OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Efficient Sparse Coding in Early Sensory Processing: Lessons from Signal Recovery

András Lörincz[1]*, Zsolt Palotai[2,3], Gábor Szirtes[1,4]

1 Department of Software Technology and Methodology, Eötvös Loránd University, Budapest, Hungary, **2** Sparsense Inc., Boca Raton, Florida, United States of America, **3** ELTE-Soft Ltd, Budapest, Hungary, **4** Center for Integrative Neuroscience, University of Tuebingen, Tuebingen, Germany

Abstract

Sensory representations are not only sparse, but often overcomplete: coding units significantly outnumber the input units. For models of neural coding this overcompleteness poses a computational challenge for shaping the signal processing channels as well as for using the large and sparse representations in an efficient way. We argue that higher level overcompleteness becomes computationally tractable by imposing sparsity on synaptic activity and we also show that such *structural* sparsity can be facilitated by statistics based decomposition of the stimuli into typical and atypical parts prior to sparse coding. Typical parts represent large-scale correlations, thus they can be significantly compressed. Atypical parts, on the other hand, represent local features and are the subjects of actual sparse coding. When applied on natural images, our decomposition based sparse coding model can efficiently form overcomplete codes and both center-surround and oriented filters are obtained similar to those observed in the retina and the primary visual cortex, respectively. Therefore we hypothesize that the proposed computational architecture can be seen as a coherent functional model of the first stages of sensory coding in early vision.

Citation: Lörincz A, Palotai Z, Szirtes G (2012) Efficient Sparse Coding in Early Sensory Processing: Lessons from Signal Recovery. PLoS Comput Biol 8(3): e1002372. doi:10.1371/journal.pcbi.1002372

Editor: Lyle J. Graham, Université Paris Descartes, Centre National de la Recherche Scientifique, France

Received February 3, 2011; **Accepted** December 20, 2011; **Published** March 1, 2012

Funding: The research reported in this paper is supported by the European Union and co-financed by the European Social Fund (grant agreement no. TAMOP 4.2.1/B-09/1/KMR-2010-0003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: andras.lorincz@elte.hu

Introduction

In the last decades a large body of research has been devoted to explain the nature of neural representations. Since experimental manipulation of the stimuli has the most direct impact on the sensory responses, most of our knowledge comes from studies about the early stages of sensory systems. Although we do not have a complete story yet, experimental and theoretical research did reveal important principles about the nature of neuronal representations together with specific constraints imposed by anatomy and physiology. Derived from the efficient coding theory [1,2], different popular models – emphasizing redundancy reduction (like [3,4]) or the sparsity constraint (Sparse Coding, SC, e.g. [5,6]) – can account for many, but not all relevant features of early sensory processing (e.g. [7,8]). In this article we argue that a novel computational model of neural representation can be obtained by focusing on one of those relevant features: overcompleteness. For codes with this property the number of potential coding units is larger than that of the input units thus offering increased memory capacity and enhanced robustness against noise and structural perturbations. We will argue that the formation of large and sparse representations of high level of overcompleteness requires adaptive learning which can effectively control the number of active synapses. This structural sparsification has a significant impact on the overall metabolic cost of neural activity. We then present a new sparse coding scheme which is motivated by both theories mentioned above, but is built on a non-conventional signal model assuming an *additive decomposition* of stimuli into "typical" and "atypical" constituents. We also analyze the model's filtering properties when trained on natural images. The main contribution of our study is that principled pre-filtering based on this alternative signal model can indeed facilitate overcomplete SC by supporting structural sparsity. The pre-filtering process is motivated by recent results on efficient compression, completion and decomposition of high dimensional data; computational functions equally important for artificial and natural systems. Based on the finding that our model can simultaneously explain several features of early vision we then suggest a biological implementation of the two stage algorithm.

The paper is organized as follows. In the Results section first we review the computational problem of overcomplete sparse coding and argue about the importance to control synaptic activity. Then we introduce our two stage algorithm which can achieve structural sparsity thus supporting overcomplete sparse coding. In support of our model numerical experiments on natural images are also presented. In the Discussion section we compare the computational properties and biological relevance of our model with alternative approaches. In the Methods section the details of the numerical experiments are provided together with brief descriptions, pseudocodes and references to more elaborate presentations of the algorithmic building blocks.

Results

In this section we present the problem of (overcomplete) sparse coding (SC) with an emphasis on metabolic constraints (regarding spike activity) and briefly discuss some alternative algorithmic solutions. We then consider if further reduction in computational

Author Summary

Neural systems favor overcomplete sparse codes in which the number of potential output neurons may exceed the number of input neurons, but only a small subset of neurons become actually active. We argue that efficient use of such large dimensional overcomplete sparse codes requires structural sparsity by controlling the number of active synapses. Motivated by recent results in signal recovery, we introduce a particular signal decomposition as a pre-filtering stage prior to the actual sparse coding, which efficiently supports structural sparsity. In contrast to most models of sensory processing, we hypothesize that the observed transformations may actually realize parallel encoding of the stimuli into representations that describe typical and atypical parts. When trained on natural images, the resulting system can handle large, overcomplete representations and the learned transformations seem compatible with the various receptive fields characteristic to different stages of early vision. In particular, transformations realized by the prefiltering units can be approximated as 'Difference-of-Gaussians' filters, similar to the receptive fields of neurons in the retina and the LGN. In addition, sparse coding units have localized and oriented edge filters like the receptive fields of the simple cells in the primary visual cortex, V1.

(metabolic) cost can be accomplished by targeting synaptic activity. Motivated by the insight that the presence of noise hinders the effective control of synaptic activity, we introduce a novel two stage sparse coding algorithm which facilitates structural sparsity (i.e. by keeping the number of active synapses low) and in turn supports the formation of overcomplete sparse codes. The model is then tested on natural images and the responses of the computing units are compared to neural responses in early vision.

Preliminaries

Due to the high metabolic cost of spiking activity [9–11], constraining average spiking rate (over time and population) seems to be a general principle in neural systems (but see [12]). Therefore we also consider sparsity central in our coding model. The objective of the sparse coding (SC) scheme is to find the sparsest representation of the data with low reconstruction error. It has been argued that this scheme offers a computationally and metabolically advantageous trade off between fully localized (like "grandmother"-cells) and distributed codes [13]. Sparse codes essentially try to approximate the underlying hidden structure (the generating sources) of the observed stimulus. The great advantage of SC over other coding schemes is that it directly controls energy consumption by setting the number of active coding units; k out of m coding units with $k<m$ can be active at any given time. Another important property of neural codes is overcompleteness, when the number of coding units (m) is greater than the number of input units (n, $m>n$). For example, in area 17 of cat the ratio of the output fibers versus the input fibers from the LGN is estimated about 25:1, while in macaque primary visual cortex, V1 the estimate is between 12:1 and 160:1 [14] or even 500:1 [15]. In principle, overcompleteness provides more flexibility in finding even sparser representations. However, overcompleteness presents a non-trivial challenge for computational models on neural representations. In comparison with biological data, most computational models of SC can find the optimal solution if overcompleteness is 2 to 8-fold at most [16]. Importantly, higher level of overcompleteness may increase the overall metabolic cost of neural coding for two reasons. First, non-optimal solutions require too many iterations thus generating excess spiking activity. Second, overcompleteness induces an asymmetry in the use of the encoder and decoder channels *within* one iteration: while the excitation process requires the use of all $n\times m$ encoder channels, selected subsets of k active decoding units require only $k\times n$ decoder channels. That is methods that avoid the heavy use of encoding are more favorable. The importance of controlling the number of active coding channels (that is the number of synapses which define the receptive field of a neuron) is highlighted by the fact that according to the estimates of [10], more than 50% of the metabolic cost of a single spike can be attributed to the excitatory potentials at the postsynaptic sites (EPSPs). Our goal is thus to find an algorithmic model that can explain overcomplete sparse coding in the brain.

Formally, SC can be stated as an alternating (two step) optimization problem:

$$\min_{\mathbf{D},\,\mathbf{a}\in R^m}\sum_{i=1}^{t}\frac{1}{2}\|\mathbf{x}^i-\mathbf{D}\mathbf{a}^i\|_2^2+\beta\|\mathbf{a}^i\|_0 \qquad (1)$$

where $\mathbf{x}^i\in\mathbb{R}^n\,(i=1,\ldots,t)$ is the i^{th} signal, or input to be reconstructed, t is the number of training inputs, $\mathbf{a}\in\mathbb{R}^m$ $(m\geq n)$ denotes the coefficient vector of the sparse decomposition also called (internal) representation and $\mathbf{D}\in R^{n\times m}$ is the basis, or *dictionary* of features. $\|\cdot\|_0$ denotes the ℓ_0-norm, which is the number of nonzero components. The first term minimizes the reconstruction error, while the second one penalizes solutions with many non-zero components. Sparsity of representation $\mathbf{a}$ is defined as $\kappa=k/m$ where k is the number of non-zero components. The resulting code is overcomplete, if $m>n$ and the difficulty of finding a sparse code with minimal reconstruction error depends on the level of overcompleteness (m/n) and κ. Parameter β controls the trade-off between the two terms. The reconstruction error or residual may be due to different noise sources that hide the structure of generating sources of the signal.

At one step the basis set is adjusted (*learning process*) to minimize the reconstruction error while the activity of the coding units, $\mathbf{a}$ is kept fixed. The straightforward solution would be to let evolve $\mathbf{D}$ by stochastic gradient on the cost function derived from the reconstruction error, $\mathbf{e}=\mathbf{x}-\hat{\mathbf{x}}$ where $\hat{\mathbf{x}}=\hat{\mathbf{D}}\hat{\mathbf{a}}$ and 'hat' denotes the actual estimation. Because of the role of the reconstruction error, this rule is not directly local [17], yet it can be translated [18] into a *set* of Hebbian (local) interactions realized by particular network structures with feedback.

During the selection of non-zero units (formation of the sparse code), features ($\mathbf{D}$) are fixed. However, selection by exhaustive search is a combinatorially hard problem [19]: the number of iterations becomes computationally prohibitive as m (the dimension of the internal representation) increases. For this reason several approximation method exist, but they either have slow convergence or provide non-optimal solutions. To overcome these limitations, we have chosen a heuristics that combines two approaches. The so called Subspace Pursuit (SP) method [20–23] has been chosen because of its superior speed. It is a generalization of matching pursuit [24], which finds local optima in a fast iterative fashion. Importantly, this method is able to discover the global optimum provided that certain conditions are met. Numerical experiments on natural visual stimuli indicate that methods, which assume these conditions, work surprisingly well [25], even though the conditions are unlikely to be met (but see [26] on the inherent limitations of matching pursuit like methods). In contrast to SP, the other algorithmic component – the so called

Cross Entropy method (CEM) [27] – is an optimization method designed to find the global optimum. Its main limitation is the slow convergence rate. The combination, termed Subspace Cross-Entropy (SCE) [16] method inherits the best of both worlds: it is reasonably fast and still can yield the optimal solution even at a higher level of overcompleteness. Since we are interested in the formation of sparse codes at very high level of overcompleteness, we used SCE in our numerical experiments. The appendix contains the pseudocodes of SP, CEM and SCE for the sake of reproducibility. Detailed analysis of these methods can be found in [16,28,29].

Improving Overcomplete Sparse Coding

The learning process of Eq. (1) is prune to perturbations: excess activation caused by noise may induce changes in all features thus introducing global (long-range) and low spatial frequency correlations among the features. Such unwanted increase in the number of active synapses implies increased metabolic cost.

Observation noise (e.g. induced by intrinsic neural activity) can significantly decrease the efficiency of OSC as it may easily generate access activation at the output (representation) level, which can only be mitigated by a number of further iterations in order to reduce the reconstruction error. In turn it is essential to counter this effect by actively controlling the number of non-zero components of the filters. This constraint is referred to as *structural sparsity* and implies that visual RFs with *local*, i.e., spatially restricted responses (like the high frequency, concentric RFs of the retinal ganglion cells, the relay neurons in the LGN, or the elongated oriented Gabor patch like RFs of the simple cells in V1) are metabolically more favorable over those that have large global structure with many synapses involved [30]. Approaches like weight thresholding or increasing overcompleteness (see Discussion) fail to address this issue properly. Instead, we turn to an alternative approach by directly separating global (involving many synapses), i.e., *low-frequency* or long-range components of the stimuli *before* the actual sparse coding. Considering the famous $1/f$ frequency fall of the amplitude spectrum of natural images [31], the low-frequency components carry most of the energy. Principal Component Analysis (PCA, [32], often used decorrelation method), for example, represents the signal in a way that the first component would carry the largest amount of energy, while the last one would carry the least amount. In turn, by applying PCA and then projecting the data *out of the subspace* of the first principal components would yield a representation without the unwanted low-frequency content. Let us remark that this approach is in contrast to conventional thinking which would keep exactly those components with high energy and filter out the rest. While this idea is appealing, PCA based separation of the subspaces strongly depends on the signal statistics: components (''outliers'') with heavy tailed amplitude distribution (characteristic to natural stimuli) can easily break down PCA. In the next section we review a robust alternative to PCA, which can efficiently separate these outliers from the low frequency components. We then propose an overcomplete SC model in which SCE (or any other efficient SC solution) is complemented by this alternative prefiltering as it is expected to support structural sparsity in the subsequent SC stage.

Two-stage overcomplete SC with structural sparsity

Our concept is based on recent findings of signal processing about recovering low-dimensional data from high dimensional observations [33]. In signal processing, conventional analysis of large dimensional data, such as sensory observations, is often based on the assumption that data have low intrinsic dimensionality: they lie on a low-dimensional subspace. In ℓ_2 norm (the ℓ_p-norm of vector $\mathbf{a}=(a_1,\ldots,a_m)^T\in\mathbb{R}^m$, where T stands for transposition, is defined as $\|\mathbf{a}\|_p \stackrel{\Delta}{=} (\sum_{i=1}^m \|a_i\|^p)^{1/p}$), PCA provides rank-$k$ estimate of the data by solving the following problem:

$$\mathbf{X} = \mathbf{L}_{PCA}+\mathbf{S}_{PCA} \tag{2}$$

$$\text{minimize} \qquad \|\mathbf{X}-\mathbf{L}_{PCA}\|_2 \tag{3}$$

$$\text{subject to} \qquad \text{rank}(\mathbf{L}_{PCA})\le k \tag{4}$$

where $\mathbf{X}=(\mathbf{x}^1,\ldots,\mathbf{x}^t)\in\mathbb{R}^{n\times t}$ is the matrix of observations (dimension of the observations: n, number of data points: t), rank of matrix $\mathbf{L}_{PCA}$ is k at most and $\mathbf{S}_{PCA}$ models a small noisy perturbation of each entry $\mathbf{L}_{PCA}$. If this perturbation is Gaussian noise, then PCA provides the statistically optimal estimate of the low-frequency, low dimensional subspace $\mathbf{L}_{PCA}$. However, deviation from the Gaussian (e.g. gross perturbations or components with heavy tailed distribution) can easily yield incorrect estimates.

Because of the $1/f$ frequency dependence natural stimuli often contain outliers and thus we need an alternative signal model. Let matrix $\mathbf{L}$ comprise the low frequency components (so it has low-rank as above), while $\mathbf{S}$ may have full rank, but it is a *sparse* matrix with arbitrarily large entries at random locations: $\mathbf{X}=\mathbf{L}+\mathbf{S}$. The surprising result is that under certain conditions (on the rank of $\mathbf{L}$ and on the sparsity of $\mathbf{S}$) *both* matrices can be *exactly* recovered [33]. Furthermore, it has been proved that efficient recovery is feasible by solving the following optimization problem (Robust Principal Component Analysis, RPCA):

$$\text{minimize} \qquad \|\mathbf{L}\|_*+\lambda\|\mathbf{S}\|_1 \tag{5}$$

$$\text{subject to} \qquad \mathbf{X}=\mathbf{L}+\mathbf{S} \tag{6}$$

where $\|\mathbf{L}\|_*$ denotes the *sum* of the singular values of $\mathbf{L}$, $\|\mathbf{S}\|_1$ denotes the ℓ_1 norm of matrix $\mathbf{S}$, i.e., $\|\mathbf{S}\|_1=\sum_{i=1}^t\sum_{j=1}^n|S_{ji}|$. λ is a trade-off parameter, which governs the dimension of matrix $\mathbf{L}$. On the other hand, matrix $\mathbf{S}$ may assume maximal rank, independent of λ.

In addition to robustness against perturbation, the proposed decomposition allows an alternative interpretation of the signals. Instead of treating sparse components as corrupting noise to be filtered, we may consider these outliers as *atypical signals* that carry further information about higher order correlations (like configurational information) not revealed by the low-rank estimate ($\mathbf{L}$). Note that conventional methods (like ICA) would analyze the low rank part only.

The suggested solution (the pseudocode is given in Table 1) iteratively improves the estimation of $\mathbf{L}$ and $\mathbf{S}$ and its computational complexity is only slightly larger than that of the traditional PCA [33]. Another surprising result is that under the assumptions of the theorem, a whole range of λ values can return the correct solution, no matter what $\mathbf{L}$ and $\mathbf{S}$ are. A simple reference value for λ is $\lambda_0=\sqrt{\max(n,t)}$ [33] and so we will use a normalized parameter: $\lambda^*=\lambda/\lambda_0$.

Interestingly, as numerical experiments suggest [33], RPCA delivers meaningful signal decomposition even if conditions (about the sparseness of $\mathbf{S}$) do not hold (like in the case of $1/f$ spectra). In these cases, however, different RPCA decompositions can be obtained by setting different λ^* values and $\mathbf{S}$ is not guaranteed to be sparse anymore. For this reason matrix $\mathbf{S}$ could be the subject of further sparsification. The corresponding sparse coding optimization (see Eq. (1)) in matrix form is given as

Table 1. RPCA pseudo-code.

initialize:
$\mathbf{S}_0=\mathbf{Y}_0=0$, $\mu>0$
while not converged do :
compute :
$\mathbf{L}_{k+1}=\mathcal{D}_\mu(\mathbf{X}-\mathbf{S}_k-\frac{1}{\mu}\mathbf{Y}_k)$
$\mathbf{S}_{k+1}=\mathcal{S}_{\lambda\mu}(\mathbf{X}-\mathbf{L}_{k+1}-\frac{1}{\mu}\mathbf{Y}_k)$
$\mathbf{Y}_{k+1}=\mathbf{Y}_{k+1}+\mu(\mathbf{X}-\mathbf{L}_{k+1}-\mathbf{S}_{k+1})$
end while
output : $\mathbf{L}$, $\mathbf{S}$.

$\mathcal{S}_\tau:\mathbb{R}\to\mathbb{R}$ denotes a shrinkage operator, $\mathcal{S}_\tau[x]=\mathrm{sgn}(x)\max(|x|-\tau,0)$ acting on matrices componentwise. For matrix X, $\mathcal{D}_\tau$ denotes the singular value threshold operator: $\mathcal{D}_\tau(X)=U\mathcal{S}_\tau(\Sigma)V^*$, where $X=U\Sigma V^*$ is the singular value decomposition.

doi:10.1371/journal.pcbi.1002372.t001

$$\min_{\mathbf{D},\,\mathbf{A}\in R^m}\sum_{i=1}^{t}\frac{1}{2}\|\mathbf{S}-\mathbf{D}\mathbf{A}\|_2^2+\beta\|\mathbf{A}\|_0, \tag{7}$$

where the matrix $\mathbf{S}=[\mathbf{s}^1,\ldots,\mathbf{s}^t]\in\mathbb{R}^{m\times t}$ and $\mathbf{A}=[\mathbf{a}^1,\ldots,\mathbf{a}^t]\in\mathbb{R}^{m\times t}$, denotes the matrix of the outliers and the matrix of their sparse representations, respectively. The ℓ_2 norm based residual may denote full rank observation noise, which implies the following signal model: $\mathbf{X}=\mathbf{L}+\mathbf{S}+\mathbf{N}$. According to [34], it is still possible to give stable estimates for $\mathbf{L}$ and $\mathbf{S}$, if $\mathbf{N}$ is bounded: $\|\mathbf{N}\|_F<\delta$, for some $\delta>0$ value, where $\|\cdot\|_F$ denotes the Froebenius norm. In the demonstrations we opted to use the simpler RPCA model (as in Eq. (6)) without explicit assumptions about the additive noise term.

Let us note that even though the formalism used above is based on matrices, the RPCA procedure can be applied on a single input (thus it may be realized in a neurally plausible form) once an approximation of the low-rank part $\mathbf{L}$ is available. Furthermore, – depending on the input statistics – $\mathbf{L}$ can be approximated even from partial observation by 'filling in' missing information [33,35].

Computer experiments

To test the impact of RPCA preprocessing on sparse coding, normalized natural image patches were first decomposed by RPCA at different λ^* values, then the resulting full rank representations were further encoded by SCE (16-fold over-completeness with $n=16\times16=256$ dimensional inputs and $m=4096$ dimensional representation; numerical details are in the Methods Section). We have chosen this particular input set since there already exist a number of computer vision studies on their statistics and the corresponding neural representations under different optimality criteria [14,31]. The actual overcomplete sparse representations were formed by SCE and the corresponding SC filters were tuned online via stochastic gradient learning. While this level of overcompleteness is still below what has been estimated in the neural sensory systems [15], we believe it is a reasonable choice, as training time is still manageable, yet the results are convincing enough to support the central message of our proposal.

A few basis features (for sparse coding, 10 out of 4096 columns of matrix $\mathbf{D}$) are shown on Figure 1. For visualization purposes each basis vector is scaled into the range [0,1] and displayed as a 16×16 image. Features in the first row of Figure 1A were obtained by conventional SC (applying SCE) without pre-filtering, which corresponds to the case of $\lambda^*=0$.

As we earlier argued, plain SC tends to learn large, global filters, thus preventing the reduction of synaptic cost. Figure 1B plots a few selected SC features when applied on the residuals of traditional PCA. Regarding locality we do not see much improvement: features are still global and manifest large, wavy structures. Figure 1E depicts example filters obtained by applying RPCA prior to SC. Different rows correspond to different λ^* values. The main result of these studies is that the learned basis features get cleaner and more localized, that is, filters get *structurally* sparser as the *single* global parameter increases. On Figure 1F we re-plotted features for $\lambda^*=0.8$ together with the corresponding filters approximated by reverse correlation. Not only the estimation error is smaller compared to the error of the native SC method (Figure 1A), but filters also show larger diversity in their shapes, similar to what has been found experimentally [8]. We also plotted the corresponding filters or RFs of the low-rank signal $\mathbf{L}$ in Figure 1C for $\lambda^*=0.5$, when the number of basis vectors was 17. Figure 1D shows the spatial-dependence of RFs of the sparsified signal $\mathbf{S}$ after RPCA for $\lambda^*=0.5$.

A surprising result is that the shape of all the obtained RFs for sparsified matrix $\mathbf{S}$ can be described as 'Difference of Gaussians' which is the characteristic RF shape [36] of the retinal ganglion cells and the neurons in LGN. The obtained concentric filters 1, are homogeneous and 2, uniformly tile the whole space. Due to their similarity, we show the cross-section of one unit only (Figure 1D). Note that the peaky structure is due to the small image size (discretized DoGs have similar shape at this scale) and more typical DoG shapes could be obtained for larger image patches. We found that for higher λ^* values the negative basin around the peak gets deeper. This development may correspond to the experimentally found developmental changes of the LGN filter profiles in cat [37].

Let us emphasize again that RPCA is not a projection: through an iterative process it extracts the large and sparse components and separates the low-rank part. Interestingly, for natural images, RPCA provides a basis visually almost indistinguishable from those of the PCA filters, but the corresponding representations are different. It implies that PCA may be a good first approximation or initialization for the RPCA iteration method (higher λ^* values allow more low-dimensional components).

Qualitative comparison between filters and RFs

Traditionally, a simple cell RF in V1 is often characterized as a 'Gabor-patch' [38]; Gaussian envelope around a cosine wave. To help compare the obtained filters with RFs of real neurons, we also approximated the filters as a Gabor-patch. As λ^* increases the filters become more localized and cleaner, and the Gabor-patch like appearance gets more pronounced. On the other hand, at too large values the filters become small and stereotyped with diminishing harmonic content (see Figure 1E).

The distribution of the shape parameters of the Gabor-patch approximations (Eqs. (9)–(11)) is shown in Figure 2 for $\lambda^*=0.8$. Filters localized at the edges of the 16×16 visual space were discarded as their distortion prevents proper fitting. For small filters fitting is imprecise. Filters yielding Gaussian envelope with width less then 0.3 pixel were thus also discarded. It implies that the true number of learned filters at around point $n_x=0,n_y=0$ is larger than what is shown in Figure 2. Visual inspection reveals that (i) filters become local and cleaner, (ii) the distribution deviates significantly from the bisection line, and (ii) a considerable portion

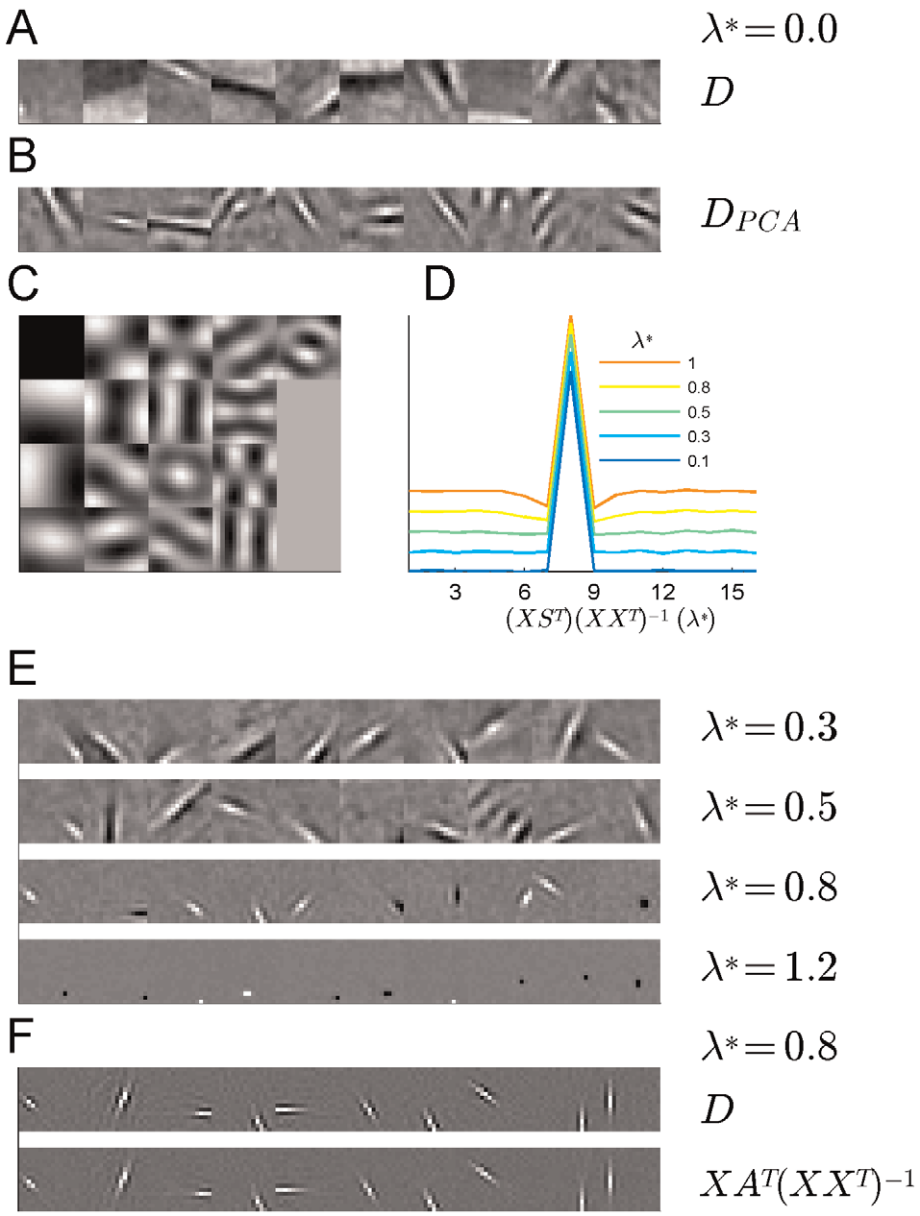

Figure 1. Different basis types of RPCA preprocessing and Sparse Coding. Sample receptive fields are scaled into range [0,1]. (A) no RPCA, columns of dictionary $\mathbf{D}$. (B) receptive fields learned after PCA pre-filtering: features show wavy, global structure. (C) Features ('global filters') of the low dimensional signal for the case $\lambda^*=0.5$ (dimension = 17). (D) reverse correlation of the full rank sparsified signal $\mathbf{S}$ yields stereotypical DoG-like filters with symmetric 2D structure. The figure shows the profile of the central section as a function of λ^*. At higher values the negative basin around the peak gets deeper. (E) Randomly selected sparse coding filter sets (over-completeness is $16\times$, $\lambda^*=0.3,0.5,0.8$ and 1.2) With increasing λ^* the filters get smaller and more localized (i.e. *cleaner*). (F) For comparison, a set of sparse coding filters ($\mathbf{D}$) and the corresponding linear approximations (normalized reverse correlation, $(\mathbf{XX^T})^{-1}\mathbf{XA}^T$) are shown at $\lambda^*=0.8$.
doi:10.1371/journal.pcbi.1002372.g001

of the filters is concentrated near the origin $n_x=0,n_y=0$. For comparison, we also plotted the distribution of the fitted shape parameters of the experimentally measured RFs of simple cells reproduced from [8]. Considering that we had to drop a number of small filters, the match between numerical and experimental data seems quite good (see, e.g., [39] for comparison), indicating that the proposed model may have biological relevance. Let us note that the observed shape distribution may depend on the level of overcompleteness, but due to the relatively small input size we suspect that further increase in the number of coding units would not result in major changes.

Numerical analysis of the prefiltering and sparse coding stages

Since the assumed signal model is only an approximation for natural image patches, different trade-offs (defined by λ^* in Eq. (5)) between the contribution of the typical and atypical features to the reconstruction influence the emerging representations after RPCA prefiltering. Figure 3A depicts the influence of λ and thus the RPCA decomposition on the statistics of the SC filter shapes as measured by the histogram of the Gabor-patch fitting error. It shows how well the linear approximation of sparse coding filters can be described with a set of oriented Gabor patches often used to

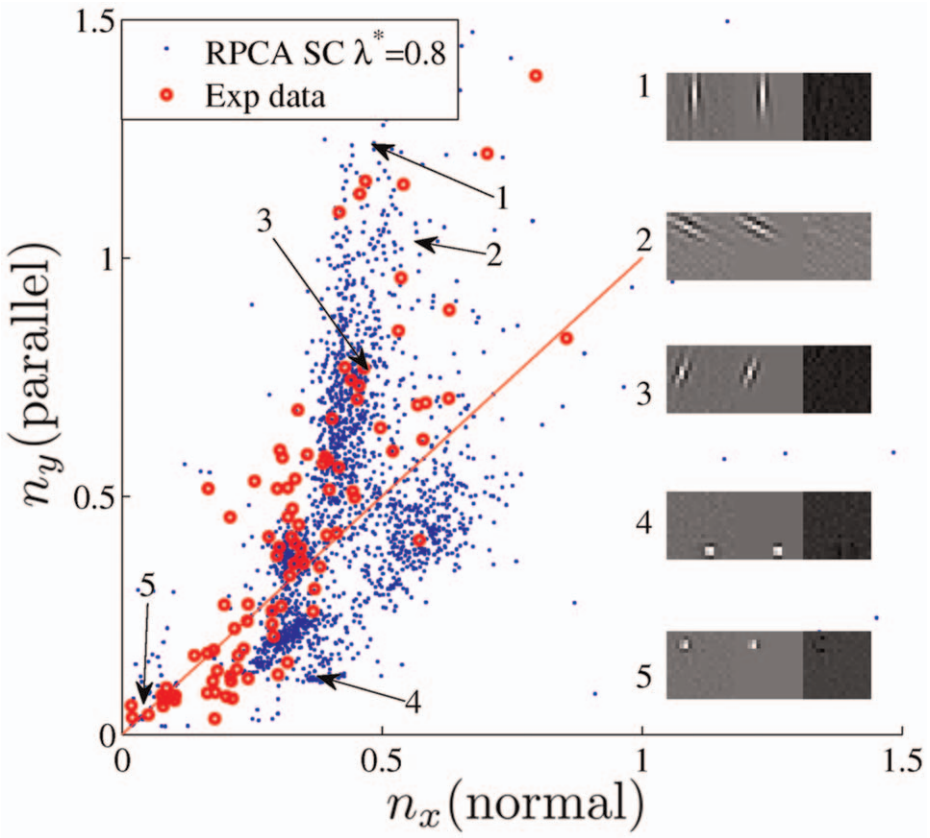

Figure 2. Distribution of the shape parameters for the model and for the experimental data. Receptive fields of simple cells in primary visual cortex, linearly approximated by spike triggered averaging. Data [8] are available at http://web.mac.com/darioringach/lab/Data.html. Our model filters show significant diversity in the fitted shapes similar to what has been found experimentally. While other models (e.g. [39,40]) are also able to partially match the filters to the observed RFs, a significant difference is that our model uses highly overcomplete representations. For other differences, see the main text.
doi:10.1371/journal.pcbi.1002372.g002

characterize experimentally measured receptive fields. If filters have 'dilated' global structure then the histogram of the fitting error is probably less peaked. And indeed this is the case: increasing λ^* results in more homogeneous, smaller and point-like filters. Let us remark that discretization has a strong contribution to the observed fitting noise.

Figure 3B displays the dependence of the dimension of the low-rank component $\mathbf{L}$ on λ^* and the relative contribution of $\mathbf{L}$ to the reconstruction of the original observations. To calculate the intrinsic dimension of $\mathbf{L}$, all singular values were zeroed out with amplitude less then 10^{-6} of the maximal amplitude. The important parameter range is where the intrinsic dimension is still low, yet $\mathbf{L}$'s role in the reconstruction is significant. Within that range, $0.5<\lambda^*\leq 0.8$ provides the best fit to the experimental data. At higher λ^* values most of the filters loose their edge-like characteristics.

We have also studied the algorithm's reconstruction ability. Due to the additive decomposition, reconstruction depends on both the "typical" part obtained by RPCA and the overcomplete sparse representation of the "atypical part". As it is demonstrated on Figure 3 the relative contribution of $\mathbf{L}$ as well as its dimension (number of coding neurons) depends on λ^*. In turn, the fidelity of reconstruction is a function of both the number of units that encode typical features and the number of nonzero entries in the sparse code. Figure 4 displays this dual dependence: reconstruction quality as a function of the total number of nonzero entries, which comprises the rank estimate of $\mathbf{L}$ at the given λ^* and the preserved number of nonzero entries in the overcomplete sparse representation (k). For $\lambda^*=0.0,0.3,0.5$ the chosen values were: $k=16,32,64,80,96$ and for $\lambda^*=0.8$, $k=8,16,32,64,80$. Reconstruction quality is measured by mean SNR: $<10\log_{10}\frac{|x^i|_2}{|x^i-l^i-\hat{s}^i|_2}>_i$. Interestingly, while SNR does not improve much when λ^* has changed from 0.3 to 0.5, the corresponding filters have significantly changed. Let us note that the overall low values of SNR are due to the fact that no high frequency components have been filtered out prior to decomposition (but see [40], where much higher SNR has been reported after filtering out those high frequency components).

So far we have dealt with static images, but temporal sequences are more realistic: sensory systems are believed to adapt to the spatio-temporal structure of the stimuli. Since RPCA does not rely on prior knowledge about the spatial or temporal arrangement of the data, one expects to see similar decomposition results for data with temporal correlation. For the sake of illustration, temporal correlation was introduced by concatenating 16 image patches of size 8×8 extracted from image sequences on natural scenes. (This was the maximum size we could handle with overcompleteness ratio 16.) Sample filters of the obtained low-rank matrix $\mathbf{L}$ for $\lambda^*=0.5$ (the corresponding rank estimate is $r=69$) are shown on the left of Figure 5. Filters are ordered by their corresponding eigenvalues. Each filter is composed of 16 frames of size 8×8 pixels. Similar to the filters shown on (Figure 1C), these

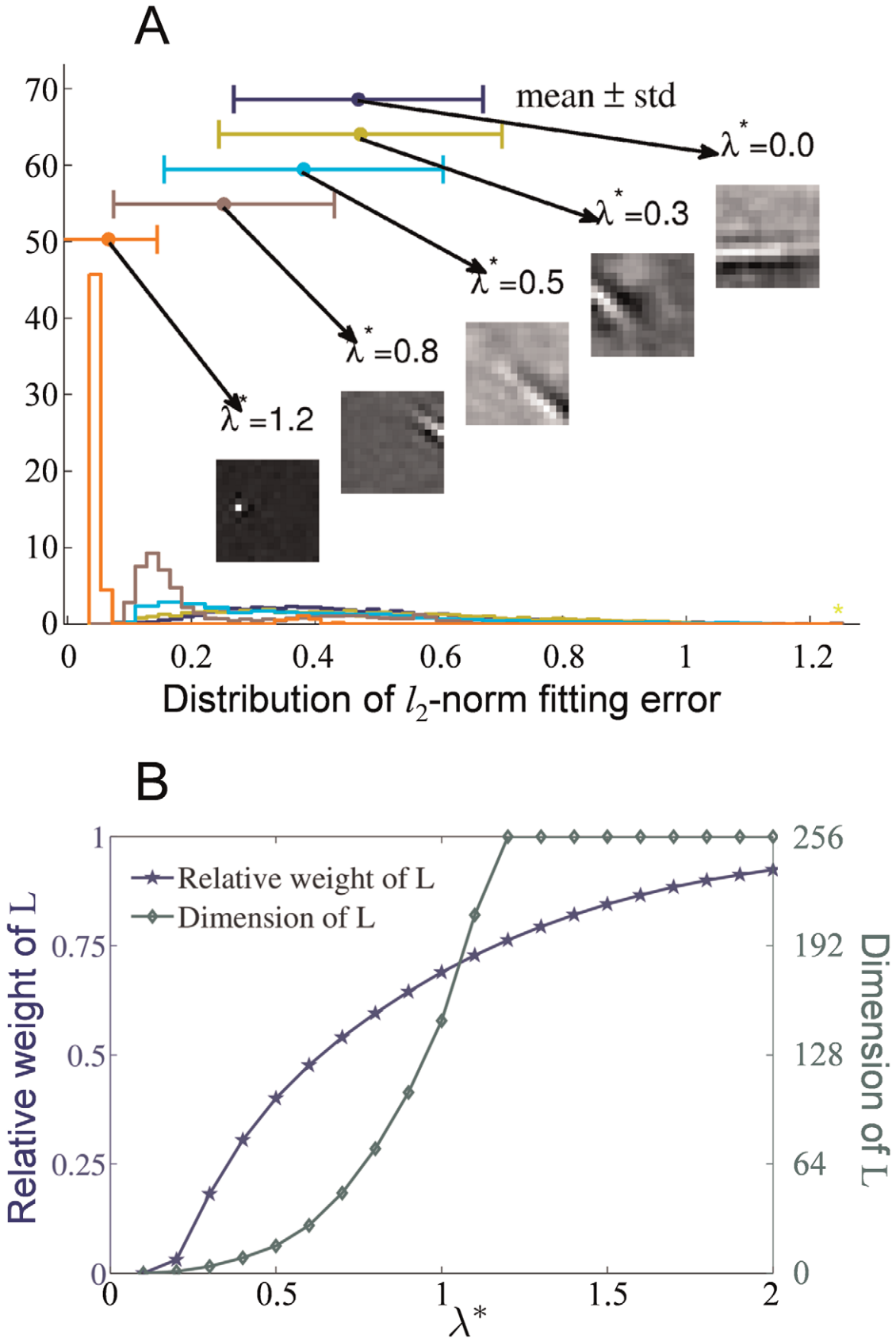

Figure 3. The impact of λ^* on the signal decomposition and the overall quality of the sparse coding filters. (A) The empirical distribution of the Gabor patch fitting error as a function of λ^*. Larger spread signifies deviation from ideal Gabor patch, often used as model shape for experimentally recorded receptive fields. The shift of the mean toward 0 as λ^* increases is a consequence of the decrease of the average filter size.

For each mean value a sample filter is shown demonstrating this shrinkage effect. (B) The dimension and the relative weight of L (the low dimensional signal) in the reconstruction as a function of λ^*. Relevant range is where the dimensionality is low, yet L is able to capture most of the original signal. For image size 16×16 this range is about 0.3–0.8.
doi:10.1371/journal.pcbi.1002372.g003

filters can also be characterized by low spatial and temporal frequency.

The corresponding filters of the atypical parts (**S**, not shown) - as in the static case- are homogeneous, localized in space and time and uniformly tile the visual space. Furthermore, they show Mexican hat like characteristics in the temporal dimension. The regularity may be due to the particular concatenation method we chose.

Sparse coding filters can also be derived from the overcomplete sparse representation of the image sequences after RPCA decomposition. As representations are temporally decorrelated, we obtained filters strongly localized in space and time which resemble to some extent to the receptive field dynamics of simple cells of V1 [41]. A sample set of the obtained sparse coding filters are shown on the right of Figure 5.

It is expected to get better match with experimentally found filters if temporal correlations are introduced into the data model by convolution [42,43] as opposed to simple concatenation and if nonlinear response properties and nonlinear dynamic interactions are included to handle time warping, for example. These studies go beyond our present goals.

Discussion

While the resemblance to the biological system is appealing, the original motivation behind applying RPCA was to find means to facilitate the formation of overcomplete sparse representations, an important feature of neural processing that significantly boosts computational efficiency. As we previously argued, *structural* sparsity is needed to control the underlying metabolic cost of the formation of large, overcomplete sparse representations. In principle this control could be realized in different ways. The most straightforward solution would be weight thresholding by zeroing out all filter components (synaptic weights) below an arbitrary threshold value. However, this intuitive regularization may cause more problems than it solves. First, it introduces error for coordinates near zero, e.g. at zero crossing of the response function of a simple cell. In addition, it does not support adaptivity as it may eliminate gradual learning of less frequently represented features. At last it strongly depends on the arbitrary threshold parameter irrespective of the actual input.

Another approach would be to further increase overcompleteness as it might implicitly reduce the number of required

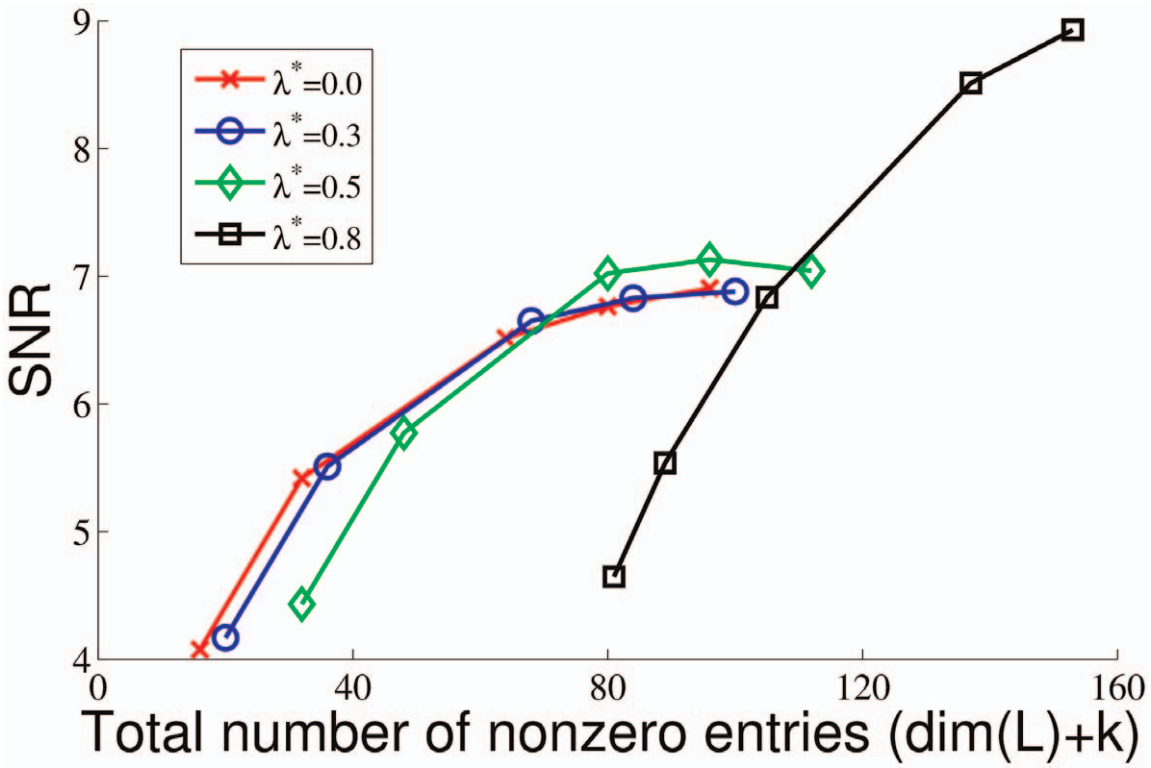

Figure 4. Reconstruction quality as a function of the number of nonzero coding units and λ^*. Reconstruction quality is measured by mean SNR: $<10\log_{10}\frac{|x^i|_2}{|e^i|_2}>_i$, where i runs over the inputs. Since RPCA is an additive decomposition, the reconstruction error is given as $e^i = x^i - l^i - \hat{s}^i$. The total number of nonzero entries is given as the sum of the rank estimate of **L** and the preserved number of nonzero units (k) in the sparse overcomplete representation of the atypical part (**S**) of the RPCA output. Since sparseness level is automatically set by SCE, the following arbitrary values for k were chosen. For $\lambda^* = 0.0,0.3,0.5$ $k = 16,32,64,80,96$ and for $\lambda^* = 0.8$, $k = 8,16,32,64,80$.
doi:10.1371/journal.pcbi.1002372.g004

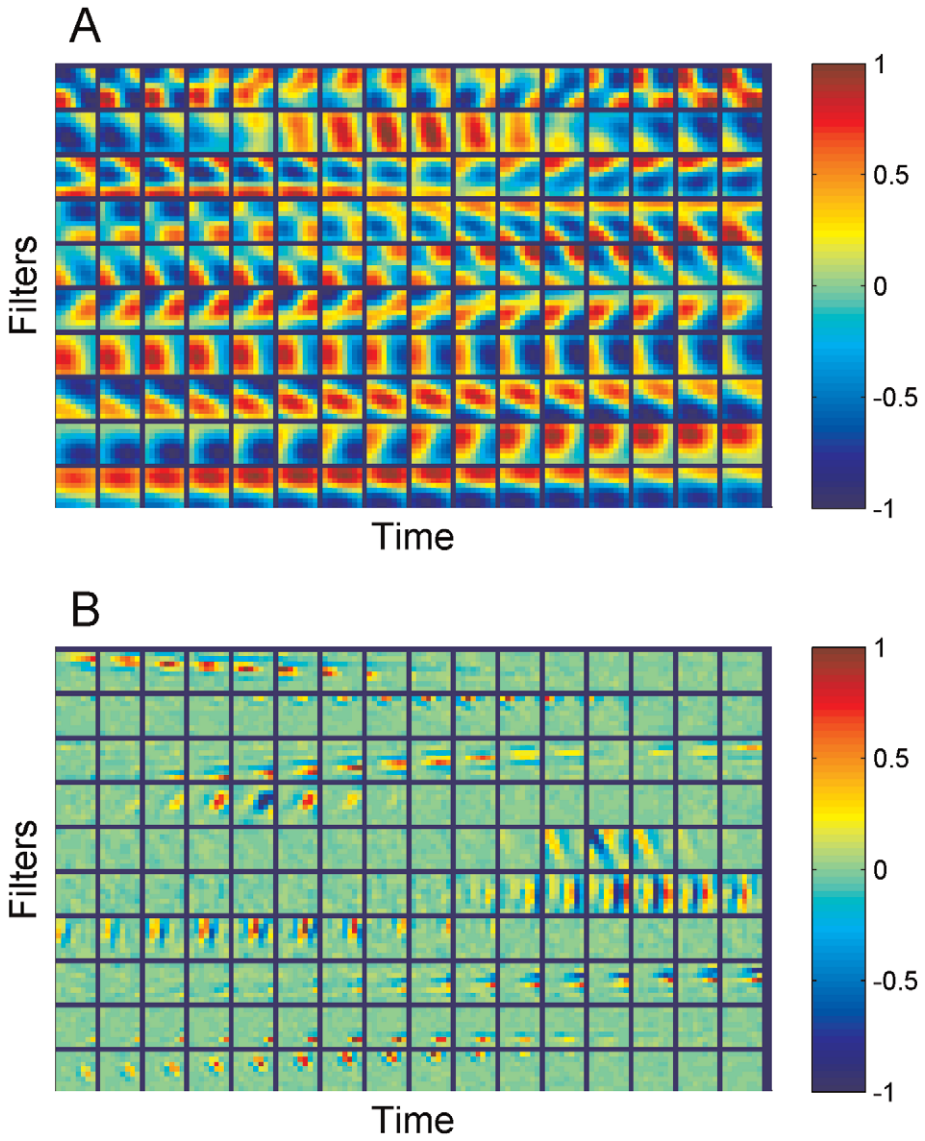

Figure 5. RPCA on concatenated image sequences. Left: The first 10 spatio-temporal filters of the low rank signal,L (rank $r = 69$) are shown. Each filter is shown as a sequence of 16 frames of size 8×8 pixels. It can be seen that there are spatio-temporally separable as well as non-separable filters. All filters correspond to low frequency temporal or spatial changes Right: 10 selected spatio-temporal filters of the corresponding overcomplete sparse codes that display different spatio-temporal localization and dynamics. While many filters are similar to the presented ones, more training would be needed to achieve similar locality for the majority of filters at this input dimensionality (8×8×16) and level of overcompleteness (16×).
doi:10.1371/journal.pcbi.1002372.g005

components (increased sparsity). However, this idea does not work [44]: when tested on natural images, many filters still show global structure.

We propose RPCA as a particular prefiltering stage prior to the actual sparse coding which indeed facilitates structural sparsity and preserves many useful properties of conventional PCA based decorrelation without its noise sensitivity. Our model may thus resolve the controversy between the hypothesis that PCA like decorrelation should precede subsequent transformations and the fact that the identified RFs cannot be generated by PCA.

Although the proposed RPCA based sparse coding mechanism does not have a biologically feasible implementation yet, its functional relevance may be supported by the following arguments.

The robustness of RPCA has been demonstrated [33] by showing that RPCA yields meaningful representations for different data sets even if the composite signal model cannot be validated (e.g. separation of background (typical) and moving objects (atypical, outstanding features) or separation of face and shadows caused by anisotropic illumination). In particular, for natural stimuli with characteristic 'scale-free' statistics (cf. '1/frequency' relation) the conditions of the RPCA theorem are definitely not met as the distinction between low-rank and sparse parts cannot be clearly defined. It may imply that a step-wise incremental separation would be better suited for the input statistics instead of the single layer iterative arrangement of RPCA.

Another important finding is that the RPCA theorem of [33] can be related to recent results on the problem of Exact Matrix Completion [35], which claims that *typical regularities* of a composite signal (represented by columns of $\mathbf{L}$) can be completed even from a *small* set of randomly sampled (or partially observed) coordinates of the input. This "sampling advantage" would also improve energy efficiency.

While our model is implicitly supported by the emerging filters, alternative models are also claimed to explain early vision by learning similar features. For this reason we briefly compare a few competing sparse coding models with our proposal.

Receptive field properties of sparse coding models

The biological relevance of neural coding models is often judged by the similarity between their filtering properties and the receptive fields of the corresponding neurons. In the case of visual stimuli, one of the criticisms against theory driven (functional) models (e.g. Independent Component Analysis [4] or Sparse Coding [5]) is the lack of diversity in the filter shapes [8]. This failure might be due to the missing prefiltering stages as seen in the visual pathway. However, nave use of different, biologically motivated prefiltering methods does not seem to offer any improvement, either. For example, applying DoG as high-pass filtering is expected to enhance edge-like features thus yielding a shift of the Gabor-patch shape parameters toward higher values, but the structure of the shape distribution barely changes. Another example is the use of PCA to filter out global features before SC (or ICA), which yields wavy SC basis (Figure 1B). Furthermore, not all filters in V1 have elongated bar shape and most models fail to yield close to concentric shapes found experimentally (for a discussion, see e.g. [39]). As the filter shape distribution on Figure 2 shows, when applied on natural images, RPCA preprocessing *together with* SC delivers the required diversity including the close to concentric shapes. It is worth noting there are other improved coding models (in particular, [40] and [39]) that also claim similarities between the observed and predicted shape distributions of the fitted filters. Our model is similar in spirit to the functional model of [40], whereas the other approach [39] describes a self-organizing system governed by complex dynamics and feedforward inhibition. While the latter one is a promising approach, its dynamics is quite involved and its parameter sensitivity is not known. The other model of [40] is also a sparse coding model and it uses greedy, iterative solutions as mentioned previously. It also uses prefiltering similar to that one used in [5]. They claim the obtained similarity is due to the particular sparsity constraint. For the similar motivations let us remark some differences between the model of [40] and the one proposed here. First, we believe their approach may not be suited to handle large overcompleteness for reasons discussed previously about greedy solutions. Second, the reported difference between the signal to noise ratio of their method and our model is likely due to two factors: we did not employ prefiltering and the overcompleteness in our case is larger. Less sparse codes can encode signals more faithfully then. A fair comparison would be to see the quality of the reconstruction of the high frequency components from sparse codes ($\mathbf{A}\rightarrow\mathbf{S}$), but such comparison would depend on both sparsity and overcompleteness. In turn, an intriguing issue is the optimality of reconstruction quality with respect to the energy consumption. Interestingly, as Figure 6 demonstrates, the linear approximation of the filtering properties of RPCA (seen as the amplitude spectrum of the ''atypical'' signal part of the RPCA output) looks quite similar to what an ideal whitening filter would yield. This similarity may have the following consequences. First, their result may be attributed both to the particular form of the filter and to the chosen form of sparse coding. Furthermore, it might be the case that such prefiltering behaves as a fast approximation to RPCA. Another difference to mention is that our two-stage model not only provides oriented band pass filters, but it also yields DoG-like filters at the RPCA pre-filtering stage thus providing a simultaneous explanation of two processing stages of early vision. Interestingly, as Figure 6 demonstrates, linear approximation of the filtering properties of RPCA (seen as the amplitude spectrum of the ''atypical'' signal part of the RPCA output) looks quite similar to what an ideal whitening filter would yield. This similarity may have the following consequences. First, results of [40] may be attributed both to the particular form of the filter *and* to the chosen form of sparse coding. Furthermore, it might be the case that such prefiltering behaves as a fast approximation to RPCA.

Biological implementation of RPCA based sparse coding

The qualitative agreement between the filtering properties of the early stages of vision and our two-stage algorithm may allow us to attempt to map the algorithm onto the neural substrate by linking the different computational functions to anatomical areas.

An important property of our model is that prefiltering requires a dual representation of the stimuli, which assumption is not in line with the current thinking of hierarchical sensory processing (e.g. [45,46]), which often comprises alternating filter and pooling operations. So how can we reconcile the assumption on dual representation with single stream models?

Since RPCA implies dynamic interaction between the two emerging representations of the typical (global) and atypical (local) features, decomposition requires either a recurrent network with distinct sub-populations of neurons or two layers with feedforward and feedback connections. As retina does not receive feedback modulations from downstream layers, DoG like filtering of the retinal ganglion cells is not a consequence of RPCA, but it may be explained as a facilitating approximation – as we argued about whitening above – before decomposition. LGN, on the other hand, receives massive amount of feedback from V1. Having learned the filters during early development, it can be assumed that LGN neurons can represent a proxy to the *atypical* features of single stimuli. This representation still contains information about the typical features (since clear decomposition of natural signals is unlikely, due to scale-free statistics). In turn, V1 has a two-fold role in processing. It holds the approximation of the global features extracted from the LGN output and it recodes or re-represents the atypical features in an overcomplete sparse form. A candidate for the first task could be a class of V1 interneurons characterized by large, global receptive fields with weak or no orientation selectivity (e.g. [47,48]). While it is possible to learn the low-frequency typical parts of new stimulus sets, RFs do not need to be continuously updated as they comprise the most typical correlations of natural images (short term adaptation to quick changes is still required). The second task of overcomplete recoding is then realized by simple cells. This setting thus allows for the alternating substraction of RPCA (Table 1) by the interaction between inhibitory neurons and simple cells in V1 and the neurons in LGN.

In summary, this paper presents a novel two-stage algorithm for efficient overcomplete sparse coding. The proposed robust extraction of low-frequency or typical correlations as a prefiltering step has a few remarkable properties that make the algorithm plausible as an important model of neural information processing. First, it supports the formation of overcomplete sparse codes by effectively controlling the transformation matrices (the synaptic weights) and reducing the number of active synapses. Second, the inclusion of RPCA could significantly facilitate perception as it allows the completion of the typical components even if a part of the stimuli is missing (undersampling, occlusion, cf. exact matrix completion). Since these properties may be beneficial for the nervous system, it would be interesting to see if our algorithm could be realized by biologically plausible neural computations.

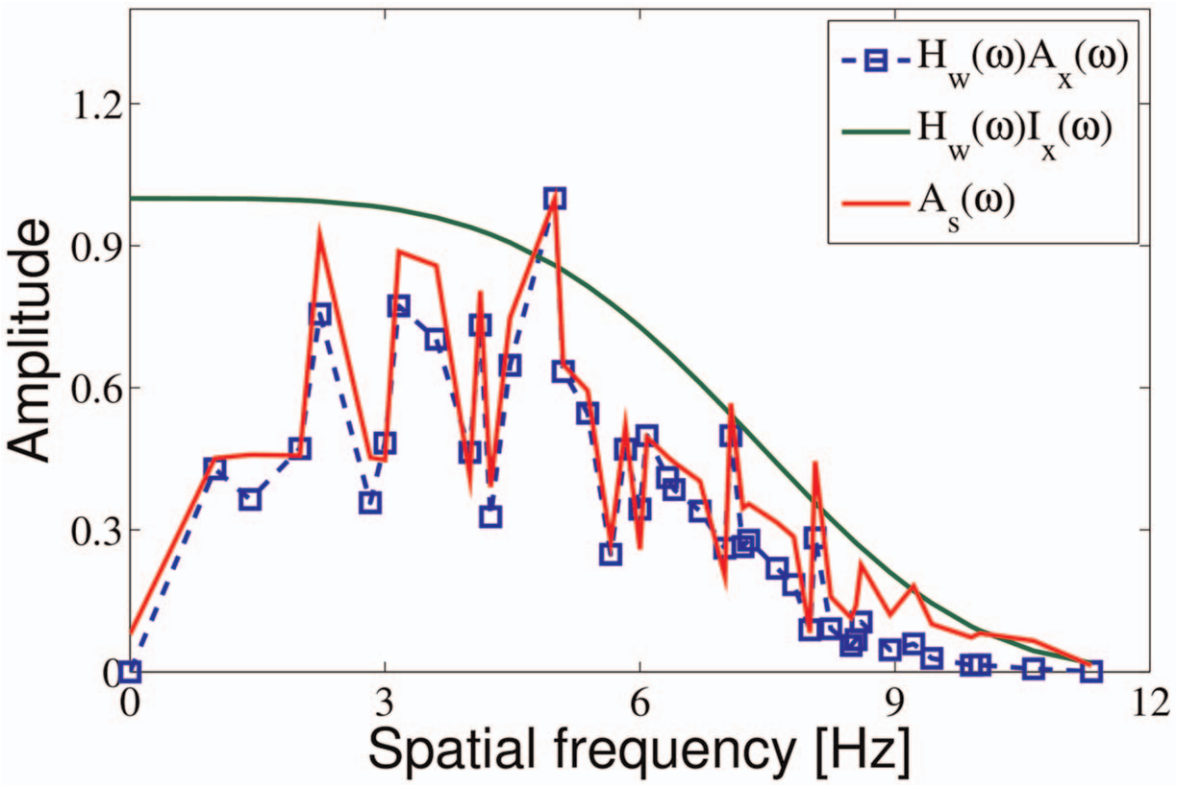

Figure 6. A comparison of the amplitude spectra of the ''atypical'' output part of RPCA, the whitened input and the whitened ideal input. This plot demonstrates that the particular whitening filter as used in [5,40] can be seen as a linear approximation of the filtering properties of RPCA when only the atypical output is considered. The thick (red) line is the amplitude spectrum of the RPCA output. The dashed (blue) line with square markers is the amplitude spectrum of the training images filtered with the whitening filter. The thin (green) line serves as a reference: this is the amplitude spectrum of whitened ideal input which has an amplitude spectrum proportional to 1/frequency. Due to the limited input size, there is a natural cutoff at higher frequencies. (Since the size of the images is 16×16, the largest frequency is $\sqrt{2}\cdot16/2$.) The whitening filter: $H(f)=fe^{-(f/f_0)^4}$, where the cutoff frequency is $f_0=8$. The variances of the plots are due the artifacts caused by the rectangular sampling lattice. For comparison purposes the plots are rescaled onto $[0,1]$.
doi:10.1371/journal.pcbi.1002372.g006

Methods

In this section we briefly present the algorithmic constituents of the subspace cross entropy method used to make overcomplete sparse codes. We also give a short algorithmic description of the RPCA implementation used in the simulations. Finally, details of the training data and the fitting methods are presented.

OSC Part I: Subspace Pursuit method, (SP)

Subspace Pursuit algorithms have been independently proposed in [23] and [49]. These methods assume that at most k components are sufficient to represent the input. The methods enlarge the subset of candidate features (''candidate subspace'') by k [23] (or $2k$ [49]) features and then decrease their number back to k at every iteration. The method of [23] is as follows (the pseudocode is given in Table 2).

First, a candidate representation is generated using all basis, then a subset of basis is selected that corresponds to the k largest components of the representation. This initial selection is then iteratively refined: the residual (that is the difference between the input and its current approximation) is calculated and mapped onto the representation space using the entire basis set again. Then – similar to the initial step – another k basis are selected based on amplitude of the corresponding components of the mapped residual. The original input is then projected again to the representation space using a $2k$ element basis set formed by fusing the two basis subsets. Finally k basis vectors are selected again that correspond to the k largest components of the projection (basis shrinkage). The iteration stops when the norm of the residual is sufficiently small. SP has superior speed, scaling and reconstruction accuracy over other iterative methods by directly refining the subset of reconstructing (active) components at *each* iteration. Its native shortcomings, though, are the heavy use of the costly encoding transformation of the residuals at each iteration and the preset number of active coding units.

OSC Part II: Cross-Entropy method, CEM

CEM is a global optimization technique [27] that finds the solution in the following form:

$$\mathbf{y}^* := \arg\min_{\mathbf{y}} f(\mathbf{y}).$$

where f is a general objective function.

While most optimization algorithms maintain a single candidate solution $\mathbf{y}(t)$ at each time step, CEM maintains a *distribution* over possible solutions. From this distribution, solution candidates are drawn at random. By continuous modification of the sampling distribution, random guess becomes a very efficient optimization method.

One may start by drawing many samples from a fixed distribution g and then selects the best samples as an estimation

Table 2. The pseudocode of the Subspace Pursuit method.

input:	
$\kappa = k/m$, $\mathbf{x} \in \mathbf{R}^n$	% sparsity and signal
t_{SP}	% max iteration number
$\mathbf{D} \in \mathbb{R}^{n \times m}$	% m column dictionary
initialization:	
$\mathcal{K} = \mathrm{MaxInd}_k(\mathbf{D}^T\mathbf{x})$	% index set of maximal amplitude elements with set size k
$\mathbf{D} = \mathbf{D}[\mathcal{K}]$	% sub-matrix belonging to index set $\mathcal{K}$
$\mathbf{r} \leftarrow \mathbf{x} - \mathbf{D}\mathbf{D}^{\dagger}\mathbf{x}$	% compute residual
optimization:	
for t from 1 to t_{SP}	% iteration main loop
compute $\mathrm{MaxInd}_k(\mathbf{D}^T\mathbf{r})$	% index set for expansion
$\mathcal{K} \leftarrow \mathcal{K} \cup \mathrm{MaxInd}_k(\mathbf{D}^T\mathbf{r})$	% increase set size to $2k$)
$\mathbf{e} \leftarrow \mathbf{D}[\mathcal{K}]^{\dagger}\mathbf{x}$	% compute projections
$\mathcal{K} \leftarrow \mathrm{MaxInd}_k(\mathbf{e})$	% new index set of size k
$\mathbf{D} \leftarrow \mathbf{D}[\mathcal{K}]$	% inserting sub-matrix of index set $\mathcal{K}$
$\mathbf{r}^t \leftarrow \mathbf{x} - \mathbf{D}\mathbf{D}^{\dagger}\mathbf{x}$	% compute residual
if $\mathbf{r}^t = 0$ then quit	% finish is residual is zero
if $\|\mathbf{r}^t\|_2 \geq \|\mathbf{r}^{t-1}\|_2$ then	% check for improvement
$t = t_{SP}$	% no new iteration
$\mathcal{K}^{t_{SP}} = \mathcal{K}^{t-1}$	% use previous index set
quit	
end loop	
output:	
$\mathcal{K}^{t_{SP}}$	% indices of optimal representation

The goal is to represent the input with minimal reconstruction error using k basis only [23]. SP differs from other iterative greedy methods in the incremental refinement of the selected basis subset. First, a representation is generated with the help of the full basis set (using pseudoinverse computations). During iteration k basis are selected based on the amplitude of the corresponding coordinates of the representation. The resulting residual (difference between the original input and the approximation obtained by projecting the representation onto the input space) is then again projected back to the representation space and another set of k basis are chosen. The two selected subsets are then fused (*expansion*) and the resulting expanded set is used again to project the original input onto the representation space. Finally a new set of k basis are selected by the amplitude of the corresponding coordinates of the projection (*shrinkage*). Iteration stops when the norm of the residual does not decrease anymore. Notation: $\mathbf{D}[\mathcal{K}]$ denotes a sub-matrix of $\mathbf{D}$ where index set $\mathcal{K}$ contains the indices of the selected columns. The index set of the first k sorted components of a vector $\mathbf{a} \in \mathbb{R}^m$ is denoted by $\mathbf{MaxInd}_k(\mathbf{a})$.
doi:10.1371/journal.pcbi.1002372.t002

of the optimum. The efficiency of this random guess depends on the distribution g from which the samples are drawn. After drawing a number of samples from distribution g, we may not be able to give an acceptable approximation of $\mathbf{y}^*$, but we may still obtain a *better sampling distribution*. The basic idea of CEM is that it selects the best few samples, and modifies g so that it becomes more similar to the empirical distribution of the selected samples. CEM resembles the estimation-of-distribution evolutionary methods (see e.g. [50]) and as a global optimization method, it provably converges to the optimal solution [27,50].

For many parameterized distribution families, the parameters of the minimum cross-entropy distribution can be computed easily from simple statistics of the elite samples. For sparse representations the Bernoulli distribution is of particular interest [51]. This particular choice may bring about bias towards solutions where sparse components are drawn independently. Derivations as well as a list of other discrete and continuous distributions with simple update rules can be found in [52]. Let us note that we have also translated CEM into an online variant in which parameter tuning is realized by neurally plausible local learning [29]. This translation then allowed us to propose a neurally plausible SC method [28] in which spikes signal the presence of active components, while rate codes encode the corresponding uncertainty of the given component. Since CEM randomly generates candidate sparse solutions hand, it uses a significantly less number of costly encoding transformations. However it updates the probability of all active components similarly, regardless their individual contributions to the actual reconstruction error.

OSC Part III Subspace Cross-Entropy method, SCE

Subspace Cross-Entropy method (SCE) is an efficient combination of CEM and SP for overcomplete sparse coding. A detailed description can be found in [16] and the pseudocode is given in Table 3. SCE inherits the flexibility and synaptic efficiency of CEM as well as the superior speed and scaling properties of SP without their shortcomings. SCE can be realized by inserting an intermediate control step in CEM to individually update the component probabilities based on their contribution to the reconstruction error. Hence the explicit refinement of the feature set via SP is replaced by an implicit modification through component probabilities.

Since the resulting algorithm is not a greedy method, the algorithm is called as Subspace Cross Entropy (SCE) method without the term 'Pursuit'.

Table 3. Pseudo-code of the subspace cross-entropy (SCE) method for Bernoulli distributions.

required:	
$\mathbf{p} = (p_1, \ldots, p_m)$	% initial distribution parameters
k	% approximate number of non-zero components
initialize : SP and CE	
for $\imath t$ from 1 to t_{SP}	% Main loop of Subspace Pursuit iteration
for τ from 0 to $t_{CE} - 1$,	% Main loop of CE iteration
execute CE iteration	
output : $\mathcal{K}$	% CE optimized index set
$\mathbf{r} \leftarrow \mathbf{x} - \mathbf{D}[\mathcal{K}]\mathbf{D}[\mathcal{K}]^{\dagger}\mathbf{x}$	% compute next residual
if $\|\mathbf{r}^t\|_2 \geq \|\mathbf{r}^{t-1}\|_2$ then quit	% check for improvement
else :	
stochastic update for CE using the residual	
$\mathbf{e} \leftarrow \mathbf{D}^T\mathbf{r}$	% BU step of Subspace Pursuit
$(i_1, \ldots, i_j \ldots, i_m) \leftarrow \mathrm{MaxInd}_m[\mathbf{e}]$	% ordered index set of $\mathbf{e}$
$p'_{i_j} \leftarrow \exp(-j/k)$	% auxiliary Bernoulli distribution
	with $\approx k$ number of 1 s on average
$\mathbf{p}' \leftarrow \mathbf{p} + \|\mathbf{r}\|_2\mathbf{p}'$	% weigh by residual's norm
	to improve distribution
$\mathbf{p} \leftarrow k\mathbf{p}'/\|\mathbf{p}'\|_1$	% normalize for k to draw
	k number of 1 s on average
end loop	

For more details, see technical reports [29] and [16].
doi:10.1371/journal.pcbi.1002372.t003

Robust Principal Component Analysis

An efficient implementation of RPCA algorithm rephrases the optimization problem of (5) by means of the augmented Lagrangian with the following objective function [33]

$$J(\mathbf{L},\mathbf{S},\mathbf{Y}) = \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \langle \mathbf{Y}, \mathbf{X}-\mathbf{L}-\mathbf{S}\rangle + \frac{\mu}{2}\|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F^2, \quad (8)$$

where $\mathbf{Y}$ denotes the current residual after subtracting $\mathbf{L}$ and $\mathbf{S}$. The efficiency stems from the fact that both $\min_{\mathbf{L}} J(\mathbf{L},\mathbf{S},\mathbf{Y})$ and $\min_{\mathbf{S}} J(\mathbf{L},\mathbf{S},\mathbf{Y})$ subproblems have simple solutions. Let $\mathcal{S}_\tau : \mathbb{R}\rightarrow\mathbb{R}$ denote $\mathcal{S}_\tau[x] = \mathrm{sgn}(x)\max(|x|-\tau,0)$, which can be applied componentwise on matrices. For matrices $\mathbf{M}$, let $\mathcal{D}_\tau(\mathbf{M})$ denote the singular value thresholding operator $\mathcal{D}_\tau(\mathbf{M}) = \mathbf{U}\mathcal{S}_\tau(\sum)\mathbf{V}^*$, where $\mathbf{M} = \mathbf{U}\sum\mathbf{V}^*$ is any singular value decomposition. The corresponding pseudocode is given in Table 1.

Training data and fitting

The algorithms were trained on 16×16, normalized (zero mean and 1 std) patches extracted from a public database (http://www.cis.hut.fi/projects/ica/data/images/). For the temporal studies, inputs were generated by concatenating 16 normalized patches of size 8×8 extracted from randomly selected parts of publicly available videos ('football(b)', 'garden', 'ice', 'tempete', 'crowd_run', 'sunflower', 'tractor'; http://media.xiph.org/video/derf/). To speed up calculations, batch learning (50000 samples for static stimuli and 25000 samples for the sequences) was applied to learn the low dimensional subspace of RPCA in the preprocessing stage. On the other hand, to learn the over-complete sparse basis (16-fold over-completeness), $2\cdot10^7$ samples have been used. RPCA was run in MATLAB. All other transformations were performed on a cluster of 17 Sony PlayStation 3 consoles in Linux environment using in-house C++ implementation of published algorithms of SVD [53] and CE [27]. The obtained filters were matched with Gabor filters [36,38] in order to characterize the spatial structures. The Gabor filter parameters are as follows:

$$x' = (x-x_0)\cos(\theta) + (y-y_0)\sin(\theta) \quad (9)$$

$$y' = (x-x_0)\sin(\theta) + (y-y_0)\cos(\theta) \quad (10)$$

$$g(x,y) = \exp(-\frac{x'^2}{n_x} - \frac{y'^2}{n_y})\cos(2\pi f + \phi) \quad (11)$$

where x_0 and y_0 denote the center of the patch, θ is the orientation of the normal to the parallel stripes of the Gabor function, f is the frequency and ϕ is the phase of the cosine factor, n_x and n_y specify the ellipticity of the Gaussian envelope. Fitting was done in MATLAB using the nonlinear least squares optimization function (nsqnonlin(.)) designed for large scale problems. For each parameter value the optimization algorithm was run 20 times with random initialization and the best solution was kept.

Acknowledgments

We are grateful to Zoltán Kisvárday for helpful discussions on the properties of the interneuron groups of the primary visual cortex. We thank the reviewers for all valuable suggestions that have greatly abetted the development of these ideas.

Author Contributions

Conceived and designed the experiments: AL ZP. Performed the experiments: ZP. Analyzed the data: ZP GS. Wrote the paper: GS AL.

References

1. Barlow HB (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, ed. Sensory Communication. CambridgeMA: MIT Press. pp 217–234.
2. Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network 3: 213–251.
3. Dong DW, Atick JJ (1995) Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. Network 6: 159–178.
4. Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Comput 7: 1129–1159.
5. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 607–609.
6. Doi E, Balcan DC, Lewicki MS (2007) Robust coding over noisy overcomplete channels. IEEE Trans Image Process 16: 442–452.
7. Graham DJ, Chandler DM, Field DJ (2006) Can the theory of "whitening" explain the centersurround properties of retinal ganglion cell receptive fields? Vision Res 46: 2901–2913.
8. Ringach DL (2002) Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. J Neurophysiol 88: 455–463.
9. A H, Otte S, Callaway E, Sejnowski TJ (2010) Metabolic cost as a unifying principle governing neuronal biophysics. Proc Natl Acad Sci U S A 107: 12329–12334.
10. Lennie P (2003) The cost of cortical computation. Curr Biol 13: 493–497.
11. Laughlin SB, de Ruyter van Steveninck RR, Anderson JC (1998) The metabolic cost of neural information. Nat Neurosci 1: 36–41.
12. Berkes P, White B, Fiser J (2009) No evidence for active sparsification in the visual cortex. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, eds. Advances in Neural Information Processing Systems 22. CambridgeMA: MIT Press. pp 108–116.
13. Földiák P (2002) Sparse coding in the primate cortex. In: Arbib MA, ed. The Handbook of Brain Theory and Neural Networks. CambridgeMA: MIT Press. second edition. pp 1064–1068.
14. Doi E, Lewicki MS (2005) Relations between the statistical regularities of natural images and the response properties of the early visual system. In: Proceedings of the Japanese Cognitive Science Society, Special interest group of Pattern Recognition and Perception Model; 28 July 2005; Kyoto Japan. pp 1–8.
15. Essen DCV, Anderson C (1995) Information processing strategies and pathways in the primate retina and visual cortex. In: Zornetzer SF, Davis JL, Lau C, eds. Introduction to Neural and Electronic Networks. Orlando: Academic Press. pp 45–76.
16. Lörincz A, Palotai Z, Szirtes G (2012) Sparse and silent coding in neural circuits. Neurocomputing 79: 115–124.
17. Widrow B, Lehr MA (1990) Thirty years of adaptive neural networks: Perceptron, madaline, and backpropagation. Proc IEEE Inst Electr Electron Eng 78: 1415–1442.
18. Lörincz A (2009) Hebbian constraint on the resolution of the Homunculus fallacy leads to a network that searches for hidden cause-effect relationships. In: Goertzel B, Hitzler P, Hutter M, eds. 2nd Conference on Artificial General Intelligence AGI-2009. pp 126–131.
19. Natarajan B (1995) Sparse approximate solutions to linear systems. SIAM J Sci Comput 24: 227–234.
20. Tropp JA (2004) Greed is good: algorithmic results for sparse approximation. IEEE Trans Inf Theory 50: 2231–2242.
21. Needell D, Vershynin R (2009) Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Found Comput Math 9: 317–334.
22. Donoho DL, Tsaig Y, Drori I, Starck J (2006) Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. Technical Report 2006-02, Stanford University.
23. Dai W, Milenkovic O (2009) Subspace pursuit for compressive sensing signal reconstruction. IEEE Tran Inf Theo 55: 2230–2249.
24. Mallat S, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process 41: 3397–3415.
25. Candès EJ, Wakin M (2008) An introduction to compressive sampling. IEEE Signal Processing Mag 25: 21–30.
26. Pati YC, Rezaiifar R, Krishnaprasad PS (1993) Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Conference Record of The Twenty- Seventh Asilomar Conference on Signals, Systems and Computers; 1–3 November, 1993; Pacific Grove, California, United States. pp 40–44.
27. Rubinstein RY (1999) The cross-entropy method for combinatorial and continuous optimization. Method Comput Appl Prob 2: 127–190.
28. Lörincz A, Palotai Z, Szirtes G (2008) Spike-based cross-entropy method for reconstruction. Neurocomputing 71: 3635–3639.
29. Szita I, Lörincz A Online variants of the cross-entropy method. http://arxiv.org/abs/0801.1988.

30. Vincent BT, Baddeley RJ (2003) Synaptic energy efficiency in retinal processing. Vision Res 43: 1283–1290.
31. Simoncelli EP, Olshausen BA (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24: 1193–1216.
32. Jolliffe IT (2002) Principal Component Analysis. New York: Springer. pp 487.
33. Candès EJ, Li X, Ma Y, Wright J (2011) Robust principal component analysis? J Assoc Comp Mach 58: 1–37.
34. Zhou Z, Wright J, Li X, Candès EJ, Ma Y (2010) Stable principal component pursuit. In: Proceedings of IEEE International Symposium on Information Theory (ISIT 2010); 13–18 June 2010; Austin, Texas, United States. pp 1518–1522. DOI:10.1109/ISIT.2010.5513535.
35. Candès EJ, Recht B (2008) Exact matrix completion via convex optimization. Found Comput Math 9: 717–772.
36. Rodieck RW (1965) Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision Res 5: 583–601.
37. Cai D, DeAngelis GC, Freeman RD (1997) Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. J Neurophysiol 78: 1045–1061.
38. Jones JP, Palmer L (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58: 1233–1258.
39. Lücke J (2007) A dynamical model for receptive field self-organization in V1 cortical columns. In: Proceedings of International Conference of Artificial Neural Networks, 13–17 September 2007; Porto Portugal. Springer. LNCS 4669. pp 389–398.
40. Rehn M, Sommer FT (2007) A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. J Comput Neurosci 22: 135–146.
41. DeAngelis GC, Ohzawa I, Freeman RD (1995) Receptive-field dynamics in the central visual pathways. Trends in Neurosci 18: 451–458.
42. Szatmáry B, Lörincz A (2001) Independent component analysis of temporal sequences subject to constraints by LGN inputs yields all the three major cell types of the primary visual cortex. J Comput Neurosci 11: 241–248.
43. Olshausen BA (2002) Sparse Codes and Spikes. In: Rao RPN, Olshausen BA, Lewicki MS, eds. Probabilistic Models of the Brain: Perception and Neural Function MIT Press. pp 257–272.
44. Chennubhotla C, Jepson AD (2001) Sparse PCA: Extracting multi-scale structure from data. In: Proceedings of IEEE International Conference on Computer Vision; 1: 641–647; 7–14 July, 2001; Vancouver, British Columbia, Canada.
45. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2: 1019–1025.
46. Cadieu C, Olshausen B (2009) Learning transformational invariants from natural movies. In: Koller D, Schuurmans D, Bengio Y, Bottou L, eds. Advances in Neural Information Processing Systems 21. CambridgeMA: MIT Press. pp 209–216.
47. Cardin JA, Palmer LA, Contreras D (2007) Stimulus feature selectivity in excitatory and inhibitory neurons in primary visual cortex. J Neurosci 27: 10333–10344.
48. Liu B, Li P, Li Y, Sun YJ, Yanagawa Y (2009) Visual receptive field structure of cortical inhibitory neurons revealed by two-photon imaging guided recording. J Neurosci 29: 10520–10532.
49. Needell D, Tropp JA (2008) Cosamp: Iterative signal recovery from incomplete and inaccurate samples. Appl Computational Harmon Anal 26: 301–321.
50. Muehlenbein H (1998) The equation for response to selection and its use for prediction. Evol Comput 5: 303–346.
51. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res 37: 3311–3325.
52. de Boer PT, Kroese DP, Mannor S, Rubinstein RY (2004) A tutorial on the cross-entropy method. Ann Oper Res 134: 19–67.
53. Golub GH, Loan CV (1996) Matrix Computation, 3rd edition. Baltimore: Johns Hopkins University Press.