DENSITY MODELING OF IMAGES USING A GENERALIZED NORMALIZATION TRANSFORMATION

Johannes Ballé, Valero Laparra & Eero P. Simoncelli *
Center for Neural Science
New York University
New York, NY 10004, USA
{johannes.balle,valero,eero.simoncelli}@nyu.edu

ABSTRACT

We introduce a parametric nonlinear transformation that is well-suited for Gaussianizing data from natural images. The data are linearly transformed, and each component is then normalized by a pooled activity measure, computed by exponentiating a weighted sum of rectified and exponentiated components and a constant. We optimize the parameters of the full transformation (linear transform, exponents, weights, constant) over a database of natural images, directly minimizing the negentropy of the responses. The optimized transformation substantially Gaussianizes the data, achieving a significantly smaller mutual information between transformed components than alternative methods including ICA and radial Gaussianization. The transformation is differentiable and can be efficiently inverted, and thus induces a density model on images. We show that samples of this model are visually similar to samples of natural image patches. We demonstrate the use of the model as a prior probability density that can be used to remove additive noise. Finally, we show that the transformation can be cascaded, with each layer optimized using the same Gaussianization objective, thus offering an unsupervised method of optimizing a deep network architecture.

1 INTRODUCTION

The learning of representations for classification of complex patterns has experienced an impressive series of successes in recent years. But these results have relied heavily on large labeled data sets, leaving open the question of whether such representations can be learned directly from observed examples, without supervision. Density estimation is the mother of all unsupervised learning problems. A direct approach to this problem involves fitting a probability density model, either drawn from a parametric family, or composed as a nonparametric superposition of kernels, to the data. An indirect alternative, which can offer access to different families of densities, and in some cases an easier optimization problem, is to seek an invertible and differentiable parametric function $\boldsymbol{y} = g(\boldsymbol{x}; \boldsymbol{\theta})$ that best maps the data onto a fixed target density $p_{\boldsymbol{y}}(\boldsymbol{y})$. The inverse image of this target density then provides a density model for the input space.

Many unsupervised learning methods may be interpreted in this context. As a simple example, consider principal component analysis (PCA; Jolliffe, 2002): we might fix $p_{\boldsymbol{y}}$ as a multivariate standard normal and think of PCA as either a linear whitening transformation, or as a density model $p_{\boldsymbol{x}}$ describing the data as a normal distribution with arbitrary covariance. Independent component analysis (ICA; Cardoso, 2003) can be cast in the same framework: In this case, the data $\boldsymbol{x}$ is modeled as a linear combination of independent heavy-tailed sources. We may fix g to be linear and $p_{\boldsymbol{y}} = \prod_i p_{y_i}$ to be a product of independent marginal densities of unknown form. Alternatively, we can apply nonparametric nonlinearities to the marginals of the linearly transformed data so as to Gaussianize them (i.e., histogram equalization). For this combined ICA-marginal-Gaussianization (ICA-MG) operation, $p_{\boldsymbol{y}}$ is again standard normal, and the transformation is a composition of a linear transform and marginal nonlinearities. Another model that aims for the same outcome is radial Gaussianiza-

*EPS is also affiliated with the Courant Institute of Mathematical Sciences at NYU; VL is also affiliated with the University of València, Spain.

tion (RG; Lyu & Simoncelli, 2009b; Sinz & Bethge, 2010), in which g is the composition of a linear transformation and a *radial* (i.e., operating on the vector length) Gaussianizing nonlinearity. The induced density model is the family of elliptically symmetric distributions.

The notion of optimizing a transformation so as to achieve desired statistical properties at the output is central to theories of efficient sensory coding in neurobiology (Barlow, 1961; Ruderman, 1994; Rieke et al., 1995; Bell & Sejnowski, 1997; Schwartz & Simoncelli, 2001), and also lends itself naturally to the design of cascaded representations such as deep neural networks. Specifically, variants of ICA-MG transformations have been applied in iterative cascades to learn densities (Friedman et al., 1984; Chen & Gopinath, 2000; Laparra et al., 2011). Each stage seeks a linear transformation that produces the "least Gaussian" marginal directions, and then Gaussianizes these using nonparametric scalar nonlinear transformations. In principle, this series of transformations can be shown to converge for any data density. However, the generality of these models is also their weakness: implementing the marginal nonlinearities in a non-parametric way makes the model prone to error and requires large amounts of data. In addition, since the nonlinearities operate only on marginals, convergence can be slow, requiring a lengthy sequence of transformations (i.e., a very deep network).

To address these shortcomings, we develop a joint transformation that is highly effective in Gaussianizing local patches of natural images. The transformation is a generalization of *divisive normalization*, a form of local gain control first introduced as a means of modeling nonlinear properties of cortical neurons (Heeger, 1992), in which linear responses are divided by pooled responses of their rectified neighbors. Variants of divisive normalization have been found to reduce dependencies when applied to natural images or sounds and to produce approximately Gaussian responses (Ruderman, 1994; Schwartz & Simoncelli, 2001; Malo & Laparra, 2010). Simple forms of divisive normalization have been shown to offer improvements in recognition performance of deep neural networks (Jarrett et al., 2009). But the Gaussianity of these representations has not been carefully optimized, and typical forms of normalization do not succeed in capturing all forms of dependency found in natural images (Lyu, 2010; Sinz & Bethge, 2013).

In this paper, we define a generalized divisive normalization (GDN) transform that includes parametric forms of both ICA-MG and RG as special cases. We solve for the parameters of the transform by optimizing an unsupervised learning objective for the non-Gaussianity of the transformed data. The transformation is continuous and differentiable, and we present an effective method of inverting it. We demonstrate that the resulting GDN transform provides a significantly better model for natural photographic images than either ICA-MG or RG. Specifically, we show that GDN provides a better fit to the pairwise statistics of local filter responses, that it generates more natural samples of image patches, and that it produces better results when used as a prior for image processing problems such as denoising. Finally, we show that a two-stage cascade of GDN transformations offers additional improvements in capturing image statistics, laying the groundwork for its use as a general tool for unsupervised learning of deep networks.

2 PARAMETRIC GAUSSIANIZATION

Given a parametric family of transformations $\boldsymbol{y} = g(\boldsymbol{x};\boldsymbol{\theta})$, we wish to select parameters $\boldsymbol{\theta}$ so as to transform the input vector $\boldsymbol{x}$ into a standard normal random vector (i.e., zero mean, identity covariance matrix). For a differentiable transformation, the input and output densities are related by:

$$p_{\boldsymbol{x}}(\boldsymbol{x}) = \left|\frac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{x}}\right| p_{\boldsymbol{y}}\big(g(\boldsymbol{x};\boldsymbol{\theta})\big), \tag{1}$$

where $|\cdot|$ denotes the absolute value of the matrix determinant. If $p_{\boldsymbol{y}}$ is the standard normal distribution (denoted $\mathcal{N}$), the shape of $p_{\boldsymbol{x}}$ is determined solely by the transformation. Thus, g *induces* a density model on $\boldsymbol{x}$, specified by the parameters $\boldsymbol{\theta}$.

Given $p_{\boldsymbol{x}}$, or data drawn from it, the density estimation problem can be solved by minimizing the Kullback–Leibler (KL) divergence between the transformed density and the standard normal, known as the *negentropy*:

$$J(p_{\boldsymbol{y}}) = \mathbb{E}_{\boldsymbol{y}}\Big(\log p_{\boldsymbol{y}}(\boldsymbol{y}) - \log \mathcal{N}(\boldsymbol{y})\Big) = \mathbb{E}_{\boldsymbol{x}}\Big(\log p_{\boldsymbol{x}}(\boldsymbol{x}) - \log\left|\tfrac{\partial g(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{x}}\right| - \log \mathcal{N}(g(\boldsymbol{x};\boldsymbol{\theta}))\Big), \tag{2}$$

where we have rewritten the standard definition (an expected value over $\boldsymbol{y}$) as an expectation over $\boldsymbol{x}$ (see appendix). Differentiating with respect to the parameter vector $\boldsymbol{\theta}$ yields:

$$\frac{\partial J(p_{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}} \left(-\sum_{ij} \left[\frac{\partial g(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{x}} \right]^{-T}_{ij} \frac{\partial^2 g_i(\boldsymbol{x}, \boldsymbol{\theta})}{\partial x_j \partial \boldsymbol{\theta}} + \sum_i g_i(\boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial g_i(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right), \tag{3}$$

where the expectation can be evaluated by summing over data samples, allowing the model to be fit using stochastic gradient descent. It can be shown that this optimization is equivalent to maximizing the log likelihood of the induced density model.

Note that, while optimization is feasible, measuring success in terms of the actual KL divergence in eq. (2) is difficult in practice, as it requires evaluating the entropy of $p_{\boldsymbol{x}}$. Instead, we can monitor the difference in negentropy between the input and output densities:

$$\Delta J \equiv J(p_{\boldsymbol{y}}) - J(p_{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{x}} \left(\tfrac{1}{2} \left\| \boldsymbol{y} \right\|_2^2 - \log \left| \tfrac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \right| - \tfrac{1}{2} \left\| \boldsymbol{x} \right\|_2^2 \right). \tag{4}$$

This quantity provides a measure of how much more Gaussian the data become as a result of the transformation $g(\boldsymbol{x}; \boldsymbol{\theta})$.

3 DIVISIVE NORMALIZATION TRANSFORMATIONS

Divisive normalization, a form of gain control in which responses are divided by pooled activity of neighbors, has become a standard model for describing the nonlinear properties of sensory neurons (Carandini & Heeger, 2012). A commonly used form for this transformation is:

$$y_i = \gamma \frac{x_i^\alpha}{\beta^\alpha + \sum_j x_j^\alpha},$$

where $\boldsymbol{\theta} = \{\alpha, \beta, \gamma\}$ are parameters. Loosely speaking, the transformation adjusts responses to lie within a desired operating range, while maintaining their relative values. A weighted form of normalization (with exponents fixed at $\alpha = 2$) was introduced in (Schwartz & Simoncelli, 2001), and shown to produce approximately Gaussian responses with greatly reduced dependencies. The weights were optimized over a collection of photographic images so as to maximize the likelihood of responses under a Gaussian model. Normalization has also been derived as an inference method for a Gaussian scale mixture (GSM) model for wavelet coefficients of natural images (Wainwright & Simoncelli, 2000). This model factorizes local groups of coefficients into a Gaussian vector and a positive-valued scalar. In a specific instance of the model, the optimal estimator for the Gaussian vector (after decorrelation) can be shown to be a modified form of divisive normalization that uses a weighted L_2-norm (Lyu & Simoncelli, 2008):

$$y_i = \frac{x_i}{\left(\beta^2 + \sum_j \gamma_j x_j^2\right)^{\frac{1}{2}}}.$$

However, the above instances of divisive normalization have only been shown to be effective when applied to spatially local groups of filter responses. In what follows, we introduce a more general form, with better Gaussianization capabilities that extend to to more distant responses, as well as those arising from distinct filters.

3.1 PROPOSED GENERALIZED DIVISIVE NORMALIZATION (GDN) TRANSFORM

We define a vector-valued parametric transformation as a composition of a linear transformation followed by a generalized form of divisive normalization:

$$\begin{aligned} \boldsymbol{y} = g(\boldsymbol{x}; \boldsymbol{\theta}) \qquad & \text{s.t.} \qquad y_i = \frac{z_i}{\left(\beta_i + \sum_j \gamma_{ij} |z_j|^{\alpha_{ij}}\right)^{\varepsilon_i}} \\ & \text{and} \qquad \boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}. \end{aligned} \tag{5}$$

The full parameter vector $\boldsymbol{\theta}$ consists of the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, as well as the matrices $\boldsymbol{H}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$, for a total of $2N + 3N^2$ parameters (where N is the dimensionality of the input space). We refer to this transformation as *generalized divisive normalization* (GDN), since it generalizes several previous models. Specifically:

- Choosing $\varepsilon_i \equiv 1$, $\alpha_{ij} \equiv 1$, and $\gamma_{ij} \equiv 1$ yields the classic form of the divisive normalization transformation (Carandini & Heeger, 2012), with exponents set to 1.
- Choosing $\boldsymbol{\gamma}$ to be diagonal eliminates the cross terms in the normalization pool, and the model is then a particular form of ICA-MG, or the first iteration of the Gaussianization algorithms described in Chen & Gopinath (2000) or Laparra et al. (2011): a linear "unmixing" transform, followed by a pointwise, Gaussianizing nonlinearity.
- Choosing $\alpha_{ij} \equiv 2$ and setting all elements of $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\gamma}$ identical, the transformation assumes a radial form:

$$\boldsymbol{y} = \frac{\boldsymbol{z}}{\left(\beta + \gamma \sum_j z_j^2\right)^{\varepsilon}} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}\, g_2\big(\|\boldsymbol{z}\|_2\big)$$

  where $g_2(r) = r/(\beta + \gamma r^2)^{\varepsilon}$ is a scalar-valued transformation on the radial component of $\boldsymbol{z}$, ensuring that the normalization operation preserves the vector direction of $\boldsymbol{z}$. If, in addition, $\boldsymbol{H}$ is a whitening transformation such as ZCA (Bell & Sejnowski, 1997), the overall transformation is a form of RG (Lyu & Simoncelli, 2009b).
- More generally, if we allow exponents $\alpha_{ij} \equiv p$, the induced distribution is an L_p-symmetric distribution, a family which has been shown to capture various statistical properties of natural images (Sinz & Bethge, 2010). The corresponding transformation on the L_p-radius is given by $g_p(r) = r/(\beta + \gamma r^p)^{\varepsilon}$.
- Another special case of interest arises when partitioning the space into distinct, spherically symmetric subspaces, with the kth subspace comprising the set of vector indices S_k. Choosing $\alpha_{ij} \equiv 2$, $\beta_i = \beta'_k$, $\varepsilon_i = \varepsilon'_k$, and $\gamma_{ij} = \gamma'_k$, all for $i, j \in S_k$ (and $\gamma_{ij} = 0$ if i or j is not in the same set), the nonlinear transformation can be written as

$$y_i = \frac{z_i}{\left(\beta'_k + \gamma'_k \sum_{j \in S_k} z_j^2\right)^{\varepsilon'_k}},$$

  where k is chosen such that $i \in S_k$. This is the Independent Subspace Analysis model (ISA; Hyvärinen & Hoyer, 2000), expressed as a Gaussianizing transformation.

The topographic ICA model (TICA; Hyvärinen et al., 2001) and the model presented in Köster & Hyvärinen (2010) are generalizations of ISA that are related to our model, but have more constrained nonlinearities. They are formulated directly as density models, which makes them difficult to normalize. For this reason, the authors must optimize approximated likelihood or use score matching (Hyvärinen, 2005) to fit these models.

3.2 Well-definedness and Invertibility

For the density function in eq. (1) to be well defined, we require the transformation in eq. (5) to be continuous and invertible. For the linear portion of the transformation, we need only ensure that the matrix $\boldsymbol{H}$ is non-singular. For the normalization portion, consider the partial derivatives:

$$\frac{\partial y_i}{\partial z_k} = \frac{\delta_{ik}}{\left(\beta_i + \sum_j \gamma_{ij} |z_j|^{\alpha_{ij}}\right)^{\varepsilon_i}} - \frac{\alpha_{ik}\gamma_{ik}\varepsilon_i z_i |z_k|^{\alpha_{ik}-1} \operatorname{sgn}(z_k)}{\left(\beta_i + \sum_j \gamma_{ij} |z_j|^{\alpha_{ij}}\right)^{\varepsilon_i + 1}} \tag{6}$$

To ensure continuity, we require all partial derivatives to be finite for all $\boldsymbol{z} \in \mathbb{R}^N$. More specifically, we require all exponents in eq. (6) to be non-negative, as well as the parenthesized expression in the denominator to be positive.

It can be shown that the normalization part of the transformation is invertible if the Jacobian matrix containing the partial derivatives in eq. (6) is positive definite everywhere (see appendix). In all practical cases, we observed this to be the case, but expressing this precisely as a condition on the parameters is difficult. A necessary (but generally not sufficient) condition for invertibility can be established as follows. First, note that, as the denominator is positive, each vector $\boldsymbol{z}$ is mapped to a vector $\boldsymbol{y}$ in the same orthant. The cardinal axes of $\boldsymbol{z}$ are mapped to themselves, and for this one-dimensional mapping to be continuous and invertible, it must be monotonic. Along the cardinal axes, the following bound holds:

$$|y_i| = \frac{|z_i|}{\left(\beta_i + \gamma_{ii}|z_i|^{\alpha_{ij}}\right)^{\varepsilon_i}} \leq \frac{|z_i|}{\gamma_{ii}^{\varepsilon_i}|z_i|^{\alpha_{ii}\varepsilon_i}} = \gamma_{ii}^{-\varepsilon_i}|z_i|^{1-\alpha_{ii}\varepsilon_i}.$$

For the magnitude of y_i to grow monotonically with $|z_i|$, the exponent $1 - \alpha_{ii}\varepsilon_i$ must be positive.

In summary, the constraints we enforce during the optimization are $\alpha_{ij} \geq 1$, $\beta_i > 0$, $\gamma_{ij} \geq 0$, and $0 \leq \varepsilon_i \leq \alpha_{ii}^{-1}$. We initialize the parameters such that $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}$ is positive definite everywhere (for example, by letting $\boldsymbol{\gamma}$ be diagonal, such that the Jacobian is diagonal, the transformation is separable, and the necessary constraint on the cardinal axes becomes sufficient).

Suppose that during the course of optimization, the matrix should cease to be positive definite. Following a continuous path, the matrix must then become singular at some point, because going from positive definite to indefinite or negative definite would require at least one of the eigenvalues to change signs. However, the optimization objective heavily penalizes singularity: The term $-\log|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}|$ in the objective (which separates into $-\log|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}}|$ and $-\log|\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}|$) grows out of bounds as the determinant of the Jacobian approaches zero. Therefore, given a sensible initialization and a sufficiently small step size, it is highly unlikely that the Jacobian should cease to be positive definite during the optimization, and we haven't observed this in practice.

Finally, we find that the GDN transformation can be efficiently inverted using a fixed point iteration:

$$z_i^{(0)} = \operatorname{sgn}(y_i)\left(\gamma_{ii}^{\varepsilon_i}|y_i|\right)^{\frac{1}{1-\alpha_{ii}\varepsilon_i}}$$
$$z_i^{(n+1)} = \left(\beta_i + \sum_j \gamma_{ij}|z_j^{(n)}|^{\alpha_{ij}}\right)^{\varepsilon_i} y_i.$$

Other iterative inverse solutions have been proposed for this purpose (Malo et al., 2006; Lyu & Simoncelli, 2008), but these only apply to special cases of the form in eq. (5).

4 EXPERIMENTS

The model was optimized to capture the distribution of image data using stochastic descent of the gradient expressed in eq. (3). We then conducted a series of experiments to assess the validity of the fitted model for natural images.

4.1 JOINT DENSITY OF PAIRS OF WAVELET COEFFICIENTS

We examined the pairwise statistics of model responses, both for our GDN model, as well as the ICA model and the RG model. First, we computed the responses of an oriented filter (specifically, we used a subband of the steerable pyramid (Simoncelli & Freeman, 1995)) to images taken from the van Hateren dataset (van Hateren & van der Schaaf, 1998) and extracted pairs of coefficients within subbands at different spatial offsets up to $d = 1024$. We then transformed these two-dimensional datasets using ICA, RG, and GDN. Figure 1 (modelled after figure 4 of Lyu & Simoncelli (2009b)) shows the mutual information in the transformed data (note that, in our case, mutual information is related to the negentropy by an additive constant.) For very small distances, a linear ICA transformation reduces some of the dependencies in the raw data. However, for larger distances, a linear transformation is not sufficient to eliminate the dependencies between the coefficients, and the mutual information of the ICA-transformed data is identical to that of the raw data. An elliptically symmetric model is good for modeling the density when the distance is small, and the RG transform reduces the mutual information to neglegible levels. However, the fit worsens as the distance increases. As described in (Lyu & Simoncelli, 2009b), RG can even lead to an increase in the mutual information relative to that of the raw data, as seen in the right hand side of the plot. The GDN transform, however, captures the dependencies at all separations, and consistently leaves a very small residual level of mutual information.

In figure 2, we compare histogram estimates of the joint wavelet coefficient densities against model-fitted densities for selected spatial offsets. The GDN fits are seen to account well for the shapes of the densities, particularly in the center of the plot, where the bulk of the data lie. Note that GDN is able to capture elliptically symmetric distributions just as well as distributions which are closer to being marginally independent, whereas RG and ICA each fail in one of these cases, respectively.

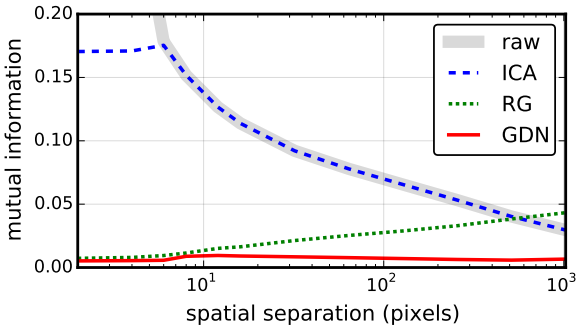


Figure 1: Mutual information in pairs of wavelet coefficients after various transformations, plotted as a function of the spatial separation between the coefficients.

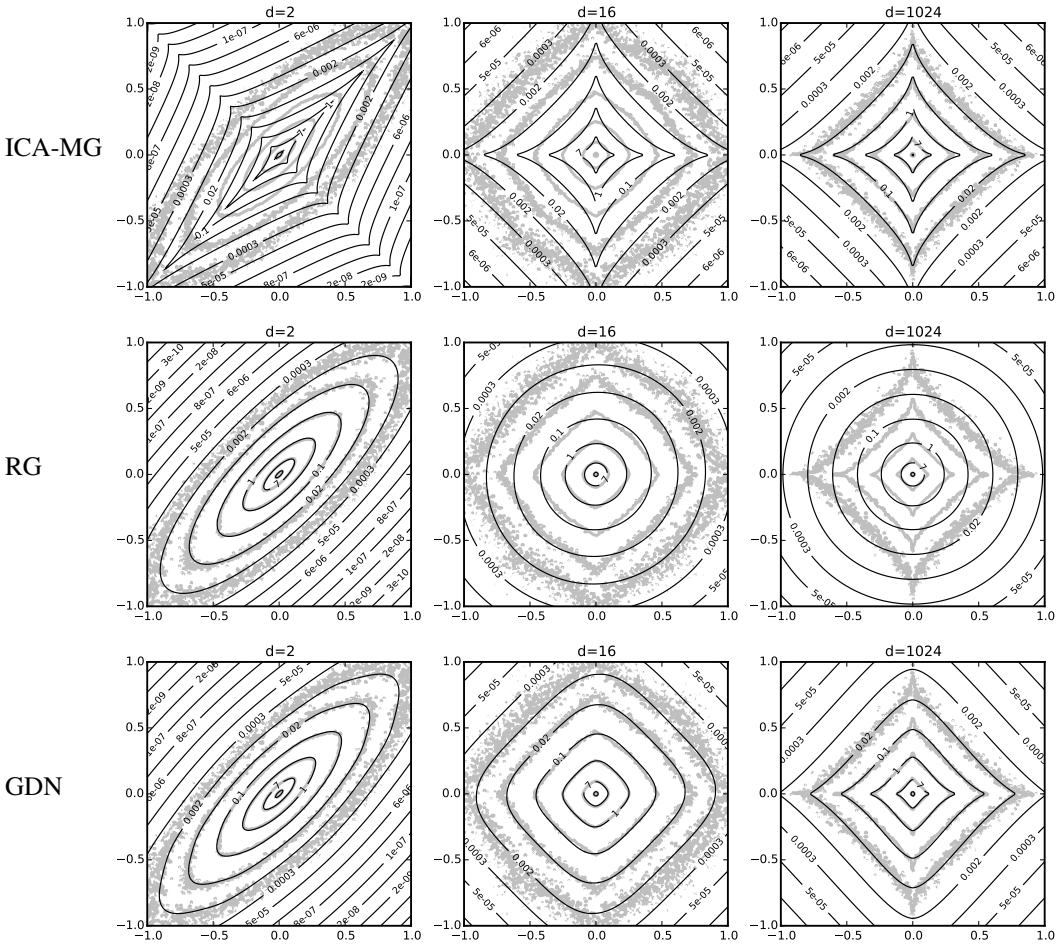


Figure 2: Contour plots of pairwise wavelet coefficient densities. Each row corresponds to a model arising from a different transformation (ICA-MG, RG, GDN). Each column corresponds to a pair of coefficients spatially separated by distance d (pixels). Gray: contour lines of histogram density estimate. Black: contour lines of densities induced by best-fitting transformations. As distance increases, the empirical density between the coefficients transitions from elliptical but correlated to separable. The RG density captures the former, and the ICA density captures the latter. Only the GDN density has sufficient flexibility to capture the full range of behaviors.

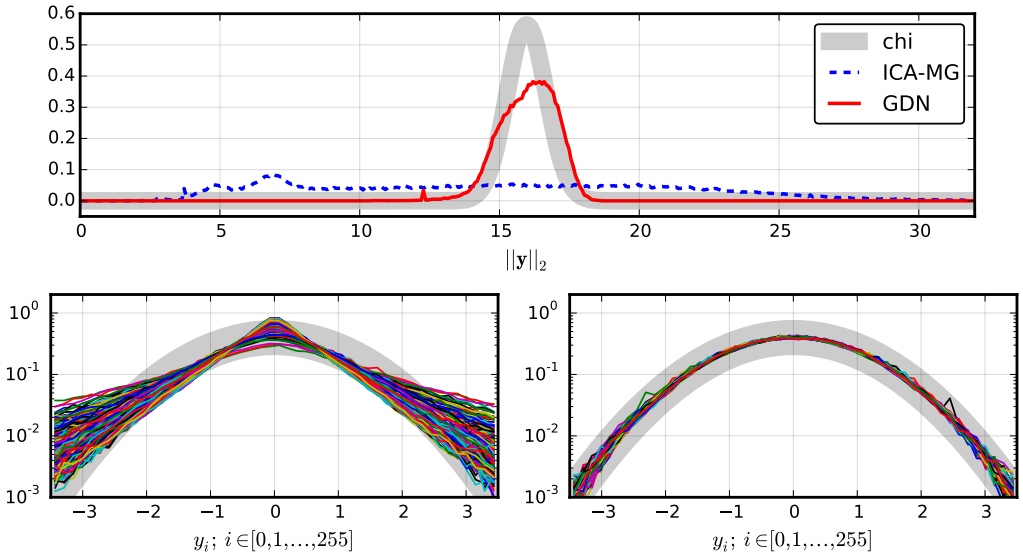


Figure 3: Histograms of transformed data. Top: Radial component for ICA-MG and GDN. Bottom left and right: Marginals for RG and GDN, respectively. Gray lines indicate the expected distributions (Chi for radial, and Gaussian for marginals).

4.2 Joint density over image patches

We also examined model behavior when applied to vectorized 16×16 blocks of pixels. We used the stochastic optimization algorithm ADAM to facilitate the optimization (Kingma & Ba, 2014) and somewhat reduced the complexity of the model by forcing $\boldsymbol{\alpha}$ to be constant along its columns (i.e., $\alpha_{ij} \equiv \alpha_j$). We also fitted versions of the model in which the normalization (denominator of eq. (5)) is constrained to marginal transformations (ICA-MG) or radial transformations (RG). For higher dimensional data, it is difficult to visualize the densities, so we use other measures to evaluate the effectiveness of the model:

Negentropy reduction. As an overall metric of model fit, we evaluated the negentropy difference ΔJ given in (4) on the full GDN model, as well as the marginal and radial models model (ICA-MG and RG, respectively). We find that ICA-MG and RG reduce negentropy by 2.04 and 2.11 nats per pixel, respectively, whereas GDN reduces it by 2.43 nats.

Marginal/radial distributions of transformed data. If the transformed data is multivariate standard normal, its marginals should be standard normal, as well, and the radial component should be Chi distributed with degree 256. Figure 3 shows these distributions, in comparison to those of ICA-MG and RG. As expected from (Lyu & Simoncelli, 2009b), RG fails to Gaussianize the marginals, and ICA-MG fails to transform the radial component into a Chi distribution. GDN comes close to achieving both goals.

Sampling. The density model induced by a transformation can also be visualized by examining samples drawn from a standard normal distribution that have been passed through the inverse transformation. Figure 4 compares sets of 25 image patches drawn from the GDN model, the ICA-MG model, and randomly selected from a database of images. GDN notably captures two features of natural images: First, a substantial fraction of the samples are constant or nearly so (as in the natural images, which include patches of sky or untextured surfaces). Second, in the cases with more activity, the samples contain sparse "organic" structures (although less so than those drawn from the natural images). In comparison, the samples from the ICA-MG model are more jumbled, and filled with random mixtures of oriented elements.

Denoising. The negentropy provides a particular metric for assessing the quality of our results, but it need not agree with other measures (Theis et al., 2015). Another test of a probability model comes from using it as a prior in a Bayesian inference problem. The most basic example is that of removing additive Gaussian noise. For GDN, we use the empirical Bayes solution of Miyasawa (1961), which

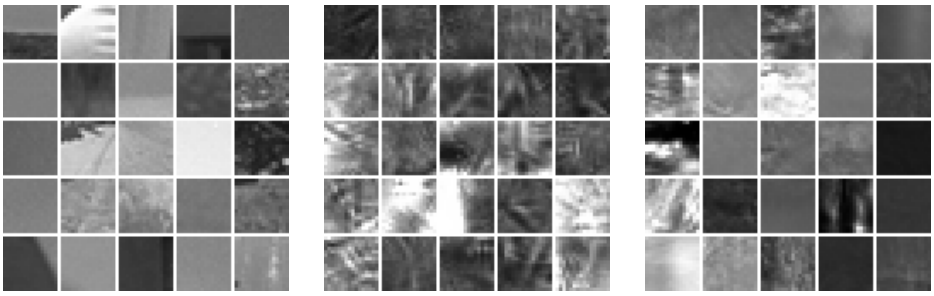


Figure 4: Sample image patches. From left to right, 25 samples drawn from: the image training set; the ICA-MG model; the GDN model.

expresses the least-squares optimal solution *directly* as a function of the distribution of the noisy data:

$$\hat{\boldsymbol{x}} = \tilde{\boldsymbol{x}} + \sigma^2 \nabla \log p_{\tilde{\boldsymbol{x}}}(\tilde{\boldsymbol{x}}), \tag{7}$$

where $\tilde{\boldsymbol{x}}$ is the noisy observation, $p_{\tilde{\boldsymbol{x}}}$ is the density of the noisy data, σ^2 is the noise variance, and $\hat{\boldsymbol{x}}$ is the optimal estimate. Note that, counterintuitively, this expression does not refer directly to the prior density, but it is nevertheless exactly equivalent to the Bayesian least squares solution (Raphan & Simoncelli, 2011). Although the GDN model was developed for modeling the distribution of clean image data, we use it here to estimate the distribution of the *noisy* image data. We find that, since the noisy density is a Gaussian-smoothed version of the original density, the model fits the data well (results not shown).

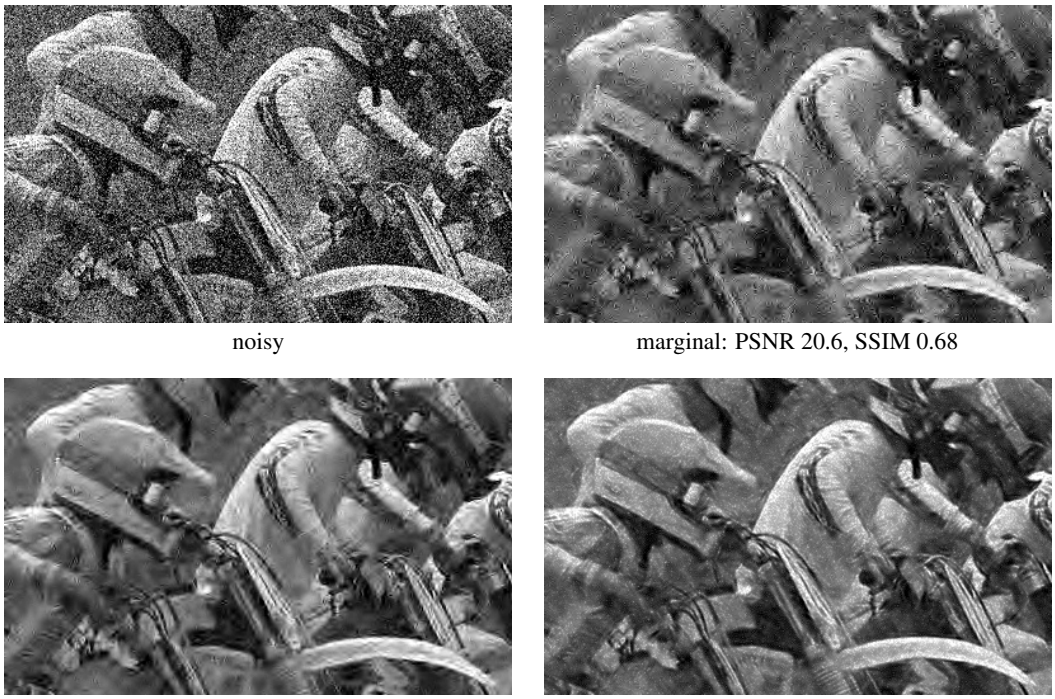
For comparison, we implemented two denoising methods that operate on orthogonal wavelet coefficients, one assuming a marginal model (Figueiredo & Nowak, 2001), and the other an elliptically symmetric Gaussian scale mixture (GSM) model (Portilla et al., 2003). Since the GDN model is applied to 16×16 patches of pixels and is restricted to a complete (i.e., square matrix) linear transformation, we restrict the wavelet transform employed in the other two models to be orthogonal, and to include three scales. We also report numerical scores: the peak signal to noise ratio (PSNR), and the structural similarity index (SSIM; Wang et al., 2004) which provides a measure of perceptual quality. Fig. 5 shows the denoising results. Both marginal and spherical models produce results with strong artifacts resembling the basis functions of the respective linear transform. The GDN solution has artifacts that are less perceptually noticeable, while at the same time leaving a larger amount of background noise.

Average model likelihood. To further assess how our model compares to existing work, we trained the model on image patches of 8×8 pixels from the BSDS300 dataset which had the patch mean removed (see Theis & Bethge, 2015, left column of table 1). We followed the same evaluation procedures as in that reference, and measured a cross-validated average log likelihood of 126.8 nats for ICA-MG and 151.5 nats for GDN, similar to the values reported there for the RIDE model, but worse than the best-performing MCGSM and RNADE models. On the other hand, GDN achieves a model likelihood of 3.47 bits/pixel for image patches of 8×8 pixels without mean removal, which is essentially equal to the best reported performance in the middle column of table 1 (ibid.).[1]

4.3 TWO-STAGE CASCADED MODEL

In the previous section, we show that the GDN transformation works well on local patches of images. However, this cannot capture statistical dependencies over larger spatial distances (i.e., across adjacent patches). One way of achieving this is to cascade Gaussianizing transformations (Chen & Gopinath, 2000; Laparra et al., 2011). In previous implementations of such cascades, each stage of the transformation consists of a linear transformation (to rotate the previous responses, exposing additional non-Gaussian directions) and a Gaussianizing nonlinear transformation applied to the marginals. We have implemented a cascade based on GDN that benefits from two innovations. First,

[1]Note, however, that the middle column in table 1 of Theis & Bethge (2015) was generated under the assumption that the patch mean is statistically independent from the rest of the data, which artificially impedes the performance of the reported models.



noisy | marginal: PSNR 20.6, SSIM 0.68

GSM: PSNR 22.4, SSIM 0.75 | GDN: PSNR 22.6, SSIM 0.78

Figure 5: Bayesian least squares denoising using different prior models. Top: noise-corrupted original; denoised with marginal model in an orthonormal wavelet decomposition. Bottom: denoised with GSM model in an orthonormal wavelet decomposition; denoised with GDN-induced density model. Below each image, errors against the original image are quantified with PSNR in dB, and the perceptual SSIM metric (for both measures, bigger is better).

by jointly Gaussianizing groups of coefficients (rather than transforming each one independently), GDN achieves a much more significant reduction in negentropy than MG (see Figure 1), thereby reducing the total number of stages that would be needed to fully Gaussianize the data. Second, we replace the ICA rotations with convolutional ICA (CICA; Ballé & Simoncelli, 2014). This is a better solution than either partitioning the image into non-overlapping blocks (which produces artifacts at block boundaries) or simply increasing the size of the transformation, which would require a much larger number of parameters for the linear transform than a convolutional solution (which allows "weight sharing" (LeCun et al., 1990)).

The central question that determines effectiveness of a multi-layer model based on the above ingredients is whether the parametric form of the normalization is suitable for Gaussianizing the data after it has been transformed by previous layers. According to our preliminary results, this seems to be the case. We constructed a two-stage model, trained greedily layer-by-layer, consisting of the transformations CICA–GDN–CICA–GDN. The first CICA instance implements a complete, invertible linear transformation with a set of 256 convolutional filters of support 48×48, with each filter response subsampled by a factor of 16 (both horizontally and vertically). The output thus consists of 256 reduced-resolution feature maps. The first GDN operation then acts on the 256-vectors of responses at a given spatial location across all maps. Thus, the responses of the first CICA–GDN stage are Gaussianized across maps, but not across spatial locations. The second-stage CICA instance is applied to vectors of first-stage responses across all maps within a 9×9 spatial neighborhood – thus seeking new non-Gaussian directions across spatial locations *and* across maps. Histogram estimates of the marginals of these directions are shown in figure 6. The distributions are qualitatively similar to those found for the first stage CICA operating on image pixels, although their heavy-tailedness is less pronounced. The figure also shows histograms of the second-stage GDN marginals, indicating that the new directions have been effectively Gaussianized.

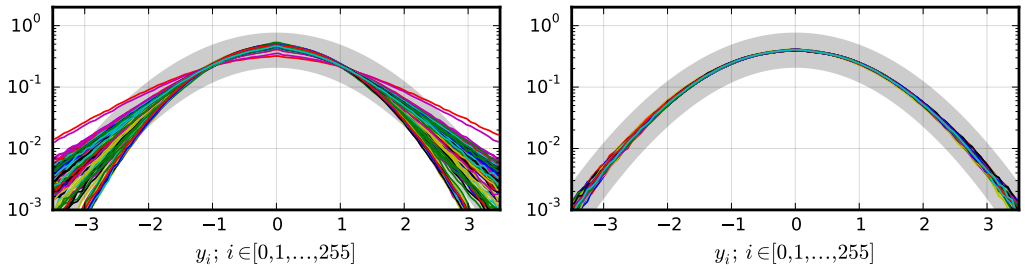


Figure 6: Marginal histograms of two-stage model responses. Left: marginals of 256 features, obtained by performing linear CICA on first stage GDN responses. Right: marginals after the second stage GDN. The thick gray line corresponds to a Gaussian distribution.

5 Conclusion

We have introduced a new probability model for natural images, implicitly defined in terms of an invertible nonlinear transformation that is optimized so as to Gaussianize the data. This transformation is formed as the composition of a linear operation and a generalized form of divisive normalization, a local gain control operation commonly used to model response properties of sensory neurons. We developed an efficient algorithm for fitting the parameters of this transformation, minimizing the KL divergence of the distribution of transformed data against a Gaussian target. The resulting density model is not closed-form (because we need to evaluate the determinant of the Jacobian matrix), but it does allow direct computation of probability/likelihood, and is readily used for sampling and inference.

Our parametric transformation includes previous variants of divisive normalization as special cases, and the induced density model generalizes forms of ICA/ISA and elliptically symmetric models. We show that the additional complexity of our generalized normalization transform allows a significant increase in performance, in terms of Gaussianization, denoising, and sampling. In addition, we found that the fitted parameters of our model (in particular, the interactions governed by γ) do not resemble any of these special cases (not shown), and we expect that their detailed structure will be useful in elucidating novel statistical properties of images. It will also be important to compare this induced density model more thoroughly to other model forms that have been proposed in the literature (e.g., finite mixtures of Gaussians or GSMs (Guerrero-Colón et al., 2008; Lyu & Simoncelli, 2009a; Zoran & Weiss, 2012; Theis et al., 2012), and sparse factorization (Culpepper et al., 2011)).

Our method arises as a natural combination of concepts drawn from two different research endeavors. The first aims to explain the architecture and functional properties of biological sensory systems as arising from principles of coding efficiency (Barlow, 1961; Rieke et al., 1995; Bell & Sejnowski, 1997; Schwartz & Simoncelli, 2001). A common theme in these studies is the idea that the hierarchical organization of the system acts to transform the raw sensory inputs into more compact, and statistically factorized, representations. Divisive normalization has been proposed as a transformation that contributes to this process. The new form we propose here is highly effective: the transformed data are significantly closer to Gaussian than data transformed by either marginal or radial Gaussianization, and the induced density is thus a more factorized representation of the data.

The second endeavor arises from the statistics literature on projection pursuit, and the use of Gaussianization in problems of density estimation (Friedman et al., 1984). More recent examples include marginal and radial transformations (Chen & Gopinath, 2000; Lyu & Simoncelli, 2009b; Laparra et al., 2011), as well as rectified-linear transformations (Dinh et al., 2014). Our preliminary experiments indicate that the fusion of a generalized variant of the normalization computation with the iterated Gaussianization architecture is feasible, both in terms of optimization and statistical validity. We believe this architecture offers a promising platform for unsupervised learning of probabilistic structures from data, and are currently investigating techniques to jointly optimize the stages of more deeply stacked models.

6 APPENDIX

6.1 NEGENTROPY

To see that the negentropy J of the transformed data $\boldsymbol{y}$ can be written as an expectation over the original data, consider a change of variables:

$$\begin{aligned} J(p_{\boldsymbol{y}}) &= \mathbb{E}_{\boldsymbol{y}}\Big(\log p_{\boldsymbol{y}}(\boldsymbol{y}) - \log \mathcal{N}(\boldsymbol{y})\Big) \\ &= \int p_{\boldsymbol{y}}(\boldsymbol{y})\Big(\log p_{\boldsymbol{y}}(\boldsymbol{y}) - \log \mathcal{N}(\boldsymbol{y})\Big)\,\mathrm{d}\boldsymbol{y} \\ &= \int p_{\boldsymbol{x}}(\boldsymbol{x})\Big|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|^{-1}\left(\log\Big(p_{\boldsymbol{x}}(\boldsymbol{x})\Big|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|^{-1}\Big) - \log \mathcal{N}(\boldsymbol{y})\right)\Big|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|\,\mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_{\boldsymbol{x}}\left(\log p_{\boldsymbol{x}}(\boldsymbol{x}) - \log\Big|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big| - \log \mathcal{N}(\boldsymbol{y})\right) \end{aligned}$$

6.2 INVERTIBILITY

Here, we show that a transformation $g : \boldsymbol{x} \mapsto \boldsymbol{y}$ is invertible if it is continuous and its Jacobian $g' : \boldsymbol{x} \mapsto \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$ positive definite everywhere. First note that g is invertible if and only if any two nonidentical inputs $\boldsymbol{x}_a$, $\boldsymbol{x}_b$ are mapped to nonidentical outputs $\boldsymbol{y}_a$, $\boldsymbol{y}_b$, and vice versa:

$$\forall \boldsymbol{x}_a, \boldsymbol{x}_b : \quad \boldsymbol{x}_a \neq \boldsymbol{x}_b \;\Leftrightarrow\; \boldsymbol{y}_a \neq \boldsymbol{y}_b.$$

Since g is a function, the left-hand inequality follows trivially from the right. To see the converse direction, we can write the inequality of the two right-hand side vectors as

$$\exists \boldsymbol{u} : \boldsymbol{u}^\top \Delta \boldsymbol{y} \neq 0,$$

where $\Delta \boldsymbol{y}$ is their difference. Second, we can compute $\Delta \boldsymbol{y}$ by integrating the Jacobian along a straight line L_{ab} between $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$:

$$\Delta \boldsymbol{y} = \int_{L_{ab}} g'(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Writing the integral as a Riemann limit, invertibility can be stated as:

$$\boldsymbol{x}_a \neq \boldsymbol{x}_b \;\Leftrightarrow\; \exists \boldsymbol{u} : \boldsymbol{u}^\top \Delta \boldsymbol{y} = \lim_{T\to\infty} \sum_{t=0}^{T-1} \boldsymbol{u}^\top g'\Big(\boldsymbol{x}_a + t\,\frac{\Delta \boldsymbol{x}}{T}\Big)\frac{\Delta \boldsymbol{x}}{T} \neq 0.$$

If $\Delta \boldsymbol{x} \neq \boldsymbol{0}$ (the left-hand inequality is true) and g' is positive definite everywhere, all terms in the sum can be made positive by choosing $\boldsymbol{u} = \frac{\Delta \boldsymbol{x}}{T}$. Hence, for g' positive definite, the right-hand inequality follows from the left.

6.3 PREPROCESSING

We performed two preprocessing steps on the van Hateren dataset before fitting our model: removal of images with saturation artifacts and passing the intensity values through a nonlinearity.

To remove heavily saturated images, we computed a histogram of intensity values for each of the images. If more than 0.1% of the pixel values were contained in the highest-valued histogram bin, we removed the image from the dataset.

We passed the remaining 2904 images through a pointwise nonlinearity. In the literature, a logarithm is most commonly used, although there is no particularly convincing reason for using precisely this function. Since the GDN densities are zero-mean by definition, mean removal is necessary to fit the density. Instead of the log, we used the inverse of a generalized logistic function, which is very similar, but can be chosen to marginally Gaussianize the intensity values, which is in line with our objective and also removes the mean.

ACKNOWLEDGMENTS

JB and EPS were supported by the Howard Hughes Medical Institute. VL was supported by the APOSTD/2014/095 Generalitat Valenciana grant (Spain).

REFERENCES

Ballé, Johannes and Simoncelli, Eero P. Learning sparse filterbank transforms with convolutional ICA. In *2014 IEEE International Conference on Image Processing (ICIP)*, 2014. doi: 10.1109/ICIP.2014.7025815.

Barlow, Horace B. Possible principles underlying the transformations of sensory messages. In *Sensory Communication*, pp. 217–234. M.I.T. Press, 1961. ISBN 978-0-262-51842-0.

Bell, Anthony J. and Sejnowski, Terrence J. The independent components of natural scenes are edge filters. *Vision Research*, 37(23), 1997. doi: 10.1016/S0042-6989(97)00121-1.

Carandini, Matteo and Heeger, David J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13, January 2012. doi: 10.1038/nrn3136.

Cardoso, Jean-François. Dependence, correlation and Gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003. ISSN 1533-7928.

Chen, Scott Saobing and Gopinath, Ramesh A. Gaussianization. In *Advances in Neural Information Processing Systems 13*, pp. 423–429, 2000.

Culpepper, B. J., Sohl-Dickstein, J., and Olshausen, Bruno. Building a better probabilistic model of images by factorization. In *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011. doi: 10.1109/ICCV.2011.6126473.

Dinh, Laurent, Krueger, David, and Bengio, Yoshua. NICE: Non-linear independent components estimation. *arXiv e-prints*, 2014. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Figueiredo, M. A. T. and Nowak, R. D. Wavelet-based image estimation: an empirical bayes approach using Jeffrey's noninformative prior. *IEEE Transactions on Image Processing*, 10(9), September 2001. doi: 10.1109/83.941856.

Friedman, Jerome H., Stuetzle, Werner, and Schroeder, Anne. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79(387), 1984. doi: 10.1080/01621459.1984.10478086.

Guerrero-Colón, J. A., Simoncelli, Eero P., and Portilla, Javier. Image denoising using mixtures of Gaussian scale mixtures. In *15th IEEE International Conference on Image Processing, 2008*, 2008. doi: 10.1109/ICIP.2008.4711817.

Heeger, David J. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2), 1992. doi: 10.1017/S0952523800009640.

Hyvärinen, Aapo. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005. ISSN 1533-7928.

Hyvärinen, Aapo and Hoyer, Patrik. Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7), 2000. doi: 10.1162/089976600300015312.

Hyvärinen, Aapo, Hoyer, Patrik, and Inki, Mika. Topographic independent component analysis. *Neural Computation*, 13(7), 2001. doi: 10.1162/089976601750264992.

Jarrett, Kevin, Kavukcuoglu, Koray, Ranzato, Marc'Aurelio, and LeCun, Yann. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, 2009. doi: 10.1109/ICCV.2009.5459469.

Jolliffe, I. T. *Principal Component Analysis*. Springer, 2 edition, 2002. ISBN 978-0-387-95442-4.

Kingma, Diederik P. and Ba, Jimmy Lei. Adam: A method for stochastic optimization. *arXiv e-prints*, 2014. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Köster, Urs and Hyvärinen, Aapo. A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, 22(9), 2010. doi: 10.1162/NECO_a_00010.

Laparra, Valero, Camps-Valls, Gustavo, and Malo, Jesús. Iterative Gaussianization: From ICA to random rotations. *IEEE Transactions on Neural Networks*, 22(4), April 2011. doi: 10.1109/TNN.2011.2106511.

LeCun, Yann, Matan, O., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., and Baird, H. S. Handwritten zip code recognition with multilayer networks. In *Proceedings, 10th International Conference on Pattern Recognition*, volume 2, 1990. doi: 10.1109/ICPR.1990.119325.

Lyu, Siwei. Divisive normalization: Justification and effectiveness as efficient coding transform. In *Advances in Neural Information Processing Systems 23*, pp. 1522–1530, 2010.

Lyu, Siwei and Simoncelli, Eero P. Nonlinear image representation using divisive normalization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008. doi: 10.1109/CVPR.2008.4587821.

Lyu, Siwei and Simoncelli, Eero P. Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4), April 2009a. doi: 10.1109/TPAMI.2008.107.

Lyu, Siwei and Simoncelli, Eero P. Nonlinear extraction of independent components of natural images using radial Gaussianization. *Neural Computation*, 21(6), 2009b. doi: 10.1162/neco.2009.04-08-773.

Malo, Jesús and Laparra, Valero. Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images. *Neural Computation*, 22(12), 2010. doi: 10.1162/NECO_a_00046.

Malo, Jesús, Epifanio, I., Navarro, R., and Simoncelli, Eero P. Non-linear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, 15(1), January 2006. doi: 10.1109/TIP.2005.860325.

Miyasawa, K. An empirical bayes estimator of the mean of a normal population. *Bulletin de l'Institut international de Statistique*, 38:181–188, 1961.

Portilla, Javier, Strela, Vasily, Wainwright, Martin J., and Simoncelli, Eero P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image Processing*, 12 (11), November 2003. doi: 10.1109/TIP.2003.818640.

Raphan, Martin and Simoncelli, Eero P. Least squares estimation without priors or supervision. *Neural Computation*, 23(2), February 2011. doi: 10.1162/NECO_a_00076.

Rieke, F., Bodnar, D. A., and Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London B: Biological Sciences*, 262(1365), 1995. doi: 10.1098/rspb.1995.0204.

Ruderman, Daniel L. The statistics of natural images. *Network: Computation in Neural Systems*, 5, 1994. doi: 10.1088/0954-898X_5_4_006.

Schwartz, Odelia and Simoncelli, Eero P. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8), August 2001. doi: 10.1038/90526.

Simoncelli, Eero P. and Freeman, William T. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *1995 IEEE International Conference on Image Processing (ICIP)*, volume 3, 1995. doi: 10.1109/ICIP.1995.537667.

Sinz, Fabian and Bethge, Matthias. L_p-nested symmetric distributions. *Journal of Machine Learning Research*, 11:3409–3451, 2010. ISSN 1533-7928.

Sinz, Fabian and Bethge, Matthias. What is the limit of redundancy reduction with divisive normalization? *Neural Computation*, 25(11), 2013. doi: 10.1162/NECO_a_00505.

Theis, Lucas and Bethge, Matthias. Generative image modeling using spatial LSTMs. In *Advances in Neural Information Processing Systems 28*, pp. 1918–1926, 2015.

Theis, Lucas, Hosseini, Reshad, and Bethge, Matthias. Mixtures of conditional Gaussian scale mixtures applied to multiscale image representations. *PLoS one*, 7(7), 2012. doi: 10.1371/journal.pone.0039857.

Theis, Lucas, van den Oord, Aäron, and Bethge, Matthias. A note on the evaluation of generative models. *arXiv e-prints*, 2015. Under review as a conference paper at the 4th International Conference for Learning Representations, San Juan, 2016.

van Hateren, J. H. and van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1394), 1998. doi: 10.1098/rspb.1998.0303.

Wainwright, Martin J. and Simoncelli, Eero P. Scale mixtures of Gaussians and the statistics of natural images. In *Advances in Neural Information Processing Systems 12*, pp. 855–861, 2000.

Wang, Zhou, Bovik, Alan Conrad, Sheikh, H. R., and Simoncelli, Eero P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), April 2004. doi: 10.1109/TIP.2003.819861.

Zoran, Daniel and Weiss, Yair. Natural images, Gaussian mixtures and dead leaves. In *Advances in Neural Information Processing Systems 25*, pp. 1736–1744, 2012.