Learning Visual Spatial Pooling by Strong PCA Dimension Reduction

Haruo Hosoya
hosoya@atr.jp
Computational Neuroscience Laboratories, ATR International, Kyoto 619-0288, Japan, and Presto, Japan Science and Technology Agency, Saitama 332-0012, Japan

Aapo Hyvärinen
aapo.hyvarinen@helsinki.fi
Department of Computer Science and HIIT, University of Helsinki, Helsinki 00560, Finland

In visual modeling, invariance properties of visual cells are often explained by a pooling mechanism, in which outputs of neurons with similar selectivities to some stimulus parameters are integrated so as to gain some extent of invariance to other parameters. For example, the classical energy model of phase-invariant V1 complex cells pools model simple cells preferring similar orientation but different phases. Prior studies, such as independent subspace analysis, have shown that phase-invariance properties of V1 complex cells can be learned from spatial statistics of natural inputs. However, those previous approaches assumed a squaring nonlinearity on the neural outputs to capture energy correlation; such nonlinearity is arguably unnatural from a neurobiological viewpoint but hard to change due to its tight integration into their formalisms. Moreover, they used somewhat complicated objective functions requiring expensive computations for optimization. In this study, we show that visual spatial pooling can be learned in a much simpler way using strong dimension reduction based on principal component analysis. This approach learns to ignore a large part of detailed spatial structure of the input and thereby estimates a linear pooling matrix. Using this framework, we demonstrate that pooling of model V1 simple cells learned in this way, even with nonlinearities other than squaring, can reproduce standard tuning properties of V1 complex cells. For further understanding, we analyze several variants of the pooling model and argue that a reasonable pooling can generally be obtained from any kind of linear transformation that retains several of the first principal components and suppresses the remaining ones. In particular, we show how the classic Wiener filtering theory leads to one such variant.

Neural Computation **28, 1249–1264** (2016)
doi:10.1162/NECO_a_00843

1 Introduction

In visual cortex, individual cells often show invariance properties, where the cells maintain their responses when the stimulus is varied in certain stimulus parameters. This property is considered to be the neural basis for the remarkable robustness of our visual system when recognizing objects and scenes even under significant variation and deformation (DiCarlo & Cox, 2007). In computational studies, invariance properties are often modeled by a mechanism called *pooling*. This mechanism assumes a set of underlying model units selective to some stimulus parameters (called *subunits*) and integrates the outputs of those subunits into a higher-level unit, which gains some invariance to other stimulus parameters. The notion of pooling had appeared in early physiological work that informally presented a model of phase-invariant V1 complex cells that pool phase-dependent simple cells (Hubel & Wiesel, 1962), which was later formalized as the classical energy model (Adelson & Bergen, 1985). Pooling then became standard in more sophisticated multilayered vision models, which stipulated repetition of V1-like computation in each layer, with pooling included for achieving complicated invariance properities (Fukushima, 1980; Riesenhuber & Poggio, 1999). This idea was refined in state-of-the-art computer vision models based on deep learning that include pooling in a form similar to energy models (Le et al., 2012) or in conjunction with convolutional architectures (Krizhevsky, Sutskever, & Hinton, 2012).

What kind of learning principle could underlie such pooling? Prior studies showed that certain spatial statistics of subunit outputs leads to a pooling of those subunits that attains phase-invariance properties similar to V1 complex cells (Hyvärinen & Hoyer, 2000, 2001; Karklin & Lewicki, 2003, 2009; Köster & Hyvärinen, 2010; Osindero, Welling, & Hinton, 2006). The basic idea used in these models (explicitly or implicitly) is to combine subunits whose squared outputs are highly correlated (such correlation is often called *energy correlation*). However, such squaring nonlinearity is arguably unnatural from a biological point of view due to its symmetry, while it cannot easily be changed to a more realistic one, for example, half-rectification ($y = \max(0, x)$) or half-squaring ($y = \max(0, x)^2$), since the nonlinearity is tightly integrated into the formalisms. Moreover, the previous models required rather expensive computations for optimizing their objective functions. (See section 4 for more specific discussion, as well as other approaches using temporal structure: Földiák, 1991; Wiskott & Sejnowski, 2002; Hurri & Hyvärinen, 2003; Berkes & Wiskott, 2005; Einhäuser, Kayser, König, & Kording, 2002; Kayser, Kording, & König, 2003.)

In this study, we revisit the learning problem for pooling and propose a simple alternative using strong dimension reduction based on principal component analysis (PCA). The idea here is to eliminate fine-grained spatial statistical structures in the signals from the subunits, which results in the integration of subunits with correlated outputs, in a manner that generalizes over different output nonlinearities. Indeed, we show that when this approach is used, a model of V1 complex cells can be constructed as a linear pooling of model V1 simple cells with similar position and orientation preferences and that the resulting model complex cells exhibit standard tuning properties similar to monkey V1 complex cells, even when using a half-rectifying nonlinearity on the subunit outputs. For further understanding, we analyze several variations of the pooling model and argue that a similar pooling model can generally be obtained from any form of linear transformation that retains several of the first principal components and suppresses the remaining ones. In particular, a Wiener-filter-like linear denoising model, which optimally achieves our principle of eliminating fine-grained structures, gives a reasonable pooling model.

This study is a follow-up of our previous publication on a model of V2 (Hosoya & Hyvärinen, 2015), in which we observed that PCA-based strong dimension reduction performed prior to overcomplete independent component analysis (ICA) was crucial to obtain model V2 cells with reasonably large receptive field sizes. Here, we focus on the pooling model with strong dimension reduction and present a series of theoretical analyses and comparisons with experimentally known properties of V1 complex cells.

2 Strong PCA Dimension Reduction

Consider an N-dimensional data set and suppose that we perform PCA on it. We assume that the eigenvalues of the covariance matrix are sorted in descending order and the eigenvectors are sorted in accordance with the eigenvalues. Thus, let $\mathbf{e}_k$ be the kth (normalized, row) eigenvector. Our proposal is, for a given smaller dimension K, to construct the following $N \times N$ matrix,

$$P = \bar{E}^{\mathsf{T}}\bar{E}, \tag{2.1}$$

where $\bar{E}$ is the $K \times N$ matrix whose rows are top K eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K$. That is, the matrix transforms a given N-dimensional vector into the reduced K-dimensional eigenspace (space of principal components) and transforms the result back into the original space, whose dimensionality is thus effectively reduced. We are particularly interested in the case of *strong dimension reduction*, where K is far smaller than N (e.g., $K \sim N/10$). We call the matrix P *pooling matrix* and its each row vector a *pooling filter*.

To show that the strong dimension reduction can give a reasonable pooling model, we constructed a model of V1 complex cells. First, we learned a set of 192 Gabor filters with various orientations and frequencies by standard ICA of natural image patches (see appendix A for details). For later comparison, 49 example filters are shown in Figure 1A. Our model simple cells (subunits) were defined as the outputs of those filters with a

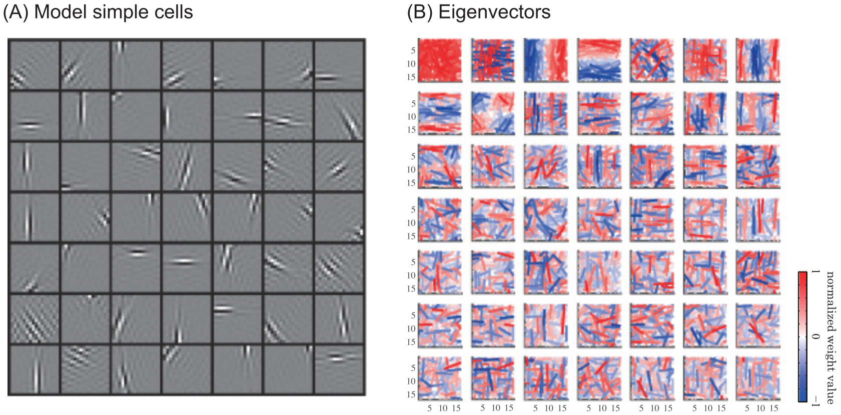
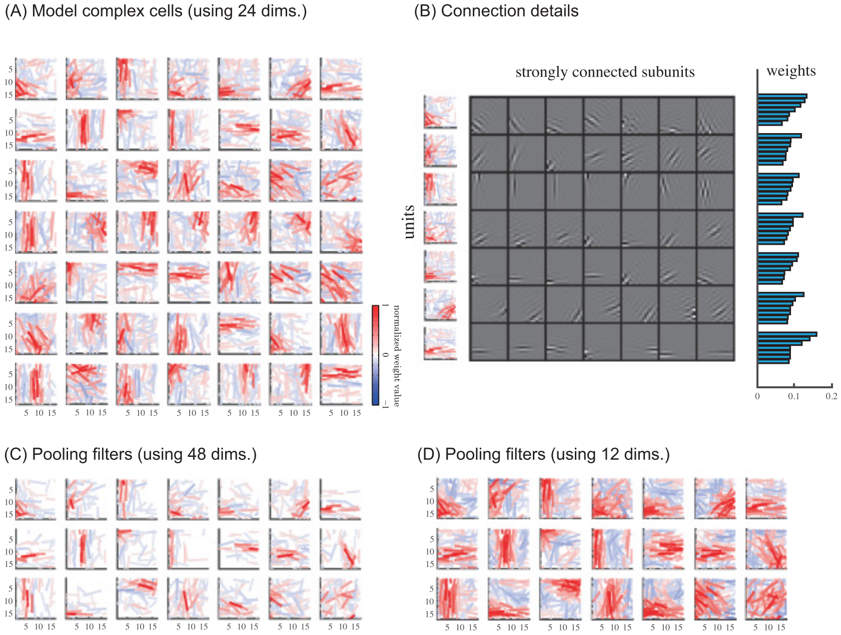


Figure 1: Simple cell model and its PCA. (A) The linear filters of 49 example model simple cells (ICA filters estimated from natural image patches). (B) Top 49 eigenvectors for the outputs of model simple cells. Each eigenvector is drawn as a set of line segments, where each line segment indicates the orientation, position, and size of a model simple cell and the color indicates the normalized value of the element of the eigenvector corresponding to the model simple cell (color bar). The weights with absolute value smaller than 10% of the maximum are dropped for readability.

half-rectifying nonlinearity. We then performed PCA on the outputs of the subunits (with the same inputs of natural image patches), whose top 49 eigenvectors are shown in Figure 1B. Here, each panel shows an eigenvector by a set of colored line segments, where each line segment indicates the orientation, position, and size of a model simple cell and the color of the line segment indicates the value of the element in the eigenvector corresponding to the model simple cell. As can be seen in the figure, earlier eigenvectors represented more coarse-grained structures, collecting subunits with similar orientation preferences from a broad spatial region. In contrast, later eigenvectors represented more fine-grained structures, describing a specific orientation combination at every spatial location.

We then performed strong dimension reduction using only the top 24 eigenvectors (out of 192 in total), ignoring most of the fine-grained structures; our model complex cells were defined as the pooling filters as given in formula 2.1. Figure 2A shows 49 pooling filters in the same display format as the eigenvectors. We can clearly recognize pooling of model simple cells with similar orientations and nearby positions. Note that since the pooling matrix does nothing but eliminate fine structures, the ordering of the input dimensions is preserved: the model simple cells in Figure 1A and the model complex cells in Figure 2A correspond to each other. For a closer look, the first seven of the model complex cells are illustrated in more


(A) Model complex cells (using 24 dims.)

(B) Connection details

(C) Pooling filters (using 48 dims.)

(D) Pooling filters (using 12 dims.)

Figure 2: Complex cell model. (A) Forty-nine pooling filters (model complex cells) obtained by strong dimension reduction using the 24 top eigenvectors. The display format is similar to Figure 1B. (B) The first seven pooling filters shown in panel A (left), with the seven most strongly connected model simple cells in each row (middle) and their (unnormalized) connection weights (right). (C–D) Twenty-one pooling filters in the case of using (C) 48 or (D) 12 top eigenvectors.

detail in Figure 2B, displaying the seven most strongly connected subunits to each model complex cell (middle) and their connection weights (right), which highlights the localization and orientation similarity of the integrated model simple cells. The strength of dimension reduction controls the spatial extent of pooling: Figures 2C, 2A, and 2D reveal that model complex cells had progressively larger shapes as the reduced dimensionality K decreased ($K = 48, 24, 12$).

To quantitatively assess the sensitivity of each model cell to the phase, orientation, and frequency, we followed the standard protocol used in electrophysiology, analyzing the responses to whole-field grating stimuli with varied phases, orientations, and frequencies (see appendix B for details). Figures 3A and 3B show the phase tuning curves of the first eight model simple cells in Figure 1A and the first eight complex cells in Figure 2B, respectively. The latter were generally more insensitive to the phase. Figures 3D and 3E show the distributions of F1/F0 ratios (the first Fourier

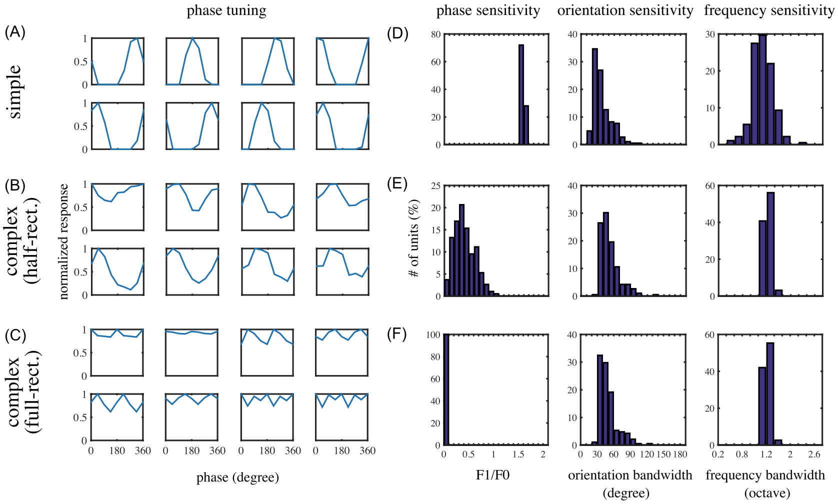


Figure 3: Analysis of model simple and complex cells. Each row is one population: top row (A,D), half-rectified model simple cells shown in Figure 1A; middle row (B,E), model complex cells based on half-rectified simple cells shown in Figure 2A; bottom row (C,F), model complex cells using full-rectified model simple cells. The quantities computed are (A–C) the phase tuning curves for eight example model cells from each population and (D–F) the distributions of F1/F0 values, orientation bandwidths, and frequency bandwidths for each population. The orientation and frequency bandwidths are determined in half-maximum full width.

component divided by the DC component) for model simple cells and for model complex cells. All of the model complex cells were relatively insensitive to the phase (F1/F0 smaller than 1), although completely phase-insensitive ones (F1/F0 close to 0) were rare. The same figures show the distributions of orientation and frequency bandwidths (in half-maximum full width) for model simple cells and model complex cells. The model complex cells had somewhat similar sensitivities, albeit slightly weaker, to orientations and frequencies compared to the model simple cells. Although not shown in the figures, the phase sensitivities became weaker (broader pooling, smaller F1/F0 values) when a complex cell model was obtained by stronger dimension reduction (smaller K), while the orientation and frequency sensitivities were barely affected. For comparison, Figure 4 shows the analogous distributions for simple and complex cells, respectively, of monkey V1 replotted from past experimental studies (Skottun et al., 1991; De Valois, Albrecht, & Thorell, 1982; De Valois, Yund, & Hepler, 1982). The peaks in the corresponding distributions in Figures 3D and 3E are more or less similar, while the distributions tend to have larger variances in monkey

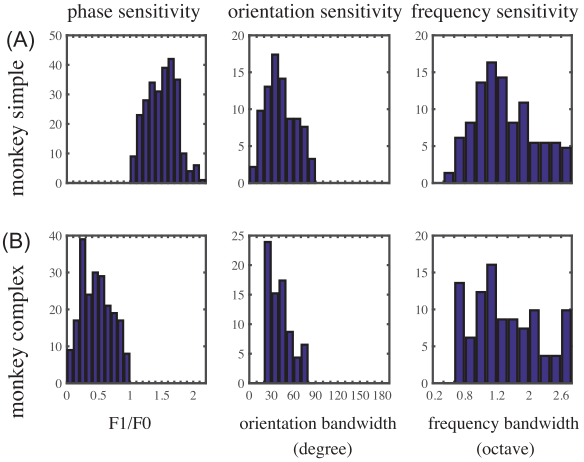


Figure 4: Replots of experimental data of F1/F0 values and orientation and frequency bandwidths of (A) simple cells and (B) complex cells in monkey V1 (Skottun et al., 1991; De Valois, Albrecht et al., 1982; De Valois, Yund et al., 1982).

(in particular, F1/F0 values of simple cells and frequency bandwidths of simple and complex cells).

In the model, some extent of phase insensitivity was obtained by strong dimension reduction on half-rectified outputs of model simple cells. Perfect phase insensitivity can be attained by changing the nonlinearity from half rectification to full rectification ($y = |x|$), capturing energy correlations. Figure 3C shows the phase tuning curves of examples of the model complex cells constructed in this way; all the model complex cells were in fact completely phase insensitive in the sense that their F1/F0 values were almost zero (see Figure 3F). The orientation and frequency bandwidths were similar to the case of half-rectifier (see Figure 3F). Although not shown here, using squaring and half-squaring gave distributions similar to the case of full rectification and half rectification, respectively, but with narrower orientation and frequency bandwidths due to the stronger nonlinear effect. Finally, it is essential to have some output nonlinearity; the pooling effect vanished when the linear output was directly fed to strong dimension reduction (data not shown).

3 Variations

So far, we have shown that strong dimension reduction using PCA has the effect of spatial pooling. In this section, we first argue that in general, any linear transformation that retains a few top eigenvectors and cuts off the remainder has a similar effect. We then derive an optimal form of pooling based on a denoising model.

3.1 General Form of Pooling. We focus on transformations in the following diagonalized form,

$$P = E^{\intercal} \operatorname{diag}(\mathbf{h}) E, \tag{3.1}$$

where E is the $N \times N$ matrix of all row eigenvectors and $\mathbf{h}$ is an N-dimensional vector with nonnegative elements. For example, strong dimension reduction using K components as in formula 2.1 can be written in the above diagonal form using

$$\mathbf{h} = (\underbrace{1, \ldots, 1}_{K}, \underbrace{0, \ldots, 0}_{N-K}),$$

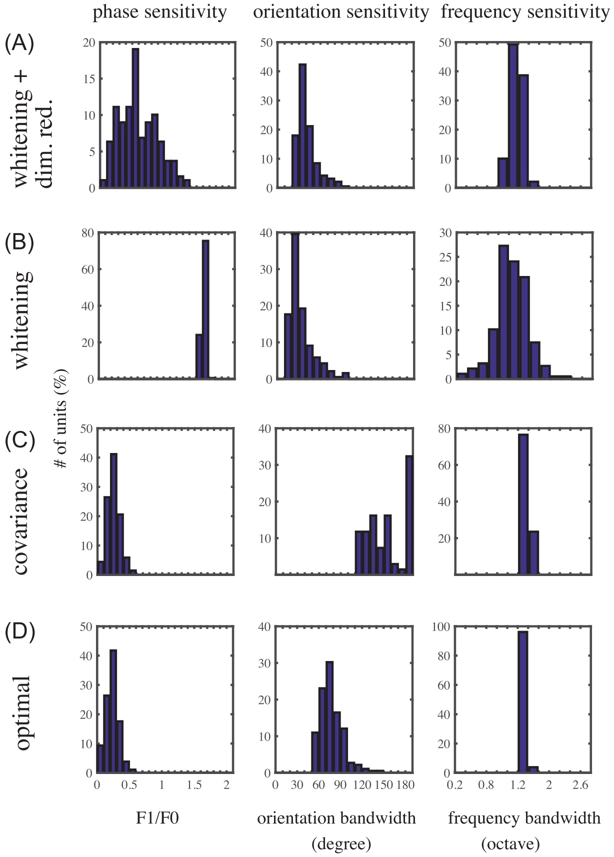
which can be read directly as retaining K top eigenvectors and removing the remaining ones.

A simple variation is the combination of strong dimension reduction and whitening. Such an operation can be formulated as our diagonalized form, equation 3.1, with

$$\mathbf{h} = (d_1^{-1/2}, \ldots, d_K^{-1/2}, \underbrace{0, \ldots, 0}_{N-K}),$$

where d_i is the ith eigenvalue. Since this operation also retains a few top components and eliminates the remaining components, it has a similar pooling effect. Indeed, model complex cells constructed with this operation properly gave low phase sensitivities and high orientation sensitivities (see Figure 5A). The diagonal values h_i's are plotted in Figure 5E (red). Whitening with strong dimension reduction may be useful when some feature extraction method is applied after pooling in a hierarchical learning model. Indeed, in our previously published model of V2, this operation was performed prior to overcomplete ICA (Hosoya & Hyvärinen, 2015) (although the use was implicit in the sense that the backward transformation $E^{\intercal}$ was omitted in that paper).

Note that performing only whitening, that is, $\mathbf{h} = (d_1^{-1/2}, \ldots, d_N^{-1/2})$, demolishes the pooling effect with a high phase sensitivity (see Figure 5B), since later components are retained (see Figure 5E, yellow). Also, simply using the covariance, that is, $\mathbf{h} = (d_1, \ldots, d_N)$, overly promotes the pooling effect so that it prevents meaningful orientation or frequency selectivities (see Figure 5C), since very few top components are retained (see Figure 5E, violet).



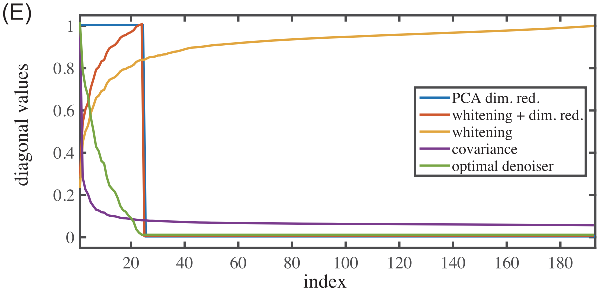


Figure 5: Variations of the model. (A–D) The distributions of F1/F0 values and orientation and frequency bandwidths in the case of using (A) whitening and dimension reduction, (B) whitening, (C) covariance, and (D) optimal pooling based on denoising model, where σ^2 is set to the 24th eigenvalue. (E) The diagonal values (normalized by the maximum) in each case, as well as PCA-based strong dimension reduction (see Figure 3E).

3.2 Optimal Pooling Based on Denoising Model. We have seen a few examples where a sensible pooling results from a choice of diagonal values h_i, which can each be seen as an instance of our principle of elimination of fine-grained structures. Is it possible to derive such diagonal values more directly from this principle? Indeed, this corresponds to a simple Wiener-filter-like denoising model formalized below, where fine-grained structures in the outputs of model simple cells are regarded as noise and the role of model complex cells is to remove them.

Suppose that model complex cells $\mathbf{c}$ attempt to eliminate additive noise $\mathbf{n}$ from outputs of model simple cells $\mathbf{s}$. Thus, we model the simple cell outputs $\mathbf{s}$ as a sum of the complex cell outputs $\mathbf{c}$ and independent noise $\mathbf{n}$:

$$\mathbf{s} = \mathbf{c} + \mathbf{n}. \tag{3.2}$$

Assume that the noise is white with variance σ^2 and the model complex cells attempt to denoise this approximately by linear filters P in our diagonal form, equation 3.1:

$$\hat{\mathbf{c}} = P\mathbf{s}. \tag{3.3}$$

Now, we want to minimize the squared error,

$$F = \mathbb{E}[\|\mathbf{c} - \hat{\mathbf{c}}\|^2], \tag{3.4}$$

under the nonnegativity constraint $h_i \geq 0$ $(i = 1, \dots, N)$, which yields the following solution:

$$h_i = \begin{cases} 1 - \sigma^2/d_i & (\sigma^2 < d_i) \\ 0 & (\sigma^2 \geq d_i) \end{cases}. \tag{3.5}$$

The proof is a simple variant of well-known theory (e.g., Doi & Lewicki, 2011), given in appendix C for completeness. The values of h_i's are plotted in Figure 5E (green). Figure 5D shows the phase, orientation, and frequency sensitivities of the optimal denoising model cells when σ^2 is set to the 24th eigenvalue. The pooling effect is evident, although the orientation sensitivities are rather low compared to strong dimension reduction. Note that a broader pooling can be achieved by a larger assumed noise variance σ^2 (data not shown).

4 Discussion

In this study, we proposed a simple learning model of spatial pooling of visual cells based on strong dimension reduction using PCA. This model learns to ignore fine-grained structures from the signals of subunits and thus discovers a linear pooling of highly correlated subunits. We demonstrated that when this approach was applied to the outputs from V1 simple cell models, the resulting pooling model exhibited response properties to oriented grating stimuli that were similar to experimental data of monkey V1 complex cells. We also showed that, more generally, any kind of linear operations that retain a few top principal components and reduce the remaining components can be a sensible pooling model; in particular, a Wiener-filter-like optimal denoiser model can give one such example.

In our previous publication (Hosoya & Hyvärinen, 2015), we implicitly used strong dimension reduction as a "preprocessing" of overcomplete ICA to construct a model of V2 cells. There, we found that this operation was crucial since otherwise the resulting model V2 cells had overly small receptive fields that could not be compared with actual V2. However, since the previous study focused on empirical properties of the V2 model itself, it did not theoretically clarify why strong dimension reduction gave such a pooling property. Thus, this study concentrates on strong dimension reduction itself, conducting a series of theoretical analyses and simulations, in particular in the archetypal case of pooling given by V1 complex cells.

While a number of previous studies proposed learning models for pooling from input spatial structure (Hyvärinen & Hoyer, 2000, 2001; Karklin & Lewicki, 2003, 2009; Köster & Hyvärinen, 2010; Osindero et al., 2006), our approach differs in that it does not need to use a squaring nonlinearity to capture energy correlation of subunit outputs. (The squaring nonlinearity is not explicit in some previous studies—for example, Karklin & Lewicki, 2003, 2009—but their probabilistic formalizations lead to similar computations.) Taking squares of linear filters has a similar effect to pooling of half-rectified outputs of the linear filters with opposite phase preferences, which can achieve perfect phase invariance. However, squaring nonlinearity is generally thought to be rather unnatural from a neurophysiological point of view; moreover, perfect phase invariance cannot always be found in actual V1 (see Figure 4B). In particular, squaring is unnatural since a large negative input produces a large positive output, which is at least not implementable by a single neuron and would require slightly complicated circuitry. However, half-squaring could be a plausible nonlinearity (Anzai, Ohzawa, & Freeman, 1999; Lau, Stanley, & Dan, 2002).

However, if the issue is the choice of nonlinearity, why can't we simply modify the nonlinearity used in the previous methods? In fact, the nonlinearity there is tightly built into the method so that such modification would disable proper functioning. For example, independent subspace analysis (ISA) works by optimizing linear filters so that their sums of squares maximize sparsity (Hyvärinen & Hoyer, 2000). However, if we modify this squaring to half-rectifier, then meaningless linear filters that maximize the occurrence of negative values (whose half-rectified values are zero) would be learned since such nonlinearity can no longer measure sparsity correctly. Therefore, an entirely new development seems to be necessary if types of nonlinearity other than squaring are used. Similar arguments can be applied to other methods as well (Hyvärinen & Hoyer, 2001; Karklin & Lewicki, 2003, 2009).

How could our PCA-based strong dimension reduction be implemented in a neural system? Presumably the pooling system could be learned by a network with two layers, each of which learns using well-known rules that have some biological plausibility. The first layer could learn PCA by a number of such neural learning rules (e.g., Oja & Karhunen, 1985). The second layer should then learn to optimally reconstruct the original data based on the principal components, the optimal reconstruction being given by the matrix of the dominant eigenvectors transposed. The second layer thus essentially implements a heteroassociative memory, such as Kohonen's correlation matrix memory (Kohonen, 1972), whose learning is well known to be achievable by very simple Hebbian schemes (Rojas, 1996, chapter 12). Whether such network can indeed achieve a pooling behavior as intended here is left for future research.

An alternative principle for learning invariances is temporal coherence (Földiák, 1991; Wiskott & Sejnowski, 2002; Hurri & Hyvärinen, 2003), which extracts most slowly changing features from time series. A few studies have demonstrated the effectiveness of this approach by reproducing phase invariances similar to V1 complex cells (Berkes & Wiskott, 2005; Einhäuser et al., 2002; Kayser et al., 2003) (although learning only from slowness may lead to overly global features—Hashimoto, 2003; Lies, Häfner, & Bethge, 2014; Hyvärinen, Hurri, & Hoyer, 2009, section 16.8), as well as viewpoint invariances as in inferotemporal cortex (Einhäuser, Hipp, Eggert, Körner, & König, 2005). Temporal coherence has also been supported by an experimental study that attempts to break invariant object recognition by unnatural sequences of visual stimuli (Cox, Meier, Oertelt, & DiCarlo, 2005). While exploiting temporal structure is a fascinating approach, it could be argued from the viewpoint of Ockham's razor that purely spatial modeling should be preferred whenever possible since it is more parsimonious.

Finally, while strong dimension reduction explained properties of V1 complex cells, could it be a general computation principle that might as well be employed in higher visual areas? This possibility seems quite attractive since the retinal inputs on the huge receptive fields of higher visual cells are tremendously high-dimensional, and therefore strong dimension reduction might help to alleviate the potential computational burden of processing such inputs. Further, it would be reasonable to think that the role of a higher visual cell may not be to look at details of a visual input (which lower cells can do), but rather to find its rough structure on a more abstract level, ignoring fine details as if they were "noise."

Appendix A: Simple Cell Model

The simple cell model in section 2 was constructed by a standard method performing ICA (or sparse coding) on natural image patches (Olshausen &

Field, 1996; Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998). Concretely, a set of 100,000 natural image patches of 16 × 16 pixels was randomly extracted from ImageNet (Deng, Berg, Li, & Fei-Fei, 2010), with the DC component removed from each image. The image data set was then whitened, with the dimension lightly reduced from 256 to 192 to eliminate aliasing artifacts (Hyvärinen et al., 2009, section 5.3.3.3). Thereafter, ICA was applied using FastICA (Hyvärinen, 1999).

Appendix B: Measuring Response Properties

Basic tuning properties were measured from the responses of each model simple or complex cell to whole-field grating stimuli with various orientations, frequencies, and phases. The pair of optimal orientation and frequency was determined as that giving the maximum of responses averaged over phases. The orientation tuning function was the responses averaged over phases at the optimal frequency. The frequency tuning function was the responses averaged over phases at the optimal orientation. The phase tuning function was the responses at the optimal orientation and frequency. The orientation tuning function was fitted with 180°-cycled von Mises functions, and the frequency tuning function was fitted with gaussian functions. The orientation and frequency bandwidths were the full width at the half maximum as determined from the fitted function. The F1/F0 value was calculated as the ratio of the first Fourier component and the DC component of the phase tuning function.

Appendix C: Deriving the Optimal Denoising Model

Our goal is to minimize the squared error function,

$$F = \mathbb{E}[\|\mathbf{c} - \hat{\mathbf{c}}\|^2], \tag{C.1}$$

which can be transformed, using assumptions 3.2 and 3.3, to

$$F = \operatorname{Tr} \Sigma_c - 2 \operatorname{Tr} \Sigma_c P^\top + \operatorname{Tr} P \Sigma_s P^\top, \tag{C.2}$$

where Σ_s and Σ_c are the covariance matrices of $\mathbf{s}$ and $\mathbf{c}$, respectively. Note that, from equation 3.2,

$$\Sigma_c = \Sigma_s - \sigma^2 I. \tag{C.3}$$

Thus, together with the assumptions 3.1 and 3.3, this can be rewritten as

$$\begin{aligned} F = &\operatorname{Tr} E^\top \left[\operatorname{diag}(\mathbf{d}) - \sigma^2 I\right] E \\ &- 2 \operatorname{Tr} E^\top \left[\operatorname{diag}(\mathbf{d}) - \sigma^2 I\right] \operatorname{diag}(\mathbf{h}) E \\ &+ \operatorname{Tr} E^\top \operatorname{diag}(\mathbf{h}) \operatorname{diag}(\mathbf{d}) \operatorname{diag}(\mathbf{h}) E, \end{aligned}$$

where $\mathbf{d} = (d_1, \ldots, d_N)$. Since the trace is preserved under transformation into the eigenspace, this can be transformed to

$$F = \sum_i (d_i - \sigma^2) - 2(d_i - \sigma^2)h_i + d_i h_i^2. \tag{C.4}$$

Solving the unconstrained minimization by equating the derivatives to zero,

$$\frac{\partial F}{\partial h_i} = 2d_i h_i - 2d_i + 2\sigma^2 = 0, \tag{C.5}$$

and taking the the constraint $h_i \geq 0$ into account, we obtain the solution in equation 3.5.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299.

Anzai, A., Ohzawa, I., & Freeman, R. D. (1999). Neural mechanisms for processing binocular information I. Simple cells. *Journal of Neurophysiology*, *82*(2), 891–908.

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

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, *5*(6), 579–602.

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). "Breaking" position-invariant object recognition. *Nature Neuroscience*, *8*(9), 1145–1147.

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.

De Valois, R. L., Yund, E. W., & Hepler, N. (1982). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 531–544.

Deng, J., Berg, A. C., Li, K., & Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision* (pp. 71–84). New York: Springer.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

Doi, E., & Lewicki, M. S. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Computation*, *23*(10), 2498–2510.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological Cybernetics*, *93*(1), 79–90.

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

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

Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems*, *14*(4), 765–788.

Hosoya, H., & Hyvärinen, A. (2015). A hierarchical statistical model of natural images explains tuning properties in V2. *Journal of Neuroscience*, *35*(29), 10412–10428.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*(1), 106.

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

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634.

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

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics: A probabilistic approach to early computational vision.* New York: Springer.

Karklin, Y., & Lewicki, M. S. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, *14*(3), 483–499.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *457*(7225), 83–86.

Kayser, C., Kording, K. P., & König, P. (2003). Learning the nonlinearity of neurons from natural visual stimuli. *Neural Computation*, *15*(8), 1751–1759.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, *c-21*, 353–359.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *25*. Red Hook, NY: Curran.

Lau, B., Stanley, G. B., & Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(13), 8974–8979.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., & Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning*. N.p.: International Machine Learning Society.

Lies, J.-P., Häfner, R. M., & Bethge, M. (2014). Slowness and sparseness have diverging effects on complex cell learning. *PLoS Computational Biology*, *10*(3), e1003468.

Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, *106*(1), 69–84.

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.

Osindero, S., Welling, M., & Hinton, G. E. (2006). Topographic product models applied to natural scene statistics. *Neural Computation*, *18*(2), 381–414.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Rojas, R. (1996). *Neural networks: A systematic introduction*. New York: Springer.

Skottun, B. C., De Valois, R. L., Grosof, D. H., Movshon, J. A., Albrecht, D. G., & Bonds, A. B. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research*, *31*(7–8), 1078–1086.

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 B: Biological Sciences*, *265*(1394), 359–366.

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

Received November 26, 2015; accepted February 19, 2016.