Factorial coding of natural images: how effective are linear models in removing higher-order dependencies?

Matthias Bethge

Redwood Neuroscience Institute, 1010 El Camino Real, Menlo Park, CA 94025

Received May 12, 2005; revised November 30, 2005; accepted December 7, 2005; posted December 9, 2005 (Doc. ID 61982)

The performance of unsupervised learning models for natural images is evaluated quantitatively by means of information theory. We estimate the gain in statistical independence (the multi-information reduction) achieved with independent component analysis (ICA), principal component analysis (PCA), zero-phase whitening, and predictive coding. Predictive coding is translated into the transform coding framework, where it can be characterized by the constraint of a triangular filter matrix. A randomly sampled whitening basis and the Haar wavelet are included in the comparison as well. The comparison of all these methods is carried out for different patch sizes, ranging from 2×2 to 16×16 pixels. In spite of large differences in the shape of the basis functions, we find only small differences in the multi-information between all decorrelation transforms (5% or less) for all patch sizes. Among the second-order methods, PCA is optimal for small patch sizes and predictive coding performs best for large patch sizes. The extra gain achieved with ICA is always less than 2%. In conclusion, the edge filters found with ICA lead to only a surprisingly small improvement in terms of its actual objective.

OCIS codes: 000.5490, 100.2960, 100.3010.

1. INTRODUCTION

Many image processing tasks rely either explicitly or implicitly on modeling the statistical dependencies between pixel intensities in images.[1,2] Within a given class of image models, unsupervised learning can be used to find an optimal candidate. Independent component analysis (ICA) is an unsupervised learning method that optimizes over a class of multivariate distributions that can be derived from a linear mapping of a reference random variable with factorial distribution (for an early review and a recent textbook, see Refs. 3 and 4, respectively). Over the last decade, ICA has become a very successful tool, and it is now used in hundreds of different applications by a diverse range of disciplines. In addition, ICA initiated a large movement in developing new unsupervised learning techniques.

In the context of visual neuroscience, the extraction of statistically independent components has been proposed as an objective of early sensory processing, shaping the receptive fields of neurons in the retina, LGN, and primary visual cortex. Referring to the standard view of neurons in the early stages of visual processing, the concept of a receptive field is well described by linear–nonlinear cascade models of neurons that include a linear filter at their first stage (see Ref. 5 for an overview). This filter computes the correlation $s_k=\langle \mathbf{w}_k,\mathbf{x}\rangle$ between the pixel intensities $\mathbf{x}$ of the stimulus and a filter kernel $\mathbf{w}_k$, which is referred to as the receptive field of neuron k. The second stage of this model describes how spikes are generated from the filter outputs. In the simplest and most widespread case, the spike generation is modeled as a Poisson process whose intensity y_k (the expected firing rate of neuron k) is computed via a nonlinear half-rectifying and saturating activation function $y_k=f_k(s_k)$. For a set of neurons $k=1,\ldots,N$, it is convenient to summarize the filter kernels into a single matrix $W=[\mathbf{w}_1,\ldots,\mathbf{w}_N]^{\mathrm{T}}$, which allows one to write compactly

$$\mathbf{s}=W\mathbf{x}, \tag{1}$$

so that the kth row of the filter matrix W determines the receptive field properties of neuron k.

As long as the number of neurons N is smaller than or equal to the dimensionality of $\mathbf{x}$, second-order correlations between the rate responses of all neurons can always be removed completely via linear filtering, and there is some evidence that retina and LGN indeed act as whitening filters in response to natural stimuli.[6] The objective of second-order decorrelation by itself, however, is not sufficient to predict the receptive field properties, because additional constraints or demands are necessary to determine the filter kernels uniquely. For illustration, the six different bases shown in Figs. 1 and 2 are all equivalent with respect to second-order statistics.

In ICA, this ambiguity is resolved by seeking to remove higher-order correlations in the input as well. A striking result of ICA image models when applied to natural images is the emergence of edge filters,[7,8] which resemble important aspects of simple cell receptive fields in the primary visual cortex[9]: The basis images are localized, oriented, and bandpass. This finding suggests that the primary visual cortex seeks to remove higher-order dependencies, while the retina and LGN are concerned with second-order decorrelation only.

The components found with ICA algorithms can be independent only if the data distribution is indeed a linear mixture of independent sources. For the statistics of natu-

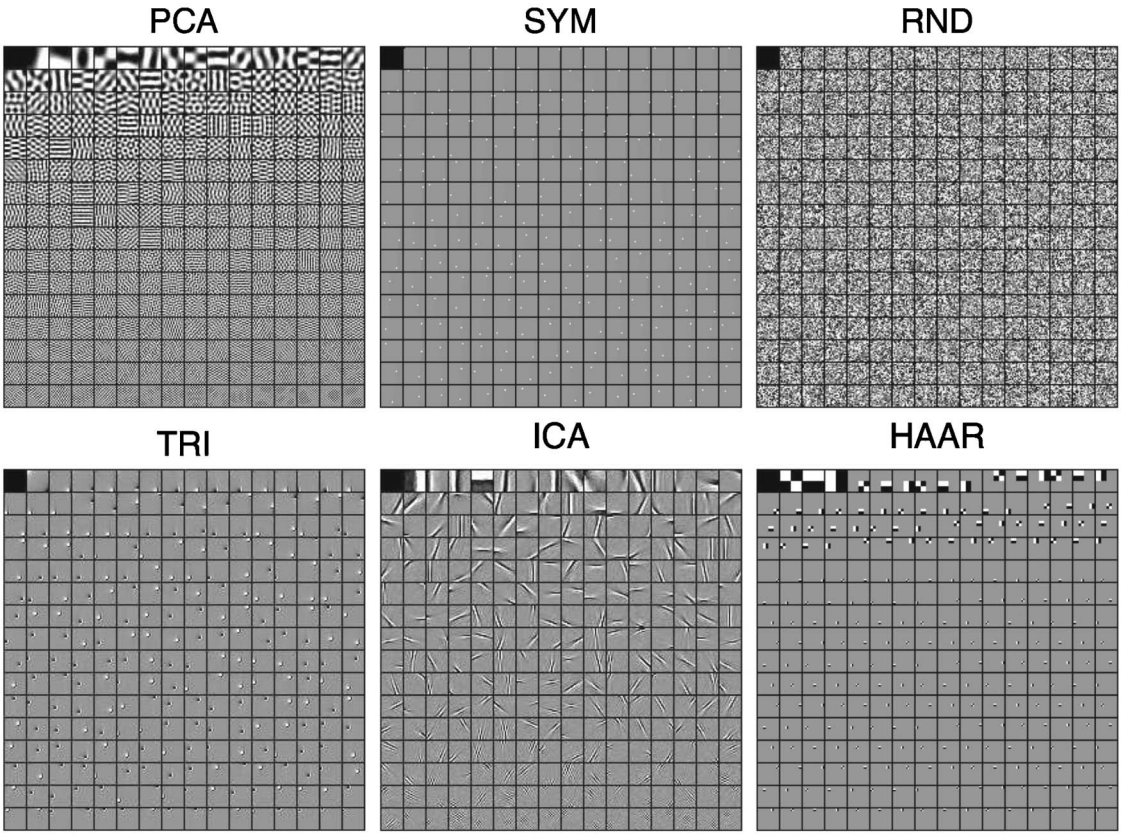

Fig. 1. Comparison of the basis image patches for the six different decorrelation transforms.

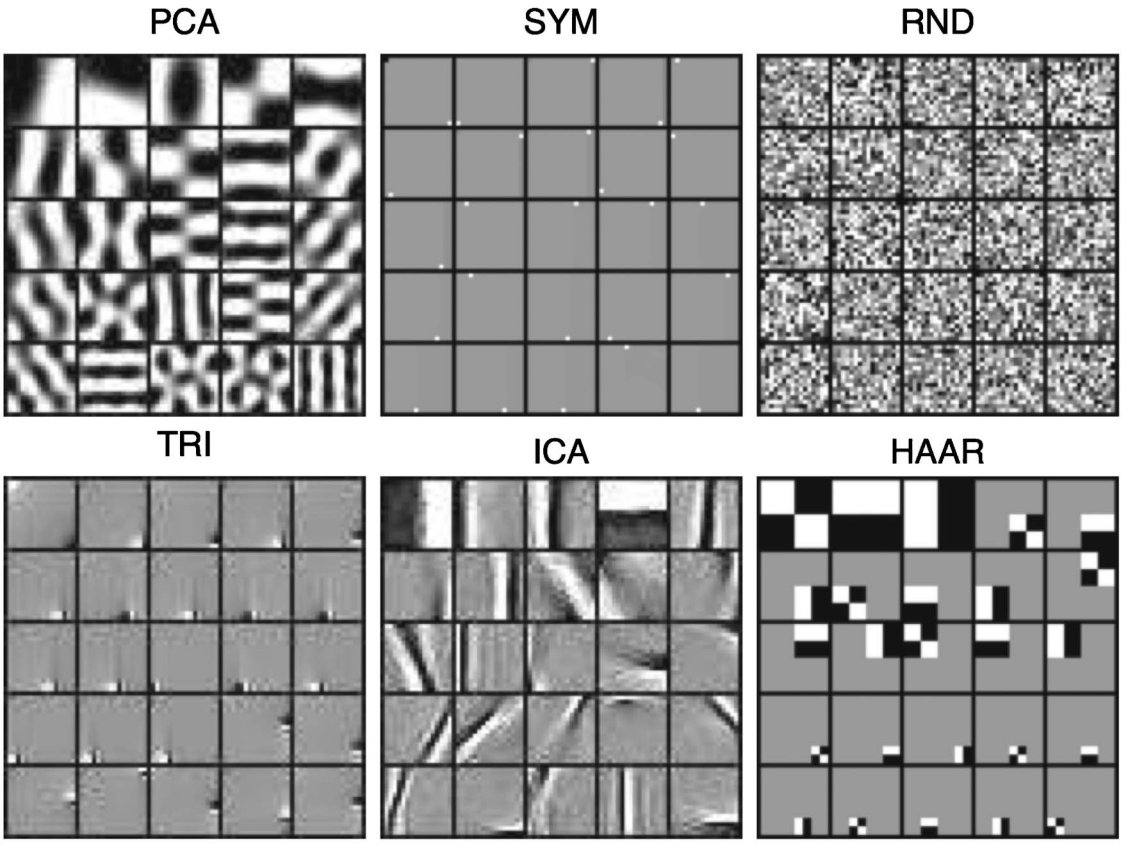

Fig. 2. The first 25 basis functions other than the DC component are shown for each method for better visibility.

ral images, this is not the case. If one examines the edge filters learned by ICA, for instance, nearby filters of similar orientation exhibit correlations in the magnitude.[1,10] Since images are not a linear mixture of independent sources, ICA seeks to make the filter outputs as independent as it can within the restrictions of the linear model. More precisely, the objective function of ICA is the multi-information, which can be defined as the Kullback–Leilbler (KL) divergence between the joint distribution and the product of its marginals:

$$I_{multi}[\mathbf{S}] = D_{KL}\left[p(\mathbf{s})\middle\|\prod_k p_k(s_k)\right] = \sum_k h[S_k] - h[\mathbf{S}], \quad (2)$$

where $h[S]=-\int p(s)\log p(s)\mathrm{d}s$ denotes the differential entropy with the understanding that S can be either one of the scalar-valued random variables S_k or a vector-valued random variable $\mathbf{S}$. Throughout the paper, we adopt the convention of using uppercase letters to refer to random variables, and we use bold font to distinguish vector variables from scalar variables.

In the special case when $\mathbf{S}$ is only two dimensional, I_{multi} is equal to the mutual information between the two components. Therefore I_{multi} itself is often called mutual information in the ICA literature according to the idea that I_{multi} may be seen as a generalization of the mutual information to the case of more than two dimensions. To avoid confusion with the mutual information between two subspaces in higher-dimensional spaces, however, we adopt the less ambiguous (but less established) terminology of Ref. 11.

While for natural images the minimization of the multi-information reliably results in the well-known image ICA basis, it has never been tested quantitatively how much this representation actually reduces the multi-information in comparison with plain second-order methods for natural images. However, it has been tested in Ref. 12 how large the gain in coding efficiency is for a certain mean square error. The important difference between efficient coding and the objective function of ICA will be addressed in the discussion. For now, we emphasize that here we do not evaluate coding efficiency but rather test the gain of (noiseless) ICA with respect to its own objective function. Using precise estimates of *changes* in the multi-information, we find that the distinct receptive fields found with ICA lead to only a very small improvement in the reduction of statistical dependencies compared with that of other linear decorrelation filters.

2. SEARCHING FOR THE LEAST DEPENDENT COMPONENTS

ICA, principal component analysis (PCA), zero-phase whitening, and predictive coding algorithms all have been extensively used with a diverse range of variations as adaptive models of sensory coding. In this section, we provide a short overview of the different assumptions that they make.

A. Second-Order Optimization

Instead of minimizing the multi-information (2) directly, second-order methods minimize the upper bound[13]

$$I_{multi}[\mathbf{S}] \leqslant \frac{1}{2}\sum_{k=1}^{n} \log_2(2\pi e(C_{\mathbf{S}})_{kk}) - h[\mathbf{S}] \quad (3)$$

via diagonalization of the covariance matrix $(C_{\mathbf{S}})_{ij} = E[S_iS_j] - E[S_i]E[S_j]$. It is assumed throughout the paper that all eigenvalues of the covariance matrix are positive. While PCA is the only *orthogonal* transform for which all second-order correlations vanish, there are many *nonorthogonal* transforms that diagonalize the covariance matrix as well. The set of all decorrelation transforms can be written as

$$\{W: W = D_2VD_1U_{PCA}\}, \quad (4)$$

where U_{PCA} is the orthogonal matrix used in PCA, whose rows are the eigenvectors of the covariance matrix. D_1 is a diagonal matrix with the square roots of the inverse eigenvalues as nonzero entries, V is an arbitrary orthogonal transform, and D_2 is an arbitrary diagonal matrix. The whitening transform represented by D_1U_{PCA} makes sure that the covariance matrix remains diagonal for all possible choices of V and D_2. Nevertheless, the multi-information in general depends on V, whereas the choice of D_2 has no effect on the multi-information.

PCA can be motivated as the special case of isometric decorrelation, where the term "isometric" refers to the additional constraint that the total filter matrix W must not change the metric of the input space.[14] Since the metric is conserved only if W is orthogonal, the optimum with respect to isometric decorrelation is uniquely determined regardless of any higher-order correlations whenever the eigenvectors of the covariance matrix of the data are all mutually different. If additionally the orthogonal mixture model is correct, PCA recovers the independent sources (even if they are not Gaussian).

The statistics of natural images, however, cannot be modeled correctly as an orthogonal mixture of independent sources. Instead of postulating orthogonality, one can require the mixing matrix to be patterned in a different way. If the data can be described by a symmetric mixing of independent sources, symmetric decorrelation (or zero-phase whitening if D_2 is the identity) is known to recover the independent axes. If the mixing matrix is triangular, triangular decorrelation will achieve this goal. Both methods constitute nonorthogonal second-order decorrelation transforms, which are simple to compute and naturally lead to highly localized receptive fields. In contrast to symmetric decorrelation,[8] triangular decorrelation has not been compared with ICA before. In Appendix A, we show that triangular decorrelation is the transform coding version of optimal linear predictive coding. Predictive coding not only plays an important role in lossless image compression[15,16] but has also been proposed early on for the information-theoretical function of the retina.[17]

In principle, any set of $d(d-1)/2$ linearly independent constraints could be used to determine V in a unique way. This can be shown, for instance, by using the Cayley parameterization of orthogonal matrices $V=(\mathbb{1}+A)^{-1}(\mathbb{1}-A)$, where the antisymmetric matrix A has only $d(d-1)/2$ free parameters. For the sake of comparison, we also include a random whitening transform, which is defined by the following choice of V: First, a random matrix $\tilde{V}$ is con-

structed by randomly drawing its column vectors from an isotropic Gaussian distribution. Subsequently, V_{RND} is obtained from $\tilde{V}$ via symmetric orthogonalization, that is, $V_{RND} \equiv \tilde{V}(\tilde{V}^{\mathrm{T}}\tilde{V})^{-1/2}$.

B. Higher-Order Optimization

Instead of minimizing the second-order upper bound on the multi-information, one can additionally or alternatively seek to minimize the multi-information directly. In "prewhitened ICA," a higher-order correlation contrast function is used to pick an optimal orthogonal transform V after the whitening step $D_1 U_{PCA}$. FastICA (Ref. 18) belongs to this class of ICA algorithms and is the one that we present in our comparison. Another well-known ICA algorithm, which does not restrict the solution to be a whitening transform, is Bell–Sejnowski ICA.[19] Its search space is a proper superset of that of FastICA, so that in principle one might find a better solution with Bell–Sejnowski ICA. This is not guaranteed, however, because the performance of any ICA algorithm substantially depends on how well it estimates the multi-information. Bell–Sejnowski ICA has been applied to natural images before,[8] and we also included it in our study. For the sake of brevity, however, we do not show the results for this algorithm, as it performs very similarly to FastICA. It exhibits slightly weaker performance than FastICA if one uses the tanh activation function as used in Ref. 8. It may perform better, however, if one uses the cumulative distribution function of the exponential power family as activation function.[20,21]

Finally, we also included the Haar wavelet[22] as a parametric basis in the comparison. More specifically, we set the rows of V to be equal to the basis vectors of the orthogonal two-dimensional Haar basis, so that the total filter matrix W is still a decorrelation transform. It is instructive to see that despite its simplicity and its blockiness the Haar wavelet turns out to perform almost as well as the ICA basis.

In Section 3, we give a short description of the data set and the variety of methods used for the quantitative comparison of the multi-information gain. Details will be explained in the appendices. The results obtained with the different transforms applied to natural image patches of different sizes are presented in Section 4. The insights about neural representations of natural images gained from this comparison are discussed in Section 5.

3. QUANTITATIVE ANALYSIS OF MULTI-INFORMATION REDUCTION

A. Description of the Data Sets

The natural image patch ensembles analyzed in this paper are constructed by sampling from the first ten images of the van Hateren data base[23] (center parts, 1024 $\times$ 1024 pixel, strictly linear intensity scale, image content dominated by woods and greens). Following Ruderman and Bialek,[24] we decided to use the pixel contrast $[\log(I(x)/I_0)]$ instead of linear intensities, which are much more similar to the common gray-level scale used in electronic image data formats. The log-intensity scale seems to be a good compromise between modeling the contrast sensitivity profile of the retina and simplicity. In Appendix B, we discuss why this choice may also enhance the robustness of the multi-information estimates. From control studies, however, we know that the multi-information reduction obtained with a linear intensity scale is almost identical to the multi-information reduction in the case of log intensities presented in this paper.

As another preprocessing step, we applied a simple dynamic range adaptation for each image, such that its overall log-intensity distribution is centered around zero and rescaled to unit variance before any patches have been sampled. In addition, we added an invisible amount of Gaussian noise with standard deviation 2^{-8} in order to compensate for the artificial alignment of intensities due to the analog/digital conversion. Again, we found that this preprocessing does not have a substantial effect on the measured multi-information. Only PCA performs slightly better relative to the other decorrelation transforms after this dynamic range adaptation.

To control for overfitting, we generated training and test sets for all studied patch sizes (39,690 image patches each), as described in detail in Appendix C. In addition, we enhanced the robustness of our results against the particular choice of the image data set by separating the DC component before adaptation of the basis images from any of the decorrelation transforms. While the histogram of the DC component of local image patches can change dramatically from image to image, the histograms of the pixel intensities after subtraction of the DC component is much more stable. In agreement with earlier observations,[25,26] the marginals of any randomly picked component in this space exhibit a kurtotic shape. An interesting, nonorthogonal basis that spans the DC0 space (i.e., the space of all zero-mean signals) is given by the $d-1$ difference values between neighboring pixels. For symmetry reasons, they all have the same distribution, which is shown for our data set in Fig. 3. The histogram closely resembles a Laplacian distribution, which helps to improve the reliability of the necessary entropy estimations.

It is important to note that the outcome of PCA, symmetric, and triangular whitening depends on the basis of

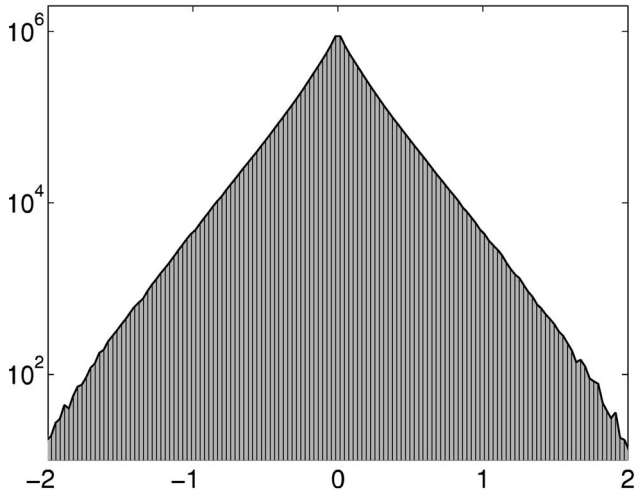

Fig. 3. A log histogram of the log-intensity differences approximates the shape of a Laplacian distribution in the case when the content of the images is dominated by woods and greens.

the input space. In fact, any whitening basis can be obtained by any second-order method if the input space is transformed appropriately beforehand. Therefore, to keep the input basis as close as possible to the pixel basis, we decided to use an orthogonal basis for the separation of the DC component, as it does not change the metric of the usual pixel basis. More specifically, an orthogonal DC0 basis that preserves the localization of the pixel basis is obtained by Gram–Schmidt orthogonalization of the modified identity transform, for which the first basis vector has been replaced by the DC vector whose entries are all identical.

B. Estimating the Multi-Information Gain

Direct estimation of the multi-information for high-dimensional random variables is very difficult due to the curse of dimensionality. Although an explicit multi-information estimator has very recently been presented[27] based on order statistics, we found that its precision is not sufficient for our purposes. Instead, we will resort to the same technique commonly exploited in ICA, where only the *difference* in the multi-information between two different transforms $\mathbf{Y}=\mathbf{f}_1(\mathbf{X})$ and $\tilde{\mathbf{Y}}=\mathbf{f}_2(\mathbf{X})$ is estimated:

$$\begin{aligned}\Delta I &= I_{multi}[\mathbf{Y}] - I_{multi}[\tilde{\mathbf{Y}}] \\ &= \sum_k h[Y_k] - h[\mathbf{Y}] - \left(\sum_k h[\tilde{Y}_k] - h[\tilde{\mathbf{Y}}]\right) \\ &= \sum_k h[Y_k] - \sum_k h[\tilde{Y}_k] + E\left[\log\left|\det\left(\frac{d\mathbf{f}_1}{\partial\mathbf{x}}\right)\right|\right] \\ &\quad - E\left[\log\left|\det\left(\frac{d\mathbf{f}_2}{\partial\mathbf{x}}\right)\right|\right]. \end{aligned} \tag{5}$$

Any possible mapping $\mathbf{y}=\{\mathbf{f}(\mathbf{x})\}$ can be modified such that either the first two terms or the last two terms on the right-hand side vanish while the multi-information stays the same. The two distinct choices of this gauge invariance are known as Bell–Sejnowski ICA[19] and volume-conserving ICA,[28] respectively (see Appendix D). In the case of volume-conserving ICA, the evaluation of the multi-information difference requires one to estimate only the marginal entropies:

$$\Delta I = I^{VC}_{multi}[\mathbf{Y}] - I^{VC}_{multi}[\tilde{\mathbf{Y}}] = \sum_k h[Y_k] - \sum_k h[\tilde{Y}_k]. \tag{6}$$

Negative values of ΔI correspond to statistical dependency reduction, while positive values reflect an increase in multi-information. In the special case when $\mathbf{Y}=\mathbf{S}$ and $\tilde{\mathbf{Y}}=\mathbf{X}$, we obtain the multi-information change achieved by a particular filter $\mathbf{s}=W\mathbf{x}$.

In the remainder of this section, we present four entropy estimators that are later used to evaluate the multi-information reduction [Eq. (6)] for the different image patch transforms. The first two estimators, labeled MAL and OPT in the following, make the assumption that the coefficient distributions can be well fitted by the exponential power family[29]

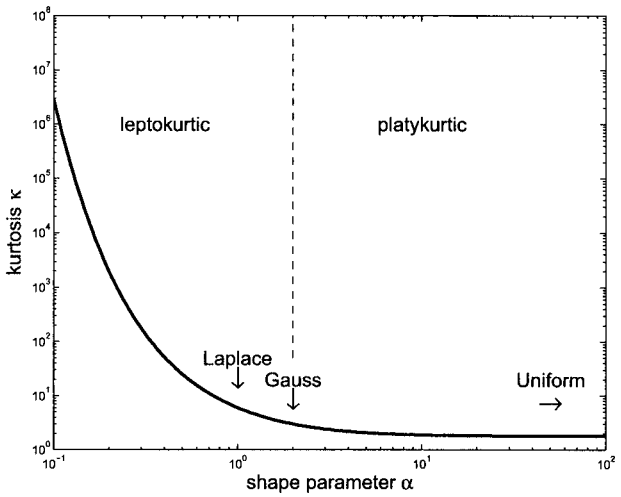

Fig. 4. Kurtosis κ of the exponential power family depending on the shape parameter α.

$$p_{\alpha,\sigma}(y) = \frac{\alpha A}{2\Gamma(1/\alpha)} \exp[-(A|y|)^\alpha], \tag{7}$$

which is also called a generalized Gaussian, or generalized Laplacian, distribution. This family has 2 degrees of freedom: The decay constant A can be expressed as a function of the variance $\sigma^2=\mathrm{Var}[Y]$ and the shape parameter α, i.e.,

$$A = A(\alpha,\sigma) = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}}. \tag{8}$$

The shape parameter α of this family makes it possible to tune the kurtosis $\kappa=E[Y^4]/E^2[Y^2]$ of the distributions from platykurtic ($\kappa<3$ for $\alpha>2$) to leptokurtic ($\kappa>3$ for $\alpha<2$) in a monotonic fashion (see Fig. 4):

$$\kappa = \frac{\Gamma(1/\alpha)\Gamma(5/\alpha)}{[\Gamma(3/\alpha)]^2}. \tag{9}$$

In fact, most ICA models for natural images optimize the kurtosis or a similar measure of the sparseness or peakedness of a distribution as a contrast function. For our quantitative analysis, however, we will take a different, more robust approach to determine the shape parameter of the exponential power family.

In addition to the parametric approach, we also use two nonparametric estimators to control for the bias caused by the choice of the exponential power family. These estimators are labeled NPL and VAS in the following. Subsection 3.C describes all four estimators in more detail. The reader who is interested to take a shortcut may jump to Subsection 3.D right away, where the performance of the estimators is compared for artificially generated samples from a Laplacian distribution.

C. Description of the Entropy Estimators

A simple upper bound for the entropy of such random variables can be obtained from the variances by using the maximum entropy property of the normal distribution:

$$h[Y_k] \leqslant \tfrac{1}{2}\log_2(2\pi e Var[Y_k]) =: G[Y_k]. \tag{10}$$

We will call $G[X_k]$ the Gaussian entropy bound. More precisely, one can decompose the entropy into two terms:

$$h[Y_k] = G[Y_k] - \underbrace{D_{KL}[Y_k \,\|\, Y_k^{Gauss}]}_{=:J[Y_k]}, \tag{11}$$

where $J[Y_k]$ denotes the negentropy,[3,30] that is, the KL divergence of the distribution of Y_k from a normal distribution of same variance. For our data set, we find that after removal of the DC component all filter outputs can be well approximated by the exponential power family (in agreement with earlier studies[31]).

For the exponential power family, it is possible to compute the negentropy explicitly as a function of the shape parameter α (see Fig. 5):

$$J[Y_k] = \frac{1}{2} - \frac{1}{\alpha} + \log\left(\alpha\sqrt{\frac{\pi\Gamma(3/\alpha)}{2[\Gamma(1/\alpha)]^3}}\right)\frac{\text{bits}}{\log(2)}. \tag{12}$$

The negentropy takes its minimum at $\alpha=2$ (the Gaussian case) and is monotone increasing toward both directions away from $\alpha=2$. In the sub-Gaussian case, J converges from below to $\log_2\sqrt{\pi e/6}$ bit≈ 0.255 bit for $\alpha\to\infty$ (uniform distribution), so that the gain in negentropy is rather limited. For the sparse, or leptokurtic, branch of the exponential power family, the negentropy increases without bound, diverging as $\alpha\to 0$. In the special case of the Laplacian distribution ($\alpha=1$), the negentropy equals $[\log(\pi)-1]/[2\log(2)]$ bit≈ 0.1 bit.

The first absolute moment of the exponential power family is given by

$$E_{\alpha,\sigma}[|Y|] = \sigma\sqrt{M(\alpha)}, \tag{13}$$

where

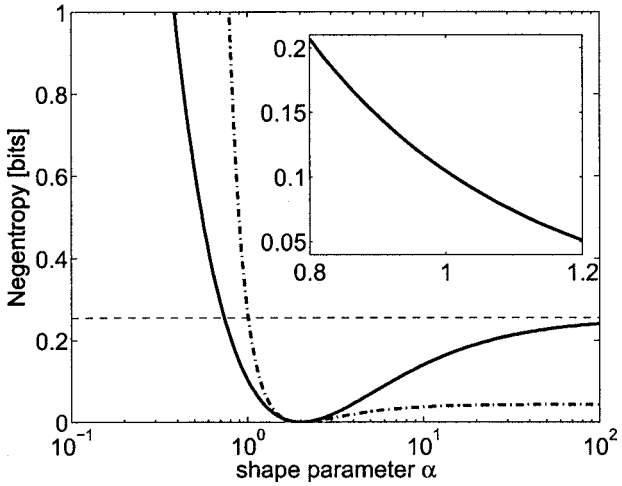

Fig. 5. The solid curve indicates the true negentropy of the exponential power family depending on the shape parameter α. The inset shows a magnification of the region around $\alpha=1$, which is most relevant to the modeling of natural image statistics when using log intensities. The dotted–dashed curve is the quadratic approximation of the negentropy utilized in the FastICA algorithm using $g(u)=|u|$, which has been shown to be asymptotic optimal in the case of a Laplacian distribution.[32]

$$M(\alpha) = \frac{E_{\alpha,\sigma}[|Y|]^2}{\sigma^2} = \frac{\Gamma^2(2/\alpha)}{\Gamma(1/\alpha)\Gamma(3/\alpha)}. \tag{14}$$

Together with the sample estimators for the first absolute moment and the variance, one can use Eq. (14) to compute an estimate for the shape parameter.[33] Having determined the shape parameter α, we use Eqs. (11) and (12) to obtain a consistent plug-in estimator for the entropy. In the following, this estimator will be labeled the MAL estimator because it was proposed first by Mallat.[33]

A simple way to check and to visualize how well the assumption of the exponential power family is met by the data[31] is to consider $\log\log[p(0)/p(Y_k)]$ as a function of $\log(|Y_k|)$, where p denotes the density or, in practice, a density estimate of $|Y_k|$. For the exponential power family, this function has to be linear, and, in principle, one could also get an estimate of the shape parameter α via linear regression between $\log\log(p(0)/p(Y_k))$ and $\log(|Y_k|)$.

Here, we will pursue another, more accurate strategy to test the goodness of fit, using the cumulative distribution function of the exponential power family,

$$\frac{1}{2} + \frac{\text{sgn}(y)}{2\Gamma(1/\alpha)}\Gamma\left([A(\alpha,\sigma)|y|]^\alpha, \frac{1}{\alpha}\right), \tag{15}$$

as a squashing function, similar to the practice in Bell–Sejnowski ICA. $\Gamma(u,a)=\int_0^u t^{a-1}\exp(-t)\mathrm{d}t$ is known as the (lower) incomplete gamma function, and $\Gamma(a)=\lim_{u\to\infty}\Gamma(u,a)$ denotes the (complete) gamma function. If the fit is correct, the output should be uniformly distributed between zero and one. In fact, we found a very good agreement between the histograms of 39,690 uniformly distributed random numbers generated with MATLAB and the histograms of the empirical marginal distributions after squashing with the fitted cumulative distribution functions. An appropriate way to quantify the goodness of fit in this context is to compare the plug-in entropy estimates for both histograms. Since the uniform distribution has maximum entropy for all distributions with bounded support, any misfit would lead to a smaller entropy. In addition, it holds that the entropy of the squashed distribution equals the negative KL divergence of the true distribution $\rho(y)$ from the model distribution whose cumulative distribution function equals the chosen squashing function[34]:

$$z = \hat{\mathcal{F}}(y) \Rightarrow h[Z] = -D_{KL}[\rho(y) \,\|\, \hat{\mathcal{F}}'(y)]. \tag{16}$$

This method has the advantage that it can easily deal with the problem of estimating densities with unbounded support, such as the exponential power family. This is difficult otherwise for large $|y_k|$, where the density converges to zero. For this reason, expression (16) may also provide an attractive alternative for fitting the parameter α of the exponential power family: The OPT estimator determines the optimal α, for which the entropy of the squashed distribution takes a maximum. After that, again, Eqs. (11) and (12) are used to compute the entropy.

The definition of the OPT estimator still requires one to specify how to estimate the empirical distribution of Z_k, $k=1,\ldots,N$. To get a robust nonparametric density estimate, it is desirable to make the outcome equally sensitive to all data points. This can be achieved with the fol-

lowing consistent estimate of the empirical distribution function, which is based on the order statistics. That is, we assume in the following that the samples $z_1 < z_2 < \cdots < z_N$ are sorted in ascending order. Similar to the sample median, we define

$$\beta_j = \frac{z_{jm} + z_{jm+1}}{2} \tag{17}$$

as the "inner" $(N/m-1)$ binning borders (assuming here for simplicity that N/m is an integer). Furthermore, the support of the distribution is confined to the output range of the squashing function, given by the interval (0, 1). Therefore we can take $\beta_0 = 0$ and $\beta_{N/m} = 1$ as the left and right "outer" binning borders. In this way, we have the same number of m data points within each bin, and hence the density estimate within each bin reads as

$$\hat{\rho}(z) = \frac{m}{N} \frac{1}{\beta_j - \beta_{j-1}} \qquad \text{for } \beta_{j-1} < z < \beta_j. \tag{18}$$

The plug-in entropy estimate follows immediately:

$$\hat{h}[Z] = \log_2 \frac{N}{m} + \frac{m}{N} \sum_{j=1}^{N/m} \log_2(\beta_j - \beta_{j-1}) \text{ bits}. \tag{19}$$

In the data analysis presented below, we set $m = 210$, so that the corresponding histogram with variable bin width has $N/m = 39{,}690/210 = 189$ bins. Taken together, the OPT estimator uses the MAL estimate as initial guess and then minimizes Eq. (19) via optimization of α using a standard line search algorithm.

The nonparametric entropy estimator just described for the estimation of Z_k, $k = 1, \ldots, N$, can also be applied directly to the coefficients Y_k, $k = 1, \ldots, N$. This nonparametric plug-in estimator is called the NPL estimator.

Finally, we apply the nonparametric m-spacing estimator, which was introduced by Vasicek.[35] The VAS estimator does not require estimating the density first, but it reduces the asymptotic variance of the estimator, loosely speaking, by averaging over $(m-1)$ shifted versions of the m-spacing estimator presented above. More specifically, we use the bias-corrected version

$$\hat{h}[Y] = \frac{1}{N \log(2)} \sum_{i=1} N - m \log\left(\frac{N}{m}[y^{(i+m)} - y^{(i)}]\right) - \psi(m) + \log(m) \text{ bits},$$

where $\psi = -\partial_x \log(\Gamma(x))$ denotes the digamma function. This estimator is part of the MATLAB toolbox of the ICA algorithm RADICAL presented in Ref. 36, and the estimator is explained in the review in Ref. 37 as well.

D. Comparison of the Four Estimators

To get an idea of how well these different estimators perform, we compare all of them on artificially generated data. Using the MATLAB random number generator for the uniform distribution, we generated 10^4 trials of 39,690 samples from a Laplacian distribution of variance 2 [for t=1:10000, x=log(rand(39690,1)).*sign ×(randn(39690,1)),..., end]. We chose the Laplacian distribution because the empirical distributions look very similar (some of them are slightly sparser, and some others are also a bit less kurtotic). Since the true entropy of a Laplacian distribution is determined by its variance to be $(1+\log\sqrt{2\ \mathrm{Var}[s_k]})/\log(2)$ bits, we can estimate not only the variance but also the bias of the different entropy estimators used. The results of this test are summarized in Table 1. It turns out that the OPT estimator performs best while the VAS estimator is the least favorable.

Table 1. Bias and Variance of the Four Different Entropy Estimators in the Case of a Laplacian Random Variable

Parameter (bits)	Estimator			
	MAL	OPT	NPL	VAS
Bias	0.0026	0	0.0023	−0.0371
$\sqrt{\text{Variance}}$	0.0072	0.0073	0.0073	0.0072
$\sqrt{\text{Total squared error}}$	0.0077	0.0073	0.0077	0.0378

4. RESULTS

Now that we have explained the details of the individual transforms and the different ways of estimating the multi-information, we are ready to compare them. Specifically, we consider PCA, zero-phase whitening, triangular whitening, ICA, the Haar wavelet, and a random decorrelation filter. The nonlinearity that we used in the contrast function of FastICA was $g(u) = 1 - \exp(-u^2)$, and the optimization was done by using the symmetric approach. All basis functions of the different transforms are shown in Figs. 1 and 2.

The results of the OPT estimator (the one that performed best on artificial data) are summarized in Fig. 6(a). Each curve shows $I[Y] - I[Y_{RND}]$ as a function of patch size, and the different curves correspond to the different transforms that generated Y. From this figure, one can directly read out the absolute *differences* in the multi-information reduction between the different transforms. As expected, the random decorrelation filter achieves the least reduction in the multi-information, and ICA achieves the maximal reduction. The slightly worse performance of symmetric decorrelation relative to the random decorrelation basis for patch sizes smaller than or equal to 4×4 can be seen as an artifact due to the separation of the DC component. While this preprocessing helps to make the results more stable against the particular choice of the data set, it affects the shape of the basis functions, especially for small patch sizes. Finally, it is interesting to note that the nonadaptive Haar wavelet decorrelation basis performs only slightly weaker than ICA despite the blockiness of the basis functions.

Apart from the fact that the absolute differences between the different transforms are small, it is interesting that these differences between all transforms stay constant over different patch sizes except for those with PCA. Apparently, PCA is a good choice for small image patches, but it is likely to perform worse than triangular whitening and zero-phase whitening for large image patches. A heuristic explanation for this finding is the lack of localization of the PCA basis in the spatial domain. The performance gap between zero-phase whitening and triangu-

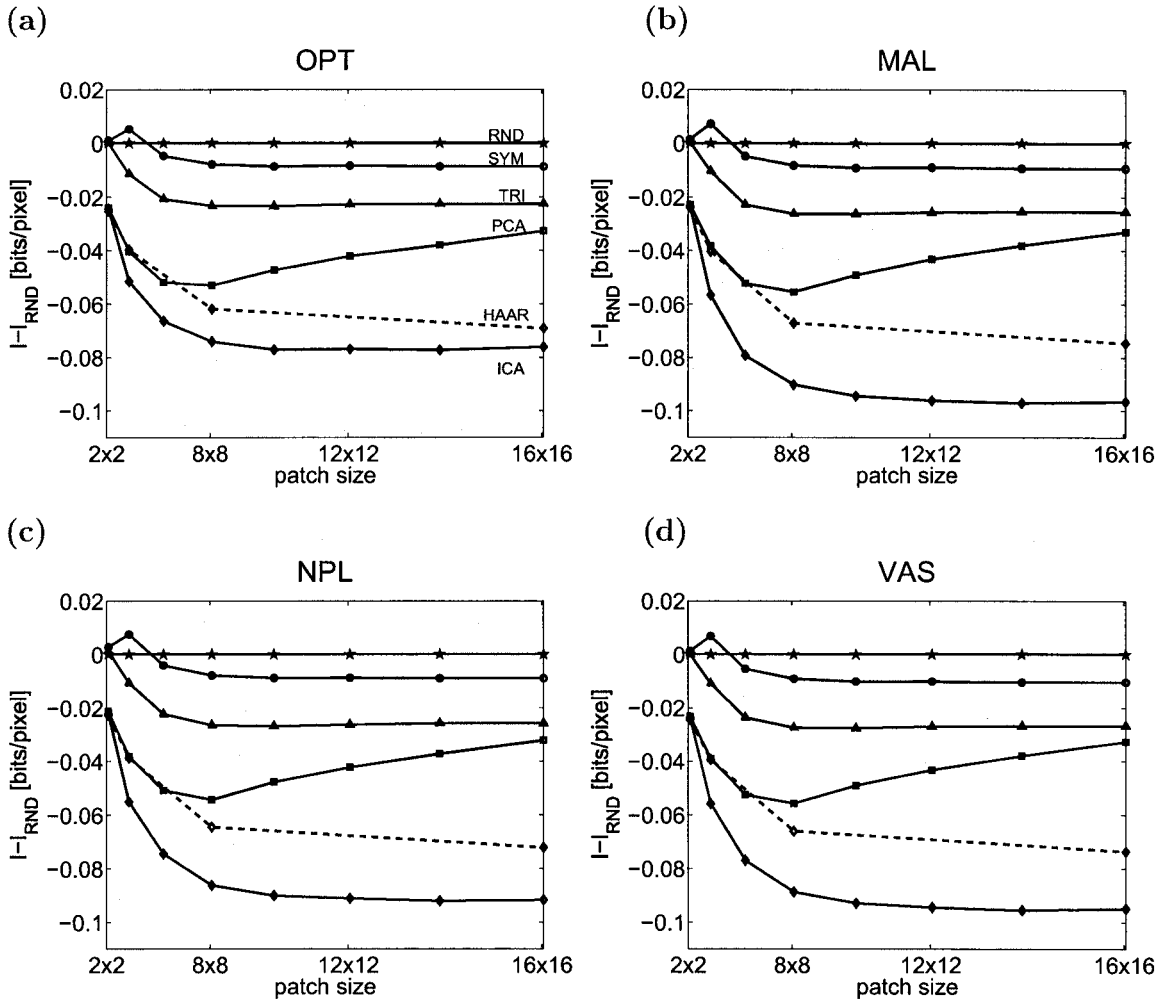

Fig. 6. (a) Multi-information estimates obtained from the OPT estimator. The six curves represent the absolute difference in multi-information relative to random whitening for random whitening (stars), symmetric whitening (circles), triangular whitening (triangles), PCA (squares), Haar wavelet (dashed curve with diamond), and ICA (diamonds) respectively, as a function of patch size. Due to construction, the difference has to vanish for random whitening. (b)–(d) Same as (a) but for results from different estimators: (b) MAL estimator, (c) NPL estimator, (d) VAS estimator. For all estimators, the maximum difference is smaller than 0.1 bits/pixel.

lar whitening might be due to the fact that the receptive fields of triangular whitening are more anisotropic (see Fig. 2). Referring to the predictive coding interpretation the reason for this anisotropy is the asymmetric sequential raster scheme with which triangular whitening predicts each pixel from the previous ones.

For control, the results of the three other estimators as well are shown in Figs. 6(b)–6(d). Additionally, the goodness of fit of the OPT estimator is shown in Fig. 7 by using the KL divergence between the optimized fit with the exponential power family and the nonparametric distribution estimate. Finally, we also inspected the fits by eye. All three control methods indicate that the presented estimates are highly reliable.

To appreciate how small the relative differences in multi-information reduction are between the different transforms, it is necessary to determine the total dependency reduction $I[Y]-I[X]$ relative to the pixel representation X including the DC component. The inclusion of the DC component was not necessary in the previous comparison, where we considered only differences between the *outputs* of the different transforms, because by con-

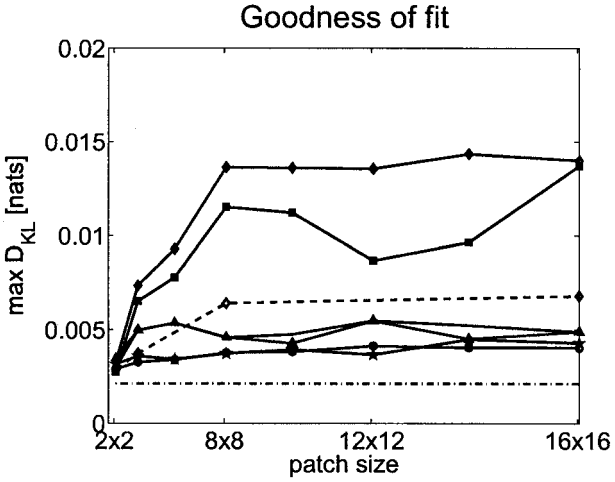

Fig. 7. Maximum empirical KL divergence over all dimensions after optimization (i.e., worst case). The theoretical optimum is indicated by the dotted–dashed line, which gives the empirical KL divergence for an artificially generated sample from a Laplacian distribution of the same size ($N=39{,}690$).

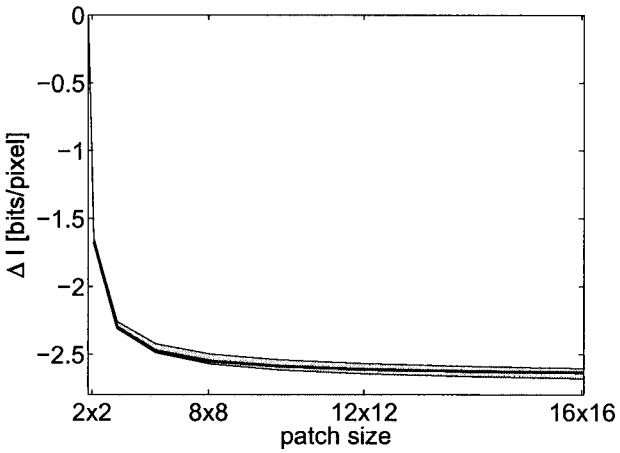

Fig. 8. Multi-information reduction on the $(m \times m)$-dimensional space including the DC component. The black line corresponds to the NPL estimator of the actual multi-information gain ΔI given by Eq. (19) in the case of PCA. The gray region around the solid curve indicates the range within which the multi-information gain varies for the different decorrelation methods. The upper bound of the gray region is given by random whitening, while the lower bound is given by ICA.

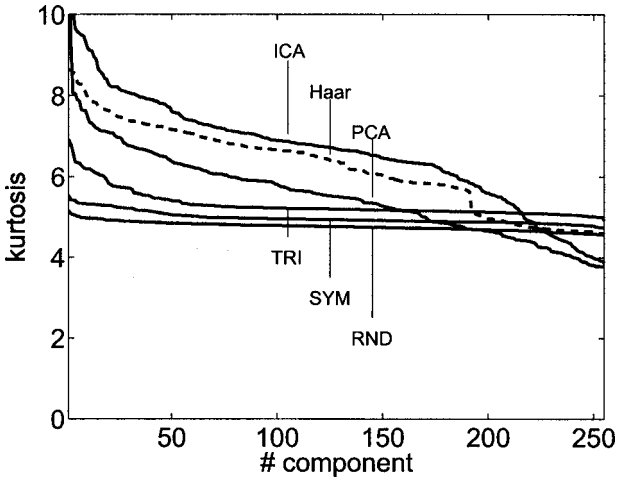

Fig. 9. Comparison of the excess kurtosis spectra for all methods in the case of 16×16 patches.

struction all transforms separate the same DC component with the same marginal entropy. To determine $I[Y]-I[X]$ for all transforms, we used the NPL estimator to evaluate the sum of the marginal entropies of the pixel representation because those cannot be fitted so well with the exponential power family. It is clear that this estimate cannot be of the same precision as that of the estimates for the output entropies, and it will also be more dependent on the particular data set used. Nevertheless, we may assume the precision to be of the order of 0.1 bits/pixel, which is sufficient to give a good ballpark figure. The total gain in the case of PCA, which includes the decorrelation of the DC component, is shown in Fig. 8. The shaded region indicates the tight range within which ΔI varies for the entire spectrum of decorrelation transforms, where the upper bound coincides with the curve of random decorrelation and the lower bound with that of ICA. As one can appreciate by eye from this graph, the relative differences in performance are very small, in fact, always smaller than 5%. Triangular whitening, in particular, achieves 98% of the multi-information reduction achieved with FastICA for all patch sizes.

Finally, to see how differently the individual components contribute to the multi-information gain, we show in Fig. 9 the excess kurtosis spectra for all transforms in the case of 16×16 patches. The bases of random whitening, zero phase, and triangular whitening exhibit very flat spectra, indicating that all components are equally sparse. Intuitively, this is to be expected, since all the components look pretty much the same apart from the location of their center peaks. In contrast, the sorted kurtosis spectra of the anisotropic bases of PCA, ICA, and the Haar wavelet are steadily decaying. For PCA, the low spatial frequencies exhibit the highest kurtosis. In the case of ICA, the most elongated edge filters are the most kurtotic. For the Haar wavelet, the kurtosis is roughly correlated with the scale of the basis functions such that the components at the smallest scale have the least kurtosis. In addition, the coefficients of the diagonal elements of the Haar basis exhibit less kurtosis than the vertical and horizontal components.

5. DISCUSSION

This study provides the first quantitative analysis of the multi-information reduction achieved with different linear filtering models of natural image statistics. Special care has been taken to make the required estimates as reliable as possible. The main result is that after second-order decorrelation, higher-order decorrelation with linear transforms amounts to a surprisingly small extra gain in terms of multi-information for natural images. As a consequence, this finding challenges the functional interpretation of V1 simple cell receptive fields as *linear* higher-order decorrelation filters.

We should be careful about the interpretation of this result. Foremost, this study seeks to be more precise about what we can conclude from the similarities between V1 simple cell receptive fields and the shape of linear ICA filters. The lack of a distinct advantage for the ICA edge filters in terms of statistical independence should not be taken as evidence against the approach of using the statistics of natural images to find better image representations. It rather demonstrates that the basic model of V1 simple cells as linear Gabor-like filters is not very effective as a means of factorial coding for natural images.

In general, the restriction to linear processing heavily constrains the range of possible computations. Given that the linear independent components still exhibit higher-order correlations, it is likely that more flexible, nonlinear mappings may achieve a much larger gain in the multi-information reduction. In terms of image analysis, the computational limitations of linear signal processing also give reason to be skeptical about the common view of V1 function. Oftentimes, the investigation of V1 simple cells builds on the notion that filtering with Gabor-like receptive fields effectively encodes for the presence of edges.

This idea, however, ignores the fact that the detection of the outline of an object in natural images is an unsolved problem in computer vision, which crucially relies on the appropriate choice of nonlinearities. Moreover, the computational limitations of linear image analysis are complemented by the physiological fact that a large fraction of the variance in V1 simple cell responses cannot be explained with the classical linear response model (for a recent critique of the standard model of early visual processing, see Ref. 38).

In addition to the need for more flexible nonlinear image models, it is also necessary to reexamine the assumptions underlying factorial coding. In particular, the presented quantitative evaluation of image models in terms of *statistical independence* should not leave us with the impression that factorial coding is the only thing that we need to consider in order to build better image models or to come up with better hypotheses about neural image representations. Traditionally, ideas about coding efficiency borrowed from information theory played a strong component in the motivation of factorial coding in neuronal representations.[13,39–41] It is important to note, however, that maximal statistical independence is not necessarily optimal for coding efficiency,[42] and coding efficiency is not sufficient as an ultimate design principle for useful image representations.

In the following, we will first discuss the principal caveats of factorial coding in terms of coding efficiency. In particular, we will discuss ICA from the rate-distortion theoretical perspective of transform coding.[42,43] Next, we will explain the conceptual limitations of blind source separation (BSS) in the context of natural images. Finally, we will explore how unsupervised learning provides a viable approach to the problem of optimal representation learning.

A. Factorial Coding Is Not Sufficient for Coding Efficiency

ICA is equivalent with the task of finding a lossless mapping of a given multivariate random variable such that the new output random variable is uniformly distributed (see Appendix D). This task coincides with the problem that one has to solve in redundancy reduction in the case of discrete sources. In the case of continuous sources, however, the reduction of statistical dependencies *per se* does not imply any compression, because the lossless description length of real numbers is always divergent.[44]

Clearly, it is easy to construct a discrete code from ICA via quantization of the output coefficients. It is also easy to construct a discrete-valued maximum entropy code with uniform distribution if the density over the continuous-valued ICA coefficients is uniform. Similarly, one may include output noise instead of quantization to turn the ICA transform into a channel with finite capacity. As has been pointed out in Refs. 19 and 45, the goal of a uniform output distribution in the presence of additive noise is equivalent with a maximization of the mutual information between input and output. In fact, if the neural noise model does not depend on the filter matrix W, then maximizing the mutual information between input and output is equivalent with maximizing the entropy of the output of the channel, as is done in Bell–Sejnowski ICA. Things become more involved, of course, if the channel noise cannot be assumed to be W-independent. In the case of independent Poisson noise, for instance, the optimal neural activation functions with respect to information maximization are staircase functions[46] (also cf. Refs. 47 and 48). The optimization for W-dependent noise, however, may be regarded as a subsequent step in the context of information maximization.

As long as one is concerned only about error-free transmission of a discrete signal (i.e., channel coding), information maximization is a valid design principle. However, information maximization *per se* is meaningless if the task is to find an efficient description of continuous-valued data (i.e., source coding). Intuitively speaking, the representation of a continuous source via a channel with finite capacity always requires one to discard infinitely many bits because only a finite number of bits can be transmitted. Therefore the most important question in source coding is how to decide *which* bits or, more precisely, which changes of the signal are most worthy to be represented. This bit-selection problem can be decided on the basis of perceptual relevance or task relevance only. Hence, to judge a coding scheme, one always has to verify how good the perceptual or performance quality is for a given information rate.

In the context of ICA, this problem can be demonstrated if we actually compare the perceptual distortion of the different linear transforms in the presence of noise or after quantization of the output. As an illustrative example, we show the Lena image for the pixel basis, the orthogonal PCA basis (that is, $\mathbf{s}=U_{PCA}\mathbf{x}$), the PCA whitening basis (that is, $\mathbf{s}=D_1U_{PCA}\mathbf{x}$), and the ICA whitening basis (that is, $\mathbf{s}=V_{ICA}D_1U_{PCA}\mathbf{x}$) after independent and equidistant quantization of the output coefficients. As one can see in Fig. 10, the perceptual quality of orthogonal PCA is by far the best although its information rate has been chosen to be the smallest. The comparison of orthogonal PCA and PCA whitening bases shows that the whitening step has a large drawback in terms of perceptual quality. So orthogonal PCA can be interpreted as the better compromise between the advantageous pixel metric and statistical independence.[49]

The obvious alternative to PCA would be to try orthogonal ICA, where the multi-information is minimized under the same constraint that W is orthogonal. However, the rate-distortion gain will be very small, since the restriction that W be orthogonal implies that PCA is the only transform for which all second-order correlations vanish. The power spectrum obtained from the pixel intensities of natural images is not flat.[50] In addition, the present study has demonstrated that the difference in negentropy between ICA and PCA is small. Taking these two facts together, it is likely that the optimal transform in the rate-distortion sense is much closer to the PCA filters than to the ICA filters: If the input signal of orthogonal ICA is not white to begin with, then any rotation away from the PCA basis in order to increase the total negentropy comes at the cost of an increase in the Gaussian entropy bound.

The fact that maximal statistical independence is not necessarily optimal for coding efficiency is well-known in transform coding research.[42] For independent coefficient quantization, there is no competitor to discrete cosine

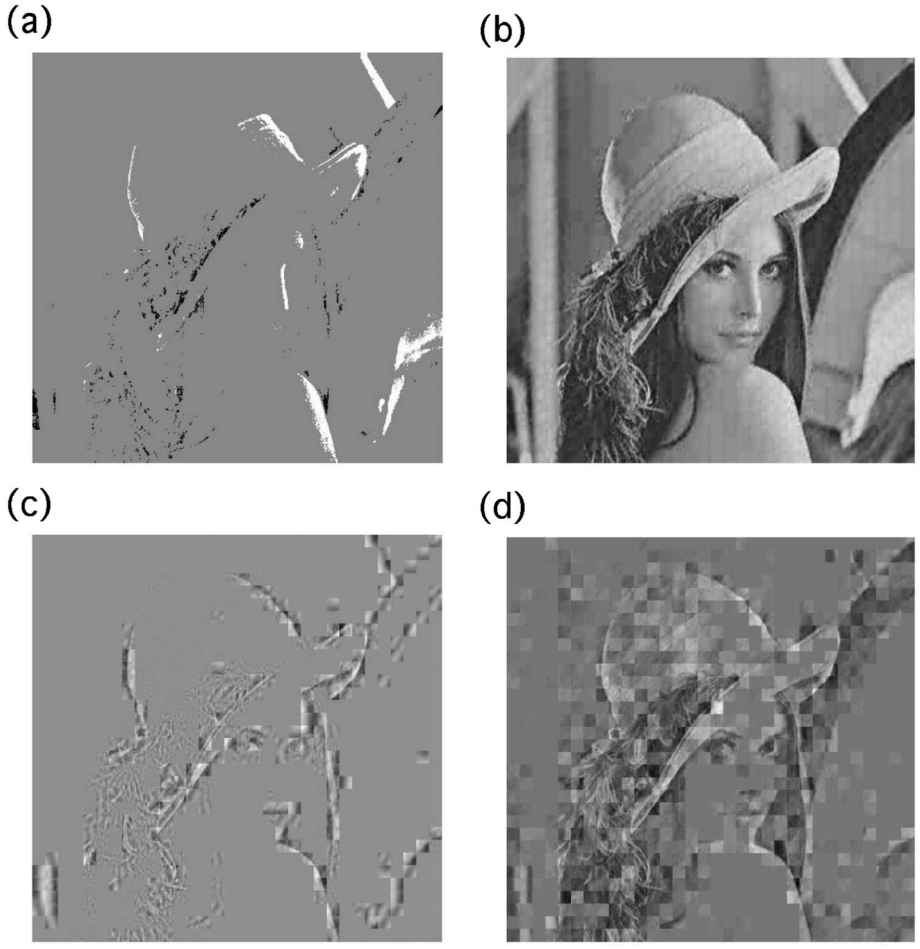

Fig. 10. Comparison of perceptual distortion between different transforms using uniform quantization of the output coefficients: (a) quantization in the original pixel basis with maximum rate (0.23 bits/pixel), (b) quantization in the orthogonal PCA basis with minimum rate (0.13 bits/pixel), (c) quantization in the PCA whitening basis with second largest rate (0.20 bits/pixel), (d) quantization in the ICA whitening basis with second smallest rate (0.17 bits/pixel).

transform coding and most people today are still using the old JPEG still image compression standard. The slight advantage of wavelet coding used in the more recent JPEG 2000 standard[51] can be achieved only as embedded zero-tree wavelet coding,[52] which gives up the independence assumption of the transform coefficients. In conclusion, the results of transform coding suggest that plain ICA is rather less efficient than PCA in terms of coding efficiency. This underlines the basic fact of rate-distortion theory that factorial coding and information maximization are not sufficient for efficient coding.

B. Blind Source Separation Is Not Sufficient for Optimal Representation Learning

Although coding efficiency is frequently used to motivate factorial coding and unsupervised learning, we agree with the view in Refs. 53 and 54 that compressive coding is not actually the ultimate goal of early vision. While neuronal representations of sensory inputs in the brain are required to avoid a waste of the physiological resources, it would be very limiting if one sought to understand neural representations from this constraint only. Foremost, we need to answer the following question: How does neural processing transform the retinal image into useful representations that make explicit the behaviorally relevant structure and geometry of the environment?

An important motive underlying the use of ICA-like algorithms in image coding is the goal of extracting meaningful parameters by means of statistical learning. Originally, ICA had been developed in the context of BSS.[3] The attribute "blind" stands for the fact that within the class of linear models no further assumptions are required to identify non-Gaussian source signals up to scaling factors from the statistics. For the purpose of image representation learning, however, the concept of BSS needs to be modified: Strictly speaking, BSS makes statements only about the case when the generative model used is correct. The optimization in BSS is used only to find a unique answer. The theory of BSS does not really care about gradual improvements in the objective function, because it does not require that the objective function express a desirable feature. It is just an arbitrary contrast function.

In the case of ICA, this means that the multi-information does not necessarily express the most desirable objective and that other criteria might be used equivalently. A way to illustrate this fact is to consider data generated by the following linear time-invariant generative model:

$$\mathbf{x}_t = A\mathbf{s}_t. \tag{20}$$

With the assumption that all source signals are mutually independent, one can recover the matrix A of basis functions either with FastICA or with Molgedey and Schuster ICA,[55] which use different objective functions. More specifically, Molgedey and Schuster ICA does not use higher-order correlations but decorrelates the time-delayed cross covariance in order to find a unique answer. If both methods are applied to wildlife movies, one still finds the edge filters with FastICA while the basis functions determined with Molgedey and Schuster ICA look very different from those (see Fig. 11). From the BSS point of view, the discrepancy in the answer between the two methods simply means that the linear time-invariant model (20) is wrong for time-varying natural images. BSS does not tell us why we may prefer the answer given by FastICA over the answer given by Molgedey and Schuster ICA or vice versa.

In image representation learning, it is not assumed that the generative models are actually correct. Therefore, the solutions cannot be interpreted in terms of BSS. The only way to assess a given answer meaningfully in this case is how well it performs as measured by the stated objective function.

C. Optimal Representation Learning: Unsupervised Learning Meets Efficient Coding

In this final part of the discussion, we name some examples of unsupervised learning models that are more closely related to the goal of efficient coding than plain ICA is. An important extension of ICA is independent factor analysis[7,12,56,57] "(IFA)" (originally called sparse coding, sometimes also called noisy ICA). Like ICA, it uses a generative model that assumes non-Gaussian sources. In contrast to ICA, however, IFA allows one to describe the input as a superposition of an arbitrary number of sources *plus noise*. The use of the noise model is not limited to the case where noise is actually present in the data. Intuitively speaking, a noise model can also be used as a means to specify the importance of different bits. In fact, the frequently chosen isotropic Gaussian noise model corresponds to the assumption of the Euclidean metric as distortion measure. Consequently, the choice of the variance corresponds to a parameter related to the rate-distortion trade-off.

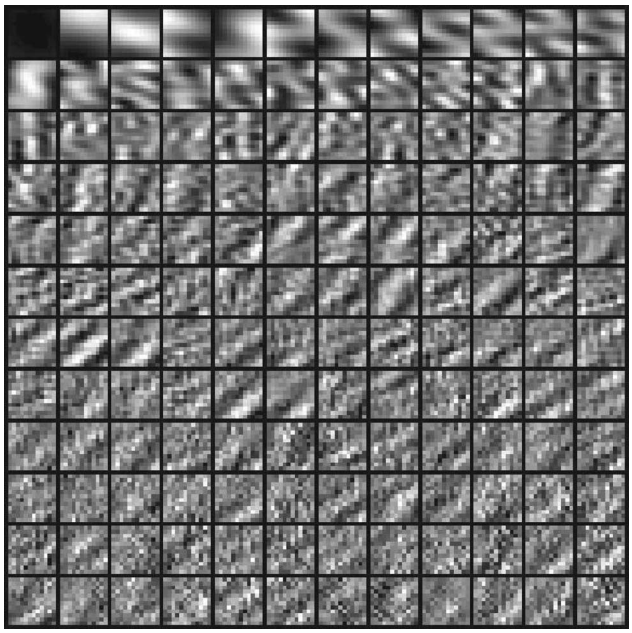

Fig. 11. Basis functions for 16×16 image patches learned with Molgedey and Schuster ICA from a wildlife movie.

A quantitative analysis of coding efficiency has been carried out by Lewicki and Olshausen in Ref. 12. They compared the discrete coefficient histogram entropy of the image basis learned with IFA with that of other image bases for a given mean square reconstruction error, which attested its good performance. However, the learning as well as the performance evaluation was carried out with respect to the Euclidean metric in the whitened space. It would be very interesting to know the results of applying IFA to images in the pixel metric.

Another possible extension is to combine ICA with a multiscale representation such as the Laplacian pyramid.[58] Intuitively speaking, we may think of prewhitened ICA as the optimal transform with respect to the Euclidean metric in the whitened space. It might well be possible that plain ICA can successfully be used to encode the individual levels of the Laplacian pyramid because within each level the covariance matrix of the image patches has a rather flat spectrum to begin with. There is some recent work along these lines, where people started to optimize wavelets with respect to the statistics of the signal to be represented.[59,60]

As a final example, the issue of efficient coding in neural representations can be addressed most explicitly by utilizing a joint-source channel coding approach. In addition to a distortion measure, it also assumes a specific neural noise model and a certain type of decoder. In a minimalist model, for instance, we may assume that each V1 simple cell belongs to a group of neurons whose coding objective is to minimize the mean square error reconstruction of a certain image patch in pixel space. This model can be seen as a combination of current neural image coding models with current models of optimal neural population codes: Neural image coding models put strong emphasis on the input statistics to inform the model but rarely address the effect of neural noise on the optimal code. Instead, most models of optimal neural population coding have been used mainly to investigate the effect of different neural noise models, while the input signal is simply assumed to be a random variable with a convenient distribution without any further specifications.

All aspects of efficient coding, stimulus statistics, neural noise, and perceptual distortion can be combined in such an optimal joint-source channel coding approach.[61] As an interesting example, one may study the minimum mean square error reconstruction for Gaussian noise achieved with a linear readout mechanism,[62] which can significantly change the shape of optimal image representations. This setting is very close to the standard transform coding setting,[42] as it essentially replaces the quantization by Gaussian noise. From previous work on optimal population coding, we can expect even larger changes in the shape of optimal neural image representations by choosing a Poisson noise model and a nonlinear

minimum mean square estimator for the reconstruction.[47,48,63]

In the above list of examples, we have focused on the aspect of coding efficiency, but other objectives besides coding efficiency can be optimized as well. The crucial point in optimal representation learning is that the objective function really define the criterion according to which one would like to judge the performance of the representation. Quantitative comparisons such as the one presented in this paper can then be used to clarify how sensitive the representation is to the goal defined by the objective function.

APPENDIX A: TRIANGULAR DECORRELATION

We included triangular decorrelation in our comparison because it can be seen as the transform coding version of linear predictive coding. A convenient way to determine a triangular decorrelation transform is to apply the Cholesky decomposition to the covariance matrix, $C_{\mathbf{x}} = LL^{\mathrm{T}}$, where L is lower triangular. Since the Cholesky decomposition is unique, it recovers the true mixing matrix $A=L$ whenever the assumption of a triangular mixing matrix is correct. Furthermore, the inverse matrix of a triangular matrix is again triangular. Thus $W_{TRI}=A^{-1}=L^{-1}$ defines the filter matrix of the triangular whitening transform. In the next section we provide a motivation showing that triangular decorrelation can be seen as the transform coding version of predictive coding.

Triangular decorrelation and predictive coding. To make the interpretation of triangular decorrelation transforms in terms of predictive coding most straightforward, we now require all entries on the main diagonal of the triangular decorrelation matrix W to be equal to *minus one*. This choice is possible because, after whitening, an arbitrary diagonal transform D_2 can be applied that does not change the off-diagonal elements of the covariance matrix. Then each output coefficient is given by

$$y_k = \underbrace{\sum_{j<k} W_{kj} x_j}_{=\hat{x}_k} - x_k. \tag{A1}$$

Because of the triangular structure, one can minimize the component variances in a greedy fashion without loss of optimality. The minimization of each individual $Var[yk]$ corresponds to the problem of optimal linear prediction of x_k from the previous $k-1$ components $(x_1, \ldots, x_{k-1})$:

$$\hat{x}_k = \sum_{j<k} W_{kj} x_j. \tag{A2}$$

Therefore the triangular decorrelation matrix can also be determined by using the linear minimum mean square estimator. If $\mathbf{x}$ is a random variable in R^n with $E[\mathbf{x}]=0$, then the linear minimum mean square estimator of the kth component as a function of the previous $k-1$ components is given by[64]

$$\hat{x}_k = \underbrace{E[x_k(x_1, \ldots, x_{k-1})]C_{k-1}^{-1}}_{=(W_{k1},\ldots,W_{k(k-1)})}(x_1, \ldots, x_{k-1})^{\mathrm{T}}. \tag{A3}$$

where C_m denotes the covariance matrix of the first m components:

$$C_m = E[(x_1, \ldots, x_m)^{\mathrm{T}}(x_1, \ldots, x_m)]. \tag{A4}$$

Consequently, triangular decorrelation can be expected to work well whenever predictive coding is expected to work well. That is, triangular decorrelation will lead to good results in the case of autoregressive processes. Note that in the case of spatially predictive coding, as used for still image coding, it is not obvious what the optimal ordering of the pixels is. More generally, the performance of triangular whitening depends on the particular basis of the input space. In the case when the input space is given by the pixel basis, one has the freedom only to choose a permutation of the ordering of the dimensions. In this paper, we use a simple rowwise raster scan through the patch to specify the ordering of the dimensions.

APPENDIX B: LOG INTENSITIES VERSUS LINEAR INTENSITIES

The multi-information is not invariant under nonlinear transforms of the pixel intensities. A principled approach to finding the "right" pixel intensity representation is to model the data as a postnonlinear (PNL) mixture.[65] A fit of the PNL model, however, is beyond the scope of this paper. We decided to use log intensities mainly because it is a common way to model the dynamic range adaptation of photoreceptors in the retina. In comparison with the occasionally used linear intensities, the choice of log intensities leads to less kurtotic pixel intensity distributions. Small kurtosis is not only more plausible in terms of a postnonlinear mixture model[66] but it also enhances the reliability of the marginal entropy estimates: After "DC0 filtering" (that is, removing the DC component), we find that the shape of all marginal distributions becomes very close to that of a double-sided exponential (or Laplacian) distribution and the fit with the exponential power family (which contains the Laplacian distribution as a special case) is excellent when using the log-intensity scale. This finding is typical for images whose content is dominated by greens and woods (see also Fig. 1(c) in Ref. 24). Images with a different content typically lead to marginal distributions with higher kurtosis.

APPENDIX C: SAMPLING SCHEME

For the sake of maximal reproducibility, we decided to use a deterministic sampling scheme, but the results do not rely on this choice. To generate the training set, we first sampled, from each of the ten images, 63^2 patches of the maximal size (16×16 pixels). These patches were sampled without overlap, such that tiling them together in the correct order would recover the entire given image apart from the lowest 16 rows and the rightmost 16 columns (see Fig. 12). In this way, we obtained $10\times63^2 = 39{,}690$ samples. For the test set, we got the same number of samples from the same ten images by choosing those 63^2 patches that tile the center part of each image instead of the upper left. That is, a margin of 8 pixel width is left out at all four edges of the given image in this case. The training and test sets for smaller image patches

are obtained from these, simply by taking only a certain fraction of each 16×16 patch (we always took the upper left part).

The basis functions are learned by using only the data from the training set. The multi-information gain is evaluated not only for the training set but also for the test set. A large difference between both measurements would indicate overfitting. In all measurements carried out in this study, the difference between both measurements is negligible.

APPENDIX D: GAUGING THE MULTI-INFORMATION GAIN

1. Bell–Sejnowski Gauge

In the Bell–Sejnowski gauge, a pointwise nonlinear mapping (a "squashing function") is applied to each individual output channel such that $f_k(\mathbf{x})=g_k(s_k)$, where

$$y_k := g_k(s_k) = \int_{-\infty}^{s_k} p(\tilde{s}_k)\mathrm{d}\tilde{s}_k \tag{D1}$$

is set to be the cumulative distribution function of each individual component S_k. In general, pointwise nonlinear mappings of the source variables S_k do not affect the multi-information. Since the squashing functions are here chosen to be the distribution functions of the S_k, it follows that the individual output components Y_k are uniformly distributed in the unit interval (0,1), so that their differential entropies vanish. In this case, we have

$$I^{BS}_{multi}[\mathbf{Y}] = -h[\mathbf{Y}] = -E\left[\log|\det(W)| + \sum_{k=1}^{d}\log\left|\frac{\partial g_k}{\partial s_k}\right|\right] - h[\mathbf{X}], \tag{D2}$$

which formally translates the minimization of multi-information into a maximum entropy problem. Since the uniform distribution is the maximum entropy distribution on the unit interval, it is also possible to estimate the distribution functions simultaneously by maximizing the entropy.

2. Volume-Conserving Gauge

For the purpose of precise estimation of the achieved statistical dependency reduction, the choice of volume-conserving ICA has some advantages. In linear volume-conserving ICA, we do not have the pointwise nonlinearities; rather, the entire mapping $\mathbf{f}$ is linear:

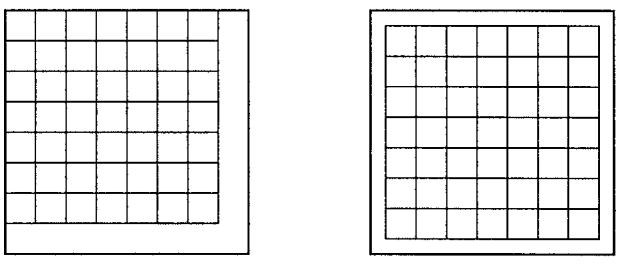

Fig. 12. Graphic demonstrating the scheme used to sample the image patches from the van Hateren images: training set (left), test set (right).

$$\mathbf{y} = W\mathbf{x}. \tag{D3}$$

The constraint of volume conservation implies that $|\det(W)|=1$, so that the log determinant is always zero. Consequently, the joint output entropy equals the joint input entropy, and thus the multi-information between the output components is

$$I^{VC}_{multi}[\mathbf{Y}] = \underbrace{\sum_k h[Y_k]}_{=c[Y_1,\ldots,Y_n]} - h[\mathbf{X}]. \tag{D4}$$

This means that differences in the efficiency between different codes require only the evaluation of differences in $c[Y_1,\ldots,Y_n]$, the sum of marginal entropies.

The restriction to volume-conserving transforms does not affect the ICA solution up to a global rescaling factor, because the multi-information function is invariant with respect to a global rescaling of the filter matrix. In other words, for any given filter matrix W, the rescaled matrix $W/\sqrt{\det(W)}d$, where d denotes the dimensionality of X and Y, is equivalently optimal with respect to the output multi-information. Conversely, Eq. (D2) can also be interpreted as the (marginally stable) Lagrangian way to solve the problem of minimizing $c[Y_1,\ldots,Y_n]$ under the side constraint $\det(W)=\text{constant}$.

ACKNOWLEDGMENTS

I thank Patrik Hoyer for pointing me at an early stage of the work to the relationship between triangular decorrelation and the QR decomposition (which I later replaced with the Cholesky decomposition). Furthermore, we had several e-mail discussions that helped a lot to improve both the content and style of the paper. I am indebted to Bruno Olshausen, who continually contributed helpful advice at all stages of this work. Special thanks go to Christian Machens, whose suggestions made the manuscript more readable. Finally, I received helpful remarks from Kilian Koepsell, David Rotermund, Jakob Macke, and Eero Simoncelli.

Author contact information: mbethge@tuebingen.mpg.de, Redwood Center of Theoretical Neuroscience, Helen Wills Neuroscience Institute, 132 Barker, MC #3190, Berkeley, California 94720-3190.

REFERENCES

1. E. P. Simoncelli, "Statistical models for images: compression, restoration, and synthesis," in *31st Asilomar Conference on Signals Systems, and Computers*, Pacific Grove, Calif. (IEEE Computer Society, 1997), pp. 673–678.
2. S. C. Zhu, "Statistical modeling and conceptualization of visual patterns," IEEE Trans. Pattern Anal. Mach. Intell. **25**, 691–712 (2003).
3. P. Comon, "Independent component analysis, a new concept?" Springer Proc. Phys. **36**, 287–314 (1994).
4. A. Hyvrinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley, 2001).
5. E. P. Simoncelli, J. Pillow, L. Paninski, and O. Schwartz, "Characterization of neural responses with stochastic stimuli," in *The Cognitive Neurosciences, III*, M. Gazzaniga, ed. (MIT Press, 2004), Chap. 23, pp. 327–338.
6. Y. Dan, J. J. Atick, and R. C. Reid, "Efficient coding of natural scenes in the lateral geniculate nucleus:

experimental test of a computational theory," J. Neurosci. **16**, 3351–3362 (1996).
7. B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature (London) **381**, 560–561 (1996).
8. A. Bell and T. Sejnowski, "The 'independent components' of natural scenes are edge filters," Vision Res. **37**, 3327–3338 (1997).
9. J. Jones and L. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," J. Neurophysiol. **58**, 1187–1211 (1987).
10. C. Zetzsche, B. Wegmann, and E. Barth, "Nonlinear aspects of primary vision: entropy reduction beyond decorrelation," in *International Symposium* (Society for Information Display, 1993), Vol. 24, pp. 933–936.
11. M. Studeny and J. Vejnarova, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in Graphical Models*, M. I. Jordan, ed. (MIT Press, 1998), pp. 261–297.
12. M. Lewicki and B. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," J. Opt. Soc. Am. A **16**, 1587–1601 (1999).
13. J. Atick, "Could information theory provide an ecological theory of sensory processing?" Network **3**, 213–251 (1992).
14. H. Kramer and M. Mathews, "A linear coding for transmitting a set of correlated signals," IEEE Trans. Inf. Theory **2**(3), 41–46 (1956).
15. A. Netravali and J. Limb, "Picture coding: a review," Proc. IEEE **68**, 366–406 (1980).
16. B. Carpentieri, M. Weinberger, and G. Seroussi, "Lossless compression of continuous-tone images," Proc. IEEE **88**, 1797–1809 (2000).
17. M. Srinivasan, S. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," Proc. R. Soc. London, Ser. B **216**, 427–459 (1982).
18. A. Hyvarinen, "Survey on independent component analysis," Neural Comput. Surv. **2**, 94–128 (1999).
19. A. Bell and T. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," Neural Comput. **7**, 1129–1159 (1995).
20. T. Lee and M. Lewicki, "The generalized Gaussian mixture model using ICA," in *ICA '00*, P. Pajunen and J. Karhunen, eds., 2000), pp. 239–244. Available at http://www.cis.hut.fi/ica2000/proceedings/proceedings.html.
21. L. Zhang, A. Cichocki, and S. Amari, "Self-adaptive blind source separation based on activation functions adaptation," IEEE Trans. Neural Netw. **15**, 233–244 (2004).
22. A. Haar, "Zur Theorie der orthogonalen Funktionensysteme," Math. Ann. **69**, 331–371 (1909–1910).
23. J. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," Proc. R. Soc. London, Ser. B **265**, 1724–1726 (1998).
24. D. Ruderman and W. Bialek, "Statistics of natural images: scaling in the woods," Phys. Rev. Lett. **73**, 814–817 (1994).
25. C. Zetzsche, "Polyspectra of natural images," presented at the Natural Scene Statistics Meeting, September 11–14, 1997, Hancock, Mass.
26. R. Baddeley, "An efficient code in V1," Nature (London) **381**, 560–561 (1996).
27. A. Kraskov, H. Stogbauer, and P. Grassberger, "Estimating mutual information," Phys. Rev. E **69**, 066138 (2004).
28. G. Deco and W. Brauer, "Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures," Neural Networks **8**, 525–535 (1995).
29. N. Farvardin and J. Modestino, "Optimum quantizer performance for a class of non-Gaussian memoryless sources," IEEE Trans. Inf. Theory **IT-30**, 485–497 (1984).
30. J.-F. Cardoso, "Dependence, correlation and non Gaussianity in independent component analysis," J. Mach. Learn. Res. **4**, 1177–1203 (2003).
31. A. Srivastava, A. Lee, E. P. Simoncelli, and S. Zhu, "On advances in statistical modeling of natural images," J. Math. Imaging Vision **18**, 17–33 (2003).
32. A. Hyvarinen, "One-unit contrast functions for independent component analysis: a statistical analysis," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing VII*, Amelia Island, Fla. (IEEE, 1997), pp. 388–397.
33. S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989).
34. Z. Roth and Y. Baram, "Multidimensional density shaping by sigmoids," IEEE Trans. Neural Netw. **7**, 1291–1298 (1996).
35. O. Vasicek, "A test for normality based on sample entropy," J. R. Stat. Soc. Ser. B. Methodol. **38**, 54–59 (1976).
36. E. Learned-Miller and J. Fisher, "ICA using spacings estimates of entropy," J. Mach. Learn. Res. **4**, 1271–1295 (2003).
37. J. Beirlant, E. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," Int. J. Math. Stat. Sci. **6**, 17–39 (2001).
38. B. Olshausen, "What is the other 85% of V1 doing?" in *23 Problems in Systems Neuroscience*, T. Sejnowski and L. van Hemmen, eds. (Oxford U. Press, 2004).
39. F. Attneave, "Informational aspects of visual perception," Psychol. Rev. **61**, 183–193 (1954).
40. H. Barlow, "Sensory mechanisms, the reduction of redundancy, and intelligence," in *The Mechanisation of Thought Processes* (Her Majesty's Stationery Office, 1959), pp. 535–539.
41. H. Barlow, "Unsupervised learning," Neural Comput. **1**, 295–311 (1989).
42. V. Goyal, "Theoretical foundations of transform coding," IEEE Signal Process. Mag. **18**(5), 9–21 (2001).
43. P. Wintz, "Transform picture coding," Proc. IEEE **60**, 809–820 (1972).
44. R. Gray, *Entropy and Information Theory* (Springer, 1990).
45. J. Nadal and N. Parga, "Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer," Network Comput. Neural Syst. **5**, 565–581 (1994).
46. N. Brunel and J.-P. Nadal, "Optimal tuning curves for neurons spiking as a Poisson process in response to a scalar stimulus," in *Proceedings of the European Symposium on Artificial Neural Networks* (D facto publicats, Brussels, 1997), pp. 163-168.
47. M. Bethge, D. Rotermund, and K. Pawelzik, "Optimal neural rate coding leads to bimodal firing rate distributions," Network Comput. Neural Syst. **14**, 303–319 (2003).
48. M. Bethge, D. Rotermund, and K. Pawelzik, "A second order phase transition in neural rate coding: binary encoding is optimal for rapid signal transmission," Phys. Rev. Lett. **90**, 088–104 (2003).
49. V. Goyal, *Simple and Multiple Description Transform Coding with Bases and Frames* (SIAM, 2001).
50. A. Pentland, "Fractal-based description of natural scene," IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 661–674 (1984).
51. B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: foundations of JPEG 2000," IEEE Signal Process. Mag. **18**(5), 22–35 (2001).
52. J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," IEEE Trans. Signal Process. **41**, 3445–3462 (1993).
53. H. Barlow, "Redundancy reduction revisited," Network **12**, 241–253 (2001).
54. E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," Annu. Rev. Neurosci. **24**, 1193–1216 (2001).
55. L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," Phys. Rev. Lett. **72**, 3634–3637 (1994).
56. A. Hyvarinen, "Gaussian moments for noisy independent component analysis," IEEE Signal Process. Lett. **6**, 145–147 (1999).
57. H. Attias, "Independent factor analysis," Neural Comput. **11**, 803–851 (1999).
58. P. Burt and E. Adelson, "The Laplacian pyramid as a

compact image code," IEEE Trans. Commun. **COM-31**, 532–540 (1983).
59. B. A. Olshausen, P. Sallee, and M. S. Lewicki, "Learning sparse image codes using a wavelet pyramid architecture," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, eds. (MIT Press, 2001), pp. 887–893.
60. A. Gupta, S. Joshi, and S. Prasad, "A new approach for estimation of statistically matched wavelet," IEEE Trans. Signal Process. **53**, 1778–1793 (2005).
61. M. Bethge, "Codes and goals of neuronal representations," Ph.D. thesis (University of Bremen, Germany, 2003).
62. E. Doi and M. S. Lewicki, "Sparse coding of natural images using an overcomplete set of limited capacity units," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, eds. (MIT Press, 2005), pp. 377–384.
63. M. Bethge, D. Rotermund, and K. Pawelzik, "Optimal short-term population coding: when Fisher information fails," Neural Comput. **14**, 2317–2351 (2002).
64. S. Kay, *Estimation Theory*, Vol. I of Fundamentals of Statistical Signal Processing (Prentice Hall, 1993).
65. A. Taleb and C. Jutten, "Nonlinear source separation: the post-nonlinear mixtures," in *European Symposium on Artificial Neural Networks* (D facto publicats, Brussels, 1997), pp. 279–284.
66. A. Ziehe, M. Kawanabe, S. Harmeling, and K. Muller, "Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation," J. Mach. Learn. Res. **4**, 1319–1338 (2003).