begin{aligned}
f_1 &= \frac{1}{2}\sum_i \left(\sum_k t_{ik}t_{ki}\right) + \sum_i \left(\sum_j w_{ij} y_j (\boldsymbol{B}^T\boldsymbol{x})_i (\boldsymbol{B}^T\boldsymbol{x})_i\right) + \gamma 1^T\boldsymbol{y} + \operatorname{const}(\boldsymbol{y}) \\
&= \frac{1}{2}\sum_{j_1,j_2} y_{j_1} y_{j_2} \left(\sum_{i,k} w_{kj_1} w_{kj_2} \left(\boldsymbol{b}_i^T\boldsymbol{b}_k\right)^2\right) \\
&\quad + \sum_j y_j \left(\sum_i w_{ij} \left((\boldsymbol{B}^T\boldsymbol{x}) \circ (\boldsymbol{B}^T\boldsymbol{x})\right)_i\right) + \gamma 1^T\boldsymbol{y} + \operatorname{const}(\boldsymbol{y}) \\
&= \frac{1}{2}\boldsymbol{y}^T \left(\boldsymbol{W}^T \left((\boldsymbol{B}^T\boldsymbol{B}) \circ (\boldsymbol{B}^T\boldsymbol{B})\right)\boldsymbol{W}\right)\boldsymbol{y} + \left(\boldsymbol{W}^T \left((\boldsymbol{B}^T\boldsymbol{x}) \circ (\boldsymbol{B}^T\boldsymbol{x})\right) + \gamma 1\right)^T \boldsymbol{y} \\
&\quad + \operatorname{const}(\boldsymbol{y}).
\end{aligned}
$$

Therefore, problems 3.1 and 3.2 are equivalent.

Appendix B: Proof of Theorem 1

The proof of theorem 1 follows Lee et al. (2007). For better readability, we say that $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are sign consistent if for each dimension i, $\operatorname{sign}(y_i) \times \operatorname{sign}(\hat{y}_i) \geq 0$, and denote $f_1' = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{y}$, and $\hat{f}_1' = \frac{1}{2}\hat{\boldsymbol{y}}^T\hat{\boldsymbol{A}}\hat{\boldsymbol{y}} + \hat{\boldsymbol{b}}^T\hat{\boldsymbol{y}}$, where $\hat{\cdot}$ is the submatrix or subvector corresponding to the *active set*.

Lemma 1. *If the solution is feasible (i.e., nonnegative), the optimality conditions (steps 4a and 4b) ensure that the algorithm finds the optimal solution of the quadratic programming problem.*

Proof. The Karush-Kuhn-Tucker (KKT) conditions for problem 3.2 are

$$\boldsymbol{y}^T \nabla f_1'(\boldsymbol{y}) = 0, \quad \boldsymbol{y} \geq 0, \quad \nabla f_1'(\boldsymbol{y}) \geq 0.$$

Note that problem 3.2 is a convex optimization problem. Therefore, the KKT conditions are both necessary and sufficient for the optimality. It is clear that conditions a and b in algorithm 1 are equivalent to the above conditions.

Lemma 2. *Each feature-sign step (step 3) guarantees strict improvement of the objective function f_1' without violating the nonnegativity constraints.*

Proof. Evidently for any $\boldsymbol{y}$ and its corresponding subvector given any configuration of the active set, the following functions take equal values,

$$f_1'(\boldsymbol{y}) \equiv \hat{f}_1'(\hat{\boldsymbol{y}}),$$

as the values apart from the submatrices and subvectors $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{b}}$, and $\hat{\boldsymbol{y}}$ have no effect on the function value. Therefore, it is intuitive that if $\hat{\boldsymbol{y}}_{new}$ is sign consistent with $\hat{\boldsymbol{y}}$ (i.e., $\hat{\boldsymbol{y}}_{new} \geq 0$), $f_1'(\boldsymbol{y}_{new}) \leq f_1'(\boldsymbol{y})$ since $\hat{\boldsymbol{y}}_{new}$ is the optimal solution for the quadratic programming problem $\min_{\hat{y}} \hat{f}_1'(\hat{\boldsymbol{y}})$.

If $\hat{\boldsymbol{y}}_{new}$ is not sign consistent with $\boldsymbol{y}$, a line search must be performed to update $\hat{\boldsymbol{y}}$ while preserving the nonnegativity constraint, as the L1 norm can be simplified to a first-order term only when nonnegativity is satisfied. Assume that the line search step is α ($0 \leq \alpha < 1$). Then from the convexity of the quadratic programming problem, the new point $\hat{\boldsymbol{y}}' = \hat{\boldsymbol{y}} + \alpha(\hat{\boldsymbol{y}}_{new} - \hat{\boldsymbol{y}})$ satisfies $\hat{f}_1'(\hat{\boldsymbol{y}}') \leq \alpha \hat{f}_1'(\hat{\boldsymbol{y}}_{new}) + (1-\alpha)\hat{f}_1'(\hat{\boldsymbol{y}}) \leq \hat{f}_1'(\hat{\boldsymbol{y}})$, that is, $f_1'(\boldsymbol{y}') \leq f_1'(\boldsymbol{y})$.

In fact, it is evident that neither equality holds unless the optimal solution has been achieved. Therefore, each feature-sign step strictly improves the objective function while preserving nonnegativity.

Given the above lemmas, the proof of theorem 1 is straightforward.

Proof of Theorem 1. Since the sign configurations of $\boldsymbol{y}$ are finite, following lemma 2, the algorithm cannot repeat a previous configuration as the objective function is strictly decreasing. Therefore, the algorithm must converge within a finite number of iterations, and lemma 1 ensures the optimality of such solutions.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful comments. This work was supported by the National Basic Research Program (973 Program) of China (grants 2013CB329403 and 2012CB316301), the National Natural Science Foundation of China (grant 61273023), the Beijing Natural

Science Foundation (grant 4132046), and the Tsinghua University Initiative Scientific Research Program (grant 20121088071).

References

Bonds, A. B. (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, *2*, 41–55.

Cai, D., He, X., & Han, J. (2008). SRDA: An efficient algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, *20*, 1–12.

Coates, A., Lee, H., & Ng, A. Y. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545–559.

Doi, E., Inui, T., Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Computation*, *15*(2), 397–417.

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

Hyvärinen, A., Hoyer, P. O., & Inki, M. O. (2001). Topographic independent component analysis. *Neural Computation*, *13*, 1527–1558.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics*. New York: Springer-Verlag.

Jones, H. E., Wang, W., & Sillito, A. M. (2002). Spatial organization and magnitude of orientation contrast interactions in primate V1. *Journal of Neurophysiology*, *88*(5), 2796–2808.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in non-stationary natural signals. *Neural Computation*, *17*(2), 397–423.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *457*, 83–86.

Köster, U., & Hyvärinen, A. (2010). A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, *22*(9), 2308–2333.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems*, *19* (pp. 801–808). Cambridge, MA: MIT Press.

Lee, H., Ekanadham, C., & Ng, A. (2008). Sparse deep belief net model for visual area V2. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems*, *20* (pp. 873–880). Cambridge, MA: MIT Press.

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology*, *283*, 53–77.

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

Ranzato, M., & Hinton, G. E. (2010). Modeling pixel means and covariances using factorized third-order Boltzmann machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2551–2558). Piscataway, NJ: IEEE.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, *18*, 2680–2718.

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

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

Received March 7, 2013; accepted October 23, 2013.