Visual Information flow in Wilson-Cowan networks

A. Gomez-Villa,^a M. Bertalmío,^a J. Malo^{b,1}

^a*Dept. Inf. Comm. Tech. Universitat Pompeu Fabra, Barcelona, Spain*

^b*Image Processing Lab, Universitat de València, Valencia, Spain*

E-mail: alexander.gomez@upf.edu, marcelo.bertalmio@upf.edu, jesus.malo@uv.es

ABSTRACT:

In this work we study the communication efficiency of a psychophysically-tuned cascade of Wilson-Cowan and Divisive Normalization layers that simulate the retina-V1 pathway. This is the first analysis of Wilson-Cowan networks in terms of multivariate total correlation. The parameters of the cortical model have been derived through the relation between the steady state of the Wilson-Cowan model and the Divisive Normalization model.

The communication efficiency has been analyzed in two ways: First, we provide an analytical expression for the reduction of the total correlation among the responses of a V1-like population after the application of the Wilson-Cowan interaction. Second, we empirically study the efficiency with visual stimuli and statistical tools that were not available before: (1) we use a recent, radiometrically calibrated, set of natural scenes, and (2) we use a recent technique to estimate the multivariate total correlation in *bits* from sets of visual responses which only involves univariate operations, thus giving better estimates of the redundancy.

The theoretical and the empirical results show that although this cascade of layers was not optimized for statistical independence in any way, the redundancy between the responses gets substantially reduced along the neural pathway. Specifically, we show that (1) the efficiency of a Wilson-Cowan network is similar to its equivalent Divisive Normalization model, (2) while initial layers (Von-Kries adaptation and Weber-like brightness) contribute to univariate equalization, the bigger contributions to the reduction in total correlation come from the computation of nonlinear local contrast and the application of local oriented filters, and (3) psychophysically-tuned models are more efficient (reduce more total correlation) in the more populated regions of the luminance-contrast plane. These results are an alternative confirmation of the *Efficient Coding Hypothesis* for the Wilson-Cowan systems. And from an applied perspective, they suggest that neural field models could be an option in image coding to perform image compression.

¹Corresponding author

Contents

1	Introduction	1
2	Models, materials and methods	2
2.1	Model: A psychophysically-tuned Wilson-Cowan network	2
2.2	Material: radiometrically calibrated stimuli and color adaptation	8
2.3	Method: measuring Total Correlation with Gaussianization transforms	10
3	Results: Redundancy reduction via the Wilson-Cowan interaction	11
3.1	Theoretical analysis	11
3.1.1	Expression for Total Correlation in Wilson-Cowan systems.	11
3.1.2	Experiment in a reduced-scale system.	12
3.2	Empirical analysis of the full-scale model	16
3.2.1	Marginal equalization	16
3.2.2	Mutual Information along the network	18
3.2.3	Total Correlation along the network	21
4	Discussion	24
A	Appendix: Psychophysically inspired reduced-scale model	26

1 Introduction

The Wilson-Cowan equations [1] and Divisive Normalization [2] are alternative influential models of interaction between cortical neurons. While the Divisive Normalization computation has been extensively studied from an information theoretic perspective [3–7], the efficiency of Wilson-Cowan networks has not been analyzed in such detail through accurate redundancy measures.

Such analysis is interesting not only because of the classical *Efficient Coding Hypothesis* [8–10], which is the information-theoretic version of the *free-energy principle* [11], but also because of practical applications. Note that equivalent analysis in the Divisive Normalization case led to substantial improvements in image compression using brain inspired architectures [12, 13]. In a low noise context [14–16] optimal systems are those that minimize the redundancy among the components of the signal representation [5, 17, 18].

In this paper we quantify how effective is each layer of the considered network in achieving the density factorization goal. This information theoretic analysis is applied to a recent psychophysically-tuned Wilson-Cowan model [19, 20], using a recent radiometrically calibrated database of visual scenes [21, 22], and using a recent accurate statistical tool to measure redundancy [23, 24].

The appropriate (multivariate) concept to describe the redundancy in vectors (or neural populations) is the *total correlation* [25, 26]. Direct computation of total correlation from its definition is not straightforward because it involves multivariate probability density estimation. That is why, in the past, different computationally convenient (bivariate) surrogates of the total correlation have been used. For instance, the conditional probability between pairs of responses [3, 4, 27–29], the correlation between the energies of the responses [30–34], or the mutual information between pairs of responses [5, 12, 35].

Now, by using a recent multivariate Gaussianization technique [23, 24], total correlation can be estimated from (easy to compute) marginal equalizations, so we can use the proper measure of communication efficiency in psychophysically-tuned networks. This study, which measures total correlation on Wilson-Cowan responses for the first time, generalizes previous analysis that used total correlation but were restricted to linear models to avoid estimation problems [17], or addressed simpler (just-one-layer) linear+nonlinear models with bivariate mutual information measures [5].

2 Models, materials and methods

In this section we first describe the considered model for the retina-V1 pathway that consists of a cascade of equivalent Divisive Normalization and Wilson-Cowan modules. Then, we present the visual stimuli over which the considered model will be applied. And finally, we will review the statistical methods to assess the communication efficiency of the model.

2.1 Model: A psychophysically-tuned Wilson-Cowan network

In this work the theory is illustrated in the context of models of the retina-cortex pathway. The considered framework follows the program suggested in [36] and implemented in [7]: a cascade of four isomorphic *linear+nonlinear* modules. These modules address brightness, contrast, frequency filtered contrast masked in the spatial domain, and orientation/scale masking. An example of the transforms of the input in such models is shown in Fig. 1.

In this illustration the input is the spatial distribution of the *spectral irradiance* at the retina. This input undergoes the following transforms: (1) The linear part of the first layer consist of three positive LMS spectral sensitivities and a linear recombination of the LMS values with positive/negative weights. This leads to three tristimulus values in each spatial location: one of them is proportional to the luminance, and the other two have opponent chromatic meaning (red-green and yellow-blue). The linear tristimulus values are normalized by the corresponding values of the *white* in the scene (Von-Kries adaptation), and then, the normalized opponent responses undergo parallel saturation transforms. Perception of *brightness* is mediated by an adaptive Weber-like nonlinearity applied to the luminance at each location. This nonlinearity enhances the response in the regions with small linear input (low luminance). (2) The linear part of the second layer computes the deviation of the brightness at each location from the local brightness. Then, this deviation is nonlinearly normalized by the local brightness to give the local nonlinear contrast. (3) The responses to local contrast are convolved by center surround receptive fields (or filtered by the Contrast Sensitivity Function). Then the linearly filtered contrast is nonlinearly normalized by the

Figure 1: Cascade of Linear+Nonlinear layers. The network addresses in turn (1) tristimulus-from-irradiance integrals and color opponency + Von-Kries adaptation together with Weber-like brightness and nonlinear opponent channels, (2) local contrast (subtraction of local mean + division by local mean), (3) CSF filter + masking in the spatial domain, and (4) Wavelet filters + masking in the wavelet domain. See details on the formulation of each Divisive Normalization layer in [7]. In this work we substitute the last nonlinearity by the equivalent Wilson-Cowan model.

local contrast. Again normalization increases the response in the regions with small input (low contrast). (4) After a linear wavelet transform modelling the response of simple cells in V1, each response is normalized by the activity of the neurons in the surround. Again, the activity relatively increases in the regions with low input. As shown in the marginal and joint PDFs below, the common effect of the nonlinear modules throughout the network is response equalization.

Divisive Normalization is the conventional model used to describe the nonlinearities in contrast perception psychophysics [37, 38], but here we will also explore the equivalent Wilson-Cowan model in the last layer.

Below we introduce the notation of both interaction models (the Divisive Normalization and the Wilson-Cowan models), and the relation between them so that we can infer psychophysically plausible parameters for the Wilson-Cowan model from the parameters already tuned for Divisive Normalization.

Modelling cortical interactions. In the case of the V1 cortex, we refer to the set of responses of a population of simple cells as the vector $\mathbf{r}$. The considered models (Divisive Normalization and Wilson-Cowan) define a nonlinear mapping, $\mathcal{N}$, that transforms the input vector $\mathbf{r}$ (before the interaction among neurons) into the output vector $\mathbf{x}$ (after the interaction),

$$\mathbf{r} \xrightarrow{\mathcal{N}} \mathbf{x} \quad (2.1)$$

In this setting, responses are called *excitatory* or *inhibitory*, depending on the corresponding *sign* of the signal: $\mathbf{r} = \text{sign}(\mathbf{r})|\mathbf{r}|$, and $\mathbf{x} = \text{sign}(\mathbf{x})|\mathbf{x}|$. The map $\mathcal{N}$ is an adaptive saturating transform, but it preserves the sign of the responses (i.e. $\text{sign}(\mathbf{x}) = \text{sign}(\mathbf{r})$). Therefore, the models care about cell activation (the modulus $|\cdot|$) but not about the excitatory or inhibitory nature of the sensors (the $\text{sign}(\cdot) = \pm$).

We will refer to as the *energy* of the input responses to the vector $\mathbf{e} = |\mathbf{r}|^\gamma$, where this is an element-wise exponentiation of the amplitudes $|r_i|$. Given the sign-preserving nature of the nonlinear mapping, for the sake of simplicity in notation, in the rest of the paper the variables $\mathbf{r}$ and $\mathbf{x}$ refer to the activations $|\mathbf{r}|$ and $|\mathbf{x}|$.

The Divisive Normalization model. *Forward transform:* The input-output transform in the Divisive Normalization is (in matrix notation [7]),

$$\mathbf{x} = \mathbb{D}_{\mathbf{k}} \cdot \mathbb{D}_{(\mathbf{b} + \mathbf{H} \cdot \mathbf{e})}^{-1} \cdot \mathbf{e} \quad (2.2)$$

where the output vector of nonlinear activations in V1, $\mathbf{x}$, depends on the energy of the input linear wavelet responses, $\mathbf{e}$, which are dimension-wise normalized by a sum of neighbor energies. Note that in this matrix notation, $\mathbb{D}_{\mathbf{v}}$, stands for a diagonal matrix with the vector, $\mathbf{v}$, in the diagonal. The non-diagonal nature of the interaction kernel $\mathbf{H}$ in the denominator, $\mathbf{b} + \mathbf{H} \cdot \mathbf{e}$, implies that the i -th element of the response may be attenuated if the activity of the neighbor sensors, e_j with $j \neq i$, is high. Each row of the kernel $\mathbf{H}$ describes how the energies of the neighbor simple cells attenuate the activity of each simple

cell after the interaction. The each element of the vectors $\mathbf{b}$ and $\mathbf{k}$ respectively determine the semisaturation and the dynamic range of the nonlinear response of each sensor.

Inverse transform: The relation between the two models is easier to obtain by identifying the corresponding decoding transforms in both models. In the case of the Divisive Normalization, the analytical inverse is [7, 12],

$$\mathbf{e} = (I - \mathbb{D}_{\mathbf{k}}^{-1} \cdot \mathbb{D}_{\mathbf{x}} \cdot \mathbf{H})^{-1} \cdot \mathbb{D}_{\mathbf{b}} \cdot \mathbb{D}_{\mathbf{k}}^{-1} \cdot \mathbf{x} \quad (2.3)$$

The Wilson-Cowan model. *Dynamical system:* In the Wilson-Cowan model the variation of the activation vector, $\dot{\mathbf{x}}$, increases with the energy of the input, $\mathbf{e}$, but, for each sensor, this variation is also moderated by its own activity and by a linear combination of the activities of the neighbor sensors,

$$\dot{\mathbf{x}} = \mathbf{e} - \mathbb{D}_{\alpha} \cdot \mathbf{x} - \mathbf{W} \cdot f(\mathbf{x}) \quad (2.4)$$

where $\mathbf{W}$ is the matrix that describes the damping factor between sensors, and $f(\mathbf{x})$ is a dimension-wise saturating nonlinearity (see Fig. 2). Note that in Eq. 2.4 both the inhibitory and the excitatory responses are considered just as negative and positive components of the same vector. Therefore, following [39], the two equations in the traditional Wilson-Cowan formulation are represented here by a single expression.

Steady state and inverse. The stationary solution of the above differential equation, $\dot{\mathbf{x}} = 0$ in Eq. 2.4, leads to the following decoding (input-from-output) relation:

$$\mathbf{e} = \mathbb{D}_{\alpha} \cdot \mathbf{x} + \mathbf{W} \cdot f(\mathbf{x}) \quad (2.5)$$

The identification of the decoding equations in both models, Eq. 2.3 and Eq. 2.5, is the key to obtain simple analytical relations between their parameters.

Equivalence of models. The analytical relation between the steady state of the Wilson-Cowan model and the Divisive Normalization was originally proposed at the Conference

Figure 2: Saturation in the Wilson-Cowan model. *Left:* Illustrative saturating function in blue and linear approximation around the origin in green. *Right:* Derivative of the saturating function decreases with amplitude.

celebrating the 50th anniversary of Prof. Cowan at the University of Chicago [19], but detailed demonstration and discussion of the properties are given in [20].

In summary, just to get a simpler analytical relation between the parameters, in [20], an assumption was done in each model: (1) a first order approximation of the nonlinear saturation of the Wilson-Cowan system (green line in Fig. 2), and (2) a first order expansion of the inverse in decoding the Divisive Normalization (in Eq. 2.3). In this way, it is easy to see that the parameters of both models introduced above are related as:

$$\alpha = \frac{b}{k}$$

$$\mathbf{W} = \mathbb{D}_{\left(\frac{x}{k}\right)} \cdot \mathbf{H} \cdot \mathbb{D}_{\left(\frac{k}{b} \odot \frac{df}{dx}\right)}^{-1} \quad (2.6)$$

where the $\odot$ stands for the dimension-wise Hadamard product, and the divisions are also Hadamard quotients. This expression allows us to obtain the interaction kernel and the attenuation of a Wilson-Cowan model which leads to a steady state compatible with the Divisive Normalization response.

The considered Wilson-Cowan model: parameters and performance. In this work we take an architecture as the one considered in Fig. 1, in which different Divisive Normalization layers were obtained through different experimental methods. For instance, Maximum Differentiation psychophysics was used to get the 2nd and 3rd layers [40], the 1st layer was obtained by fitting human opinion in subjective image distortion [7], and the last layer was tuned to reproduce contrast response curves [41].

In this psychophysically-tuned network, the nonlinearity in the cortical layer (the 4th layer consisting of a linear wavelet transform followed by Divisive Normalization) was substituted by the equivalent Wilson-Cowan interaction. Following [1, 42] we assumed a Gaussian interaction between the sensors tuned to different locations, scales and orientations, and it makes sense to have different attenuation coefficients per subband. However, instead of performing additional psychophysics to determine these parameters, here we took the Divisive Normalization layer optimized in [7, 41], and we applied the relation proposed in [19, 20], i.e. Eq. 2.6. The resulting kernel and attenuation for the Wilson-Cowan interaction are shown in Fig. 3.

Using these parameters, simple Euler integration from an initial output given by the input energy converges to the Divisive Normalization solution, see Fig. 4. A more detailed discussion on the Divisive Normalization as a stable node of this Wilson-Cowan system (theoretical stability, phase diagrams) is given in [20], together with evidences of their perceptual equivalence beyond the mathematical equivalence.

Perceptual evidences in [20] were only focused on the *visual consistence* between the original psychophysically-tuned kernel, $\mathbf{H}$, and the signal-dependent kernel obtained from Eq. 2.6, but *always* within the Divisive Normalization context. Specifically, it was shown that Divisive Normalization with these two different kernels has similar contrast response curves, and achieves similar correlation with human opinion in subjective image quality.

In this introductory section we present a new (more direct) evidence of the perceptual plausibility of the Wilson-Cowan model derived from the psychophysically-tuned Divisive

Figure 3: Parameters in the Wilson-Cowan cortical layer derived from the Divisive Normalization cortical layer. *Left:* The wavelet diagram represents the auto-attenuation values, α_i , for each sensor of the V1-like population. Gray values are linearly scaled to represent attenuation values from the minimum (black, for low frequencies) to the maximum (white, for high frequencies). *Right:* Illustrative interaction kernel for a specific coefficient (the one highlighted in red). Here lighter gray represents stronger interaction between the considered sensor and the rest of the sensors in the wavelet-like structure. Darker gray represents zero interaction. This diagram represents the corresponding row of $\mathbf{W}$ arranged as a wavelet vector. In both cases, α and $\mathbf{W}$, were computed from the average response, $\hat{x}$, in a large image database using the equivalence relation in Eq. 2.6. The width of the Gaussian kernel in space is 0.08 deg, in scale is 1.1 octave, and in orientation is $\pi/6$ rad, and the values of the attenuation vector are in the range $[3 \times 10^3, 1 \times 10^5]$.

Figure 4: Convergence of the Wilson-Cowan model to the Divisive Normalization solution. *Left:* evolution of the relative energy of the update of the solution along the integration. *Right:* evolution of the relative energy of the difference between the Wilson-Cowan solution and the Divisive Normalization response along the integration. The curves in blue are the average of the update and difference over 35 natural images of the Van Hateren database [43], and the intervals in red represent 3 standard deviations below and above the mean.

Figure 5: Wilson-Cowan improves the description of image distortion. *Left:* Alignment between subjective opinion and distance in the input representation. *Right:* Alignment between subjective opinion and distance in after the proposed Wilson-Cowan interaction. Stimuli and ground truth were taken from the subjectively rated image quality database TID [44].

Normalization. In the image quality context, here we explicitly compute the response of the dynamic Wilson-Cowan model (integrating the equation until the steady state is achieved) for the original and the distorted images. And then we check if the proposed Wilson-Cowan representation is more perceptually Euclidean than the input representation (the response after the three initial layers plus the linear wavelet transform).

In Fig. 5 we compare the correlation between the experimental subjective visibility of distortions and the Euclidean distance between responses computed in the wavelet domain, $\mathbf{r}$, and after the convergence of the Wilson-Cowan network, i.e. in the representation $\mathbf{x}$. In our implementation (restricted to patches subtending 0.63 degrees of visual angle), this illustration including 68 distortions applied to one full-size image of the database, implied the integration of the Wilson-Cowan equation in 8970 image regions. The result shows that the considered Wilson-Cowan model does improve the description of the perceived distortion in naturalistic environments with regard to the previous layer of the model.

In summary, the proposed Wilson-Cowan model, with parameters obtained according to Eq. 2.6, has the proper mathematical and perceptual behavior: (a) it converges to the Divisive Normalization solution (Fig. 4), and (b) it describes the visibility of distortions better than the image representation given by the linear simple cells (Fig. 5).

2.2 Material: radiometrically calibrated stimuli and color adaptation

Scenes. In this work we use stimuli from the databases of Foster and Nascimento [21, 22] because (a) the scenes consist of radiometrically calibrated spectra in each pixel, and (b) they also include information about the illumination through gray spheres of controlled reflectance. While the spectral nature of the scenes is useful to simulate the perception process from the initial integration over wavelengths [45, 46], the information about the

Figure 6: Examples of Von-Kries adaptation. *Top:* Representative scenes from [21, 22] under the original illumination. Spectral integration and *tristimulus to digital counts* rendering was done with Colorlab [52]. Note that (in the common illumination setting of the figure) the appearance of the reference gray is quite variable in the different scenes (different brightness, hue, and even saturation). *Right:* Same scenes after Von-Kries adaptation. All the gray references are normalized so their appearance is more stable and also the gamut of colors in the scenes.

illumination allows a straightforward implementation of Von-Kries adaptation with no extra gray-world assumptions [47].

Alternative possibilities to perform Von-Kries adaptation in sensible LMS spaces include the Barcelona database [48], in which CIE XYZ images also include gray spheres; and the IPL database [49, 50], in which CIE XYZ scenes were illuminated using standard CIE D65 and CIE A spectra.

Standard chromatic adaptation transforms. We applied the Stockman and Sharpe LMS fundamentals [51] to 5700 spectral image patches of size $40 \times 40 \times 33$ (including 33 wavelengths in the [380,700] nm range). In this way we got 5700 tristimulus image patches of size $40 \times 40 \times 3$. We assumed that these patches subtend 0.625 degrees of visual angle. This implies assuming certain observation distance and spatial sampling frequency (in our case 64 cycles/deg), which is important to apply spatially calibrated models. Each of these patches was associated to a specific scene, and the *white point* of each scene was computed from the average LMS tristimulus values in the gray spheres of the scene (segmented by hand). Each patch in the original LMS representation was Von-Kries normalized [47] by the tristimulus values of the corresponding *white point* leading to 5700 Von-Kries-adapted LMS images of size $40 \times 40 \times 3$. Fig. 6 shows the transform of different scenes to a common (Von-Kries adapted) canonical chromatic representation.

Information-theoretic considerations on the color manifold. The above chromatic adaptation stage reduces the variability and simplifies the structure of the color manifold in the LMS space because tristimulus measurements corresponding to objects seen under different illuminations are aligned and represented in a common scale [18, 49, 53, 54]. As a result, Von-Kries necessarily reduces the entropy of the color manifold. Additionally, the

considered linear opponent transform [55] is similar to PCA [56], so it strongly reduces the redundancy between the color components. Even though the actual nonlinearities applied to the linear opponent spaces involve the three chromatic channels [47, 49], here we took a simplified (dimension-wise) approximation [46]: we applied saturating nonlinearities only depending on the average tristimulus value in the scene in each separated channel to get the brightness and the corresponding nonlinear RG and YB values. These additional nonlinearities should not modify the redundancy shared among the linear opponent tristimulus values because the total correlation is invariant under dimension-wise transforms [26, 57].

Despite the deep impact of these transforms on the redundancy among the components of the spectra and the components of the colors, in this work we restrict ourselves to the effects of the transforms on the spatial information since the Wilson-Cowan interaction considered above acts on spatial features. Further work may use the data available in the supplementary material and the statistical methods reviewed below to compare the efficiency of the (relatively simple) chromatic transforms used here with regard to more sophisticated color appearance models [47], or other color representations specifically designed for information maximization or error minimization [18, 49, 58].

2.3 Method: measuring Total Correlation with Gaussianization transforms

Information transference from stimuli to response is maximized if the components of the inner representation are statistically independent [14, 15, 59]. Therefore, the appropriate description of the communication efficiency of a perception system consist of tracking the amount of information shared by the different components of the signal along the neural pathway. This redundancy (or shared information) is the *total correlation* [25] or *multi-information* [26]. The total correlation, T , is the Kullback-Leibler divergence between the joint PDF and the product of its marginal PDFs. Unfortunately, direct computation of T from its definition is not straightforward because it involves multivariate PDF estimation in spaces with a huge number of dimensions.

The problems for direct computation of T imply that the quantification of the efficiency of image representations, $\mathbf{x} = \mathcal{N}(\mathbf{r})$, is done using the variations ΔT from the input, $\mathbf{r}$, to the output, $\mathbf{x}$ [26]:

$$\Delta T = T(\mathbf{r}) - T(\mathbf{x}) = \sum_{i=1}^d (h(r_i) - h(x_i)) + \mathbb{E}(\log|\nabla_{\mathbf{r}} \mathcal{N}(\mathbf{r})|) \quad (2.7)$$

where the term with the sum over the dimensions is easy to compute because the marginal entropies, $h(\cdot)$, only depend on univariate PDF estimations, but the expected value of the Jacobian of the transform is required. This second term complicates the estimation and, as a result, sometimes the analysis is restricted to linear transforms [17], where this second term is just a constant; or surrogates of total correlation have to be used, as for instance multiple measures of mutual information between pairs of responses [4, 5, 12], which only involve bivariate PDF estimations.

In this work we solve the above problems by using a novel estimator of T which only relies on univariate estimations: the Rotation-Based Iterative Gaussianization (RBIG) [24].

The RBIG is a cascade of nonlinear+linear layers, each one made of (easy) marginal Gaussianizations followed by an (easy) rotation. This invertible architecture is able to transform any input PDF into a zero-mean unit-covariance multivariate Gaussian even if the chosen rotations are random [23]. This ability to completely remove the structure of any PDF is useful to estimate T of arbitrary vectors $\mathbf{x}$: as the redundancy of a Gaussianized signal is zero, $T(\mathbf{x})$ corresponds to the sum of the individual variations, ΔT_l , that take place along the layers of RBIG while Gaussianizing $\mathbf{x}$. Interestingly, the individual variation in each RBIG layer only depends on (easy to compute) marginal negentropies [23]:

$$T^{RBIG}(\mathbf{x}) = \sum_{l=1}^L \Delta T_l = \sum_{l=1}^L J_m(\mathbf{x}^{(l)}) \quad (2.8)$$

because the marginal negentropies, $J_m(\mathbf{x}^{(l)})$, are just the sum of divergences between the marginal PDFs of the signal that is being Gaussianized at each layer (the vectors $\mathbf{x}^{(l)}$) and a univariate Gaussian.

In the *Results* section the theoretical predictions on efficiency (obtained from the analytical Jacobian of a Wilson-Cowan system plugged into Eq. 2.7) are empirically confirmed by the computationally-convenient RBIG total correlation estimate, Eq. 2.8.

3 Results: Redundancy reduction via the Wilson-Cowan interaction

Here we study the evolution of the statistical dependence between the responses along the psychophysically-tuned neural pathway described in *Models, Materials and Methods*.

First, we present an analytical result for the reduction of total correlation due to a Wilson-Cowan interaction. We check the validity of this theoretical result with an illustrative reduced-scale system. This reduced-scale example is useful not only to confirm the theory but also to illustrate the accuracy of the RBIG estimates of total correlation.

Then, we empirically analyze the behavior of the full-scale model in different ways: (1) by analyzing the shape of the marginal PDFs for different kinds of sensors at different layers of the network, (2) by computing the *mutual information*, I , between the responses of multiple pairs of sensors at different layers, and finally, (3) by computing the (more appropriate) *total correlation*, T , among the responses at the different layers.

3.1 Theoretical analysis

Here, we first present an analytical expression for the most interesting (multivariate) term in the reduction of total correlation for the specific case of the Wilson-Cowan interaction. Afterwards, we exhaustively check the validity of this expression in a reduced-scale example, and we compare the efficiency of the Wilson-Cowan system with the efficiency of the equivalent Divisive Normalization system.

3.1.1 Expression for Total Correlation in Wilson-Cowan systems.

The problem in estimating the variation of total correlation under arbitrary transforms (Eq. 2.7) is the term depending on the Jacobian. The determinant $|\nabla_{\mathbf{r}} \mathcal{N}(\mathbf{r})|$ represents

the compression or expansion of the response space at the point $\mathbf{r}$. Therefore, the last term is the *average* variation of the volume of the response space over the PDF of natural signals. As shown in the experiments below, this term determines the basic trends of the redundancy reduction.

In this section we give an analytical expression for this Jacobian-dependent term for a Wilson-Cowan interaction. We use this expression to predict how the efficiency of the system is going to be for natural images of different luminance/contrast.

Note that in the steady state of the Wilson-Cowan system there is a straightforward expression for the inverse (eq. 2.5). Taking into account that $\nabla_{\mathbf{r}}\mathcal{N}(\mathbf{r}_0) = (\nabla_{\mathbf{x}}\mathcal{N}^{-1}(\mathbf{x}_0))^{-1}$, $\forall \mathbf{r}_0$, where $\mathbf{x}_0 = \mathcal{N}(\mathbf{r}_0)$, the critical term of Eq. 2.7 is:

$$\mathbb{E}(\log|\nabla_{\mathbf{r}}\mathcal{N}|) = \log(\gamma) - \mathbb{E}\left(\log(|\mathbb{D}_{\alpha} + \mathbf{W} \cdot \mathbb{D}_{\frac{df}{dx}}|)\right) - (\gamma^{-1} - 1) \sum_{i=1}^d \mathbb{E}(\log(\alpha_i x_i + v_i)) \quad (3.1)$$

where $\mathbb{D}_{\frac{df}{dx}}$ is a diagonal matrix with the derivative of the sigmoidal functions in each element of the diagonal, and $\mathbf{v} = \mathbf{W} \cdot \mathbf{f}(\mathbf{x})$.

Eq. 3.1 is interesting because one can infer the redundancy reduction as a function of relevant visual features such as luminance and contrast. First, note that the activity of cortical sensors tuned to DC (flat patterns) increase with luminance and the activity of those tuned to AC (textured patterns) increase with contrast. Then, note that the first term in Eq. 3.1 is a constant, but the derivative in the second term decreases with contrast (see Fig. 2), so this *negative* term subtracts less for bigger contrasts. On the contrary, the last term (also negative because typically $\gamma^{-1} > 1$) quickly increases with luminance and contrast, note that both components of this term increase with $\mathbf{x}$.

The above considerations suggest that, for constant contribution of the marginal entropies (which is the case, as shown below), the efficiency of a Wilson-Cowan network should be bigger in the low-luminance / low-contrast region of the image space. This trend suggested by the analysis of the terms in Eq. 3.1 is interesting because natural images typically have low luminance and low contrast [16, 28, 60, 61].

3.1.2 Experiment in a reduced-scale system.

Here we consider a simplified example with the basic elements of the considered network but in a reduced-scale scenario: 3-pixel images. In this situation the Jacobian matrices are small so the theory can be visualized and systematically checked. The structure of this reduced-scale perception system is as follows: first, 3-pixel luminance images are transformed into brightness through a pixel-wise (dimension-wise) Weber-like saturation [47]. Then, 3 Fourier-like analyzers extract the DC, the low-frequency, and the high-frequency components of the brightness simulating frequency-selective cortical filters [62]. This frequency representation is weighted by a low-pass transfer function that simulates contrast sensitivity [63]. Finally, this set of responses interact either according to a static Divisive Normalization transform, or dynamically through the equivalent Wilson-Cowan equation. The specific parameters of this reduced-scale system are given in Appendix A.

Figure 7 shows how this system transforms the manifold of 3-pixel luminance images of the considered database. First, note how the frequency sensors (second scatter plot) look

Figure 7: Natural images through the reduced-scale system. The scatter plots (2×10^3 samples randomly chosen from a set of 5×10^6 samples used in the experiments) display the changes on the PDF of the responses of the system to 3-pixel natural images at different layers of the network. *From left to right:* (1) linear responses to luminance at the retina, (2) responses of frequency analyzers applied to brightness (saturated luminance), and (3) two versions of the nonlinear interaction at the cortex: the Divisive Normalization and the Wilson-Cowan responses. Absolute luminance (in cd/m^2) has been normalized by the 95th percentile luminance value.

for the axes of symmetry of the manifold of natural images in the input domain, similarly to PCA [64]. However, the luminance-to-brightness saturating transform (scatter plot not shown) expands the low-luminance region and compresses the high-luminance region so the tail of brighter stimuli (high values in r_1^2 in the second scatter plot) is shorter. This will be discussed in further detail in analyzing the behavior of the full-scale model. Finally, note how the two interaction schemes considered here lead to a sort of factorization of the PDF in four separate blobs in the AC frequency components. This effect is also discussed in the next section with the marginal PDFs of the full-scale model in the wavelet domain.

The deformation of the manifolds shown in Fig. 7 is an interesting illustration of the *multivariate* equalization/factorization effect of the model. This suggests the model(s) are actually operating under an information maximization goal. However, this qualitative intuition has to be quantified. To do so, we computed the average luminance and the RMSE contrast of 5×10^6 three-dimensional samples extracted from the considered natural scenes, and we estimated their distribution in the luminance/contrast plane. Then, these images were injected through the model: on the one hand the static Divisive Normalization response was computed, and, on the other hand, the equivalent Wilson-Cowan system was stimulated with the linear frequency representation of the brightness of each sample until it converged to a steady state (after 500 Euler integration steps). Then, we computed the reduction in total correlation for the stimuli at different locations of the luminance/contrast plane. We did this in two ways: (1) theoretically, through the proposed analytical expression, Eq. 3.1, and (2) empirically, by using the RBIG method reviewed above, Eq. 2.8. In each case, we considered up to 2.5×10^4 samples per location in the luminance/contrast plane to do the estimations.

Fig. 8 shows the elements of the theoretical computation: the final result comes from (i) a Jacobian-dependent term (Eq. 3.1 in the Wilson-Cowan case, and Eqs. 42 and 43 in

Figure 8: Theoretical efficiency of Wilson-Cowan and Divisive Normalization. The theoretical reduction of total correlation for natural images of different luminance/contrast (surfaces at the left) is computed from a term that depends on the Jacobian of the transform (surfaces at the center), and a term that depends on the sum of marginal entropies of the input and output (surfaces at the right). The dark surfaces at the left are the (small) standard deviations of the estimates over 10 different realizations.

[7] in the case of Divisive Normalization), and (ii) and a term that depends on the marginal entropies. As discussed above, the Jacobian-dependent term decreases with luminance and contrast, while the variation of the entropy-dependent term is smaller across the image space. Therefore, the Jacobian determines the general behavior of the system.

Results of the theoretical and the empirical descriptions of the efficiency are compared in Fig. 9 together with the estimated PDF in the luminance/contrast plane. The general trend of the efficiency surfaces for Divisive Normalization and for the Wilson-Cowan model are the same (with both estimation approaches): first, the redundancy is reduced all over the image space (the reduction of total correlation is positive almost everywhere), but more importantly, the efficiency is clearly bigger for the low-luminance / low-contrast region,

Figure 9: Efficiency is consistent with natural image statistics. The general trend of the efficiency surfaces in the luminance/contrast plane for Divisive Normalization and Wilson-Cowan are the same. Moreover, both are consistent with natural image statistics since efficiency is bigger in the most populated region of the image space. The dark surfaces are the (small) standard deviations of the estimates over 10 different realizations.

which also is the most populated region (see the PDF of the same calibrated training set at the left).

These results confirm the accuracy of the empirical estimation of T and the correctness of the theoretical expression for the Wilson-Cowan nonlinearity (or more precisely, they show the consistency between the theoretical and the empirical estimation). Moreover, results show the equivalent behavior of Divisive Normalization and Wilson-Cowan in coding efficiency: both models focus on the same region of the image space. And, more interestingly, they show that psychophysically-inspired nonlinearities have their peak performance in the proper region of the image space even though they were not optimized in any way to that end.

3.2 Empirical analysis of the full-scale model

In this section we study the efficiency of the full-scale model using a purely empirical approach. This is because the size of the Jacobian for each sample is huge and it is difficult to collect a big enough set for a reliable estimation of Eq. 3.1. For instance, in the current implementation of the model (using a 3-scale, 4-orientation, steerable transform [65]), discrete luminance patches of size 40×40 turn into response vectors $\mathbf{x} \in \mathbb{R}^{10025}$, so the Jacobian matrices are 10025×10025 . One could restrict the interest to a selected set of sensors (as done in some empirical analysis below). However, note that this restriction does not imply a reduced-size Jacobian in the theoretical expression. No rows/columns in the Jacobian can be neglected since the evolution of a subset of responses in the Wilson-Cowan integration depends on *all* the responses because, in principle, the matrix $\mathbf{W}$ is dense. And this dense nature makes perceptual sense because (stronger or weaker) there is interaction between all scales/orientations/positions [37, 38].

Here we analyze the behavior of the full-scale model in three different empirical ways: (1) by discussing the shape of the marginal PDFs for the responses of sensors at different layers of the network, (2) by computing the *mutual information*, I , between the responses of multiple pairs of sensors at different layers (as in [5]), and finally, and more interestingly, (3) by computing the *total correlation*, T , among the responses at the different layers. In the latter, different spatial sampling schemes are explored to capture how visual information depends on the field of view.

In all the experiments in this section we start from the responses of the considered model to the 5700 spectral image patches described in the *Materials* section. In each case, the responses at the different layers are further subsampled according to the goal of the specific experiment and to get a representative set for the considered estimation.

3.2.1 Marginal equalization

Descriptions of natural image statistics usually start from the marginal PDFs because they show the basic complexity of the signal (e.g. naive coding would only consider zero-order entropy [57]). Marginal PDFs are relevant because once 2nd order correlation has been removed, marginal non-Gaussianity (or sparsity) is an appropriate description of the total correlation in certain cases [66]. It is also interesting to see how a perception system modifies the marginal PDFs (e.g. eventual equalization, increased sparsity, Gaussianization [16, 28, 60, 67, 68]) because these changes may reveal an information maximization goal [67, 68] and the general multivariate factorization goal [16] may be achieved through more complicated marginal PDFs [5].

The initial layers of the system we are considering (1st to 3rd layer) consist of sensors tuned to specific spatial locations, while the 4th layer is formed by sensors with wavelet-like receptive fields. In the marginal approach considered here we assume the signal is stationary across space and orientation. Therefore we pool together samples from different spatial locations and (in the case of wavelets) also corresponding to different orientations. In all the results in this section we collected 1×10^7 randomly chosen samples to estimate each marginal PDF.

Figure 10: Marginal PDFs of responses of sensors tuned to spatial locations. *Left:* PDFs for the original luminance, Von-Kries luminance, and Brightness. *Center:* Adaptive *luminance to brightness* transform. *Right:* PDF of nonlinear contrast. See [7] for details on specific Divisive Normalization expressions of the transforms.

Figure 10 shows the effect of the layers of sensors tuned to spatial locations over the original PDF of luminance at the retina. We see that the linear luminance representation (blue line in Fig. 10-left) is strongly biased and has several peaks in the high-luminance tail. These different peaks correspond to high reflectance objects seen in different illumination conditions. Note how these peaks disappear after the Von-Kries normalization that puts every scene in the same relative-luminance range. Then, the brightness transform that depends on the background luminance (Fig. 10-center) tends to expand the low-luminance range so that the resulting marginal PDF is relatively more flat.

All the above transforms make no use of spatial or contextual information (apart from the global scaling factors in Von-Kries obtained from average illumination). The variation of the PDFs in these point-wise layers is consistent with equalization goals [67]. This completely changes in the 2nd and 3rd layers that compute contrast from local normalizations of brightness and apply linear center-surround receptive fields whose response is subsequently normalized by the local activity.

After contrast computation the mean is removed (see that the peak in Fig. 10-right is in zero). The distribution is not symmetric around zero because darker regions (now below zero) are more frequent in natural images. Interestingly, the contrast masking through Divisive Normalization in the third layer generates a *bimodal distribution* above and below zero. This effect in Divisive Normalization has been interpreted as predictive coding, where the numerator is predicted from the neighbors in the denominator leading to peaks above and below zero where this prediction is successful. The predictive coding interpretation has been suggested many times [12, 27, 33, 69], and peaks of this kind have been consistently found in [5, 7] using different models and scenes.

Figure 11 shows the marginal PDFs of the responses of sensors tuned to oriented features

Figure 11: Marginal PDFs of responses of sensors tuned to local oriented features. *Left:* PDFs of the responses to distributions of nonlinear contrast of V1-like linear sensors. *Center:* PDFs after Divisive Normalization. *Right:* PDFs after Wilson-Cowan interaction.

of different scales. These sensors are the linear wavelet-like filters at the 4th layer that simulate V1 simple cells, and the corresponding mechanisms after the nonlinear interactions (either Divisive Normalization or Wilson-Cowan). The linear wavelet-like sensors at the 4th layer display heavy tailed PDFs with decreasing variance for finer scales. However the sparsity of these responses is lower than the sparsity of the same filterbank applied on luminance images (results not shown). This lower sparsity maybe because wavelet filters certainly lead to highly sparse (non-Gaussian) responses when applied to luminance images, but the nonlinear-contrast images are substantially different from luminance images (see for example the output of 3rd layer at Fig. 1).

More interestingly, these (moderately) sparse response distributions turn into *bimodal distributions* after the nonlinear interactions, both for Divisive Normalization and Wilson-Cowan. Again (as in Fig. 10-right) the masking interaction generates bimodal PDFs. However, this is not an exclusive feature of Divisive Normalization: it happens in equivalent Wilson-Cowan systems as well. Note that this marginal behavior (two modes around a depression in zero) can be understood in multivariate terms using the visualization of the reduced-scale example in Fig. 7: the projection of the blobs on the different axes leads to the peaks in the marginals.

3.2.2 Mutual Information along the network

Criticisms to linear Independent Component Analysis (ICA) pointed out that sparsity maximization in the marginal PDFs does not guarantee complete statistical independence [16, 70]. This is obvious from the mathematical point of view [66], however, in practice, a substantial amount of work was devoted to point out the existence of residual statistical relations after ICA-like filters had been applied to images. The problems for direct estimation of total correlation mentioned in the *Methods* section, implied a variety of surrogates to measure this remaining redundancy, as for instance the analysis of conditional probabilities

Figure 12: Mutual information between a spatial sensor and its neighbors. Different layers along the network are considered from left to right. The value of auto-mutual information has been set to zero for better visualization. We did that because this value is arbitrarily large (every response contains arbitrarily large amount of information about itself), and it should not be considered in describing the interaction with the neighbors.

of neighbor responses, the so-called bow-ties [3–5, 27–29], the analysis of the correlation between the energies of pairs of responses as a way to identify subspaces with residual relations [30–34], or the measure of mutual information between pairs of responses [5, 12, 35].

Here we quantify the relation between pairs of responses using mutual information, I , as a function of the separation between the sensors in the corresponding feature space (spatial departure for the first three layers and departure in space, scale and orientation for the wavelet-like layer). We applied the straightforward definition of mutual information because each computation reduces to the estimation of a joint (bivariate) PDF and two marginal PDFs. We did that by gathering neighbors from sliding windows of size smaller than the available visual field. In that way we got 8×10^4 samples for each pair of sensors. We computed 10 estimations of I using a randomly chosen subset with 80% of those samples in each estimation. We chose the bin size according to the Silverman rule of thumb [71]. The results in the plots are the average of those estimations.

Figure 12 shows the evolution of the interactions as the signal goes through the three initial layers. In every case, the relation is higher with closer neighbors and decays with distance. However, in the explored spatial range (which is limited, about 0.5 degrees) these pair-wise relations do not drop to zero in layers before contrast computation. The I would eventually arrive to zero for sufficiently large separation, but we didn’t have access to big enough distances. Von-Kries adaptation and brightness computation reduce the amount of paired relations about 0.5 bits, but the substantial change again comes with the introduction of spatial processing in the third layer: the relations with the closer neighbors are smaller, they drop faster, and actually arrive to zero for small departures. The division by the local activity removes the relation with the spatial neighborhood: it seems that the predictive coding interpretation of normalization mentioned above [12, 27, 33, 69] also applies here.

Figure 13 shows the mutual information values for the responses of two sensors (those highlighted in red) with all their neighbors in the response vector at the V1 layer before

Figure 13: Mutual information between two wavelet-like sensors and all their neighbors. *Left:* interactions of a sensor tuned to vertical patterns of middle frequency (highlighted in red). *Right:* interactions of a sensor tuned to diagonal patterns of middle frequency (highlighted in red). *Top:* interactions between the linear sensors. *Top:* interactions between the sensors after the Wilson-Cowan recurrence. Lighter gray indicate higher mutual information. All figures are scaled in the same way. Comparing *top* and *bottom* one can see that the Wilson-Cowan recurrence reduces the statistical relations a little bit (darker diagrams at the bottom).

and after the application of the Wilson-Cowan interaction. Finally, Fig. 14 shows a subset of the results shown in Fig. 13 represented as a surface as in Fig. 12. As previously reported for linear wavelet domains [5, 35, 72], here mutual information also decays with distance in space scale and orientation. After the nonlinear interactions (we only show the Wilson-Cowan result) the relations between the coefficients seem to be reduced.

However, the reduction obtained here is substantially smaller than the one obtained with previous Divisive Normalization models where I was computed in the same way [5]. It is not a specific matter of Wilson-Cowan, but it is related to the multi-layer architecture: results with the current architecture with Divisive Normalization (not shown) are similar

Figure 14: Mutual information between a wavelet-like sensor and its neighbors in space and orientation. The considered sensor is at the center of the domains represented at the left column. *Left:* Decay of the interaction in space (for the same orientation -or subband-). *Right:* Decay of the relation in space for a different orientation. *Top:* result before the Wilson-Cowan recurrence. *Bottom:* result after the Wilson-Cowan recurrence.

to the Wilson-Cowan results in Figs. 13 and 14, i.e. only moderate gain in I . It is not the database either: we made preliminary experiments with the current architecture and Divisive Normalization using the Van Hateren stimuli [43] (the data used in [5]), and the results (not shown) are similar: also a very moderate gain.

The difference in the architecture explored here with regard to the one in [5] is that the current one is *deeper*: the current wavelet-like layer comes after previous layers that already are doing spatially significant Divisive Normalization (in our case the 3rd layer). The presence of previous layers which already reduce the redundancy, may limit the ability of the considered nonlinearities at this later stage. Nevertheless, note that the set of mutual information measures is only a surrogate of the conceptually appropriate measure, the total correlation, which is addressed in the next section.

3.2.3 Total Correlation along the network

In this section we measure the shared information (total correlation) in the responses of sensors covering progressively bigger portions of the visual field, and we see how this shared information modifies along the neural pathway. We take two different spatial sampling strategies: (1) *no-subsampling*, i.e. considering *all* the sensors tuned to the spatial region, so bigger regions imply a bigger number of sensors, and (2) *subsampling*, i.e. taking a fixed number of sensors progressively more separated, so the bigger the visual field, the poorer the sampling of the visual patterns.

An illustration of the dimensions of the response vector considered at the wavelet-like layer in both experiments is shown in Fig. 15. The corresponding selection for the layers

where sensors have spatial meaning is straightforward.

Taking such subsets of responses (or sensors) keeps the dimension small and makes the empirical estimation easier. This specific sampling schemes were proposed to allow gathering response vectors by sliding windows like those in Fig. 15. Collecting big datasets is particularly important since the dimensions of the considered vectors range from 5 (in the *subsampling experiment* in the layers with spatial meaning) up to 321 (in the *no-subsampling experiment* with bigger field of view in the wavelet-like layers). In every case, the total correlation is reported *by coefficient*. All the total correlation measures in this section were computed using 2.5×10^4 samples and RBIG with 600 layers to ensure the convergence to the Gaussian. In each case, 10 realizations of the estimates were computed using 80% of the samples.

The sampling strategies were selected because they lead to a clear pattern in the total correlation in the input representation. This pattern is a safety check in this full-scale situation where the theoretical computation is not feasible due to the size of the Jacobians. Note that given the smoothness of objects in natural scenes, the consideration of bigger visual fields in the *no-subsampling experiment* should include more coherent structures (e.g. bigger portions of objects in the images). As a result, one would expect increased redundancy given the bigger coherence of the visual structures. On the other hand, if a fixed number of sensors is progressively separated covering bigger visual fields, as in the *subsampling experiment*, the statistical relation between the sensors in the extremes of the visual field should decrease with separation.

This is exactly the pattern that is found in the results, see Fig. 16: (1) in luminance

Figure 15: Wavelet sensors in total correlation experiments. Visual angle and spatial sampling strategy. From left to right: increasing visual angle covered by the sensors. Top and bottom: no sub-sampling, versus sparse sampling.

images in the *no-subsampling experiment*, total correlation increases with the size of the visual field, and (2) in luminance images in the *subsampling experiment*, total correlation decreases with the size of the visual field. Note also that the stimuli that can be properly called *images* are those of the *no-subsampling experiment*. In the *subsampling experiment* we just take samples at the corners and at the center of the visual field.

Regarding the efficiency of the signal representations, Fig. 16-left shows how the total correlation is reduced for *images* (note that the signals in the *subsampling experiment* are not strictly *images*). Results show that the initial layers (Von-Kries chromatic adaptation, and Weber-like adaptive brightness) do not contribute to redundancy reduction in images. This is because these operations introduce no fundamental spatial information, so the redundancy between neighbor spatial locations is basically preserved. Redundancy starts to be removed in the third layer (*Nonlinear Contrast*) where the Contrast Sensitivity Function is applied to the local contrast and then the result is nonlinearly transformed using Divisive Normalization. In this layer, bigger reductions are obtained in smaller visual fields due to the (relatively small) size of the masking kernel of Divisive Normalization in the spatial domain (about 0.02 degrees). On the contrary, the bigger the visual field (which involves more complicated visual structures), the more effective is the linear filterbank of wavelet-like receptive fields (*Linear V1* layer). Finally, the nonlinear interaction at the cortical level (after all the previous stages) does not imply substantial reductions of redundancy in terms of total correlation. The values at the *Nonlinear V1* layer in Fig. 16 display the specific result for the Wilson-Cowan interaction, but Table 1 shows that the removed redundancy for the equivalent Divisive Normalization model is virtually the same: in the biggest visual

Figure 16: Reduction of Total Correlation along the layers of the network. Visual signals of different spatial size are considered. *Left:* no spatial subsampling (actual images). *Right:* with spatial subsampling (samples at the corners and the center of the visual field).

Table 1: Removed Redundancy (in %). Percentage of Total Correlation removed at the inner representation with regard to the values at the input representation. Visual fields of different size (in degrees) with no subsampling (left) and with subsampling (right).

	NO-SUBSAMPL.			SUBSAMPL.		
	$\Delta x = 0.24$	$\Delta x = 0.12$	$\Delta x = 0.06$	$\Delta x = 0.44$	$\Delta x = 0.32$	$\Delta x = 0.18$
Divisive Norm.	84 ± 3	67 ± 3	54 ± 2	87 ± 2	84 ± 4	79 ± 3
Wilson-Cowan	85 ± 4	67 ± 3	54 ± 1	86 ± 3	84 ± 3	79 ± 3

field explored for images the inner representation removes about 85% of the total correlation in the input representation.

Fig. 16-right shows the equivalent results for visual fields in which spatial samples were taken from the corners and the center of the considered regions. In this configuration the inner representation also removes about 85% of the total correlation at the input. However, in this case, most of the redundancy reduction comes from the nonlinear contrast computation.

Similarly to what was found in the mutual information analysis, in this cascade (for images) the biggest reduction in redundancy comes from the wavelet-like transform and the contribution of the nonlinear interaction between simple cells is relatively small.

4 Discussion

Alternative confirmation of Barlow’s hypothesis for Wilson-Cowan systems. In this work we performed the first analysis of the communication efficiency of Wilson-Cowan networks in accurate information theoretic terms using the appropriate (multivariate) description of redundancy: the total correlation. Pointing out the efficiency of a psychophysically tuned system is an alternative confirmation of Barlow’s *Efficient Coding Hypothesis*.

This hypothesis states that natural perception systems evolved to efficiently encode natural signals [8–10]. The conventional way to check this hypothesis is training artificial systems with natural scenes assuming some communication efficiency goal (as for instance redundancy reduction), and see the emergence of natural properties from the statistical learning. This logic (*from-statistics-to-perception*) has been applied to explain the linear image representation in V1 [32, 50, 73], the nonlinear interactions [3, 4, 49, 58], and even the origin of visual illusions [18, 74].

However, note that here we reasoned in the opposite direction: we took a psychophysically plausible network and analyzed its behavior in information-theoretic terms after no statistical learning whatsoever. We found that the considered systems have bigger efficiency (bigger reductions in T) in the most populated regions of the image space, see Fig. 9, and they remove about 85% of the total correlation of natural scenes in the input representation, see Fig. 16 and Table 1. The approach used here (*from-perception-to-statistics*) is an alternative way to confirm the Efficient Coding Hypothesis, as suggested in [5].

Efficiency of Wilson-Cowan is similar to Divisive Normalization. Several evidences suggest that the Wilson-Cowan interaction can be as effective in information theo-

retic terms as the Divisive Normalization model. First, consider that the interaction kernel $\mathbf{W}$ derived here from the relation proposed in [19, 20], roughly decays with space, orientation, and scale. The similarity between the interaction kernel and the mutual information in the linear wavelet-like representation (compare Figs. 3-right and 13-left) has been related to the effectiveness of this kind of interactions to capture the structure of the signal [5, 72]. Second, recurrent subtraction of the activity of the neighbors weighted in this way in the integration of the Wilson-Cowan equation lead to similar *bimodal* marginal PDFs, and to similar clusters in the joint PDF as the Divisive normalization (see Figs. 11 and 7). Third, the Jacobian of the Wilson-Cowan nonlinearity induces similar trends in the reduction of total correlation as the Divisive Normalization (see Figs. 8 and 9): both models have better efficiency in the low-luminance / low-contrast region. And finally, note that the redundancy reduction in the full-scale model is roughly the same using the Wilson-Cowan interaction and the Divisive Normalization (see Table 1).

The above statistical similarities between the Wilson-Cowan and the Divisive Normalization models are consistent with the ability of Wilson-Cowan interaction to improve the explanation of pattern visibility, which is similar to Divisive Normalization. Note that the results in Fig. 5 with Wilson-Cowan are consistent with better explanations of subjective image quality reported for Divisive Normalization [7, 75, 76]. These information theoretic and perceptual similarities suggest the Divisive Normalization actually is the steady state of the Wilson-Cowan dynamics, as assumed in [20].

Accuracy of total correlation estimates. A side technical benefit of the theory presented here (the expression of the reduction of total correlation for Wilson-Cowan systems, Eqs. 2.7 and 3.1) is that this kind of theoretical results can be used to check the accuracy of empirical estimates of total correlation in neuroscience applications. This is also true for the equivalent result for Divisive Normalization presented in [7]. Here we directly compare these theoretical results with empirical estimates (in Fig. 9), which was not done in [7].

These ground truth results are relevant because information measures using Gaussianization [23, 24] had been used to identify interesting patterns in other spatio-spectral and spatio-temporal data [77, 78], but in those cases there was no theoretical bound to compare with.

The success of the empirical measure of total correlation [23, 24] shown in this work suggests that it can be used as a goal function or as an analysis tool to optimize artificial systems, or to check natural systems according to the information bottleneck principle [79].

Differences with shallower networks. Despite here we quantify redundancy through a more appropriate concept, the analysis presented in this work is similar to the one done in [5] for a shallower network based on another psychophysically tuned Divisive Normalization. The model consisted of a single linear+nonlinear layer: a wavelet transform plus a linear weighting to simulate the contrast sensitivity followed by a Divisive Normalization of the energies of the weighted wavelet.

However, as noted in section 3.2, the results obtained here differ from those reported in [5]: the reduction of the mutual information with regard to the linear wavelet representation obtained here is substantially smaller. This is consistent with the small reduction in

total correlation with regard to the linear wavelet representation shown in Fig. 16, which was not measured in [5] because the statistical tool was not available. This could mean that the efficiency of this interaction (either implemented through Wilson-Cowan or Divisive Normalization) is not as big as suggested in [5].

On the contrary, we argue that the small gain with regard to the linear wavelet is not a fundamental limitation of the Divisive Normalization or the Wilson-Cowan models, but a simple consequence of the fact that we are looking at a *deeper* layer: the presence of previous layers which already reduce the redundancy limits the ability of the considered nonlinearities a later stage. In fact, the reduced sparsity displayed by our linear wavelet representation in Fig. 11-left has the same origin: the modified effect of wavelets on contrast images (instead of the conventional luminance images).

Reproducible results. Code to reproduce all the empirical results on the behavior of the full-scale model as well as the experiment of the reduced-scale model that includes the theoretical result are available at https://github.com/alviur/information_wc.git

Final remarks. In summary, the considered psychophysically tuned 4-layer network removes 85% of the total correlation from achromatic (luminance) images subtending 0.24 degrees. For this field of view the Wilson-Cowan model has the same communication efficiency as the equivalent Divisive Normalization model. In a reduced-scale scenario, the theoretical and the empirical results on redundancy reduction are consistent, and both show that the psychophysically inspired Wilson-Cowan and Divisive Normalization networks are more efficient in the regions of the image space where natural scenes are more frequent. The above results represent a confirmation of Barlow’s Efficient Coding Hypothesis for Wilson-Cowan models in the *perception-to-statistics* direction (alternative to the conventional *statistics-to-perception* approach).

The similarities in the statistical effect of the Wilson-Cowan and the Divisive Normalization suggest that recurrent neural field models could be an alternative in image coding applications where Divisive Normalization has beaten both JPEG and JPEG2000 [12, 13]. From a more fundamental point of view, future work should use the proposed analysis before the stationary state is reached. This will show how the information efficiency may evolve while adaptation takes place.

A Appendix: Psychophysically inspired reduced-scale model

Overview. The reduced-scale model consist of two *linear+nonlinear* layers: (1) a linear *radiance-to-luminance* transform using a standard Spectral Sensitivity Function, V_λ , in the spectral integration [45], followed by a simple exponential for the *luminance-to-brightness* nonlinearity applied pixel-wise in the spatial domain, that simulates the Weber-Fechner response to luminance [47], and (2) a *linear+nonlinear* layer in which the linear transform is a discrete cosine transform (a orthonormal rotation) followed by a low-pass weighting function that simulate frequency-tuned sensors and the Contrast Sensitivity Function (CSF) [63]. Then, the outputs of the frequency sensors undergo a nonlinear interaction that may be

a Divisive Normalization [2, 7, 41], or its equivalent Wilson-Cowan network, with parameters computed according to Eq. 2.6 [20].

$$\mathbf{x}^0 \xrightarrow{\mathcal{L}^{(1)}} \mathbf{r}^1 \xrightarrow{\mathcal{N}^{(1)}} \mathbf{x}^1 \xrightarrow{\mathcal{L}^{(2)}} \mathbf{r}^2 \xrightarrow{\mathcal{N}^{(2)}} \mathbf{x}^2 \quad (\text{A.1})$$

Transform. The actual inputs of our code are the responses of the linear photoreceptors: 3-pixel image vectors with normalized luminance values, i.e. $\mathbf{r}^1 \in \mathbb{R}^3$. The normalized luminance was computed dividing the absolute luminance in cd/m^2 by the value corresponding to the 95% percentile of the luminance, in our case 260 cd/m^2 .

- The *luminance-to-brightness* transform, $\mathcal{N}^{(1)}$, is just:

$$\mathbf{x}^1 = (\mathbf{r}^1)^\gamma \quad \text{where } \gamma = 0.6 \quad (\text{A.2})$$

- The linear transform of frequency-tuned sensors with CSF gain, $\mathcal{L}^{(2)}$, is:

$$\mathbf{r}^2 = G_{\text{CSF}} \cdot F \cdot \mathbf{x}^1 \quad \text{where} \quad (\text{A.3})$$

$$F = \begin{pmatrix} \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} \\ \sqrt{\frac{1}{2}} & 0 & -\sqrt{\frac{1}{2}} \\ -\sqrt{\frac{1}{6}} & \sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{6}} \end{pmatrix}$$

$$G_{\text{CSF}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.3 \end{pmatrix}$$

- The *Divisive Normalization* of the frequency-tuned sensors, $\mathcal{N}_{\text{DN}}^{(2)}$, is:

$$\mathbf{x}^2 = \text{sign}(\mathbf{x}^2) \odot \mathbb{D}_{\mathbf{k}} \cdot \mathbb{D}_{(\mathbf{b} + \mathbf{H} \cdot |\mathbf{r}^2|^\gamma)}^{-1} \cdot |\mathbf{r}^2|^\gamma \quad \text{where } \gamma = 0.7, \quad \text{and}, \quad (\text{A.4})$$

$$\mathbb{D}_{\mathbf{k}} = \begin{pmatrix} 0.18 & 0 & 0 \\ 0 & 0.03 & 0 \\ 0 & 0 & 0.01 \end{pmatrix}$$

$$H = \mathbb{D}_l \cdot W \cdot \mathbb{D}_r = \begin{pmatrix} 0.06 & 0 & 0 \\ 0 & 0.35 & 0 \\ 0 & 0 & 0.27 \end{pmatrix} \cdot \begin{pmatrix} 0.93 & 0.06 & 0.01 \\ 0.04 & 0.93 & 0.05 \\ 0 & 0.02 & 0.98 \end{pmatrix} \cdot \begin{pmatrix} 0.95 & 0 & 0 \\ 0 & 0.27 & 0 \\ 0 & 0 & 0.13 \end{pmatrix}$$

and the vector of semisaturations, $\mathbf{b}$, is:

$$\mathbf{b} = \begin{pmatrix} 0.08 \\ 0.03 \\ 0.01 \end{pmatrix}$$

- The equivalent Wilson-Cowan interaction, $\mathcal{N}_{\text{WC}}^{(2)}$, is defined by the differential equation 2.4, where the auto-attenuation, $\boldsymbol{\alpha}$, and the interaction matrix, $\mathbf{W}$, are:

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.41 \\ 1.10 \\ 1.30 \end{pmatrix} \quad (\text{A.5})$$

$$\mathbf{W} = \begin{pmatrix} 0.93 & 0.06 & 0.01 \\ 0.04 & 0.93 & 0.05 \\ 0 & 0.02 & 0.98 \end{pmatrix} \quad (\text{A.6})$$

and the saturation function is:

$$f(\mathbf{x}) = c \mathbf{x}^\gamma \quad \text{where} \quad \gamma = 0.4, \quad \text{and}, \quad (\text{A.7})$$

the scaling constant is, $c = \hat{\mathbf{x}}^{1-\gamma}$, and $\hat{\mathbf{x}}$ is the average response over natural images (for the Divisive Normalization transform):

$$\hat{\mathbf{x}} = \begin{pmatrix} 1.12 \\ 0.02 \\ 0.01 \end{pmatrix}$$

This exponent is also used for the definition of energy in Wilson-Cowan, $\mathbf{e} = |\mathbf{r}|^\gamma$.

Note that the interaction neighborhoods have unit volume, $\sum_j W_{ij} = 1 \quad \forall j$, as suggested in [38], and then, the Divisive Normalization kernel is given by the product of this unit-volume neighborhood and two left and right filters in the diagonal matrices, $\mathbb{D}_l$ and $\mathbb{D}_r$ [41]. The values for the semisaturation, $\mathbf{b}$, and the diagonal matrices $\mathbb{D}_l$ and $\mathbb{D}_r$ were inspired by the contrast response results in [41]: we set the semisaturation according to the average response of natural images (low-pass in nature), and we initialized the left and right filters to high-pass. However, afterwards, in order to make $\mathcal{N}_{\text{DN}}$ and $\mathcal{N}_{\text{WC}}$ consistent, we applied the Divisive Normalization over natural images and we iteratively updated the values of the right and left filters according to Eq. 2.6. In the end, we arrived to the values in the above expressions (where the filter at the left is high-pass, but the filter at the right is not). Note that the attenuation in Wilson-Cowan is computed using Eq. 2.6.

Jacobian. The information theoretic computations strongly depend on how the system (locally) deforms the signal representation (e.g. Eq. 3.1). This is described by the Jacobian of the transform with regard to the signal, $\nabla_{\mathbf{r}^1} S = \nabla_{\mathbf{r}^2} \mathcal{N}^{(2)} \cdot \nabla_{\mathbf{x}^1} \mathcal{L}^{(2)} \cdot \nabla_{\mathbf{r}^1} \mathcal{N}^{(1)}$. In this reduced-scale model, this Jacobian (for the Wilson-Cowan case) is:

$$\nabla_{\mathbf{r}^1} S = \gamma_2 \left(\mathbb{D}_{\boldsymbol{\alpha}} + \mathbf{W} \cdot \mathbb{D}_{\frac{df}{dx}} \right)^{-1} \cdot \left(\mathbb{D}_{(\mathbb{D}_{\boldsymbol{\alpha}} \cdot \mathbf{x}^2 + \mathbf{W} \cdot f(\mathbf{x}^2))} \right)^{1-\frac{1}{\gamma_2}} \cdot G_{\text{CSF}} \cdot F \cdot \mathbb{D}_{\gamma_1(\mathbf{r}^1)^{\gamma_1-1}} \quad (\text{A.8})$$

Acknowledgments

This work was partially funded by the Spanish Government and EU FEDER fund through the MINECO grants TIN2015-71537-P and DPI2017-89867-C2-2-R; and by the European Union's Horizon 2020 research and innovation programme under grant agreement number 761544 (project HDR4EU) and under grant agreement number 780470 (project SAUCE).

References

- [1] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, *Kybernetik* **13** (1973), no. 2 55–80.
- [2] M. Carandini and D. Heeger, *Summation and division by neurons in visual cortex*, *Science* **264** (1994), no. 5163 1333–6.
- [3] O. Schwartz and E. Simoncelli, *Natural signal statistics and sensory gain control*, *Nature Neurosci.* **4** (2001), no. 8 819–825.
- [4] J. Malo and J. Gutiérrez, *V1 non-linear properties emerge from local-to-global non-linear ICA*, *Network: Computation in Neural Systems* **17** (2006), no. 1 85–102.
- [5] J. Malo and V. Laparra, *Psychophysically tuned divisive normalization approximately factorizes the pdf of natural images*, *Neural computation* **22** (2010), no. 12 3179–3206.
- [6] R. Coen-Cagli, P. Dayan, and O. Schwartz, *Cortical surround interactions and perceptual salience via natural scene statistics*, *PLOS Computational Biology* **8** (03, 2012) 1–18.
- [7] M. Martínez-García, P. Cyriac, T. Batard, M. Bertalmío, and J. Malo, *Derivatives and inverse of cascaded linear+nonlinear neural models*, *PLOS ONE* **13** (10, 2018) 1–49.
- [8] H. Barlow, *Sensory mechanisms, the reduction of redundancy, and intelligence*, *Proc. of the Nat. Phys. Lab. Symposium on the Mechanization of Thought Process* (1959), no. 10 535–539.
- [9] H. Barlow, *Possible principles underlying the transformation of sensory messages*, in *Sensory Communication* (W. Rosenblith, ed.), pp. 217–234. MIT Press, Cambridge, MA, 1961.
- [10] H. Barlow, *Redundancy reduction revisited*, *Network: Comp. Neur. Syst.* **12** (2001), no. 3 241–253.
- [11] K. Friston, *The free-energy principle: a rough guide to the brain?*, *Trends in cognitive sciences* **13** (2009), no. 7 293–301.
- [12] J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli, *Nonlinear image representation for efficient perceptual coding*, *IEEE Transactions on Image Processing* **15** (2006), no. 1 68–80.
- [13] J. Ballé, V. Laparra, and E. P. Simoncelli, *End-to-end optimized image compression*, in *5th Int. Conf. Learn. Repres., ICLR 2017*, 2017.
- [14] A. J. Bell and T. J. Sejnowski, *An information maximization approach to blind separation and blind deconvolution.*, *Neural Computation* **7** (1995) 1129–1159.
- [15] A. J. Bell and T. J. Sejnowski, *The independent components of natural scenes are edge filters*, *Vision Research* **37** (1997), no. 23 3327 – 3338.
- [16] E. P. Simoncelli and B. A. Olshausen, *Natural image statistics and neural representation*, *Annual Review of Neuroscience* **24** (2001), no. 1 1193–1216.
- [17] M. Bethge, *Factorial coding of natural images: how effective are linear models in removing higher-order dependencies?*, *JOSA A* **23** (2006), no. 6 1253–1268.
- [18] V. Laparra and J. Malo, *Visual aftereffects and sensory nonlinearities from a single statistical framework*, *Frontiers in Human Neuroscience* **9** (2015) 557.
- [19] J. Malo and M. Bertalmio, *Appropriate kernels for divisive normalization explained by Wilson-Cowan equations*, *Celebration of Cowan’s 50th anniv. at Univ. Chicago. ArXiv: Quant. Biol.* <https://arxiv.org/abs/1804.05964> (2018).

- [20] J. Malo, J. Esteve-Taboada, and M. Bertalmío, *Divisive normalization from Wilson-Cowan dynamics*, *ArXiv: Quant. Biol.* <https://arxiv.org/abs/1906.08246> (2019).
- [21] D. H. Foster, K. Amano, and S. M. Nascimento, *Time-lapse ratios of cone excitations in natural scenes*, *Vision research* **120** (2016) 45–60.
- [22] S. M. Nascimento, K. Amano, and D. H. Foster, *Spatial distributions of local illumination color in natural scenes*, *Vision Research* **120** (2016) 39–44.
- [23] V. Laparra, G. Camps-Valls, and J. Malo, *Iterative gaussianization: from ICA to random rotations*, *IEEE Trans. Neural Networks* **22** (2011), no. 4 537–549.
- [24] J. Johnson, V. Laparra, R. Santos, G. Camps, and J. Malo, *Information theory in density destructors*, in *7th Int. Conf. Learn. Repres., ICLR 2019, Workshop on Invertible Normalization Flows*, 2019.
- [25] S. Watanabe, *Information theoretical analysis of multivariate correlation*, *IBM Journal of research and development* **4** (1960), no. 1 66–82.
- [26] M. Studeny and J. Vejnarova, *The Multi-information function as a tool for measuring stochastic dependence*, pp. 261–298. Kluwer, January, 1998.
- [27] R. W. Buccigrossi and E. P. Simoncelli, *Image compression via joint statistical characterization in the wavelet domain*, *IEEE Transactions on Image Processing* **8** (Dec, 1999) 1688–1701.
- [28] Jinggang Huang and D. Mumford, *Statistics of natural images and models*, in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 1, pp. 541–547 Vol. 1, June, 1999.
- [29] A. Hyvärinen, J. Hurri, and J. Väyrynen, *Bubbles: a unifying framework for low-level statistical properties of natural image sequences*, *JOSA A* **20** (2003), no. 7 1237–1252.
- [30] A. Hyvärinen and P. O. Hoyer, *A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images*, *Vision Research* **41** (2001), no. 18 2413–2423.
- [31] A. Hyvärinen and P. Hoyer, *Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces*, *Neural Computation* **12** (2000), no. 7 1705–1720.
- [32] A. Hyvärinen, J. Hurri, and P. Hoyer, *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer, 2009.
- [33] J. Malo, R. Navarro, I. Epifanio, F. Ferri, and J. Artigas, *Non-linear invertible representation for joint statistical and perceptual feature decorrelation*, *Lect. Not. Comp. Sci.* **1876** (2000) 658–667.
- [34] I. Epifanio, J. Gutierrez, and J. Malo, *Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding*, *Pattern Recognition* **36** (2003), no. 8 1799–1811.
- [35] J. Liu and P. Moulin, *Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients*, *IEEE Trans. Im. Proc.* **10** (2001), no. 11 1647–1658.
- [36] M. Carandini and D. J. Heeger, *Normalization as a canonical neural computation*, *Nature Rev. Neurosci.* **13** (2012), no. 1 51–62.

- [37] J. Foley, *Human luminance pattern mechanisms: Masking experiments require a new model*, *Journal of the Optical Society of America A* **11** (1994), no. 6 1710–1719.
- [38] A. B. Watson and J. A. Solomon, *Model of visual contrast gain control and pattern masking*, *JOSA A* **14** (1997), no. 9 2379–2391.
- [39] P. C. Bressloff and J. D. Cowan, *The functional geometry of local and horizontal connections in a model of v1*, *Journal of Physiology-Paris* **97** (2003), no. 2 221 – 236. Neurogeometry and visual perception.
- [40] J. Malo and E. Simoncelli, *Geometrical and statistical properties of vision models obtained via maximum differentiation*, in *SPIE Electronic Imaging*, pp. 93940L–93940L, International Society for Optics and Photonics, 2015.
- [41] M. Martinez, M. Bertalmío, and J. Malo, *In praise of artifice reloaded: Caution with natural image databases in modeling vision*, *Front. Neurosci.* doi: 10.3389/fnins.2019.00008 (2019).
- [42] P. Chossat and O. Faugeras, *Hyperbolic planforms in relation to visual edges and textures perception*, *PLoS Computational Biology* **5** (12, 2009) e1000625.
- [43] J. H. v. Hateren and A. v. d. Schaaf, *Independent component filters of natural images compared with simple cells in primary visual cortex*, *Proceedings: Biological Sciences* **265** (Mar, 1998) 359–366.
- [44] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, *Color image database for evaluation of image quality metrics*, *Proc. Int. Workshop on Multimedia Signal Processing* (Oct., 2008) 403–408.
- [45] G. Wyszecki and W. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & Sons, New York, 1982.
- [46] A. Stockman and D. Brainard, *OSA Handbook of Optics (3rd. Ed.)*, ch. Color vision mechanisms, pp. 147–152. McGraw-Hill, NY, 2010.
- [47] M. Fairchild, *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, 2013.
- [48] J. Vazquez-Corral, C. Párraga, R. Baldrich, and M. Vanrell, *Color constancy algorithms: Psychophysical evaluation on a new dataset*, *Journal of Imaging Science and Technology* **53** (2009), no. 3 31105–1–31105–9.
- [49] V. Laparra, S. Jiménez, G. Camps-Valls, and J. Malo, *Nonlinearities and adaptation of color vision from sequential principal curves analysis*, *Neural Computation* **24** (2012), no. 10 2751–2788.
- [50] M. U. Gutmann, V. Laparra, A. Hyvärinen, and J. Malo, *Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images*, *PloS ONE* **9** (2014), no. 2 e86481.
- [51] A. Stockman and L. T. Sharpe, *The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype*, *Vision Research* **40** (2000), no. 13 1711 – 1737.
- [52] J. Malo and M. Luque, “ColorLab: A Matlab Toolbox for Color Science and Calibrated Color Image Processing.” <http://isp.uv.es/code/visioncolor/colorlab.html>, 2002.
- [53] M. A. Webster and J. D. Mollon, *Colour constancy influenced by contrast adaptation*, *Nature* **373** (1995) 694–698.

- [54] M. A. Webster and J. Mollon, *Adaptation and the color statistics of natural images*, *Vision Research* **37** (1997), no. 23 3283 – 3298.
- [55] C. R. Ingling Jr and B. H.-P. Tsou, *Orthogonal combination of the three visual channels*, *Vision research* **17** (1977), no. 9 1075–1082.
- [56] G. Buchsbaum and A. Gottschalk, *Trichromacy, opponent colours coding and optimum colour information transmission in the retina*, *Proc. Roy. Soc. Lond. B Biol. Sci.* **220** (1983), no. 1218 89–113.
- [57] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Edition. Wiley-Interscience, 2 ed., July, 2006.
- [58] D. MacLeod and T. von der Twer, *The pleistochrome: optimal opponent codes for natural colors*, in *Color Perception: From Light to Object* (D. Heyer and R. Mausfeld, eds.). Oxford Univ. Press, Oxford, UK, 2003.
- [59] A. Hyvärinen and E. Oja, *Independent component analysis: algorithms and applications*, *Neural networks* **13** (2000), no. 4-5 411–430.
- [60] E. Simoncelli, *Statistical models for images: Compression, restoration and synthesis*, in *Conference Record of the Asilomar Conference on Signals, Systems and Computers* (M. Farques and R. Hippenstiel, eds.), vol. 1, pp. 673–678, IEEE Comp Soc, 1998.
- [61] J. Malo, F. Ferri, J. Albert, J. Soret, and J. Artigas, *The role of perceptual contrast non-linearities in image transform quantization*, *Image and Vision Computing* **18** (2000), no. 3 233–246.
- [62] A. Watson, *The cortex transform: Rapid computation of simulated neural images*, *Computer Vision, Graphics and Image Processing* **39** (1987) 311–327.
- [63] F. Campbell and J. Robson, *Application of Fourier analysis to the visibility of gratings*, *Journal of Physiology* **197** (1968) 551–566.
- [64] P. Hancock, R. Baddeley, and L. Smith, *The principal components of natural images*, *Network* **3** (1991) 61–70.
- [65] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, *Shiftable multi-scale transforms*, *IEEE Trans Information Theory* **38** (Mar, 1992) 587–607. Special Issue on Wavelets.
- [66] J. Cardoso, *Dependence, correlation and gaussianity in independent component analysis*, *J. Mach. Learn. Res.* **4** (2003) 1177–1203.
- [67] S. B. Laughlin, *Matching coding to scenes to enhance efficiency*, in *In Braddick, O.J. & Sleight, A.C. (Eds) Physical and Biological Processing of Images*, pp. 42–52, Springer, 1983.
- [68] D. L. Ruderman, *The statistics of natural images*, *Network: computation in neural systems* **5** (1994), no. 4 517–548.
- [69] J. Ballé, V. Laparra, and E. P. Simoncelli, *Density modeling of images using a generalized normalization transformation*, in *Int. Conf. Learn. Rep. (ICLR 16’)*, 2016.
- [70] S. Lyu and E. P. Simoncelli, *Nonlinear extraction of independent components of natural images using radial gaussianization*, *Neural Comp.* **21** (2009), no. 6 1485–1519.
- [71] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

- [72] V. Laparra, J. Gutiérrez, G. Camps-Valls, and J. Malo, *Image denoising with kernels based on natural image relations*, *J. Mach. Learn. Res.* **11** (2010) 873–903.
- [73] B. Olshausen and D. Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, *Nature* **281** (1996) 607–609.
- [74] H. Barlow, *Vision: Coding and Efficiency*, ch. A theory about the functional role and synaptic mechanism of visual aftereffects. Cambridge Univ. Press, Cambridge, UK, 1990.
- [75] V. Laparra, J. Muñoz-Marí, and J. Malo, *Divisive normalization image quality metric revisited*, *JOSA A* **27** (2010), no. 4 852–864.
- [76] A. Berardino, J. Balle, V. Laparra, and E. Simoncelli, *Eigen-distortion of hierarchical representations*, *Adv. Neur. Inf. Proc. NIPS-17* (2017).
- [77] J. E. Johnson, E. Diaz, V. Laparra, M. Mahecha, D. Miralles, and G. Camps-Valls, *Estimating information in Earth data cubes*, *Geophysical Research Abstracts* **Vol. 20** (2018) 12702.
- [78] V. Laparra and R. Santos-Rodríguez, *Spatial/spectral information trade-off in hyperspectral images*, in *2015 IEEE Int. Geosci. Rem. Sens. Symp. (IGARSS)*, pp. 1124–1127, July, 2015.
- [79] A. Painsky and N. Tishby, *Gaussian lower bound for the information bottleneck limit*, *Journal of Machine Learning Research* **18** (2018), no. 213 1–29.