Vision Research 50 (2010) 2213–2222

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Lower bounds on the redundancy of natural images

Reshad Hosseini, Fabian Sinz, Matthias Bethge [*]

Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 23 October 2009
Received in revised form 28 July 2010

Keywords:
Redundancy
Natural image statistics
Multi-information rate
Information theory

ABSTRACT

The light intensities of natural images exhibit a high degree of redundancy. Knowing the exact amount of their statistical dependencies is important for biological vision as well as compression and coding applications but estimating the total amount of redundancy, the multi-information, is intrinsically hard. The common approach is to estimate the multi-information for patches of increasing sizes and divide by the number of pixels. Here, we show that the limiting value of this sequence—the multi-information rate—can be better estimated by using another limiting process based on measuring the *mutual information* between a pixel and a causal neighborhood of increasing size around it. Although in principle this method has been known for decades, its superiority for estimating the multi-information rate of natural images has not been fully exploited yet. Either method provides a lower bound on the multi-information rate, but the *mutual information* based sequence converges much faster to the multi-information rate than the conventional method does. Using this fact, we provide improved estimates of the multi-information rate of natural images and a better understanding of its underlying spatial structure.

1. Introduction

Natural images contain an abundance of structure and regularities which can be quantified as statistical dependencies or redundancy between image pixels. Coding and compression algorithms for photographic images exploit these dependencies for achieving a good performance. Besides technical applications, the statistical regularities in natural images also play an important role for our understanding of sensory coding in the mammalian brain. In a wide range of studies it has been shown that many response properties of neurons in the early visual system such as color opponency, bandpass filtering, contrast gain control and orientation selectivity can be interpreted as mechanisms for removing these redundancies in natural images (Atick & Redlich, 1992; Barlow, 1959; Buchsbaum & Gottschalk, 1983; Karklin & Lewicki, 2008; Linsker, 1990; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001; Simoncelli & Olshausen, 2001; Sinz & Bethge, 2009; Srinivasan, Laughlin, & Dubs, 1982). Quantitative comparisons have shown that these response properties are not all equally effective in removing statistical dependencies. Mechanisms removing second-order correlations in natural images such as color opponency and bandpass filtering yield a large reduction of redundancy. Less pronounced but still substantial is the effect of contrast gain control (Lyu & Simoncelli, 2009; Sinz & Bethge, 2009). For orientation selectivity, however, the potential for redundancy reduction turns out to be much smaller (Bethge, 2006). Since the emergence of orientation selectivity is the most prominent difference in the response properties of V1 neurons compared to the retina it can serve as an important witness on whether neural response properties in cortex can still be interpreted convincingly in terms of redundancy reduction (Eichhorn, Sinz, & Bethge, 2009).

An important unknown that is critical to judging this case is the true total amount of redundancy in natural images. A principled way of quantifying redundancy is to measure the *multi-information* of a distribution (Perez, 1977). The multi-information of a multivariate random variable is the difference between the sum of its marginal entropies and its joint entropy

$$I[X_1:\ldots:X_n]=\sum_{i=1}^{n}H[X_i]-H[X_1,\ldots,X_n].$$

It equals zero if and only if the individual components are statistically independent and is positive otherwise. It measures the information gain caused by statistical dependencies between the single variables. Unlike differential entropy, the multi-information is invariant against arbitrary component-wise transformations both for linear mappings, such as scaling, and nonlinear mappings, such as taking the logarithm.

The conventional approach for estimating the redundancy per pixel—the *multi-information rate*—is to estimate the multi-information for patches of increasing sizes and divide by the number of pixels (Bethge, 2006; Chandler & Field, 2007; Eichhorn et al., 2009; Lee, Wachtler, & Sejnowski, 2002; Lewicki & Olshausen, 1999; Lewicki & Sejnowski, 2000; Lyu & Simoncelli, 2009; Sinz & Bethge, 2009; Wachtler, Lee, & Sejnowski, 2001). In this way we

[*] Corresponding author.
E-mail address: mbethge@tuebingen.mpg.de (M. Bethge).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.07.025

obtain a monotonically increasing sequence converging to the multi-information rate

$$I_\infty = \lim_{n\to\infty} \frac{1}{n} I[X_1 : \ldots : X_n].$$

There is an important trade-off between two different kinds of errors that affect the outcome of this limiting process: On the one hand, the earlier we stop the sequence of increasing patch sizes, the more we ignore long-range dependencies between image pixels and, hence, underestimate the redundancy of natural images. On the other hand, the larger the patch sizes get, the more difficult it becomes to estimate the multi-information reliably due to the increase in dimensionality. Multi-information estimation strongly resembles the problem of estimating the joint density and similarly suffers from the curse of dimensionality: The number of states that need to be estimated grows exponentially with the number of dimensions. This means that more and more regularization is needed to avoid overfitting in high dimensions. As a consequence, with increasing dimensionality it becomes increasingly unlikely to capture all the structure of the density.

The trade-off between ignoring long range correlations for small n and the increasing difficulty to estimate $I[X_1:\ldots:X_n]$ for large n suggests that the estimation of the multi-information rate can be improved substantially if one manages to construct sequences other than $\{\frac{1}{n} I[X_1 : \ldots : X_n]\}_{n=1}^{\infty}$ which converge faster to the same limiting value I_∞.

In this paper, we show that it is possible to construct such a sequence. The basic idea can be illustrated in the case of one-dimensional stationary stochastic processes. From information theory it is known that the conditional entropy converges to the entropy rate of such processes[1] (Cover & Thomas, 2006; Shannon, 1948)

$$\lim_{n\to\infty} \frac{1}{n} H[X_1, \ldots, X_n] = \lim_{n\to\infty} H[X_n | X_{n-1}, \ldots, X_1].$$

Multiplying this equation by (-1) and adding the marginal entropy of the stationary process $H[X_1] = \frac{1}{n}\sum_{k=1}^{n} H[X_k]$ at both sides, yields an analogous relationship for the multi-information rate

$$\begin{aligned} I_\infty &= \lim_{n\to\infty} \frac{1}{n} I[X_1 : \ldots : X_n] = \lim_{n\to\infty} I[X_n : X_{n-1}, \ldots, X_1] \\ &= \lim_{n\to\infty} H[X_n] - H[X_n | X_{n-1}, \ldots, X_1]. \end{aligned} \tag{1}$$

Note that the sequence on the left hand side of Eq. (1) reflects the *multi*-information[2] between all the variables $X_1, \ldots, X_n$ while the sequence on the right hand side reflects the *mutual* information between X_n and $(X_1, \ldots, X_{n-1})$. The mutual information is the special case of the multi-information which measures the statistical dependencies between two random variables only, while it is possible that the dimensionality of the two random variables is different. For example, in our case X_n is a univariate random variable and $(X_1, \ldots, X_{n-1})$ is $(n-1)$-dimensional. The chain rule for the multi-information (Cover & Thomas, 2006)

$$I[X_1 : \ldots : X_n] = \sum_{k=2}^{n} I[X_k : X_{k-1}, \ldots, X_1],$$

shows that the multi-information can be decomposed into a sum of mutual information terms. This suggests that the mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ with $I_n^{inc} := I[X_n : X_{n-1}, \ldots, X_1]$ quantifies the asymptotic *increment* in the multi-information while the conventionally used multi-information based sequence $\left\{I_n^{cum}\right\}_{n=1}^{\infty}$ with $I_n^{cum} := \frac{1}{n} I[X_n : \ldots : X_1]$ constitutes a *cumulative* approach which averages over these increments.

Inspired by an early study in the fifties (Schreiber, 1956), an incremental approach for estimating I_∞ has already been used before in Petrov and Zhaoping (2003) but did not reveal its full potential. Our work elucidates a couple of points that have not been addressed in those papers: First, we revise the mathematical justification for using the incremental approach in case of two-dimensional random fields rather than one-dimensional processes as it is necessary for modeling images. Second, we show that the mutual information based method yields significantly better estimates of I_∞ than the conventional method does while Petrov and Zhaoping (2003) did not provide any comparisons with previous methods. Third, we show how particularly reliable multi-information estimators can be constructed for the incremental approach such that one obtains conservative lower bounds to the multi-information rate. This allows us, fourth, to systematically investigate how the two approaches perform on natural images for different number of dimensions n also far beyond the case of $n = 7$ pixels that was studied in Petrov and Zhaoping (2003). Our best lower bound on the multi-information rate for the van Hateren data set exceeds their estimate by more than 20% and slightly outperforms the bound obtained with the L_p-spherical model (Sinz & Bethge, 2009). It is obtained when using a causal neighborhood of only 25 pixels.

The remaining part of the paper is structured as follows: In Section 2, we introduce the *multi-information* based and the *mutual information* based method for estimating the multi-information rate. In particular, we present a proof for the convergence of the two methods to the same limiting value I_∞ for two-dimensional stationary stochastic processes. In Section 3, we perform experiments on artificial images in order to demonstrate the validity of the method, and apply it to natural images afterwards. Our results show that the incremental method based on conditional distributions performs significantly better and indicates that the multi-information rate of natural images contains a substantial contribution from higher-order moments. We further corroborate this finding by a second set of experiments where we first pre-whiten the images before we fit the local image statistics. In this way, we not only confirm our previous estimates for the multi-information rate but we can also show that the predominant statistical dependencies captured by current models of natural images are of very limited spatial extent. In particular, the increase in the multi-information rate observed for the cumulative method for increasing patch size does not reflect a meaningful contribution of long range correlations but rather an artifact caused by the pixels at the boundary. Finally, in Section 5, we discuss the significance of our results and compare them to existing work.

2. Methods

In order to describe the statistical regularities of natural images, they are often modeled as two-dimensional stationary random fields. For the present study, stationarity is crucial as it is provides the critical link between the *cumulative* and the *incremental* method for computing the multi-information rate. Stationarity means that the random field is invariant under translations with respect to the x- and y-coordinates of the image intensities. In the following, we will first depict the mathematical underpinnings for using the incremental approach in case of two-dimensional stationary random fields. After that we will show that the incremental method is generally superior to the cumulative method, and then we will describe how to construct reliable multi-information and mutual information estimators for the cumulative and the incremental method, respectively. In particular, we will construct conservative estimators such that also the empirical quantities become reliable lower bounds to the multi-information rate.

[1] For continuous random variables it is necessary to additionally assume that the limit exists.

[2] More precisely the multi-information divided by n.

2.1. Mathematical underpinnings

Throughout the paper, we use uppercase letters to denote random variables, bold font to indicate vectors sometimes equipped with an subindex denoting the dimensionality. In particular, we write $I[\mathbf{X}_{1:n}]$ to refer to the multi-information $I[X_1:\ldots:X_n]$ and $I[X_1:\mathbf{X}_{2:n}]$ to refer to the mutual information between X_1 and $(X_2,\ldots,X_n)$.

For the incremental method, we estimate the multi-information rate I_∞ via the mutual information between X_n and $\mathbf{X}_{1:(n-1)}$ for increasing n

$$I_n^{inc} = I[X_n : \mathbf{X}_{1:(n-1)}] = H[X_n] - H[X_n|\mathbf{X}_{1:(n-1)}]. \tag{2}$$

As mentioned in the introduction, I_n^{inc} and I_n^{cum} converge to the true multi-information rate I_∞ for one-dimensional stationary stochastic processes. One subtle complication, hidden in the expression $H[X_n|\mathbf{X}_{1:(n-1)}]$, is that the proof for the one-dimensional case (see Cover & Thomas, 2006) uses stationarity to replace all conditional entropy terms $H[X_k|\mathbf{X}_{1:k-1}]$ in the chain rule decomposition of the joint entropy

$$H[\mathbf{X}_{1:n}] = H[X_1] + \sum_{k=2}^{n} H[X_k|\mathbf{X}_{1:k-1}] = H[X_n] + \sum_{k=2}^{n} H[X_n|\mathbf{X}_{n-k+1:n-1}],$$

with shifted versions $H[X_n|\mathbf{X}_{n-k+1:n-1}]$ where the index of each component is shifted by $(n-k)$. For two-dimensional Markov chains, however, the two-dimensional shape of the causal neighborhood (see Fig. 1) implies that there are always conditional entropy terms $H[X_k|\mathbf{X}_{1:k-1}]$ that cannot be matched by index shifting. Nevertheless, it is possible to show that $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ converges to the same limiting value I_∞ as $\{I_n^{cum}\}_{n=1}^{\infty}$ for all stationary random fields of arbitrary dimensions (Föllmer, 1973). In order to make this theorem more assessable we provide a simple proof for the special case of two dimensions in Appendix A.

2.2. Superiority of the incremental approach over the cumulative approach

Both types of limiting processes, the *cumulative*, multi-information based sequence $\{I_n^{cum}\}_{n=1}^{\infty}$ and the *incremental*, mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$, grow monotonically with n and converge to the true multi-information rate from below. In other words, each sequence defines a lower bound on the multi-information rate that becomes increasingly tighter for large n and in the limit converges to the same value for the multi-information rate. Using the chain rule for the multi-information together with the fact that conditioning reduces entropy we further obtain the following relations

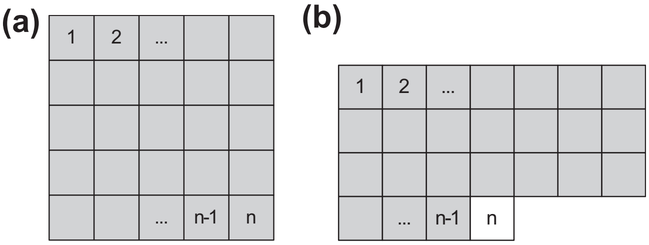

Fig. 1. Illustration of the shape of the image regions for the two different entropy estimation methods: (a) The square shaped patch used for estimating I_n^{cum}. (b) The causal neighborhood used for estimating I_n^{inc}. In this approach we compute the conditional distribution of the white pixel given the gray ones.

$$\begin{aligned} I_n^{cum} &= \frac{1}{n} I[\mathbf{X}_{1:n}] < \frac{1}{n-1} I[\mathbf{X}_{1:n}] \equiv I_n^{cum*} = \frac{1}{n-1}\sum_{k=2}^{n} I[X_k : \mathbf{X}_{1:k-1}] \\ &\leqslant \frac{1}{n-1}\sum_{k=2}^{n} I[X_n : \mathbf{X}_{1:n-1}] = I_n^{inc} \leqslant I_\infty. \end{aligned} \tag{3}$$

First, this demonstrates that I_n^{cum*}, for which the total multi-information is divided by $(n-1)$, is a uniformly better approximation to I_∞ than the conventionally used I_n^{cum}, for which the multi-information is divided by n. While the difference between the two sequences decays very fast, $(I_n^{cum*} - I_n^{cum}) \sim 1/n^2$, the difference between the cumulative and the incremental sequence

$$I_n^{inc} - I_n^{cum*} = \frac{1}{n-1}\sum_{k=2}^{n} (I[X_n : \mathbf{X}_{1:n-1}] - I[X_k : \mathbf{X}_{1:k-1}]),$$

can be quite substantial also for moderately large n. It is zero if and only if $I[X_n:\mathbf{X}_{1:n-1}] = I[X_{n-1}:\mathbf{X}_{1:n-2}] = \cdots = I[X_2:X_1]$ which is equivalent to saying that the process is a stationary Markov process of order one. For all other processes, both cumulative sequences, I_n^{cum} and I_n^{cum*}, always underestimate the true multi-information rate for any finite n. In contrast, for the incremental, mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ it holds $I_n^{inc} = I_\infty$ for any Markov chain model if only the neighborhood $\mathbf{X}_{1:n-1}$ is sufficiently large (i.e. X_n conditioned on $\mathbf{X}_{1:n-1}$ is statistically independent of all other variables). In summary, for any given number of dimensions n, the incremental, mutual information based sequence in general yields better estimates of I_∞ than the cumulative, multi-information based one.

2.3. Cumulative (multi-information based) method

The cumulative method is commonly used for estimating the multi-information rate of natural images. For the sequence of the multi-information of image patches of increasing size we have

$$\begin{aligned} I_n^{cum} &= \frac{1}{n} I[\mathbf{X}_{1:n}] = \frac{1}{n}\sum_{i=1}^{n} H[X_i] - H[\mathbf{X}_{1:n}] \\ &= H[X_1] + \frac{1}{n}\langle \log p(\mathbf{x}_{1:n})\rangle_{\mathbf{X}_{1:n}} \\ &\geqslant -\langle \log p(x_1)\rangle_{X_1} + \frac{1}{n}\langle \log \hat{p}(\mathbf{x}_{1:n})\rangle_{\mathbf{X}_{1:n}} \equiv \widehat{I}_n^{cum}, \end{aligned} \tag{4}$$

where $\hat{p}$ denotes a particular model distribution.

In order to obtain an empirical estimate of I_n^{cum} we use the lower bound given by Eq. (4). The first term is the entropy $H[X_i]$ of the univariate marginal distribution over the pixel intensities which is the same for all $i = 1, \ldots, n$ due to stationarity. Since the problem of estimating this term is identical for both cases, the cumulative as well as the incremental approach, we will discuss it separately at the end of the method section.

The second term in the definition of our estimator $\widehat{I}_n^{cum}$ reflects the *average log-loss* (Bernardo, 1979)

$$-\langle \log \hat{p}(\mathbf{x})\rangle_{\mathbf{X}_{1:n}} = H[\mathbf{X}_{1:n}] + D_{KL}[p\|\hat{p}] \geqslant H[\mathbf{X}_{1:n}],$$

where D_{KL} denotes the Kullback–Leibler divergence, a positive quantity that measures the mismatch between the true and the model distribution. Therefore, the average log-loss has the desirable property that any systematic mismatch between the model distribution $\hat{p}$ and the true distribution p will lead to overestimation of the joint entropy. In this way, we obtain a conservative estimate of the true multi-information rate I_∞.

For estimating the average log-loss, we follow (Eichhorn et al., 2009; Lewicki & Olshausen, 1999; Lewicki & Sejnowski, 2000) and use Monte-Carlo sampling

$$-\langle \log \hat{p}(\mathbf{x})\rangle_{\mathbf{X}_{1:n}} \approx -\frac{1}{m}\sum_{i=1}^{m}\log \hat{p}(\mathbf{x}_i),$$

over a large ensemble of m samples $\mathbf{x}_i$ which differs from the training set used for fitting the parameters of $\hat{p}$.

2.4. Incremental (mutual information based) method

For the incremental approach we employ the same strategy as for the cumulative method: We use the average log-loss of a parametric density for estimating the conditional entropy in Eq. (2) in order to obtain a conservative estimator for I_n^{inc}. In principle, it would be nice to rewrite the conditional entropy in terms of the joint entropy again

$$\begin{aligned} H[X_n|\mathbf{X}_{1:(n-1)}] &= H[\mathbf{X}_{1:n}] - H[\mathbf{X}_{1:(n-1)}] \\ &\approx \frac{1}{n}\langle \log \hat{p}(\mathbf{x}_{1:(n-1)})\rangle_{\mathbf{X}_{1:(n-1)}} - \frac{1}{n}\langle \log \hat{p}(\mathbf{x}_{1:n})\rangle_{\mathbf{X}_{1:n}}, \end{aligned}$$

as it would allow one to use exactly the same parametric density model like in the cumulative method to estimate the joint entropies. The caveat, however, is that the upward bias in the error induced by using the average log-loss when estimating entropies can now occur in both directions.

Therefore, we resort to a different strategy, using the average log-loss directly for estimating the conditional entropy which again yields a lower bound

$$\begin{aligned} \widehat{I}_n^{inc} &\equiv H[X_n] + \langle \log \hat{p}(x_n|\mathbf{x}_{1:(n-1)})\rangle_{\mathbf{X}_{1:n}} && (5) \\ &\leqslant H[X_n] - H[X_n|\mathbf{X}_{1:(n-1)}] = I_n^{inc}. && (6) \end{aligned}$$

Therefore, we have to fit a conditional density model $\hat{p}(x_n|\mathbf{x}_{1:(n-1)})$ rather than a joint density model $\hat{p}(\mathbf{x}_{1:n})$ like in the cumulative approach.

2.5. Parametric density model

For the sake of better comparison, we will use the same Gaussian scale mixture (GSM) model to serve as the parametric model for the average log-loss estimators in both approaches. The GSM model is a rich subfamily of elliptical contoured distributions (Wainwright & Simoncelli, 2000) which have recently been demonstrated to provide a good fit to local patches of natural images (Eichhorn et al., 2009; Lyu & Simoncelli, 2009).

We use a variant of the GSM model which is defined as a mixture of a finite number of zero mean Gaussians with differently scaled versions of the same covariance matrix Σ:

$$p(\mathbf{x}) = \mathrm{GSM}(\mathbf{x}|\boldsymbol{s},\Sigma,\boldsymbol{\lambda}) = \sum_{k=1}^{K}\lambda_k \cdot \mathcal{N}(\mathbf{x}|s_k\cdot\Sigma), \quad \boldsymbol{\lambda},\boldsymbol{s}\in\mathbb{R}^K,$$

where the class probabilities λ_k sum up to one.

For parameter fitting we use an expectation maximization (EM) algorithm. To this end, we define the hidden variable Z indicating which scale is picked for a specific data point $\mathbf{x}$:

$$p_{\mathbf{X}|Z}(\mathbf{x}|k) = \mathcal{N}(\mathbf{x}|s_k\cdot\Sigma) \quad \text{and} \quad p_Z(k) = \lambda_k.$$

For the E-step, we need to compute the probability t_i^k that $Z = k$ given the ith data point

$$t_i^k = p_{Z|\mathbf{X}}(k|\mathbf{x}_i) = \frac{\lambda_k\mathcal{N}(\mathbf{x}_i|s_k\cdot\Sigma)}{\sum_{k=1}^{K}\lambda_k\mathcal{N}(\mathbf{x}_i|s_k\cdot\Sigma)}.$$

In the M-step, for given λ_k and t_i^k, $1\leqslant k\leqslant K$, $1\leqslant i\leqslant m$ we obtain

$$\lambda_k = \frac{\sum_{i=1}^{m}t_i^k}{\sum_{i=1}^{m}\sum_{k=1}^{K}t_i^k}.$$

For computing the scales and the covariance in the M-step, we need to maximize

$$\mathcal{L}(\boldsymbol{s},\Sigma) = \sum_{i=1}^{m}\sum_{k=1}^{K}t_i^k\log\mathcal{N}(\mathbf{x}_i|s_k,\cdot\Sigma).$$

Since the maximum cannot be calculated analytically, we use a block coordinate descent approach. In the first step, we fix $\boldsymbol{s}$ and calculate Σ, in the second step, we fix Σ and calculate $\boldsymbol{s}$, using the equations

$$\Sigma = \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m}\frac{t_i^k}{s_k}\mathbf{x}_i\mathbf{x}_i^\top \quad \text{and} \quad s_k = \frac{\sum_{i=1}^{m}t_i^k\mathbf{x}_i^\top\Sigma^{-1}\mathbf{x}_i}{K\sum_{i=1}^{m}t_i^k}.$$

In our simulations, we find that one or two iteration are enough for the covariance matrix and scale parameters to converge.

In order to use the same distribution for the second method, we calculate the conditional distribution from the GSM model for fixed parameters. This can be done analytically in the GSM model: Let the covariance matrix of Σ of $\mathrm{GSM}(\mathbf{x}_{1:n}|\boldsymbol{s},\Sigma,\boldsymbol{\lambda})$ be

$$\Sigma = \begin{bmatrix} \Sigma_{1:(n-1),1:(n-1} & \Sigma_{1:(n-1),n} \\ \Sigma_{1:(n-1),n} & \Sigma_{n,n} \end{bmatrix}.$$

Marginalizing out the random variable X_n again yields a GSM with parameters

$$\mathrm{GSM}\big(\mathbf{x}_{1:(n-1)}|\boldsymbol{s}_{1:(n-1)},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda}_{1:(n-1)}\big).$$

Then the conditional distribution is just the ratio between the original joint and the marginalized distribution:

$$p_{X_n|\mathbf{X}_{1:(n-1)}}(x_n|\mathbf{x}_{1:(n-1)}) = \frac{\mathrm{GSM}(\mathbf{x}_{1:n}|\boldsymbol{s}_{1:n},\Sigma_{1:n,1:n},\boldsymbol{\lambda}_{1:n})}{\mathrm{GSM}(\mathbf{x}_{1:(n-1)}|\boldsymbol{s}_{1:(n-1)},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda}_{1:(n-1)})}.$$

2.6. Estimation of the univariate pixel entropy

In order to minimize the risk of overestimating the univariate marginal entropy in either of the two approaches, we aim at using a very precise nonparametric approach. To this end we use a histogram based jackknifed maximum likelihood estimator (see e.g. Paninski, 2003). Given m samples with a marginal standard deviation of σ we chose the bin width Δ according to the heuristic proposed by Scott (1979): $\Delta = 3.49\sigma m^{-\frac{1}{3}}$. Since the discrete entropy asymptotically equals the differential entropy plus $-\log\Delta$, we obtain an estimate of the marginal entropy by adding the log of the bin width Δ. Using that method we reliably obtain a value of 1.57 bits per pixel for the univariate pixel entropy. Note that this number like all differential entropies depends on the scale of the pixel intensities. The multi-information rate, however, is independent of the scale as it is computed from differences between differential entropies.

3. Experiments

3.1. Experiment on artificial data

In order to illustrate the two estimation methods, we first compare the cumulative and the incremental approach on an artificial stationary Gaussian random field using the autocorrelation of natural images. To this end, we generated 10.000 images of 60 × 60 pixels by applying a linear transformation A to Gaussian white noise ξ such that the covariance matrix of the resulting Gaussian distribution $\Sigma = AA^\top$ resembles the covariance matrix of the van Hateren data set. We estimated the covariance matrix from samples of 60 × 60 patches using the fact that due to stationarity the covariance between two pixels at location (x,y) and location (x',y'), respectively, must only depend on their relative distance

$(x - x', y - y')$, which results in a symmetric block-Toeplitz covariance matrix.

From those images we sampled ten pairs of training and test sets of 1.000.000 image patches each, for a range of different patch sizes. Fig. 1 shows the shape of the image patch shape used in our two approaches. For the cumulative approach, we use patch sizes $2 \times 2, \ldots, 12 \times 12$. For the incremental approach, we use causal neighborhoods of sizes 5, 13, 25, 41, 61, 85, 113, 145. The parameter estimation for the models was done in exactly the same way as for the natural images below.

As a stationary Gaussian random field is completely defined by the autocorrelation function we can compute the multi-information and the mutual information analytically from $AA^\top$. Fig. 2a shows the result for the full range from 1 to 3600 pixels.

Fig. 2b shows the empirical results obtained for $\widehat{I}_n^{cum}$, $\widehat{I}_n^{cum*}$ and $\widehat{I}_n^{inc}$ as a function of the dimension N when using the average log-loss of a Gaussian model distribution. For comparison, the dashed black lines indicate the true multi-information rate I_∞ obtained analytically from the relevant submatrices Σ_n^{cum} and Σ_n^{inc} of the covariance matrix C needed to compute the multi-information bounds

$$I_n^{cum} = \frac{1}{2n}\left(\sum_{k=1}^{n} \log_2 (\Sigma_n^{cum})_{k,k} - \log_2 \left|\det (\Sigma_n^{cum})\right|\right),$$

$$I_n^{inc} = \frac{1}{2}\left(\log_2 \sigma_n^2 - \log_2 \sigma_{n|1:(n-1)}^2\right),$$

respectively, where

$$\sigma_n^2 := \left(\Sigma_n^{inc}\right)_{n,n},$$

$$\sigma_{n|1:(n-1)}^2 := \sigma_n^2 - \left(\Sigma_n^{inc}\right)_{n,1:(n-1)} \left(\Sigma_n^{inc}\right)^{-1}_{1:(n-1),1:(n-1)} \left(\Sigma_n^{inc}\right)_{1:(n-1),n}.$$

The example visualizes the superiority of the incremental method over the cumulative method. The agreement between the analytical and empirical curves illustrates that the difference between the two methods is not caused by insufficient amount of data or by wrong model assumptions but solely by an unavoidable downward bias of the cumulative method. As apparent from Eq. (3), this downward bias originates from the fact that pixels close to the boundaries suffer from an incomplete neighborhood. Therefore, they do not contribute the full amount of redundancy to the multi-information rate and it requires very large image patches until the pixels in the interior can sufficiently outnumber the pixels at the boundaries. Even at a patch size of 60×60 the cumulative method still underestimates the asymptotic information rate of this stationary Gaussian random field by 0.02 bits per pixel. In other words, the convergence of the cumulative methods is extremely slow even though we are using the correct density model for the evaluation of the average log-loss.

3.2. Natural image dataset and parameter estimation

We perform two blocks of experiments with natural images. In the first block, we use images whose pixel values encode log-intensities. In the second block, we use pre-whitened images generated by a predictive coding scheme that subtracts from each pixel the optimal linear prediction from a causal neighborhood around it. In order to compute the multi-information for the original pixels, we have to account for the whitening transformation. As this whitening step can be described by a linear transform which has vanishing log-Jacobian in the limit, we can lower bound the multi-information rate by the difference of the marginal entropy (1.57 bits) on the pixel domain and the ALLs on the whitened domain:

$$\begin{aligned}
\widehat{I}_n^{inc} &= H[X_n] + \langle \log \hat{p}(y_n|\boldsymbol{y}_{1:(n-1)})\rangle \\
&\leqslant H[X_n] - H[Y_n|\mathbf{Y}_{1:(n-1)}] \\
&\leqslant H[X_n] - \lim_{k\to\infty} H[Y_k|\mathbf{Y}_{1:(k-1)}] \\
&= H[X_n] - \lim_{k\to\infty} \frac{1}{k} H[\mathbf{Y}_{1:k}] \\
\log|\underline{\underline{J}}| &= 0 H[X_n] - \lim_{k\to\infty} \frac{1}{k} H[\mathbf{X}_{1:k}] = I_\infty
\end{aligned}$$

where $\mathbf{Y}_{1:n}$ denotes the whitened pixels. This lower bound is equal to the multi-information estimate after whitening the images, plus the difference of original marginal pixel entropy and marginal pixel entropy of pre-whitened data, i.e.

$$\widehat{I}_n^{inc} = \underbrace{H[Y_n] - \langle \log \hat{p}(y_n|\boldsymbol{y}_{1:(n-1)})\rangle}_{\text{MI estimate in second layer}} + \underbrace{H[X_n] - H[Y_n]}_{\text{marginal entropy difference}}. \tag{7}$$

The difference between the marginal entropies for the van Hateren dataset is equal to 2.9 bits.

For the experiments on natural images we used exactly the same amount of data as in the artificial example described above. That is for each patch size we sampled ten pairs of training and test sets of 1.000.000 log-intensity image patches from the van Hateren database (van Hateren & van der Schaaf, 1998). Again, for the cumulative approach, we use patch sizes $2 \times 2, \ldots, 12 \times 12$. For the incremental approach, we use causal neighborhoods of sizes

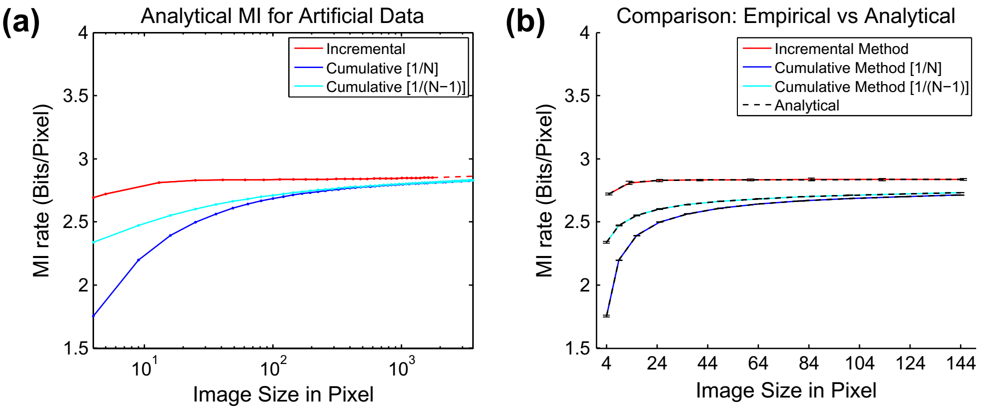

Fig. 2. Verification of the estimation methods on artificial data: Multi-information rate in bits per pixel as estimated by our two methods as a function of the number of pixels. The blue and cyan curves show the result for the cumulative method and the red curve shows the result of the incremental method which significantly outperforms the cumulative ones. The left figure (a) shows the analytic results for the full range of up to $n = 3600$ dimensions using a logarithmic x-axis. The right figure (b) shows an excellent agreement between the analytical and the empirically estimated lower bounds for both methods.

5, 13, 25, 41, 61, 85, 113, 145. For each patch size we run different versions of the GSM model with $K = 1, 4, 7, 10$ scale mixture components. All results shown for $\widehat{I}^{cum}$ and $\widehat{I}^{inc}$ are evaluations on the test set. Importantly, all evaluations on the training set yield identical results so that potential effects due to overfitting can be safely excluded. The error bars in all figures indicate *three* standard deviations over the ensemble of ten different test sets, apart from Fig. 6b where we used *two* standard deviations because of the smaller range of the *y*-axis.

4. Results

Fig. 3 shows the multi-information rate computed with the two different methods for the SGM with $K = 10$ scale mixture components. One can see from the figure that the incremental method significantly outperforms the cumulative one and provides a tighter lower bound.

Fig. 4 shows the estimated multi-information rates for the different methods and different numbers of scale mixture components. The performance seems to saturate for about seven mixture components.

For the incremental method, the lower bound takes a maximum at a neighborhood size of 25 pixels, whereas the cumulative method still exhibits a tiny increase of the lower bound at 144 pixels. This raises two questions:

(1) How can it be that the amount of dependencies captured with the incremental method is decreasing with increasing patch size?
(2) Could it be that the cumulative method is able to better capture long range interactions between pixels and hence at some point can yield a tighter lower bound when using very large image patches?

The first question is motivated by the fact that I_n^{cum} and I_n^{inc} can only increase with increasing patch or neighborhood size. As one can see from the Eqs. (4) and (5), however, the lower bounds $\widehat{I}_n^{cum}$ and $\widehat{I}_n^{inc}$ can still decrease with increasing n if the inequalities $\widehat{I}_n^{cum} \leqslant I_n^{cum}$ and $\widehat{I}_n^{inc} \leqslant I_n^{inc}$ become less and less tight. The differences between the true and the estimated quantities $I_n^{cum} - \widehat{I}_n^{cum}$ and $I_n^{inc} - \widehat{I}_n^{inc}$ equal the Kullback–Leibler distance between the true distribution and the model distribution. If the mismatch of the model distribution becomes larger for increasing patch size, this can result in a lower bound which decreases with increasing patch size.

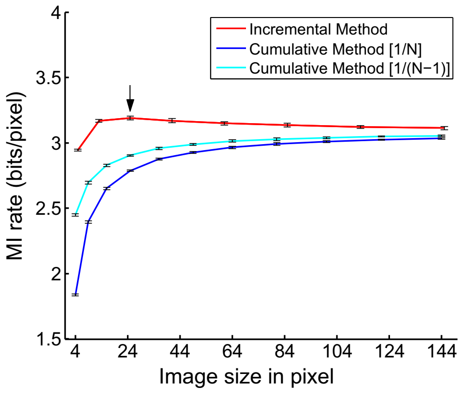

Fig. 3. Comparison of the cumulative and the incremental approach on natural images with $K = 1, 4, 7, 10$ scale mixture components. The blue and cyan curves show the result for the cumulative method and the red curve shows the result of the incremental method. Analogous to the results for artificial data, the incremental method significantly outperforms the cumulative ones. The arrow shows the maximum amount of multi-information estimated by the incremental method.

This is what we see in case of the incremental method. In case of small patch sizes, the GSM model can exploit higher-order correlations to model contrast dependencies between nearby pixels. In case of large image patches, however, the GSM model has to compromise between strong higher-order correlations between nearby pixels and weak higher-order correlations between distant pixels. Therefore, the model fit of the GSM becomes worse for larger patch sizes which causes the decrease in $\widehat{I}_n^{inc}$. In other words, the limited flexibility of the GSM model to capture the structure of higher-order correlations becomes increasingly severe with increasing dimensionality. For second-order correlations, however, this is different, because with a Gaussian distribution one can always fit any possible pattern of second-order correlations. Since in contrast to a general GSM, a single Gaussian distribution is always entirely ignorant against higher-order correlation, we do not see the effects of imperfect fitting of higher-order correlations in case of $K = 1$. For a Gaussian model, the lower bound can therefore only increase. This is nicely reflected in Fig. 4b: for $K = 1$ the lower bound always increases, whereas for $K \geqslant 4$ the lower bound decreases for large patch sizes.

Given that we explained the decrease of the lower bound for the incremental method with the limited flexibility of the GSM model, why do we not see a decrease for the cumulative method? We can explain this with the downward bias caused by the reduced contribution to the multi-information from pixels close to the patch boundaries. It is important to note that the persistent increase in case of the cumulative method does not originate from a better image model. Like in the artificial example, we fitted the same model distribution to optimally fit the *joint* distribution over the image pixels for the cumulative as well as for the incremental method. The crucial difference lies only in the way how we compute the lower bound to the asymptotic information rate from it. In one case we divide the total multi-information by the number of pixels and in the other case we compute the mutual information between one pixel and the rest by computing the conditional from the joint model. Therefore, the persistent increase up to $N = 144$ for the cumulative method does not reflect a better fit to the data but merely shows that the downward bias of the cumulative method for small image patches, for which the ratio of boundary to interior pixels is still large enough, is so substantial that it easily outbalances the decrease caused by degradation in the model fit.

Our second set of experiments on the pre-whitened images (see Section 3) further corroborates this explanation. The redundancy reduction caused by the pre-whitening is assessed as explained above and is the same for both methods. Therefore, after pre-whitening, all differences between the two methods can only originate from differences in assessing the contribution of higher-order correlations. Without the large contribution of second-order correlations, the downward bias for the cumulative method for small image patches becomes much smaller and hence, the effect of degradation in the model fit on the lower bound becomes more visible for the cumulative method as well. As can be seen in Fig. 5, the cumulative method now has a maximum as well at a patch size of 7 × 7 pixels. For the incremental method, the optimal neighborhood size is further reduced to $n = 13$. The type of higher-order correlations that can be captured by the GSM model are limited to variance (contrast) correlations between the different pixels. The fact that the lower bound takes its maximum for a very small neighborhood size shows that this type of correlations can be explained (away) by short range couplings.

Note that the curves shown include the contribution of second-order correlations that were removed during the pre-whitening step. The second-order contribution equals the lower bound obtained with the Gaussian distribution ($K = 1$) and is about 2.7 bits per pixel. Remarkably, the maximum lower bound determined with the pre-whitened images yields the same estimate for the

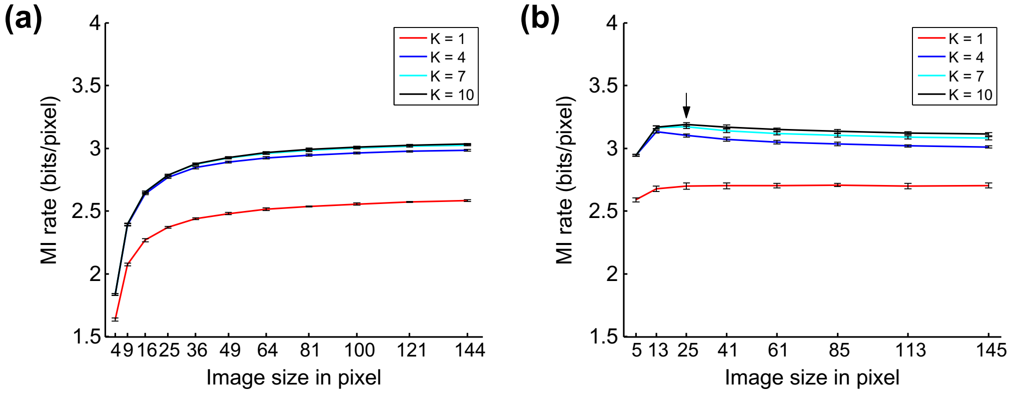

Fig. 4. Comparison of the multi-information rate estimates for different numbers of components ($K = 1, 4, 7, 10$). (a) Multi-Information rate estimated by the cumulative approach. (b) The same result using the incremental approach. In both cases the number of scale mixture components have similar effects and the performance seems to saturate for seven components. The arrow indicates the maximum amount of multi-information estimated by the incremental method.

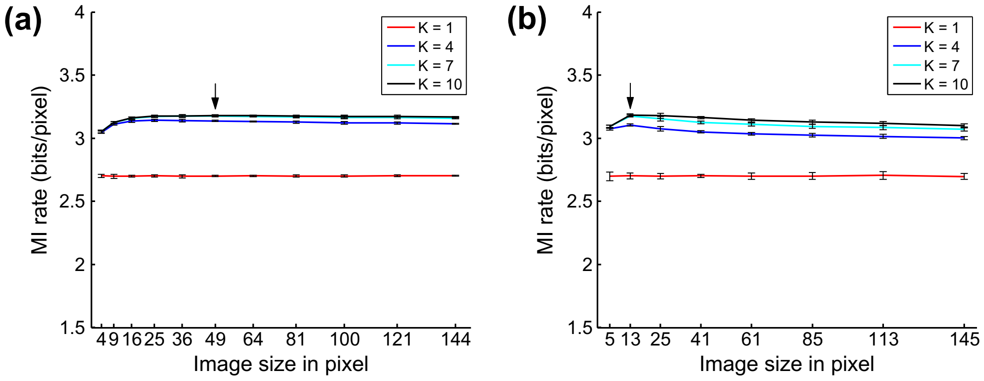

Fig. 5. Comparison of the multi-information rate estimates for different numbers of components ($K = 1, 4, 7, 10$) based on pre-whitened image data set. (a) Multi-Information rate estimated by the cumulative approach. (b) The same result using the incremental approach. Since the pre-whitening removes the downward bias of the cumulative method for the second-order contribution to the multi-information, it now has substantially improved and its lower bound—similarly to the incremental method—now takes a maximum for a relatively small patch size as well. The arrows indicate the maxima for both methods.

multi-information rate as the maximum lower bound obtained on the original images. This nicely underlines the reliability of our estimates.

As a final result we show how the incremental method can be further improved by improving the parameter fitting. As explained in Section 2, we always optimized the likelihood for the joint distribution and not for the conditional one. However, maximizing the likelihood for the joint model does not necessarily also maximize the likelihood for the conditional distribution which would be equivalent to minimizing the average log-loss of the conditional distribution. Based on Jebara's work on conditional expectation maximization (Jebara, 2002) we developed a new algorithm (see Appendix B) that we used to optimize the conditional likelihood for the GSM model. The result of this optimization is shown in Fig. 6a. In this way we obtained our best lower bound of 3.26 bits per pixel which is almost 0.6 bits larger than the multi-information rate obtained for a single Gaussian.

Fig. 6b shows the residual multi-information rate (see Eq. (7)) achieved by optimizing the conditional likelihood after pre-whitening (solid red). For comparison we also show the residual multi-information rate when optimizing for the joint likelihood (dashed) and the cumulative method (solid).

The large difference between the GSM using only a single mixture component and the GSMs with several ones is particularly interesting. Since the GSM in case of $K = 1$ is a plain Gaussian which is completely determined by its mean and its covariance matrix, the entropy rate of this GSM shows the contribution of the second-order moments to the total entropy of the signal. The fact that this difference is large shows the highly non-Gaussian behavior of natural images and, therefore, a substantial amount of higher-order correlations (Eichhorn et al., 2009; Chandler & Field, 2007; Ruderman & Bialek, 1994).

5. Summary and discussion

Measuring the total redundancy of natural images is a challenging task. In this paper we showed that the conventionally used cumulative method suffers from an unfavorable downward bias for small image patches. This problem can be avoided by using the incremental method. We compared the two methods for both artificial data and natural images, and demonstrated that the incremental method always yields a better lower bound on the multi-information rate.

As our method yields a conservative lower bound on the multi-information rate, we can safely conclude from our results that $I_\infty \geqslant 3.26$ bits per pixel for the van Hateren data set. This number is substantially larger than the 2.7 bits per pixel previously estimated by Petrov and Zhaoping (2003) who used very small neighborhoods only ($n = 7$). While they concluded that the total amount of higher-order correlations in natural images is small, the difference in the performance of the Gaussian model ($K = 1$) and the full GSM model ($K = 10$) suggests that the amount of higher-order correlations is at least 0.6 bits per pixel which we think is quite substantial.

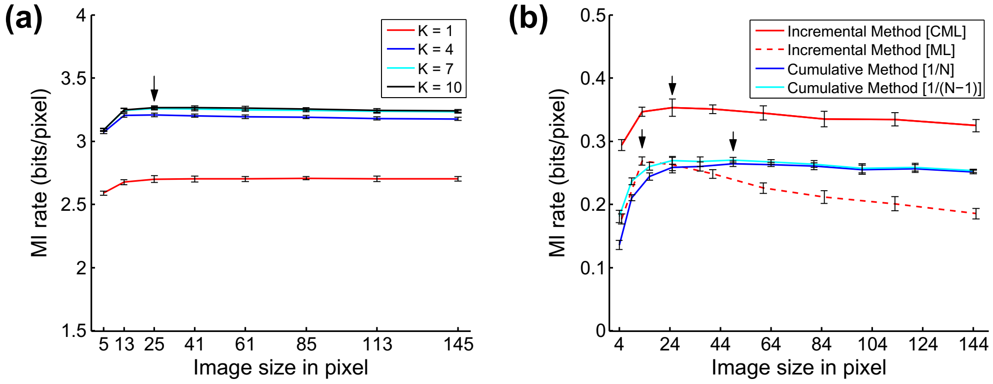

Fig. 6. Further improvement of the lower bound by optimizing the GSM model for the conditional likelihood. The arrow indicates the maximal amount of multi-information that was estimated. (a) shows the total multi-information rate while (b) shows the residual multi-information rate after the pre-whitening step (first term of Eq. (7)). The optimization of the conditional likelihood leads to a better fit of the conditional distribution and, hence, less degradation in the incremental method (solid vs. dashed red curve). It further corroborates the superiority of the incremental method above the cumulative method also for the pre-whitened data (red vs. other solid curves).

Using a less conservative nearest neighbor estimation method, Chandler and Field (2007) arrived at an information rate similar to ours. Taking the difference between the data points for Gaussian white noise and natural scenes in Fig. 14 in Chandler and Field (2007) would yield a multi-information rate estimate of about 3.1–3.3 bits per pixel. From their extrapolation in the same figure one obtains a multi-information rate of 3.7 bits per pixel in the limit.

In previous studies, we used the cumulative method together with an L_p-spherically symmetric model and an ICA model to estimate the redundancy reduction achieved by different neural response properties (Sinz & Bethge, 2009). The multi-information reported for ICA and the L_p-spherically symmetric model are 3.41 and 3.62 bits per pixel. Given that the multi-information estimates in Sinz and Bethge (2009) were obtained on a different dataset (Bristol Hyperspectral), the results are reasonably similar. We repeated the experiments of Sinz and Bethge (2009) and computed the values for the van Hateren dataset for 144 dimensions. We obtained 2.92 bits per pixel for ICA, 3.05 bits per pixels for the joint GSM, and 3.17 bits per pixel for the L_p-spherically symmetric model. This is better than the result of the cumulative method for the GSM but about 0.1 bits per pixel worse than the result of the incremental method. Thus, again the incremental method provides a better bound by using only 25 dimensions. The differences between the results for the Bristol Hyperspectral dataset and the van Hateren dataset are within the typical variations one observes for different image libraries. They mainly originate from variations in the second-order redundancies. In particular, the difference between the L_p-spherically symmetric model and ICA is very similar for both data sets: 0.21 bits per pixel for Bristol Hyperspectral and 0.25 for van Hateren.

In this study, we used the Gaussian scale mixture model for both the cumulative and the incremental approach for the sake of comparison. In the future we can make further advantage by using more sophisticated conditional density models that are optimally tailored to the incremental approach. It is interesting to note that the conditional distribution has a close link to the inverse of the auto-covariance matrix of random processes (the so called *precision matrix*). Typically, the precision matrix is much sparser and hence captures the conditional dependency structures much more efficiently than the covariance matrix (Rue & Held, 2005). In fact, for a Gaussian Markov random field, an entry of the precision matrix is non-zero if and only if the two points are conditionally dependent. When looking at the precision matrix for natural images, the number of components that have a value significantly larger than zero is typically very small and restricted to a very small neighborhood around that pixel.

In summary, we expect that the incremental method combined with an appropriate conditional density model will lead to major improvements in statistical modeling of natural images.

Acknowledgments

This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601) and a scholarship of the German National Academic Foundation to FS. We thank Sebastian Gerwinn for comments on the manuscript.

Appendix A

Definition 1. (**Causal points**). Let the *causal points* of a particular point in a random field be all points that are above that particular point or at its left in the same row.

Definition 2. (**Causal neighborhood of radius** l). Let the *causal neighbors of radius* l of a particular point be all causal points which their horizontal and vertical distance from that particular point being smaller or equal to l (see Fig. 1b for an example of a causal neighborhood of radius 3).

Theorem 1. (**Convergence of entropy rate for 2D stationary process**). *The sequence of conditional entropies with causal neighborhoods converges to the entropy rate of a stationary random process.*

Proof. Consider a sequence of sections $\mathbf{X}$ with increasing size which is taken from a 2D stationary process (see Fig. 7). Each section is parametrized by a parameter l which determines the extent of the section. The width of the section is chosen to be $w = l^3$ and its height is equal to $h = l^2 + l - 2$. □

The pixels are enumerated from top-left to button-right as it is shown in the Fig. 7. Let G and $\overline{G}$ be the sets that contain the indices which are shaded in gray and white colors, respectively. Furthermore, let n denote the total amount of pixels in the section, and let n_G and $n_{\overline{G}}$ be the number of pixels in the gray and white regions, respectively.

If we let the size of the sections go to infinity by letting l go to infinity, they will cover the whole plane and the number of white pixels will become negligible, i.e.

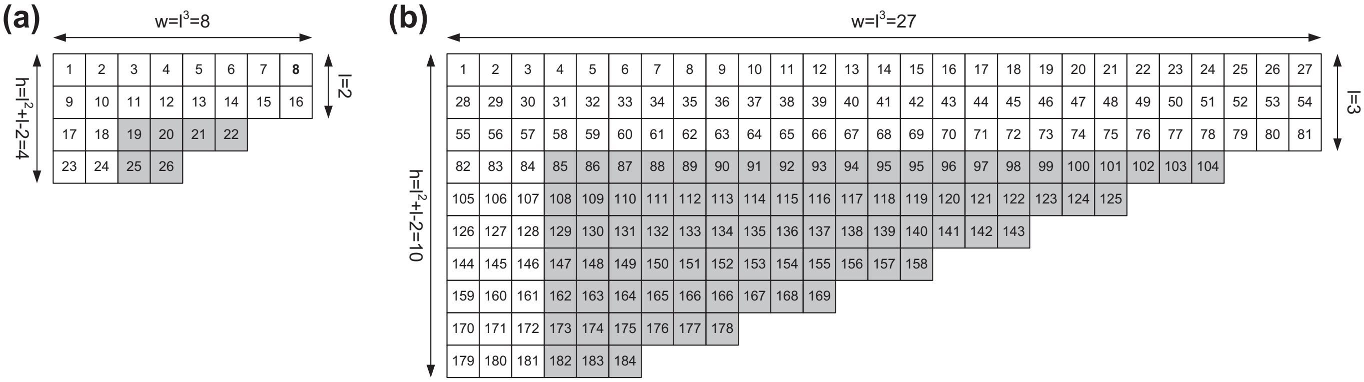

Fig. 7. Enumeration of the pixels in a 2D stationary process.

$$\lim_{l\to\infty}\frac{n_G(l)}{n(l)}=\lim_{l\to\infty}\frac{\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}{w(l)\cdot l+h(l)\cdot l-l^2+\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}=1, \qquad (8)$$

$$\lim_{l\to\infty}\frac{n_{\overline{G}}(l)}{n(l)}=\lim_{l\to\infty}\frac{w(l)\cdot l+h(l)\cdot l-l^2}{w(l)\cdot l+h(l)\cdot l-l^2+\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}=0. \qquad (9)$$

Since the sections $\mathbf{X}$ will cover the whole plane in the limit, the sequence of entropies of the single sections converges to the entropy of the stationary process:

$$h=\lim_{l\to\infty}\frac{1}{n(l)}H[\mathbf{X}_{1:n(l)}]=\lim_{l\to\infty}\frac{1}{n(l)}\sum_{k=1}^{n(l)}H[X_k|\mathbf{X}_{1:k-1}].$$

If we split the sum into two sums for the pixels in the gray, and white region, respectively, we obtain

$$h=\lim_{l\to\infty}\frac{1}{n(l)}\left(\sum_{k\in G}H[X_k|\mathbf{X}_{1:k-1}]\right)+\lim_{l\to\infty}\frac{1}{n(l)}\left(\sum_{k\in\overline{G}}H[X_k|\mathbf{X}_{1:k-1}]\right). \qquad (10)$$

Define $H_\alpha[X]$ to be the conditional entropy of X given a causal neighborhood of radius α (see Fig. 1b). Since conditioning decreases the entropy we obtain the following inequalities for stationary processes:

$$H_w\leqslant H[X_k|\mathbf{X}_{1:k-1}]\leqslant H_l,\quad \forall\ k\in G,$$
$$H_w\leqslant H[X_k|\mathbf{X}_{1:k-1}]\leqslant H_0,\quad \forall\ k\in\overline{G}.$$

Using these inequalities in Eq. (10) we obtain:

$$\lim_{l\to\infty}H_{w(l)}\leqslant h\leqslant\lim_{l\to\infty}\frac{n_G(l)}{n(l)}H_l+\lim_{l\to\infty}\frac{n_{\overline{G}}(l)}{n(l)}H_0. \qquad (11)$$

Using Eqs. (8) and (9) in Eq. (11) we get:

$$\lim_{l\to\infty}H_{w(l)}\leqslant h\leqslant\lim_{l\to\infty}H_l.$$

The sequences $H_{w(l)}$ and H_l will converge to the same limit, since $\{H_{w(l)}\}_{l=1,2,\ldots}$ is a proper subsequence of $\{H_l\}_{l=1,2,\ldots}$. Hence, using the sandwich theorem the sequence of conditional entropies $\{H_l\}_{l=1,2,\ldots}$ converges to the true entropy rate from above.

Appendix B

Minimizing the conditional average log-loss for a given model is equal to maximizing the conditional likelihood. Given the observed data $\{\mathbf{x}_i\}_{i=1}^m$, the conditional log-likelihood is given by:

$$\mathscr{L}(\mathbf{s},\Sigma,\lambda)=\underbrace{\sum_{i=1}^m\log\mathrm{GSM}(x_{1:n,i}|\mathbf{s},\Sigma,\boldsymbol{\lambda})}_{\mathscr{L}_1(\mathbf{s},\Sigma,\boldsymbol{\lambda})}-\underbrace{\sum_{i=1}^m\log\mathrm{GSM}(x_{1:(n-1),i}|\mathbf{s},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda})}_{\mathscr{L}_2(\mathbf{s},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda})}.$$

The conditional log-likelihood is the difference between the joint log-likelihood $\mathscr{L}_1$ and the marginal log-likelihood $\mathscr{L}_2$. Commonly, the EM algorithm is used to estimate mixture distributions. It constitutes a variational approach which maximizes a lower bound on the *joint* log-likelihood based on the Jensen inequality. In each iteration the maximum of the bound is computed. Since here $\mathscr{L}_2$ enters the conditional log-likelihood with a negative sign, the normal Jensen inequality is not useful to bound this function. Jebara derived a reversed form of the Jensen inequality for the exponential family (Jebara, 2002).

We used Jebara's method for deriving a conditional EM algorithm for the scale mixture of Gaussians. In the E-step the following coefficients are computed:

$$t_i^k=\frac{\lambda_k\mathscr{N}(\mathbf{x}_i|s_k\cdot\Sigma)}{\sum_{k=1}^K\lambda_k\mathscr{N}(\mathbf{x}_i|s_k\cdot\Sigma)},$$

$$r_i^k=\frac{\lambda_k\mathscr{N}(x_{1:(n-1),i}|s_k\cdot\Sigma_{1:(n-1),1:(n-1)})}{\sum_{k=1}^K\lambda_k\mathscr{N}(x_{1:(n-1),i}|s_k\cdot\Sigma_{1:(n-1),1:(n-1)})},$$

$$w_i^{k'}=\max\left[0,\frac{r_i^k}{x_{1:(n-1),i}^T s_k^{-1}\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}-1}\right],$$

$$w_i^k=2G\left(\frac{r_i^k}{2}\right)\left(\left(x_{1:(n-1),i}^T s_k^{-1}\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}-1\right)^2+n-2\right)+w_i^{k'},$$

$$G(\xi)=\begin{cases}\xi+\frac{1}{4\log(6)}+\frac{25}{36\log(6)^2}-\frac{1}{6} & \text{if } \xi\geqslant\frac{1}{6};\\ \frac{(\xi-1)^2}{\log(\xi)^2}-\frac{1}{4\log(\xi)} & \text{if } \xi\leqslant\frac{1}{6}.\end{cases}$$

Using these coefficients we get the following update rule for the scale and marginal covariance parameters.

$$\Sigma_{1:(n-1),1:(n-1)}=\frac{1}{m+\sum_{k=1}^K\sum_{i=1}^m w_i^k}\left(\sum_{k=1}^K\sum_{i=1}^m t_i^k s_k^{-1}x_{1:(n-1),i}x_{1:(n-1),i}^T-\sum_{k=1}^K\sum_{i=1}^m r_i^k s_k^{-1}x_{1:(n-1),i}x_{1:(n-1),i}^T\right)+\Sigma_{1:(n-1),1:(n-1)},$$

$$s_k=\frac{1}{K\sum_{i=1}^m t_i^k+(K-1)\sum_{i=1}^m w_i^k}\left(\sum_{i=1}^m t_i^k\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i-\sum_{i=1}^m r_i^k x_{1:(n-1),i}^T\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}\right)+\frac{(K-1)\sum_{i=1}^m(w_i^k+r_i^k)}{K\sum_{i=1}^m t_i^k+(K-1)\sum_{i=1}^m w_i^k}s_k$$

The conditional prediction matrix $\mathit{\Gamma} = \Sigma^{-1}_{1:(n-1),1:(n-1)}\Sigma_{1:(n-1),n}$ and the conditional variance $\gamma = \Sigma_{n,n} - \Sigma_{n,1:(n-1)}\Sigma^{-1}_{1:(n-1),1:(n-1)}\Sigma_{1:(n-1),n}$ only depend on the joint log-likelihood $\mathscr{L}_1$ and their estimations in the M-step are given by:

$$\mathbf{M} = \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m} t_i^k s_k^{-1}\mathbf{x}\mathbf{x}^T,$$

$$\mathit{\Gamma} = \mathbf{M}^{-1}_{1:(n-1),1:(n-1)}\mathbf{M}_{1:(n-1),n},$$

$$\gamma = \mathbf{M}_{n,n} - \mathbf{M}_{n,1:(n-1)}\mathbf{M}^{-1}_{1:(n-1),1:(n-1)}\mathbf{M}_{1:(n-1),n}.$$

Similar to the normal EM algorithm for optimizing the joint likelihood of theGSM model, one needs to iterate between estimating $\mathbf{s}$ and Σ for maximizing the bound.

The derivation before was for the case of fixed weighting coefficients $\boldsymbol{\lambda}$. For updating the weighting coefficients one can derive another EM update rule. Define a $(K-1) \times (K-1)$ matrix $\mathbf{N}$ with the following entries

$$N_{i,j} = \begin{cases} \lambda_i - \lambda_i^2 & \text{if } i = j; \\ -\lambda_i\lambda_j & \text{if } i \neq j. \end{cases}$$

Consider a $K-1$-dimensional vector $\boldsymbol{z}_k, 0 < k < K$ for which all entries are zero except the kth one, which equals one. Furthermore, let $\boldsymbol{z}_K$ a zero vector with $K-1$ elements. Using those vectors, we get the following update rules for the E-step

$$v_i^k = 4G(r_i^k/2)(\boldsymbol{z}_k - \lambda_{1:(K-1)})^T\mathbf{N}^{-1}(\boldsymbol{z}_k - \lambda_{1:(K-1)}),$$

and the M-step

$$\lambda_k = \frac{\sum_{i=1}^m t_i^k - \sum_{i=1}^m r_i^k}{m + \sum_{k=1}^K\sum_{i=1}^m v_i^k} + \lambda_k.$$

We observed that in practice the conditional EM algorithm converges very slowly. We found out that this is because the reverse Jensen inequality for the covariance is a very loose bound which becomes even looser for higher dimensions since the coefficient w increases rapidly with increasing dimensionality. As a consequence of this, we observed empirically that the EM algorithm increases the log-likelihood slower than gradient ascend with line search.

We accelerated the EM algorithm by using the Quasi-Newton method (algorithm QN2 in Jamshidian & Jennrich, 1997). The idea behind this method is to approximate the Newton update $\boldsymbol{H}^{-1}\mathbf{g}(\boldsymbol{\theta})$, where $\boldsymbol{H}$ is the Hessian and $\mathbf{g}$ is the gradient at $\boldsymbol{\theta}$ with the update $\hat{\mathbf{g}}(\boldsymbol{\theta}) - \boldsymbol{S}\mathbf{g}(\boldsymbol{\theta})$ where $\hat{\mathbf{g}}$ is EM gradient. In other words, the difference of two EM steps and $\boldsymbol{S}$ is a matrix that needs to be updated as well. The authors modify BFGS Quasi-Newton method to get the update for $\boldsymbol{S}$:

$$\Delta\boldsymbol{S} = \left(1 + \frac{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}^\star}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}}\right)\frac{\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}} - \frac{\Delta\boldsymbol{\theta}^\star\Delta\boldsymbol{\theta}^T + (\Delta\boldsymbol{\theta}^\star\Delta\boldsymbol{\theta}^T)^T}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}}.$$

where $\Delta\boldsymbol{\theta}^\star = -\Delta\hat{\mathbf{g}} + \boldsymbol{S}\Delta\mathbf{g}$ while $\Delta\boldsymbol{\theta}$ and $\Delta\mathbf{g}$ show the amount of change in variables $\boldsymbol{\theta}$ and $\mathbf{g}$ after each iteration, respectively. In the implementation one initializes with $\boldsymbol{S} = 0$ and then updates $\boldsymbol{S}$ according to the update rule. If the line search is not successful $\boldsymbol{S}$ is reset to zero.

In practice we observed that this Quasi-Newton acceleration significantly increases the convergence speed but it still remains slow. In the future, this may be substantially improved by exploiting the Quasi-Newton and Newton method directly on the log-likelihood.

References

Atick, J., & Redlich, A. (1992). What does the retina know about natural scenes. *Neural Computation, 4*, 196–210.

Barlow, H. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes* (pp. 535–539). London: Her Majesty's Stationery Office.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics, 7*(May), 686–690.

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*(6), 1253–1268.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 220*(November), 89–113.

Chandler, D. M., & Field, D. J. (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A, 24*(4), 922–941.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley & Sons. Sep.

Eichhorn, J., Sinz, F., & Bethge, M. (2009). Natural image coding in v1: How much use is orientation selectivity? *PLoS Computational Biology, 5*(4), e1000336.

Föllmer, H. (1973). On entropy and information gain in random fields. *Probability Theory and Related Fields, 26*(3), 207–217.

Jamshidian, M., & Jennrich, R. I. (1997). Acceleration of the EM algorithm by using Quasi-Newton methods. *Journal of the Royal Statistical Society. Series B (Methodological), 59*(3), 569–587.

Jebara, T. (2002). Discriminative, generative, and imitative learning. Thesis. Massachusetts Institute of Technology.

Karklin, Y., & Lewicki, M. (2008). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*, 83–86.

Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research, 42*(17), 2095–2103.

Lewicki, M. S., & Olshausen, B. A. (1999). Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A, 16*(July), 1587–1601.

Lewicki, M., & Sejnowski, T. (2000). Learning overcomplete representations. *Neural Computation, 12*, 337–365.

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience, 13*(1), 257–281.

Lyu, S., & Simoncelli, E. P. (2009). Reducing statistical dependencies in natural signals using radial Gaussianization. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. 21, pp. 1009–1016). Cambridge, MA: MIT Press.

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

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation, 15*(6), 1191–1253.

Perez, A. (1977). ε-Aadmissible simplification of the dependence structure of a set of random variables. *Kybernetika, 13*, 439–444.

Petrov, Y., & Zhaoping, L. (2003). Local correlations, information redundancy, and sufficient pixel depth in natural images. *Journal of the Optical Society of America A, 20*(1), 56–66.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters, 73*(6), 814. copyright (C) 2009 The American Physical Society. Please report any problems to prola@aps.org..

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. Chapman & Hall/CRC.

Schreiber, W. (1956). The measurement of third order probability distributions of television signals. *IRE Transactions on Information Theory, 2*(3), 94–105.

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

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*(3), 605–610.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. and 623–656.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Sinz, F., & Bethge, M. (2009). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Neural information processing systems, 2008* (p. 8).

Srinivasan, M., Laughlin, S., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 216*(1205), 427–459.

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 265*(1394), 1724–1726.

Wachtler, T., Lee, T. W., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 18*, 65–77. pMID: 11152005.

Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of gaussians and the statistics of natural images. In: *Advances in neural information processing systems* (Vol. 12, pp. 855–861).