OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Temporal Adaptation Enhances Efficient Contrast Gain Control on Natural Images

Fabian Sinz[1]*, **Matthias Bethge**[1,2,3]

1 Department for Neuroethology, University Tübingen, Tübingen, Germany, **2** Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany, **3** Bernstein Center for Computational Neuroscience, Tübingen, Germany

Abstract

Divisive normalization in primary visual cortex has been linked to adaptation to natural image statistics in accordance to Barlow's redundancy reduction hypothesis. Using recent advances in natural image modeling, we show that the previously studied static model of divisive normalization is rather inefficient in reducing local contrast correlations, but that a simple temporal contrast adaptation mechanism of the half-saturation constant can substantially increase its efficiency. Our findings reveal the experimentally observed temporal dynamics of divisive normalization to be critical for redundancy reduction.

Citation: Sinz F, Bethge M (2013) Temporal Adaptation Enhances Efficient Contrast Gain Control on Natural Images. PLoS Comput Biol 9(1): e1002889. doi:10.1371/journal.pcbi.1002889

Editor: Laurence T. Maloney, New York University, United States of America

Received May 23, 2012; **Accepted** December 4, 2012; **Published** January 31, 2013

Funding: This study was financially supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award (BMBF; FKZ: 01GQ0601). 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: fabian.sinz@bethgelab.org

Introduction

It is a long-standing hypothesis that the computational goal of the early visual processing stages is to reduce redundancies which are abundantly present in natural sensory signals [1,2]. Redundancy reduction is a general information theoretic principle that plays an important role for many possible goals of sensory systems like maximizing the amount of information between stimulus and neural response [3], obtaining a probabilistic model of sensory signals [4], or learning a representation of hidden causes [3,5]. For a population of neurons, redundancy reduction predicts that neuronal responses should be made as statistically independent from each other as possible [2].

Many prominent neural response properties such as receptive field structure or contrast gain control have been linked to redundancy reduction on natural images [2]. While an appropriate structure of linear receptive fields can always remove all redundancies caused by second order correlations, they have only little effect on the reduction of higher order statistical dependencies [6,7]. However, one of the most prominent contrast gain control mechanisms—divisive normalization—has been demonstrated to reduce higher order correlations on natural images and sound [8–10]. Its central mechanism is a divisive rescaling of a single neuron's activity by that of a pool of other neurons [8, see also Figure 1a].

Recently, *radial factorization* and *radial Gaussianization* have been derived independently by [11] and [12], respectively, based on Barlow's redundancy reduction principle [1]. Both mechanisms share with divisive normalization the two main functional components, linear filtering and rescaling and have been shown to be the unique and optimal redundancy reduction mechanism for this class of transformations under certain symmetry assumptions for the data. Radial factorization is optimal for a more general symmetry class than radial Gaussianization [11,13] and contains radial Gaussianization as a special case. As a consequence, radial factorization can achieve slightly better redundancy reduction for natural images than radial Gaussianization but the advantage is very small.

Here, we compare the redundancy reduction performance of divisive normalization to that of radial factorization in order to see to what extent divisive normalization can serve the goal of redundancy reduction. Our comparison shows that a non-adapting *static* divisive normalization is not powerful enough to capture the contrast dependencies of natural images. Furthermore, we show that (i) the shape of contrast response curves predicted by radial factorization is not consistent with that found in physiological recordings, and (ii) that for a *static* divisive normalization mechanism this inconsistency is a necessary consequence of strong redundancy reduction. Finally, we demonstrate that a *dynamic* adaptation of the half-saturation constant in divisive normalization may provide a physiologically plausible mechanism that can achieve close to optimal performance. Our proposed adaptation mechanism works via horizontal shifts of the contrast response curve along the log-contrast axis. Such shifts have been observed in experiments in response to a change of the ambient contrast level [14].

Results

Measures, Models, Mechanisms

We now briefly introduce divisive normalization, radial factorization, and the information theoretic measure of redundancy used in this study.

Redundancy reduction and multi-information. We consider a population of sensory neurons that transforms natural image patches $\mathbf{x}$ into a set of neural activities $\mathbf{y}$ or $\mathbf{z}$. We always use

Author Summary

The redundancy reduction hypothesis postulates that neural representations adapt to sensory input statistics such that their responses become as statistically independent as possible. Based on this hypothesis, many properties of early visual neurons—like orientation selectivity or divisive normalization—have been linked to natural image statistics. Divisive normalization, in particular, models a widely observed neural response property: The divisive inhibition of a single neuron by a pool of others. This mechanism has been shown to reduce the redundancy among neural responses to typical contrast dependencies in natural images. Here, we show that the standard model of divisive normalization achieves substantially less redundancy reduction than a theoretically optimal mechanism called *radial factorization*. On the other hand, we find that radial factorization is inconsistent with existing neurophysiological observations. As a solution we suggest a new physiologically plausible modification of the standard model which accounts for the dynamics of the visual input by adapting to local contrasts during fixations. In this way the dynamic version of the standard model achieves almost optimal redundancy reduction performance. Our results imply that the dynamics of natural viewing conditions are critical for testing the role of divisive normalization for redundancy reduction.

$\mathbf{y}$ to denote responses to linear filters, and $\mathbf{z}$ for the output of divisive normalization or radial factorization. The goal of redundancy reduction is to remove statistical dependencies between the single coefficients of $\mathbf{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$ or $\mathbf{z}=(z_1,\ldots,z_n)^{\mathrm{T}}$.

Redundancy is quantified by the information theoretic measure called *multi-information*

$$I[\mathbf{Y}]=D_{KL}\left(\rho(\mathbf{y})\,\|\,\prod_{i=1}^{n}\rho_i(y_i)\right)=\sum_{i=1}^{n}H[Y_i]-H[\mathbf{Y}], \qquad (1)$$

which measures how much the representation differs from having independent components. More precisely, the multi-information is the Kullback-Leibler divergence between the joint distribution and the product of its marginals or, equivalently, the difference between the sum of the marginal entropies and the joint entropy. In case of $n=2$ this equals the better known mutual information. If the different entries of $\mathbf{Y}$ are independent, then its joint distribution equals the product of the single marginals or—equivalently—the joint entropy equals the sum of the marginal entropies. Thus, the multi-information is zero if and only if the different dimensions of the random vector $\mathbf{Y}$ are independent, and positive otherwise. In summary, the multi-information measures all kinds of statistical dependencies among the single coefficients of a random vector. In the Methods Section, we describe how we estimate the multi-information for the various signals considered here.

Divisive normalization. From all existing divisive normalization models considered previously in the literature, ours is most closely related to the one used by Schwartz and Simoncelli [9]. It consists of two main components: a linear filtering step and a rescaling step based on the Euclidean norm of the filter responses

$$y_i=\mathbf{w}_i^{\top}\mathbf{x},\ for\ i=1,\ldots,n \qquad \mathbf{z}=\frac{\kappa\mathbf{y}}{\sqrt{\sigma^2+\|\mathbf{y}\|^2}}. \qquad (2)$$

While the linear filters $\mathbf{w}_i$ capture the receptive field properties, the rescaling step captures the nonlinear interactions between the single neurons. Most divisive normalization models use filters $\mathbf{w}_i$ that resemble the receptive fields of complex cells [9,15,16]. Therefore, we use filters obtained from training an *Independent Subspace Analysis (ISA)* on a large collection of randomly sampled image patches [15,16, see also Methods]. ISA can be seen as a redundancy reduction transform whose outputs are computed by the complex cell energy model [17,18]. For this study, the algorithm has the advantage that it not only yields complex cell-like filter shapes, but also ensures that single filter responses y_i are decorrelated and already optimized for statistical independence. This ensures that the redundancies removed by divisive normalization and radial factorization are the ones that cannot be removed by the choice of linear filters [7,19].

Several divisive normalization models exist in the literature. They differ, for instance, by whether a unit y_i is contained in its own normalization pool, or in the exact form of the rescaling function $g_{DN}(t)=\kappa t/\sqrt{\sigma^2+t^2}$ also known as *Naka-Rushton function*. From the viewpoint of redundancy reduction, the former distinction between models is irrelevant because the influence of a single unit on its normalization pool can always be removed by the elementwise invertible transformation $z_i\mapsto z_i/\sqrt{1-z_i^2}$ which does not change the redundancies between the responses [20] (the multi-information is invariant with respect to elementwise invertible transformations). Sometimes, a more general form of the Naka-Rushton function is found in the literature which uses different types of exponents

$$y_i\mapsto\frac{\kappa y_i}{\left(\sigma^2+\sum_j|\tilde{y}_j|^p\right)^{1/p}}=\frac{\kappa y_i}{\left(\sigma^2+\|\mathbf{y}\|_p^p\right)^{1/p}}. \qquad (3)$$

The divisive normalization model considered in this study (equation (2)) differs from this more general version by the type of the norm used for rescaling the single responses: Where equation (3) uses the L_p-norm $\|\mathbf{y}\|_p=\left(\sum_j|y_j|^p\right)^{1/p}$ we use the Euclidean norm. Because radial factorization is defined for the more general L_p-norm (see Methods), all analyses in this paper could be carried out for this more general transform. However, we instead chose to use the Euclidean norm for simplicity and to make our model more comparable to the ones most commonly used in redundancy reduction studies of divisive normalization [9,20–22].

Also note that the Naka-Rushton function is often defined as the pth power of equation (3). However, the form of equation (3) is more common in redundancy reduction studies in order to maintain the sign of y_i. We mention the consequences of this choice in the discussion.

Radial factorization. Radial factorization is an optimal radial rescaling for redundancy reduction. We will now briefly introduce radial factorization starting from divisive normalization. For more mathematical details see the Methods Section.

On a population level, the rescaling step of divisive normalization is a nonlinear mapping that changes the Euclidean radius of the filter response population. This can be seen by decomposing divisive normalization into two multiplicative terms

$$\mathbf{z}=\frac{\kappa\|\mathbf{y}\|}{\sqrt{\sigma^2+\|\mathbf{y}\|^2}}\cdot\frac{\mathbf{y}}{\|\mathbf{y}\|}=g_{DN}(\|\mathbf{y}\|)\cdot\frac{\mathbf{y}}{\|\mathbf{y}\|}. \qquad (4)$$

The second term normalizes the response vector $\mathbf{y}$ to length one while the Naka-Rushton function in the first term determines the new radius. Since the rescaling g_{DN} depends only on the norm, the new radius does not depend on any specific direction of $\mathbf{y}$.

The redundancy between the coefficients of $\mathbf{z}$ is determined by three factors: The statistics of natural image patches $\mathbf{x}$ which—together with the choice of filters $\mathbf{w}_i$—determine the statistics of $\mathbf{y}$, and the radial transformation g_{DN}. If we allow the radial transformation to be a general invertible transform $g(\|\mathbf{y}\|)$ on the Euclidean norm, we can now ask how the different model components can be chosen in order to minimize the redundancy in $\mathbf{z}$.

A substantial part of the redundancies in natural images are second order correlations, which can be removed by linear filters during *whitening* [6]. Whitening does not completely determine the filters since the data can always be rotated afterwards and still stay decorrelated. Higher order decorrelation algorithms like *independent component analysis* use this rotational degree of freedom to decrease higher order dependencies in the filter responses $\mathbf{y}$ [3]. However, there is no set of filters that could remove all statistical dependencies from natural images [6,7], because whitened natural images exhibit an approximately spherical but non-Gaussian joint distribution [7,21,23,24]. Since spherical symmetry is invariant under rotation and because the only spherically symmetric factorial distribution is the Gaussian distribution [13,25], the marginals cannot be independent.

Hence, the remaining dependencies must be removed by nonlinear mechanisms like an appropriate radial transformation g. Fortunately, the joint spherically symmetric distribution of the filter responses $\mathbf{y}$ already dictates a unique and optimal way to choose g: Since a rescaling with g will necessarily result in a spherically symmetric distribution again, g must be chosen such that $\mathbf{z}$ is jointly Gaussian distributed. Therefore, we need to choose g such that $g(\|\mathbf{y}\|)$ follows the radial distribution of a Gaussian or, in other words, a χ-distribution. This is a central point for our study: For a spherically symmetric distribution the univariate distribution on $\|\mathbf{y}\|$ determines higher order dependencies in the multi-variate joint distribution of $\mathbf{y}$. This means that if we restrict ourselves to radial transformations, it is sufficient to look at radial distributions only. The fact that the Gaussian is the only spherically symmetric factorial distribution implies that the coefficients in $\mathbf{z}$ can only be statistically independent if $\|\mathbf{z}\|$ follows radial χ-distribution. *Radial factorization* finds a transformation g which achieves exactly that by using histogram equalization on the distribution of $\|\mathbf{y}\|$ [11,12, see also Methods]. All these considerations also hold for L_p-spherically symmetric distributions [11,13].

Note that this does not imply that the neural responses $\mathbf{z}$ must follow a Gaussian distribution if they are to be independent because the distribution of the single responses z_i can always be altered by applying an elementwise invertible transformation $z_i \mapsto f_i(z_i)$ without changing the redundancy. The above considerations only mean that given the two main model components of divisive normalization (and the assumption of spherical symmetry), the best we can do is to choose the $\mathbf{w}_i$ to be whitening filters and $g(\|\mathbf{y}\|)$ according to radial factorization.

Radial factorization and divisive normalization are not equivalent. The goal of this study is to compare the redundancy reduction achieved by divisive normalization and radial factorization. Apart from all similarities between the two models, there is a profound mathematical difference showing that the two mechanisms are not equivalent (as noted by [12]).

Both mechanisms have the form

$$\mathbf{z} = g(\|\mathbf{y}\|) \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|}.$$

However, the radial rescalings of radial factorization and that of divisive normalization, g_{RF} and g_{DN}, have a different range. Since the χ-distribution is non-zero on all of $\mathbb{R}^+$ the range of g_{RF} must be $\mathbb{R}^+$ as well. However, in case of divisive normalization, the Naka-Rushton function g_{DN} saturates at κ. This means that g_{DN} can never transform a radial distribution into a χ-distribution since values beyond κ cannot be reached.

While this implies that the two mechanisms are mathematically not equivalent, it could still be that they perform similarly on data if the probability mass of the χ-distribution in the range beyond κ is small. Therefore, we choose κ to be the 99% quantile of the χ-distribution in all our experiments (see Methods).

Comparison of the redundancy reduction performance. We compared the amount of redundancy removed by divisive normalization and radial factorization by measuring the multi-information in the plain filter responses $\mathbf{y}$ and the normalized responses $\mathbf{z}$ for a large collection of natural image patches (Figure 1b). In both cases the parameters of the radial transformation were chosen to yield the best possible redundancy reduction performance (see Methods). While both divisive normalization and radial factorization remove variance correlations (Figure 1a), the residual amount of dependencies for divisive normalization is still approximately 34% of the total redundancies removed by radial factorization (Figure 1a–b). This demonstrates that divisive normalization is not optimally tailored to the statistics of natural images.

To understand this in more detail, we derived the distribution that $\|\mathbf{y}\|$ should have if divisive normalization were the optimal redundancy reducing mechanism and compared it to the empirical radial distribution of $\|\mathbf{y}\|$ represented by a large collection of uniformly sampled patches from natural images. This optimal distribution for divisive normalization can be derived by transforming a χ-distributed random variable with g_{DN}^{-1} (see Methods). Since g_{DN} has limited range $[0,\kappa)$ we actually have to use a χ-distribution which is truncated at κ. The parametric form of the resulting distribution is given in the Methods Section. We refer to is as *Naka-Rushton distribution* in the following. The parameters of the Naka-Ruston distribution are κ and σ^2. Since κ is already determined by fixing the range of g_{DN} to the 99% quantile of the χ-distribution, the remaining free parameter is σ^2. In the Naka-Rushton function g_{DN} this parameter is called half-saturation constant and controls the horizontal position of the contrast response curve in model neurons.

We fitted σ^2 via maximum likelihood (see Methods) and found that even for the best fitting σ^2 there is a pronounced mismatch between the Naka-Rushton distribution and the empirical distribution given by the histogram (Figure 1c). This explains the insufficient redundancy reduction because the Naka-Rushton distribution expects most of the responses $\|\mathbf{y}\|$ to fall into a much narrower range than responses to natural images do in reality. The Naka-Rushton function g_{DN} would map the red radial density in Figure 1c perfectly into a truncated χ-distribution. However, it maps a profound part of the true radial distribution of $\|\mathbf{y}\|$ (gray histogram) close to κ, since this part is located to the right of the mode of the Naka-Rushton distribution where it expects almost no probability mass. Additionally, the Naka-Rushton distribution exhibits a small gap of almost zero probability around zero. This gap, however, also contains a portion of empirical distribution. This part gets mapped close to zero. To understand why this leaves significant redundancies, imagine the most extreme case in which all the probability mass of $\|\mathbf{y}\|$ would either be mapped onto κ or on onto 0. The corresponding distribution on $\mathbf{z}$ would consist of a point mass at zero and a spherical shell at κ. Such a distribution would clearly exhibit strong dependencies.

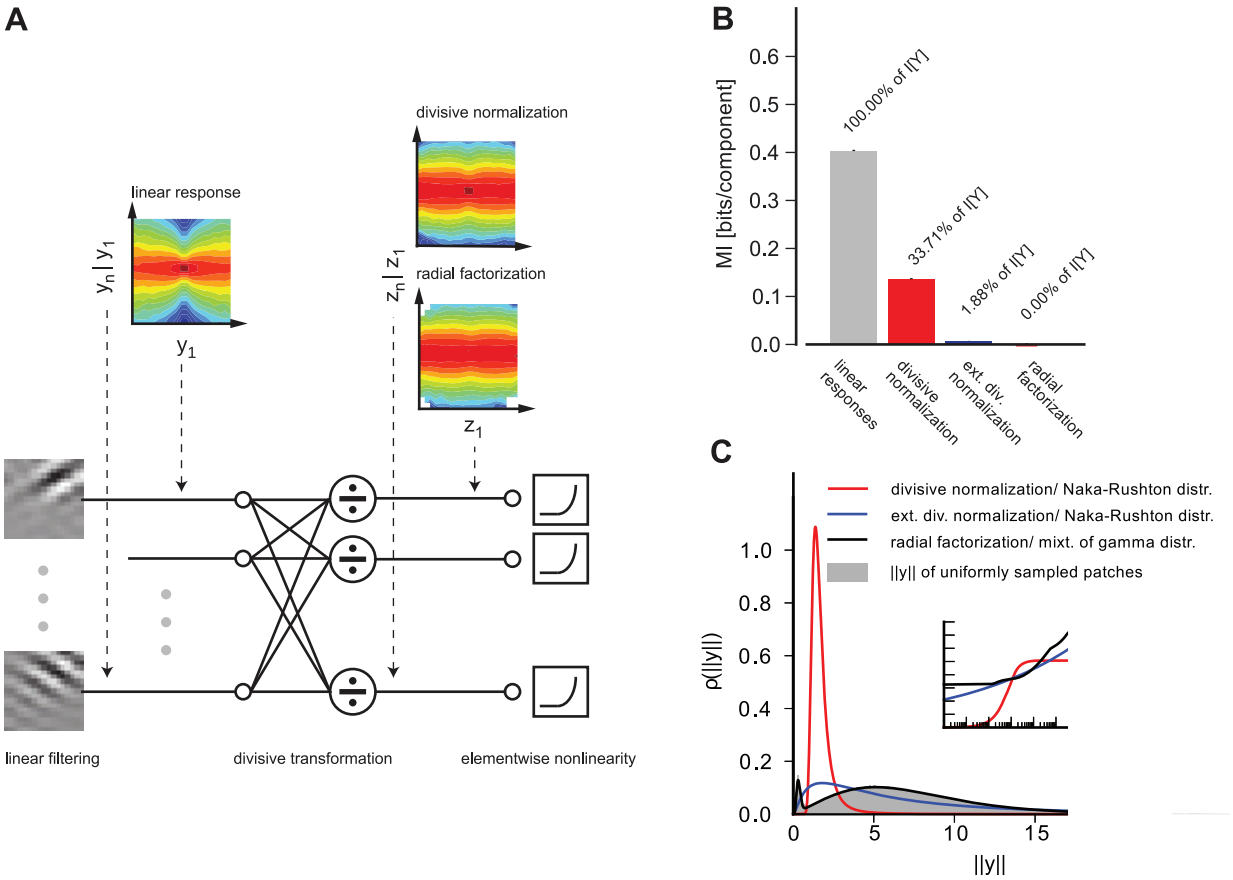

Figure 1. Redundancy reduction and radial distributions for different normalization models. A: Divisive normalization model used in this study: Natural image patches are linearly filtered. These responses are nonlinearly transformed by divisive normalization or radial factorization (see text). After linear filtering the width of the conditional distribution $p(y_j|y_i)$ of two filter responses depends on the value of y_i (conditional log-histograms as contour plots). This demonstrates the presence of variance correlations. These dependencies are decreased by divisive normalization and radial factorization. **B**: Redundancy measured by multi-information after divisive normalization, extended divisive normalization, and radial factorization: divisive normalization leaves a substantial amount of residual redundancy (error bars show standard deviation over different datasets). **C**: Distributions on the norm of the filter responses $\|\mathbf{y}\|$ for which divisive normalization (red) and extended divisive normalization (blue) are the optimal redundancy reducing mechanisms. The radial transformation of radial factorization and its corresponding distribution (mixture of five γ-distributions) is shown in black. While radial factorization (inset, black curve) and extended divisive normalization (inset, blue curve) achieve good redundancy reduction, they lead to physiologically implausibly shaped contrast response curves which are mainly determined by their respective radial transformations $g(\|\mathbf{y}\|)$ shown in the inset. The radial transformation of divisive normalization is shown for comparison (inset, red curve).
doi:10.1371/journal.pcbi.1002889.g001

Augmenting divisive normalization by more parameters. It is clear that the suboptimal redundancy reduction performance of divisive normalization is due to its restricted parametric form. Therefore, we explored two options how to increase its degrees of freedom and thereby its redundancy reduction performance: the first option endows static divisive normalization with additional parameters γ,δ, the second option allows for a dynamic temporal adaptation of σ^2.

The simplest way to increase the degrees of freedom in divisive normalization is to introduce two additional parameters in the Naka-Rushton function

$$\|\mathbf{z}\| = g_{DNE}(\|\mathbf{y}\|) = \frac{\kappa \|\mathbf{y}\|^{\frac{\gamma}{2}+\delta}}{\sqrt{\sigma^2 + \|\mathbf{y}\|^{\gamma}}}.$$

These parameters allow for more flexibility in the scale and shape of the corresponding Naka-Rushton distribution. We label all models that use this parametrization as *extended* in the following. Note that the extended Naka-Rushton function only saturates for $\delta = 0$. This means that it could in principle transform $\|\mathbf{y}\|$ into $\|\mathbf{z}\|$ such that $\|\mathbf{z}\|$ is χ-distributed. For $\delta = 0$ and $\gamma = 2$, the original Naka-Rushton function is recovered. As before, we derived the corresponding extended Naka-Rushton distribution by transforming a (truncated) χ-distributed random variable with g_{DNE}^{-1}. We fitted the resulting distribution to a large collection of $\|\mathbf{y}\|$, used the maximum likelihood parameters for extended divisive normalization, and measured the redundancy via multi-information in the resulting normalized responses $\mathbf{z}$.

We found that an extended divisive normalization transform achieves substantially more redundancy reduction and that the extended Naka-Rushton distribution on $\|\mathbf{y}\|$ fits the image data significantly better (Figure 1b–c). However, we also find that the best extended Naka-Rushton function for redundancy reduction would yield biologically implausible contrast response curves which capture the firing rate of a neuron upon stimulation with gratings of different contrast at the neuron's preferred spatial frequency and orientation.

In the divisive normalization and the radial factorization model, the shape of the contrast response curve is determined by the

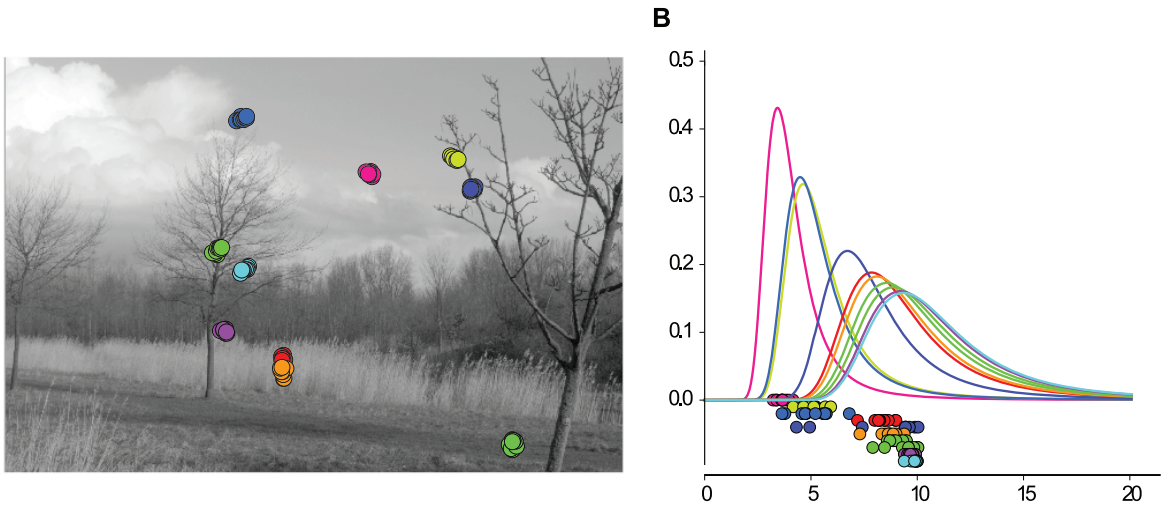

Figure 2. Simulated eye movements and adapted contrast distributions. A: Simulated eye movements on a image from the van Hateren database [31]. Local microsaccades are simulated with Brownian motion with a standard deviation of 5px. In this example, 8×8 patches are extracted around the fixation location and whitened. **B**: Values of $\|\mathbf{y}\|$ for the extracted patches plotted along the x-axis. Vertical offset was manually introduced for visibility. Colors match the ones in **A**. The different curves are the maximum likelihood Naka-Rushton distributions estimated from the data points of the same color.
doi:10.1371/journal.pcbi.1002889.g002

shape of the radial rescaling function (Figure 1c, inset) [8]. In contrast to the normal Naka-Rushton function (Figure 1c, inset, red curve), the extended version (Figure 1c, inset, blue curve) exhibits a physiologically unreasonable shape: it starts at a non-zero value, increases without saturation, and does not resemble any sigmoidal shape at all. The non-zero level for low contrasts is a direct consequence of the optimization for redundancy reduction: redundancy reduction implies that the target radial distribution is a (truncated) χ-distribution which has only very little probability mass close to zero. Therefore, the radial rescaling function must map the substantial portion of low contrast values in the empirical distribution upwards in order to match the χ-distribution. This results in the immediate non-zero onset. This is a pronounced mismatch to the typical contrast response curves measured in cortical neurons (see Figure 2 in [14]). In fact, the addition of more parameters merely leads to a contrast response curve which is more similar to radial factorization (Figure 1, inset, black) which does not have a plausible shape, too. Therefore, we dismiss the option of adding more parameters to the Naka-Rushton function and turn to the option in which σ^2 is allowed to dynamically adapt to the ambient contrast level.

Dynamic divisive normalization. Previous studies found that single neurons adapt to the ambient contrast level via horizontal shifts of their contrast response curve along the log-contrast axis [8,14]. In the divisive normalization model, this shift is realized by changes in the half-saturation constant σ^2. This means, however, that there is not a single static divisive normalization mechanism, but a whole continuum whose elements differ by the value of σ^2 (Figure 2). This is equivalent to a continuum of Naka-Rushton distributions which can be adapted to the ambient contrast level by changing the value of σ^2. Since this kind of adaptation increases the degrees of freedom, it could also lead to a better redundancy reduction performance.

In order to investigate adaptation to the local contrast in a meaningful way, we used a simple model of saccades and micro-saccades on natural images to sample fixation locations and their corresponding filter responses $\mathbf{y}$ (see Methods). Previous studies on redundancy reduction with divisive normalization [9,11,12] ignored both the structure imposed by fixations between saccades in natural viewing conditions, and the adaptation of neural contrast response curves to the ambient contrast level via the adaptation of σ^2 [14]. Figure 2 shows an example of simulated eye movements on a natural image from the van Hateren database. For each sample location, we computed the corresponding values of $\|\mathbf{y}\|$ and fitted a Naka-Rushton distribution to it. The right hand side show the resulting Naka-Rushton distributions. One can see that the mode of the distribution shifts with the location of the data, which itself depends on the ambient contrast of the fixation location.

A dynamically adapting σ^2 predicts that the distribution of $\|\mathbf{y}\|$ across time should be well fit by a mixture of Naka-Rushton distributions. Let $r = \|\mathbf{y}\|$ (we use r to emphasize that the radial distribution is a univariate density and not a multivariate density on $\mathbf{y}$), then averaged over all time points t, the distribution of r is given by

$$\varrho(r) = \int v(r|\sigma_t)\rho(\sigma_t)d\sigma_t, \tag{5}$$

where $v(r|\sigma_t)$ denotes a single Naka-Rushton distribution at a specific point in time.

We fitted such a mixture distribution to samples $\|\mathbf{y}\|$ from simulated eye movements (see Methods). Figure 3a shows that the mixture of Naka-Rushton distributions fits the empirical data very well, thus confirming the possibility that a dynamic divisive normalization mechanism may be used to achieve optimal redundancy reduction.

The next step is to find an explicit dynamic adaptation mechanism that can achieve optimal redundancy reduction. To this end, we sought for a way to adapt σ^2 such that the redundancies between the output responses $\mathbf{z}$ were small. Our temporally adapting mechanism chooses the current σ^2 based on the recent stimulation history by using correlations between the contrast values at consecutive time steps. We estimated σ^2 for the present set of filter responses $\mathbf{y}_t$ from the immediately preceding responses $\mathbf{y}_{t-1}$ by sampling σ^2 from a γ-distribution whose

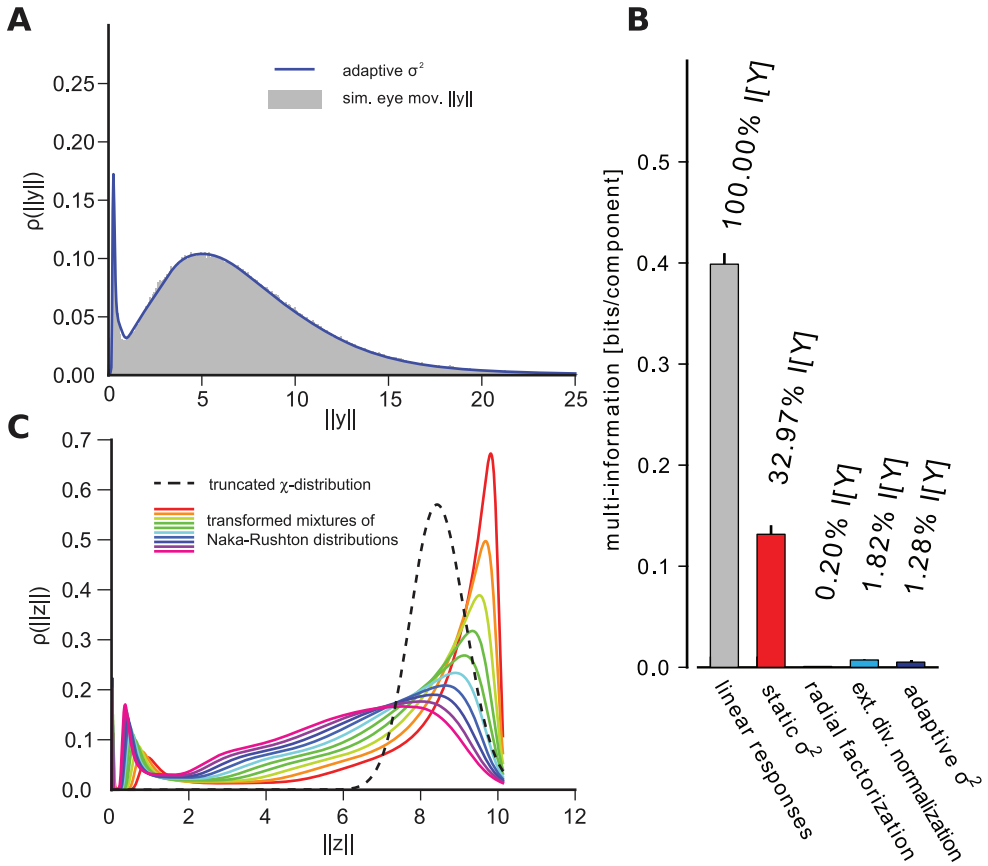

Figure 3. Radial distribution and redundancy reduction achieved by the dynamically adapting model. A: Histogram of $\|\mathbf{y}\|$ for natural image patches sampled with simulated eye movements: The distribution predicted by the dynamically adapting model closely matches the empirical distribution. **B**: Same as in Fig. 1B but for simulated eye movement data. The dynamically adapting σ^2 achieves an almost optimal redundancy reduction performance. **C**: Each colored line shows the distribution of a random variable from 3A transformed with a Naka-Rushton function. Different colors correspond to different values of σ. The dashed curve corresponds to a truncated χ-distribution. A mixture of the colored distributions cannot resemble the truncated χ-distribution since there will either be peaks on the left or the right of the dashed distribution that cannot be canceled by other mixture components.
doi:10.1371/journal.pcbi.1002889.g003

parameters were determined by the mean and the variance of the posterior $\varrho(\sigma_{t-1}|\|\mathbf{y}_{t-1}\|)$ which was derived from the mixture distribution above (see Methods). We found that this temporal adaptation mechanism significantly decreased the amount of residual redundancies to about 1.3% (Figure 3B). Note that the proposed mechanism is a simple heuristic that does not commit to a particular biophysical implementation of the adaptation, but it demonstrates that there is at least one mechanism that can perform well under realistic conditions a neural system would face.

Looking at the joint dynamics of r_t and its σ (Figure 4) we find them to be strongly and positively correlated. Therefore, a higher value of r_t is accompanied by a higher value of σ. This is analogous to the adaptation of neural contrast response curves observed in vivo where a higher contrast (higher $\|\mathbf{y}\|$) shifts the contrast response curve to the right (higher σ^2), and vice versa [14].

In order to demonstrate that improved redundancy reduction is a true adaptation mechanism which relies on correlations between temporally subsequent sample, we need to preclude the possibility that σ^2 can be sampled independently (i.e. context independent). For strong redundancy reduction, the normalized responses $\|\mathbf{z}\|$ should follow a (possibly truncated) χ-distribution (see Methods). The history-independent choice of σ^2 predicts that this truncated χ-distribution should be expressible as a mixture of distributions that result from transforming random variables, that follow a mixture of Naka-Rushton distributions from Figure 3C, with Naka-Rushton functions for different values of σ^2 (see Methods for the derivation). We transformed the input distribution with Naka-Rushton functions that differed in the value of σ^2 (Figure 3C, colored lines). Different colors in Figure 3C refer to different values of σ^2. If σ^2 was history-independent, a positively weighted average of the colored distributions should be able to yield a truncated χ-distribution (Figure 3C, dashed line). It is obvious that this is not possible. Every component will either add a tail to the left of the χ-distribution or a peak to the right of it. Since distributions can only be added with non-negative weight in a mixture, there is no way that one distribution can make up for a tail or peak introduced by another. Therefore, σ^2 cannot be chosen independently of the preceding stimulation, but critically relies on exploiting the temporal correlation structure in the input.

Discussion

In this study we have demonstrated that a *static* divisive normalization mechanism is not powerful enough to capture the contrast dependencies of natural images leading to a suboptimal

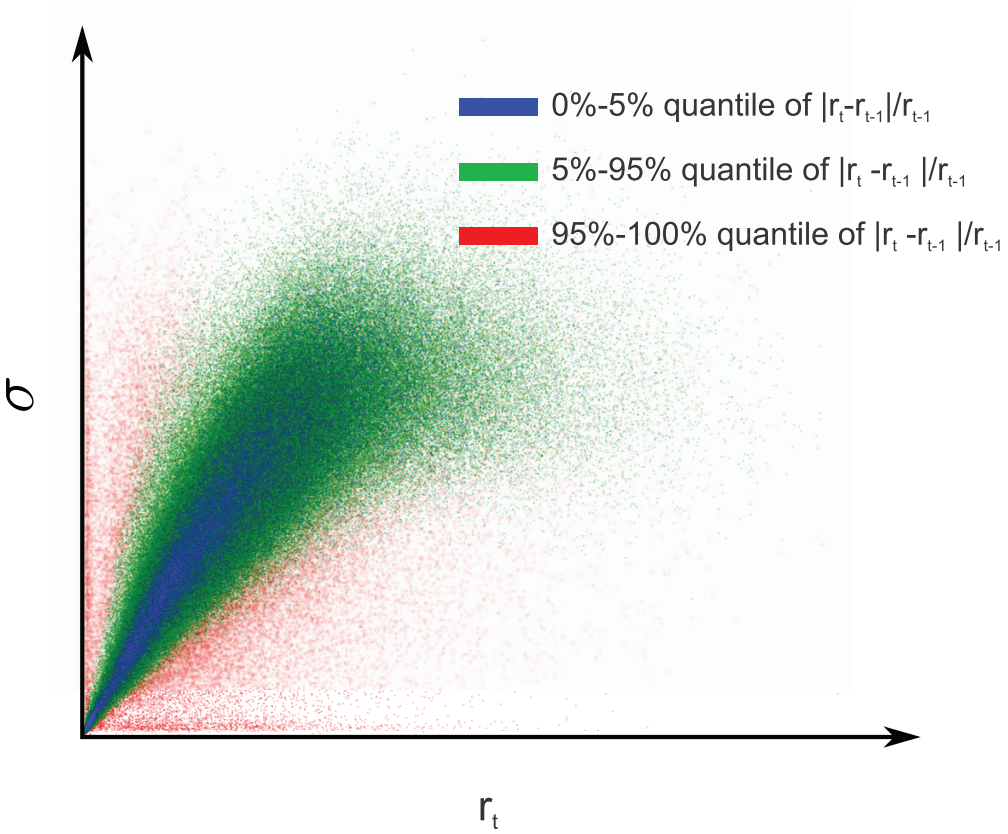

Figure 4. Dynamics of the adaptive σ. The scatter plot shows the values of r_t plotted against the σ used to transform r_t in the dynamic divisive normalization model. The two values are clearly correlated. This indicates that the shift of the contrast response curve, which is controlled by σ, tracks the ambient contrast level, which is proportional to r_t. Single elements in the plot are colored according to the quantile the value of $|r_t-r_{t-1}|/r_{t-1}$ falls in. When the ambient contrast level changes abruptly (e.g. when a saccade is made), this value is large. If the ambient contrast level is relatively stable (e.g. during fixation), this value is small. In those situations (blue dots), σ and r_t exhibit the strongest proportionality.
doi:10.1371/journal.pcbi.1002889.g004

redundancy reduction performance. Static divisive normalization could only exhibit close to optimal performance if the contrast distribution of the input data would be similar to a Naka-Rushton distribution that we derived in this paper. For the best fitting Naka-Rushton distribution, however, the interval containing most of the probability mass is too narrow and too close to zero compared to the contrast distribution empirically found for natural image patches. A divisive normalization mechanism that uses the L_p-norm as in equation (3) instead of the Euclidean norm would suffer from the same problem because the Naka-Rushton distribution for L_p-norms other than $p=2$ would have similar properties. However, the good performance of extended divisive normalization demonstrates that it is not necessary to model the contrast distribution perfectly everywhere but that it would be sufficient to match the range where most natural contrasts appear (Figure 1C).

Not every mapping on natural contrasts that achieves strong redundancy reduction is also physiologically plausible: We showed that the extended static mechanism yields physiologically implausible contrast response curves. Extending the static mechanism of divisive normalization for better redundancy reduction simply makes it more similar to the optimal mechanism and, therefore, yields implausible tuning curves as well. We thus suggested to consider temporal properties of divisive normalization and devised a model that can resolve this conflict by temporally adapting the half-saturation constant σ^2 using temporal correlations between consecutive data points caused by fixations.

Another point concerning physiological plausibility is the relationship between divisive normalization models used to explain neurophysiological observations, and those used in redundancy reduction studies like ours. One very common neurophysiological model was introduced by Heeger [8] which uses half-squared instead of linear single responses:

$$y_i = \mathbf{w}_i^{\top}\mathbf{x}, \quad \tilde{y}_i = \lfloor y_i \rfloor, \quad \tilde{z}_i^2 = \frac{\kappa \tilde{y}_i^2}{\sigma^2 + \sum_j \tilde{y}_j^2}. \tag{6}$$

In order to represent each possible image patch this model would need two neurons per filter: one for the positive part and one for the negative part $\tilde{y}_{i,\pm} = \lfloor \pm w_i^{\top}\mathbf{x} \rfloor$. Of course, these two units would be strongly anti-correlated since only one can be nonzero at a given point in time. Therefore, taking a redundancy reduction view requires considering the positive and the negative part. For this reason it is reasonable to use $y_i = \tilde{y}_{i,+} - \tilde{y}_{i,-}$ as the most basic unit and define the normalization as in equation (2). Since y_i and $\{\tilde{y}_{i,+}, \tilde{y}_{i,-}\}$ are just two different representations of the same information, the multi-information between $y_1,\ldots,y_n$ is the same as the multi-information between different tuples $\{\tilde{y}_{1,+}, \tilde{y}_{1,-}\},\ldots,\{\tilde{y}_{n,+}, \tilde{y}_{n,-}\}$. Apart from this change of viewpoint, the two models are equivalent, because the normalized half-

squared response of equation (6) can be obtained by half-squaring the normalized response of equation (2). Therefore, a model equivalent to the one in equation (6) can be obtained by using the model of equation (2) and representing its responses $\mathbf{z}$ by twice as many half-squared coefficients afterwards.

Previous work on the role of contrast gain control for efficient coding has either focused on the temporal domain [26,27], or on its role in the spatial domain as a redundancy reduction mechanism for contrast correlations in natural images [9,11,12]. Our results emphasize the importance of combining both approaches by showing that the temporal properties of the contrast gain control mechanism can have a critical effect on the redundancies that originate from the spatial contrast correlations in natural images. Our analysis does not commit to a certain physiological implementation or biophysical constraints, but it demonstrates that the statistics of natural images require more degrees of freedom for redundancy reduction in a population response than a classical static divisive normalization model can offer. Our heuristic mechanism demonstrates that strong redundancy reduction is possible with an adaptation mechanism that faces realistic conditions, i.e. has only access to stimuli encountered in the past.

As we showed above, biologically plausible shapes of the contrast response curve and strong redundancy reduction cannot be easily brought together in a single model. Our dynamical model offers a possible solution to this problem. To what extent this model reflects the physiological reality, however, still needs to be tested experimentally.

The first aspect to test is whether the adaptation of the half-saturation constant reflects the temporal structure imprinted by saccades and fixations as predicted by our study. Previous work has measured adaptation timescales for σ^2 [14,28]. However, these measurements are carried out in anesthetized animals and cannot account for eye movements. Since our adaptation mechanism mainly uses the fact that contrasts at a particular fixation location are very similar it predicts that that adaptive changes of σ^2 should be seen from one fixation location to another when measured under natural viewing conditions.

The mechanism we proposed is only one possible candidate for a *dynamic* contrast gain control mechanism that can achieve strong redundancy reduction. We conclude the paper with defining a measure that can be used to distinguish contrast gain control mechanisms that are likely to achieve strong redundancy reduction from those that do not. As discussed above, a necessary condition for strong redundancy reduction is that the the location and the width of the distribution of $\|\mathbf{y}\|$ implied by a model must match the distribution of unnormalized responses $\|\mathbf{y}\|$ determined by the statistics of natural images. In order to measure the location and the width of the distributions in a way that does not depend on a particular scaling of the data, we plotted the median against the width of the 10%–90%–percentile interval (Figure 5). For the empirical distributions generated by the statistics of the image data we always found a ratio greater than 1.5. We also included a dataset from real human eye movements by Kienzle et al. to ensure the generality of this finding [29] as real fixations could introduce a change in the statistics due to the fact that real observers tend to look at image regions with higher contrasts [30]. All models that yield strong redundancy reduction also exhibit a ratio greater than 1.5. Thus, the ratio of the median to the width of the contrast distribution is a simple signature that can be used to check whether an adaptation mechanism is potentially powerful enough for near-optimal redundancy reduction.

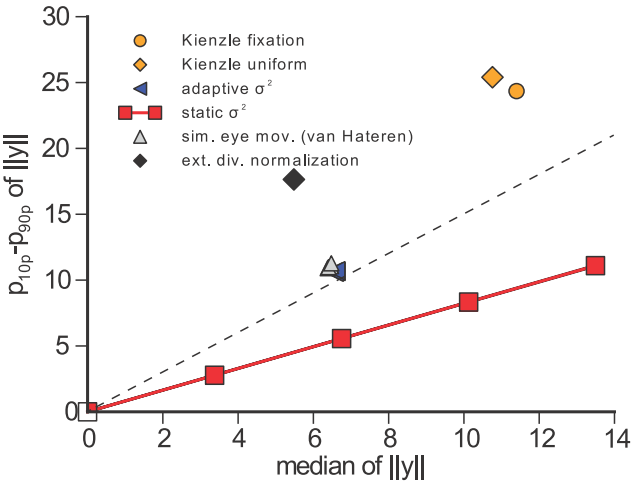

Figure 5. Median vs. width of 10% to 90% percentile interval of the models shown in Figure 3b. The red line corresponds to a static σ^2 for different values of σ^2, the blue triangles correspond to the temporally adapting σ^2, the orange markers correspond to uniformly sampled (diamond) and fixational image patches with Brownian motion micro-saccades (circle) from Kienzle et al. [29], the gray markers to simulated eye movement datasets from van Hateren image data [31], and the black marker to the optimal extended divisive normalization model. All transforms that yield a strong redundancy reduction have models that exhibit a ratio greater than 1.5 (dashed lines).
doi:10.1371/journal.pcbi.1002889.g005

Methods

The code and the data are available online under http://www.bethgelab.org/code/sinz2012.

Data

van Hateren data. For the static experiments, we used randomly sampled 17×17 patches from the van Hateren database [31]. For all experiments we used the logarithm of the raw light intensities. We sampled 10 pairs of training and test sets of 500,000 patches which we centered on the pixel mean.

For the simulated eye movements, we also used 4 pairs of training and test sets. For the sampling procedure, we repeated the following steps until 500,000 samples were drawn: We first drew an image randomly from the van Hateren database. For each image, we simulated ten saccades to random locations in that image. For each saccade location which was uniformly drawn over the entire image, we determined the number m of patches to be sampled from around that location by $m = \lceil v \cdot \tau \rceil$ where $v = 50\text{Hz}$ was the assumed sampling frequency and τ was a sample from an exponential distribution with average fixation time 0.2s (i.e. $\langle m \rangle = 10$). The actual locations of the patches were determined by Brownian motion starting at the saccade location and then propagating with a diffusion constant of $D = (30\text{px})^2/\text{sec}$. This means that each patch location was drawn relative to the previous one based on an isotropic Gaussian centered at the current location with a standard deviation of $\sigma/\sqrt{v} \approx 4.25\text{px}$.

Kienzle data. The van Hateren database is a standard dataset for static natural image statistics. To make sure that our results also hold for real fixations, we sampled data from the images used by Kienzle et al. [29]. We computed the 10% and 90% percentiles, as well as the width of the interval between them, for both datasets for Figure 5.

We constructed two datasets: One where the patches were uniformly drawn from the images, and one where we again used

Brownian motion with a similar standard deviation around human fixation spots to simulate human fixational data. We applied the same preprocessing as for the van Hateren data: centering and whitening.

Models

Both the divisive normalization model and the optimal radial factorization consist of two steps: a linear filtering step and a radial rescaling step (Table 1). In the following, we describe the different steps in more detail.

Filters. The receptive fields of our model neurons, i.e. the linear filters of our models, are given by the rows of a matrix $W = Q\Lambda^{-\frac{1}{2}}U^{\top}A$. In summary, the filters are obtained by (i) projecting the data onto the $n-1$ dimensional subspace that is insensitive to the DC component in the image patches, (ii) performing dimensionality reduction and whitening using principal component analysis, and (iii) training an independent subspace analysis algorithm (ISA) to obtain Q:

(i) The projection of the data onto the $n-1$ dimensional subspace that is insensitive to the DC component is achieved via the matrix A. This matrix is a fixed matrix for which the coefficients in each row sum to zero and all rows are mutually orthogonal. The matrix we used has been obtained via a QR-decomposition as described in the Methods Section of [7].

(ii) The dimensionality reduction and whitening is achieved by $\Lambda^{-1/2}U^{T}$. The matrix U contains the principal components of $A\mathbf{x}$ such that $A\langle\mathbf{x}\mathbf{x}^{\top}\rangle A^{\top} = U\Lambda U^{\top}$. As it is common practice, we kept only the first 72 principal components to avoid "noisy" high frequency filters. However, our analysis would also be valid and lead to the same conclusions if we kept the full set of filters.

(iii) The last matrix Q is constrained to be an orthogonal matrix because the covariance of whitened data remains white under orthogonal transformations. This additional degree of freedom is used by Independent Subspace Analysis (see below) to optimize the filter shapes for redundancy reduction beyond removing second-order correlations. While the matrix Q has a large effect on the particular filter shapes, the same results would have been obtained with any type of whitening filter, i.e. for any orthogonal matrix Q, because they only differ by an orthogonal rotation. Since we use the Euclidean norm in the divisive normalization model, the rotation would not change the norm of the filter responses and therefore all radial distributions would be the same. The only aspect in our analysis for which the filter choice would make a (small) difference is the multi-information of the raw filter responses. When using ICA filter, the multi-information could be a bit lower. However, since even for rather drastic changes of filter shapes (within the class of whitening filters) there is only a small effect on redundancy reduction [6], the particular choice of filter shapes does not affect any of our conclusions. The same is true for any choice of parametric filters as long as the covariance matrix of the filter responses is proportional to the identity matrix. Since the second-order correlations provide the dominant contribution to the multi-information any substantial deviation from the class of whitening filters is likely to yield suboptimal results.

Table 1. Model components of the divisive normalization and radial factorization model: Natural image patches are filtered by a set of linear oriented band-pass filters.

	divisive normalization model	radial factorization
filtering	$\mathbf{y} = W\mathbf{x}$	$\mathbf{y} = W\mathbf{x}$
normalization	$\mathbf{z} = \frac{\kappa\|\mathbf{y}\|_2^{\frac{\gamma}{2}+\delta}}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^{\gamma}}}\frac{\mathbf{y}}{\|\mathbf{y}\|_2}$	$\mathbf{z} = \frac{\left(\mathcal{F}_{\mathcal{I}_p}^{-1}\circ\mathcal{F}_{\rho}\right)(\|\mathbf{y}\|_p)}{\|\mathbf{y}\|_p}\mathbf{y}$
	(static case $\delta=0$ and $\gamma=2$)	

The filter responses are normalized and their norm is rescaled in the normalization step.

doi:10.1371/journal.pcbi.1002889.t001

The independent subspace analysis (with two-dimensional subspaces) used to obtain the matrix Q is based on the model by Hyvärinen [16]:

$$\rho(\mathbf{y}) = \prod_{k=1}^{n/2}\rho_k(y_{2k},y_{2k+1}|\vartheta_k) \; with \; \mathbf{y} = W\mathbf{x} \tag{7}$$

where ϑ_k denotes the list of free parameters for each ρ_k. More specifically, ϑ_k consists of the value p for the L_p-norm and the parameters of the radial distribution for each of the L_p-spherically symmetric distributions. Each single ρ_k was chosen to be a two-dimensional L_p-spherically symmetric distribution [32]

$$\rho_k(\mathbf{y}_{2k:2k+1}|\vartheta_k) = \frac{\varrho_k(\|\mathbf{y}_{2k:2k+1}\|_p|\vartheta_k)}{\|\mathbf{y}_{2k:2k+1}\|_p^{K-1}\mathcal{S}_p^2}$$

$$\|\mathbf{y}\|_p = \left(\sum_{i=1}^{2}|y_i|^p\right)^{\frac{1}{p}}, p>0$$

with a radial γ-distribution $\rho(r|u,s) = \gamma(u,s)$ with shape u and scale s. Therefore, the parameters ϑ_k were given by $\vartheta_k = (p_k,u_k,s_k)$. In the denominator, $\mathcal{S}_p^2$ denotes the surface area of the L_p-norm unit sphere in two dimensions [32]. During training, we first fixed $p = u = 1$; after initial convergence, we retrained the model with free p and u.

The likelihood of the data under equation (7) was optimized by alternating between optimizing Q for fixed ϑ_k, and optimizing the ϑ_k for fixed Q. The gradient ascent on the log-likelihood of Q over the orthogonal group used the backprojection method by Manton [19,33,34]. Optimizing over Q yields filter pairs that resemble quadrature pairs like in the energy model of complex cells [17,18].

Radial rescaling

Optimal contrast gain control: radial factorization. In the following we describe the general mechanism of radial factorization. The spherical symmetric case mostly used in this study is obtained by setting $p = 2$.

Radial factorization is the optimal redundancy reduction mechanism for L_p-spherically symmetric distributed data [11,32]. Samples from L_p-spherically symmetric distributions with identical L_p-norm $r = \|\mathbf{y}\|_p = \left(\sum_{i=1}^{n}|y_i|^p\right)^{\frac{1}{p}}$ are uniformly distributed on the L_p-sphere with that radius. A radial distribution $\varrho(r)$ determines how likely it is that a data point is drawn from an L_p-sphere with that specific radius. Since the distribution on the sphere is uniform for any L_p-spherically symmetric distribution, the radial distribution ϱ determines the specific type of distribu-

tion. For example, $p=2$ and $\varrho(r)=\chi(r)$ yields an isotropic Gaussian since the Gaussian distribution is spherically symmetric ($p=2$) and has a radial χ-distribution ($\varrho(r)=\chi(r)$). One can show that, for a fixed value of p, there is only one type of radial distribution such that the joint distribution is factorial [13]. For $p=2$ this radial distribution is the χ-distribution corresponding to a joint Gaussian distribution. For $0<p\neq 2$, the radial distribution is a generalization of the χ-distribution and the joint distribution is the so called p-generalized Normal [35].

Radial factorization is a mapping on the L_p-norm $r=\|\mathbf{y}\|_p$ of the data points that transforms a given source L_p-spherically symmetric distribution into a p-generalized Normal. To this end, it first models the distribution of r with a flexible distribution ϱ and then nonlinearly rescales r such that the radial distribution becomes a generalized χ-distribution. This is achieved via histogram equalization $\left(\mathcal{F}_{\chi_p}^{-1}\circ\mathcal{F}_{\varrho}\right)(\|\mathbf{y}\|)$ where the $\mathcal{F}$ denote the respective cumulative distribution functions. On the level of joint responses $\mathbf{y}$, radial factorization first normalizes the radius to one and then rescales the data point with the new radius:

$$\mathbf{y}\mapsto\frac{\left(\mathcal{F}_{\chi_p}^{-1}\circ\mathcal{F}_{\varrho}\right)(\|\mathbf{y}\|_p)}{\|\mathbf{y}\|_p}\mathbf{y}$$

In our case ϱ was chosen to be a mixture of five γ-distributions.

When determining the optimal redundancy reduction performance on the population response, we set $p=2$ in order to use the same norm as the divisive normalization model. Only when estimating the redundancy of the linear filter responses, we use $p=1.3$ [11].

Note that the divisive normalization model and the radial factorization model used in this study are invariant with respect to the choice of Q since the Euclidean norm ($p=2$) is invariant under orthogonal transforms. However, the choice of Q would affect the redundancies in the plain filter responses $\mathbf{y}$ in Figure 1B. But even if we had chosen a different Q, i.e. another set of whitening filters, the redundancy between the coefficients of $\mathbf{y}$ would not vary much as previous studies have demonstrated [6,7].

Divisive normalization model and Naka-Rushton distribution. We use the following divisive normalization transform

$$\|\mathbf{y}\|_2\mapsto\frac{\kappa\|\mathbf{y}\|_2}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^2}}$$

which is the common model for neural contrast gain control [8] and redundancy reduction [9].

Divisive normalization acts on the Euclidean norm of the filter responses $\mathbf{y}$. Therefore, divisive normalization can only achieve independence if it outputs a Gaussian random variable. While in radial factorization the target and source distribution were fixed, and the goal was to find a mapping that transforms one into the other, we now fix the mapping to divisive normalization, the target distribution on the normalized response $\mathbf{z}$ to be Gaussian ($\|\mathbf{z}\|_2$ to be χ-distributed) and search for the corresponding source distribution that would lead to a factorial representation when divisive normalization is applied. Since divisive normalization saturates at κ, we will actually have to use a truncated χ-distribution on $\|\mathbf{z}\|_2$. κ becomes the truncation threshold. Note that radial truncation actually introduces some dependencies, but we keep them small by choosing the truncation threshold κ to be the 99% percentile of the radial χ-distribution which is approximately $\kappa\approx 10.14$. The 99% was chosen to keep the target distribution close to a factorial Gaussian. However, it could still be that another cut-off (value of κ) leads to a better redundancy reduction even though the target distribution is less factorial for lower values of κ (quantiles lower than 99%). We made sure that this is not the case by choosing different values of κ, computing the best σ via a maximum likelihood fit of a Naka-Rushton distribution (see below), and estimating the multi-information in the transformed outputs. We found that the choice of κ has virtually no effect on the residual multi-infomation (it varies by $\pm 0.1\%$ for $\kappa\in[\mathcal{F}_{\chi}^{-1}(0.5),\mathcal{F}_{\chi}^{-1}(0.99)]$ and takes its optimum within this interval). Therefore, we kept the 99% choice as it is most similar to the target distribution of radial factorization.

Note also that choosing a Gaussian target distribution does not contradict the finding that cortical firing rates are found to be exponentially distributed [36] since each single response z_i can always be transformed again to be exponentially distributed without changing the redundancy of $\mathbf{z}$.

The distribution on $r=\|\mathbf{y}\|_2$ such that

$$\|\mathbf{z}\|_2=\frac{\kappa\|\mathbf{y}\|_2}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^2}}$$

is truncated χ-distributed can be derived by a simple change of variables. In the resulting distribution

$$\varrho(r)=\frac{2\kappa^n\sigma^2 r^{n-1}}{\mathfrak{G}\left(\frac{n}{2},\frac{\kappa^2}{2s}\right)\Gamma\left(\frac{n}{2}\right)(2s)^{\frac{n}{2}}(\sigma^2+r^2)^{\frac{n+2}{2}}}\exp\left(-\frac{\kappa^2 r^2}{2s(\sigma^2+r^2)}\right),$$

the truncation threshold κ, the half-saturation constant σ, and the scale of the χ-distribution become parameters of the model. The parameter s of the Naka-Rushton distribution controls the variance of the corresponding Gaussian and was always chosen such that the Gaussian was white with variance one. κ was determined by the 99%-percentile. The only remaining free parameter of the Naka-Rushton distribution is σ which simultaneously affects both shape and scale. $\mathfrak{G}$ is the regularized-incomplete-gamma function which accounts for the truncation at κ. We call the distribution *Naka-Rushton distribution* and denote it with $v(\kappa,\sigma,s)$.

To derive the distribution on $\|\mathbf{y}\|$ for which the extended divisive normalization transformation $\frac{\kappa\|\mathbf{y}\|^{\frac{\gamma}{2}+\delta}}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^{\gamma}}}$ yields a χ-distribution, the steps are exactly the same as for the plain divisive normalization transform above. This yields

$$\varrho(r)=\frac{p\kappa^n r^{\frac{n\gamma+2n\delta-2}{2}}\left(2\delta\left(r^{\gamma}+\sigma^2\right)+\gamma\sigma^2\right)}{\Gamma\left(\frac{n}{p}\right)s^{\frac{n}{p}}2^{\frac{n+p}{p}}\left(r^{\gamma}+\sigma^2\right)^{\frac{n+2}{2}}}\times\exp\left(-\frac{\kappa^p r^{\frac{p\gamma}{2}+p\delta}}{2s(\sigma^2+r^{\gamma})^{\frac{p}{2}}}\right)$$

for $\delta>0$. The parameters of the distribution are now $\sigma,\delta,\kappa,\gamma$ and s.

The parameters for all divisive normalization transforms were estimated via maximum likelihood of the Naka-Rushton distribution

on the Euclidean norms $\{r_i\}_{i=1}^m = \{\|\mathbf{y}_i\|_2\}_{i=1}^m$ of the filter responses to natural image patches. As before, we did not optimize for s in the extended Naka-Rushton distribution but fixed it such that the corresponding Gaussian was white.

Dynamically adapting σ^2. For the model with dynamically adapting σ^2, we first model the Euclidean norms $r_i = \|\mathbf{y}_i\|_2$ of the filter responses to the patches from the simulated eye movement data with a mixture of 500 Naka-Rushton distributions

$$\varrho(r) = \sum_{i=1}^{500} v(r|\sigma_i)\pi_i,$$

using EM [37]. π_i denotes the probability that $\sigma = \sigma_i$. The values of σ_i where chosen in 500 equidistant steps from 0.01 to 12.

How much redundancy reduction can be achieved with a dynamically adapting σ, depends on the dynamics according to which it is selected based on the recent history. While there might be many strategies, we chose a parsimonious one based on the mean and the standard deviation of the posterior over σ_{t-1}. Our heuristic consists of two steps: First the mean and the standard deviation of the posterior $\varrho(\sigma|r)$ derived from the mixture distribution is approximated with piecewise linear functions $\mu(r)$ and $\sigma(r)$, then we sample σ_t used to transform r_t from a γ-distribution with mean and standard deviation $\mu(r_{t-1})$ and $\sigma(r_{t-1})$. This strategy emphasizes that the first two moments of the posterior are the important features for obtaining a good σ_t.

In more detail, we evaluated the posterior

$$\varrho(\sigma_i|r) = \frac{\pi_i v(r|\sigma_i)}{\sum_{j=1}^{500} v(r|\sigma_j)\pi_j}.$$

of the mixture distribution at 100 equidistant locations between 10^{-12} and 35, computed the posterior mean and standard deviation at those locations, rescaled the standard deviation by $1/\sqrt{2}$, and fitted the piecewise linear functions on the intervals $[0,1),[1,2),\ldots,[30,\infty)$ to each set of values. In the first interval, the linear function was constraint to start at zero. From these two functions $\mu(r)$ and $\sigma(r)$, we computed two functions for the scale θ and the shape u of a γ-distribution

$$u(r) = \frac{\mu(r)^2}{\sigma(r)^2} \text{ and } \theta(\mathrm{r}) = \frac{\sigma(\mathrm{r})^2}{\mu(\mathrm{r})}$$

via moment matching. We obtained the value σ_t for transforming a value r_t with a Naka-Rushton function by sampling σ_t from a γ-distribution with shape and scale determined by $u(r_{t-1})$ and $\theta(r_{t-1})$.

Computation of percentiles for Figure 5. For the dynamically adapting σ^2 in Figure 5, we sampled from

$$p(r) = \int\int v(r|\sigma,\kappa,s)\gamma(\sigma|u(r_i),\theta(r_i))p(r_i)d\sigma dr_i$$

and computed the percentiles based on the sampled dataset. For the sampling procedure, we drew σ from the γ-distribution $\gamma(\sigma|u(r_i),\theta(r_i))$ with shape and scale computed from r_i and then sampled r from the Naka-Rushton distribution $v(r|\sigma,\kappa,s)$ with that σ. We repeated that for all r_i from a test set of simulated eye movement radii. This procedure was carried out for all pairs of training and test sets, and the distributions fitted to them.

For the static case, we sampled data from single Naka-Rushton distributions for different values of σ and computed the percentiles from the samples.

History-independent choice of σ^2. In the following, let $r_t = \|\mathbf{y}_t\|$ and $\zeta_t = \|\mathbf{z}_t\|$ be the unnormalized and normalized responses at time t, respectively, and $H_k = (r_{t-1},\ldots,r_{t-k})$ be the recent history of responses. The underlying generative structure of the model for temporally correlated data is the following: given a fixed history H_k, σ_t and r_t are sampled from $\rho(\sigma|H_k)$ and $\rho(r_t|H_k)$. Then, ζ_t is generated from r_t and σ_t through divisive normalization.

For strong redundancy reduction, ζ_t should follow a truncated χ-distribution, which means that for given history H_k and σ_t, the unnormalized response energy r_t must have a Naka-Rushton distribution

$$r_t|\sigma_t,H_k \sim v(r_t|\sigma_t),$$

because normalizing this response via $\kappa r_t/\sqrt{\sigma_t^2 + r_t^2}$ yields a truncated χ-distribution. Averaged over all histories H_k and half-saturation constants σ_t^2 the distribution of r_t is a mixture of Naka-Rushton distributions

$$r_t \sim \varrho(r_t) = \int\int v(r_t|\sigma_t,H_k)\rho(\sigma_t,H_k)d\sigma_t dH_k. \tag{8}$$

If σ_t depends deterministically on H_k we obtain equation (5).

If σ_t could be chosen independently of the preceding history the distribution of ζ_t would be given by

$$\begin{aligned}\varrho(\zeta_t) &= \int\int\int \varrho(\zeta_t|r_t,\sigma_t)\varrho(r_t|H_k)\varrho(\sigma_t)\varrho(H_k)dH_k d\sigma_t dr_t \\ &= \int \varrho(\zeta_t|\sigma_t)\varrho(\sigma_t)d\sigma_t,\end{aligned}$$

where $\varrho(\zeta_t|\sigma_t)$ is the marginal distribution of r_t transformed with divisive normalization and a specific value of σ_t. Since redundancy reduction requires $\varrho(\zeta_t)$ to be truncated χ-distributed, σ_t can be chosen independently only if the truncated χ-distribution can be modelled as mixture of the different $\varrho(\zeta_t|\sigma_t)$. Since we assume stationarity, we can drop the index t in the equation.

Multi-information estimation

We use the *multi-information* to quantify the statistical dependencies between the filter responses $\mathbf{y}$ [38]. The multi-information is the n-dimensional generalization of the *mutual-information*. It is defined as the Kullback-Leibler divergence between the joint distribution and the product of its marginals or, equivalently, the difference between the sum of the marginal entropies and the joint entropy

$$I[\mathbf{Y}] = D_{KL}\left(\rho(\mathbf{y}) \| \prod_{i=1}^{n} \rho_i(y_i)\right) = \sum_{i=1}^{n} H[Y_i] - H[\mathbf{Y}]. \tag{9}$$

The multi-information is zero if and only if the different dimensions of the random vector $\mathbf{Y}$ are independent. Since the joint entropy $H[\mathbf{Y}]$ is hard to estimate we employ a semi-parametric estimate of the multi-information that is conservative in the sense that it is downward biased.

For the marginal entropies $H[Y_i]$, we use a jackknifed estimator for the discrete entropy on the binned values [39]. We chose the bin size with the heuristic proposed by Scott [40]. We obtain an

estimate for the differential entropy by correcting with the logarithm of the bin width (see e.g. [7]).

In order to estimate the joint entropy, we use the average log-loss to get an upper bound

$$A[\hat{\rho}(\mathbf{y})] : = -\langle \log \hat{\rho}(\mathbf{y}) \rangle_{\mathbf{Y} \sim \rho(\mathbf{y})} = H[\mathbf{Y}] + D_{KL}(\rho(\mathbf{y}) \| \hat{\rho}(\mathbf{y})).$$

Since the average log-loss overestimates the true entropy, replacing the joint entropy by A in equation (1) underestimates the multi-information. Therefore, we sometimes get estimates smaller than zero. Since the multi-information is always positive, we set the value to zero in that case. For computing errorbars on the multi-information estimations, we use the negative values but a mean zero in such cases, which effectively increases the standard deviation of the error.

Since we want commit ourselves as little as possible to a particular model, we estimate $A[\hat{\rho}(\mathbf{y})]$ by making the assumption that $\mathbf{y}$ is L_p-spherically symmetric distributed but estimating everything else with non-parametric estimators. If $\mathbf{y}$ is L_p-spherically symmetric distributed, the radial component is independent from the directional component [32] and we can write

$$\hat{H}[\mathbf{Y}] = \hat{H}[R] + (n-1)\langle \log r \rangle_R + \log \mathcal{S}_p. \tag{10}$$

The entropy $H[R]$ of the radial component is again estimated via a histogram estimator. The term $(n-1)\langle \log r \rangle_R$ is approximated by the empirical mean.

Putting all the equations together yields our estimator for the multi-information under the assumption of L_p-spherically symmetric distributed $\mathbf{Y}$

$$\hat{I}[\mathbf{Y}] = \sum_{i=1}^{n} \hat{H}[Y_i] - \hat{H}[R] - \frac{(n-1)}{m} \sum_{j=1}^{m} \log r_j - \log \mathcal{S}_p,$$

where $\hat{H}[\cdot]$ are the univariate entropies estimated via binning.

Since the optimal value of p for filter responses $\mathbf{y}$ to natural image patches is approximately $p \approx 1.3$ we use that value to estimate the multi-information of $\mathbf{y}$.

When estimating the multi-information of the responses $\mathbf{z}$ of either divisive normalization or radial factorization, we use the fact that

$$I[\mathbf{Z}] = \sum_{i=1}^{n} H[\mathbf{Z}_i] - H[\mathbf{Z}] = \sum_{i=1}^{n} H[\mathbf{Z}_i] - H[\mathbf{Y}] - \left\langle \log \det \left| \frac{d\mathbf{z}}{d\mathbf{y}} \right| \right\rangle_{\mathbf{Y}}$$

where $\frac{d\mathbf{z}}{d\mathbf{y}}$ is the Jacobian of the normalization transformation. The mean is estimated by averaging over data points. The determinants of radial factorization, divisive normalization, and extended divisive normalization are given by

$$\det \left| \frac{d\mathbf{z}}{d\mathbf{y}} \right| = \frac{\|\mathbf{z}\|_p^{n-1} \varrho(\|\mathbf{y}\|_p)}{\|\mathbf{y}\|_p^{n-1} \chi_p(\|\mathbf{z}\|_p)}$$

$$\det \left| \frac{d\mathbf{z}}{d\mathbf{y}} \right| = \kappa^n \left(\sigma^2 + \|\mathbf{y}\|_2^2\right)^{-\frac{n+2}{2}} \sigma^2$$

$$\det \left| \frac{d\mathbf{z}}{d\mathbf{y}} \right| = \frac{\|\mathbf{z}\|_p^{n-1}}{\|\mathbf{y}\|_p^{n-1}} \frac{\kappa r^{2\frac{\gamma}{2}+\delta-1} \left(2\delta\left(r^\gamma + \sigma^2\right) + \gamma\sigma^2\right)}{2(r^\gamma + \sigma^2)^{\frac{3}{2}}}.$$

All multi-information values were computed on test data.

For the dynamically adapting model, the σ for each data point r_t is sampled from a γ-distribution whose parameters are determined from the previous value r_t and the posterior over σ obtained from the mixture of Naka-Rushton distributions. Since σ changes from step to step it becomes part of the representation and should be included when computing the multi-information (i.e. the redundancy) between the outputs $\mathbf{z}$. Therefore, the redundancy for the dynamically adapting model is measured by $I[Z_1,\ldots,Z_n,\sigma]$. For its computation, we use that $I[Z_1,\ldots,Z_n,\sigma] = I[\mathbf{Z} : \sigma] + I[\mathbf{Z}]$, where $I[\mathbf{Z},\sigma]$ is the mutual information between $\mathbf{Z}$ and σ. In the following, we write $\mathbf{X}|Y$ if $\mathbf{X} \sim \rho(\mathbf{X}|Y)$. Under the assumption that both $\mathbf{Z}$ and $\mathbf{Z}|\sigma$ are spherically symmetric distributed, we can decompose respective random variables into the uniform (on the sphere) and the radial part: $\mathbf{Z} = \mathbf{U} \cdot R$ and $\mathbf{Z}|\sigma = \mathbf{U} \cdot R|\sigma$. This yields

$$\begin{aligned} I[\mathbf{Z},\sigma] &= \left\langle \log \frac{\rho(\mathbf{Z}|\sigma)\rho(\sigma)}{\rho(\mathbf{Z})(\sigma)} \right\rangle = \left\langle \log \frac{\rho(\mathbf{Z}|Y)\rho(\sigma)}{\rho(\mathbf{Z})(\sigma)} \right\rangle \\ &= \left\langle \log \frac{\rho(\mathbf{U})\rho(R|\sigma)\rho(\sigma)}{\rho(\mathbf{U})\rho(R)\rho(\sigma)} \right\rangle = \left\langle \log \frac{\rho(R|\sigma)\rho(\sigma)}{\rho(R)\rho(\sigma)} \right\rangle \\ &= I[R;\sigma], \end{aligned}$$

which means that we can restrict ourselves to the mutual information between the two univariate signals R and σ, which we estimate from a two-dimensional histogram with 100^2 bins.

Acknowledgments

We thank P. Berens, L. Busse, S. Katzner and L. Theis for fruitful discussions and comments on the manuscript.

Author Contributions

Conceived and designed the experiments: FS MB. Performed the experiments: FS. Analyzed the data: FS. Contributed reagents/materials/analysis tools: FS MB. Wrote the paper: FS MB.

References

1. Barlow HB (1961) Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. Sensory Communication. Cambridge, MA: MIT Press. pp. 217–234.
2. Simoncelli EP, Olshausen BA (2003) Natural Image Statistics and Neural Representation. Annual Review of Neuroscience 24: 1193–1216.
3. Bell AJ, Sejnowski TJ (1997) The "independent components" of natural scenes are edge filters. Vision Research 37: 3327–3338.
4. Barlow HB (1989) Unsupervised Learning. Neural Computation 1: 295–311.
5. Lewicki MS, Olshausen BA (1999) Probabilistic framework for the adaptation and comparison of image codes. Journal of the Optical Society of America A 16: 1587–1601.
6. Bethge M (2006) Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? Journal of the Optical Society of America A 23: 1253–1268.
7. Eichhorn J, Sinz F, Bethge M (2009) Natural Image Coding in V1: How Much Use Is Orientation Selectivity? PLoS Comput Biol 5: e1000336.
8. Heeger DJ (1992) Normalization of cell responses in cat striate cortex. Vis Neurosci 9: 181–197.
9. Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control. Nat Neurosci 4: 819–825.
10. Carandini M, Heeger DJ (2011) Normalization as a canonical neural computation. Nature reviews Neuroscience 13: 51–62.

11. Sinz F, Bethge M (2009) The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in neural information processing systems 21: 22nd Annual Conference on Neural Information Processing Systems 2008. Red Hook, NY, , USA: Curran Associates. pp. 1521–1528.
12. Lyu S, Simoncelli EP (2009) Nonlinear extraction of independent components of natural images using radial gaussianization. Neural Computation 21: 1485–1519.
13. Sinz F, Gerwinn S, Bethge M (2009) Characterization of the p-generalized normal distribution. Journal of Multivariate Analysis 100: 817–820.
14. Bonds AB (1991) Temporal dynamics of contrast gain in single cells of the cat striate cortex. Vis Neurosci 6: 239–255.
15. Hyvärinen A, Hoyer P (2000) Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. Neural Computation 12: 1705–1720.
16. Hyvärinen A, Koester U (2007) Complex cell pooling and the statistics of natural images. Network: Computation in Neural Systems 18: 81–100.
17. Pollen D, Ronner S (1981) Phase relationships between adjacent simple cells in the visual cortex. Science 212: 1409–1411.
18. Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A 2: 284–299.
19. Sinz F, Simoncelli EP, Bethge M (2009) Hierarchical Modeling of Local Image Features through Lp-Nested Symmetric Distributions. In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A, editors. Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Red Hook, NY, , USA: Curran Associates. pp. 1696–1704.
20. Lyu S (2011) Dependency Reduction with Divisive Normalization: Justification and Effectiveness. Neural Computation 23: 2942–2973.
21. Wainwright MJ, Simoncelli EP (2000) Scale mixtures of Gaussians and the statistics of natural images. Neural Information Processing Systems 12: 855–861.
22. Wainwright MJ, Schwartz O, Simoncelli EP (2002) Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons. In: Statistical theories of the brain. MIT Press. pp. 203–222.
23. Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A 4: 2379–2394.
24. Ruderman DL, Bialek W (1994) Statistics of natural images: Scaling in the woods. Physical Review Letters 73: 814.
25. Kac M (1939) On a Characterization of the Normal Distribution. American Journal of Mathematics 61: 726–728.
26. Brenner N, Bialek W, De Ruyter Van Steveninck R (2000) Adaptive rescaling maximizes information transmission. Neuron 26: 695–702.
27. Wark B, Lundstrom BN, Fairhall A (2007) Sensory adaptation. Current Opinion in Neurobiology 17: 423–429.
28. Hu M, Wang Y (2011) Rapid Dynamics of Contrast Responses in the Cat Primary Visual Cortex. PLoS ONE 6: e25410.
29. Kienzle W, Franz MO, Schölkopf B, Wichmann FA (2009) Center-surround patterns emerge as optimal predictors for human saccade targets. Journal of Vision 9: 7.1–15.
30. Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. Network 10: 341–350.
31. Van Hateren JH, 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: 359–366.
32. Gupta AK, Song D (1997) Lp-norm spherical distribution. Journal of Statistical Planning and Inference 60: 241–260.
33. Manton JH (2002) Optimization algorithms exploiting unitary constraints. Signal Processing, IEEE Transactions on 50: 635–650.
34. Sinz F, Bethge M (2010) Lp -Nested Symmetric Distributions. Journal of Machine Learning Research 11: 3409–3451.
35. Goodman IR, Kotz S (1973) Multivariate theta]-generalized normal distributions. Journal of Multivariate Analysis 3: 204–219.
36. Baddeley R, Abbott LF, Booth MC, Sengpiel F, Freeman T, et al. (1997) Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proceedings of the Royal Society B Biological Sciences 264: 1775–1783.
37. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological 39: 1–38.
38. Perez A (1977) ε-admissible simplification of the dependence structure of a set of random variables. Kybernetika 13: 439–444.
39. Paninski L (2003) Estimation of Entropy and Mutual Information. Neural Computation 15: 1191–1253.
40. Scott DW (1979) On optimal and data-based histograms. Biometrika 66: 605–610.