LETTER Communicated by Odelia Schwartz

Dependency Reduction with Divisive Normalization: Justification and Effectiveness

Siwei Lyu
lsw@cs.albany.edu
Computer Science Department, University at Albany, State University of New York, Albany, NY 12222, U.S.A.

Efficient coding transforms that reduce or remove statistical dependencies in natural sensory signals are important for both biology and engineering. In recent years, divisive normalization (DN) has been advocated as a simple and effective nonlinear efficient coding transform. In this work, we first elaborate on the theoretical justification for DN as an efficient coding transform. Specifically, we use the multivariate *t* model to represent several important statistical properties of natural sensory signals and show that DN approximates the optimal transforms that eliminate statistical dependencies in the multivariate *t* model. Second, we show that several forms of DN used in the literature are equivalent in their effects as efficient coding transforms. Third, we provide a quantitative evaluation of the overall dependency reduction performance of DN for both the multivariate *t* models and natural sensory signals. Finally, we find that statistical dependencies in the multivariate *t* model and natural sensory signals are increased by the DN transform with low-input dimensions. This implies that for DN to be an effective efficient coding transform, it has to pool over a sufficiently large number of inputs.

1 Introduction

A central principle in the study of biological sensory systems is that they are adapted to match the statistical properties of the sensory signals in the natural environments to which they are exposed (Attneave, 1954). The efficient coding hypothesis (Barlow, 1961; Atick, 1992) further suggests that a sensory system might be understood as a transform that reduces redundancies in the input stimuli. Such efficient coding transforms are of importance in both biology and engineering: the reduced redundancies in the neural responses facilitate efficient representations of sensory input for ecologically relevant tasks such as novelty detection and associative learning (Barlow, 2001). In addition, with the reduced dependencies, sensory signals can be more efficiently stored, transmitted, and processed.

Starting with the use of second-order decorrelation methods in explaining functional roles of photoreceptors and retinal ganglion cells (Atick &

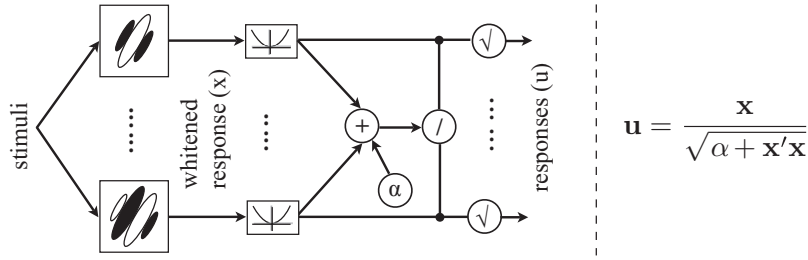

Figure 1: A schematic illustration and the definition of the DN transform.

Redlich, 1992; Atick, Li, & Redlich, 1992; Ruderman, Cronin, & Chiao, 1998), studies in linear efficient coding transforms for natural sensory signals have led to fruitful developments that culminate in the independent component analysis (ICA) methodology (Olshausen & Field, 1996; van der Schaaf & van Hateren, 1996; Bell & Sejnowski, 1997; Lewicki, 2002). These efforts were widely lauded as a confirmation of the efficient coding hypothesis in the study of biological perception, as the obtained ICA basis functions closely resemble the receptive fields of neurons in various cortex areas. In spite of these successes, early studies suggested that linear transforms may not be optimal for reducing dependencies in natural sensory signals (Zetzsche & Barth, 1990; Baddeley, 1996; Zetzsche & Krieger, 1999), which were further confirmed with observations of strong residual statistical dependencies after ICA-like linear transforms (Wegmann & Zetzsche, 1990; Simoncelli & Buccigrossi, 1997) and quantitative evaluations that ICA achieves only a marginal improvement over principal component analysis (PCA) in reducing statistical dependencies in natural images (Bethge, 2006). Indeed, there are statistical dependencies in natural sensory signals that linear transforms cannot reduce (Lyu & Simoncelli, 2009b; Eichhorn, Sinz, & Bethge, 2009). This motivates the search for effective nonlinear efficient coding transforms for natural sensory signals.

Divisive normalization (DN) is a simple nonlinear efficient coding transform that recently has been widely studied (Schwartz & Simoncelli, 2001a; Valerio & Navarro, 2003a, 2003b; Malo & Laparra, 2010; Lyu, 2010). The standard form of DN transform we adopt in this work is illustrated schematically in Figure 1. Here, $\mathbf{x} = (x_1, \ldots, x_d)'$ is a vector describing the responses of input stimuli projected onto a set of front-end linear basis functions. These linear basis functions remove first- and second-order local statistical dependencies and whiten the inputs so that they all have the same weights when squared and pooled with a semisaturation constant α. The square root of the pooling is divided from the response of each individual linear basis function to obtain the final output of the DN transform, $\mathbf{u} = (u_1, \cdots, u_d)'$.

In this work, we first elaborate on the theoretical justification of DN as an efficient coding transform. Specifically, we use the multivariate t model to represent several important statistical properties of natural sensory signals and show that DN approximates the optimal transforms that eliminate statistical dependencies in the multivariate t model. Second, using the multi-information as a quantitative measure of statistical dependency, we show that several different forms of DN are equivalent in terms of dependency reduction. Third, we provide a quantitative evaluation of the overall dependency reduction performance of DN for both the multivariate t models and natural sensory signals. Finally, we find that statistical dependencies in the multivariate t model and natural sensory signals are increased by the DN transform with low input dimensions. This implies that for DN to be an effective and efficient coding transform, it has to pool over a sufficiently large number of inputs.

The rest of this article is organized as follows. After reviewing relevant previous works in section 2, we describe in section 3 some basic statistical properties of natural sensory signals and demonstrate how these properties can be captured with the multivariate t model. In section 4, we show that DN transform approximates the optimal efficient coding transforms for the multivariate t model. Sections 5 and 6 report the experimental evaluation of the effectiveness of the DN transform as an efficient coding transform for the multivariate t models and natural sensory signal data. Section 7 concludes with discussion and future work. To make the description continuous, we defer all formal proofs to the appendixes. (A preliminary version of this work has been presented in Lyu, 2010.)

2 Related Work

In biology, DN is a popular model for many nonlinear behaviors of neural responses that cannot be well described with the classical linear-nonlinear Poisson model (Chichilnisky, 2001; Pillow & Simoncelli, 2006). Such nonlinearities can be found in the auditory (Schwartz & Simoncelli, 2001b) and the olfactory pathways (Olsen, Bhandawat, & Wilson, 2010), as well as various stages of the visual pathway, including the retina (Shapley & Enroth-Cugell, 1984; Solomon, Lee, & Sun, 2006), the lateral geniculate nucleus (Mante, Bonin, & Carandini, 2008), the primary visual cortex (Heeger, 1992; Rust, Schwartz, Movshon, & Simoncelli, 2005), and other extrastriate cortex, such as area MT (Simoncelli & Heeger, 1998) and area IT (Zoccolan, Cox, & DiCarlo, 2005). In low-level visual perception, DN has been related to various functional roles, including dynamic gain control (Shapley & Enroth-Cugell, 1984), decoding activities of neuronal populations (Deneve, Pouget, & Latham, 1999; Ringach, 2010), neural adaptation (Wainwright, Schwartz, & Simoncelli, 2002), and visual saliency (Gao & Vasconcelos, 2009). It has also been used to account for high-level perceptual phenomena such as masking (Foley, 1994; Watson & Solomon, 1997) and attention (Lee

& Maunsell, 2009; Reynolds & Heeger, 2009). DN is believed to be implementable with the cortical neurons (Carandini & Heeger, 1994; Carandini, Heeger, & Senn, 2002), though the specific neural mechanism for such an implementation is still under debate (see Holt & Koch, 1997).

Because of the prevalence of normalization-type behaviors in biological sensory systems, DN has become an integral component in theoretical models describing information encoding of cortical neurons (Ringach, 2010). In addition, in engineering fields such as image processing and computer vision, nonlinear image representations based on DN have been applied to image compression (Malo, Epifanio, Navarro, & Simoncelli, 2006), contrast enhancement (Lyu & Simoncelli, 2008), image quality metrics (Li & Wang, 2008; Laparra, Muñoz-Mari, & Malo, 2010), and object recognition (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009), all showing significant improvements in performance over linear representations.

In the context of efficient coding theory, the study in Brady and Field (2000) suggests that DN maximizes the entropy of each output component to better use the channel capacity. Based on empirical observations, the seminal work of Schwartz and Simoncelli (2001a) proposes that DN is a nonlinear efficient coding transform in biological perception that reduces statistical dependencies in the input natural sensory signals. Subsequently, this hypothesis was tested in Valerio and Navarro (2003a, 2003b). However, experiments in these works examined only pairwise dependencies with mutual information estimated from histograms, a process prone to biases due to the data binning procedure (Paninski, 2003). The more recent work of Malo and Laparra (2010) uses a form of DN whose parameters are obtained from psychophysical experiments. While providing an interesting alternative perspective, this is only an indirect account for DN as an efficient coding transform for natural sensory signals.

Recently a general methodology known as radial gaussianization (RG) has been shown to provide efficient coding transforms for sources with elliptical symmetric densities that capture local statistical dependencies of natural sensory signals (Lyu & Simoncelli, 2009b; Sinz & Bethge, 2009). In these studies, it has been shown that the transformation obtained by RG can be closely approximated by DN. On the other hand, in spite of better performance in dependency reduction, the nonlinear transform obtained from RG has not been fitted to data in biological sensory pathways.

3 Statistical Properties of Natural Sensory Signals and Multivariate *t* Model

Sensory signals in natural environments are highly structured and nonrandom. These regularities exhibit statistical properties that distinguish them from the rest of the ensemble of all possible signals. Particularly, in the bandpass-filtered domains constructed from various linear transforms consisting of basis functions with localized support in space, frequency,

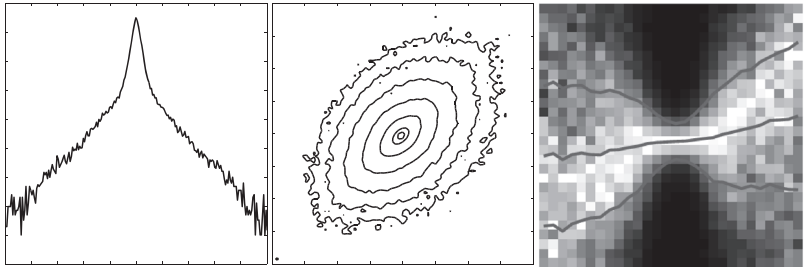

Figure 2: Statistical properties of natural images in a bandpass-filtered domain. (Left) Log marginal distribution. (Middle) Contour plot of the joint distribution of two responses from filter functions whose centers are separated by 1 pixel in space. (Right) Each column of the image corresponds to a conditional density $p(x_1|x_2)$ of different x_2 values, where x_1 and x_2 are the coordinates of the joint shown in the middle panel. The three curves correspond to $\mathrm{E}(x_1|x_2)$ (center) and $\mathrm{E}(x_1|x_2) \pm \mathrm{std}(x_1|x_2)$, respectively.

or orientations (such as PCA, ICA, wavelet transform, the receptive fields of retina gangalion cells or V1 simple cells, or even random bandpass filters), three distinct statistical characteristics have been widely observed for natural sounds and images (see Figure 2):

1. Pooling over space, the responses have symmetric supergaussian distributions with high kurtosis (Burt & Adelson, 1981; Field, 1987).
2. Joint densities of pairs of neighboring responses exhibit elliptically symmetric contours of equal probability (Wegmann & Zetzsche, 1990). Note that such joint densities can be "sphericalized" by a linear whitening operation that eliminates second-order statistical dependency.
3. The conditional distributions of one response given the values of a neighboring response, $p(x_1|x_2)$, have a bow-tie shape (Simoncelli & Buccigrossi, 1997), which can be described using the conditional means and variances (Lyu, 2009), as

$$\mathrm{E}(x_1|x_2) \approx a\,x_2, \quad \text{and} \quad \mathrm{var}(x_1|x_2) \approx b + c\,x_2^2, \tag{3.1}$$

   where a, b, c are parameters obtained from data.

These statistical dependencies are beyond second order and cannot be effectively reduced by any linear transform (Lyu & Simoncelli, 2009b; Eichhorn et al., 2009). But they can be approximately and concisely captured with the multivariate t model (Kotz & Nadarajah, 2004), which has been used in modeling local statistics of natural images (Welling, Hinton, & Osindero, 2002; Roth & Black, 2005; Chantas, Galatsanos, Likas, & Saunders, 2008).

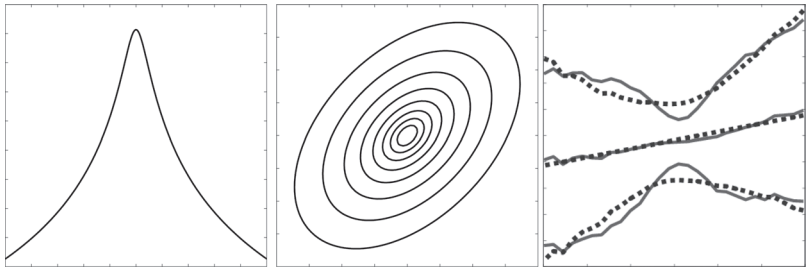

Figure 3: Properties of the multivariate t models. (Left) Log marginal distribution. (Middle) Contour plot of the pairwise joint distribution. (Right) Dashed curves correspond to $\mathrm{E}(x_1|x_2)$ (center) and $\mathrm{E}(x_1|x_2) \pm \mathrm{std}(x_1|x_2)$ of the optimally fitted multivariate t-model to pairs of adjacent bandpass-filtered responses of a natural image. The solid curves are the same as in the right panel of Figure 2.

Formally, assuming zero mean, the probability density function of a d-dimensional multivariate t vector is

$$p_t(\mathbf{x}; \alpha, \beta) = \frac{\alpha^\beta \Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)\sqrt{\det(\Sigma)}} \left(\alpha + \mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-\beta - d/2},$$

where $\alpha > 0$ and $\beta \geq 1$ are known as the scale and shape parameters, respectively. $\Gamma(\beta) = \int_0^\infty u^{\beta-1} \exp(-u)\, du$ is the standard gamma function. Σ is a symmetric and positive definite matrix, proportional to the covariance matrix when $\mathbf{x}$ has finite second-order statistics (which is not true when $\beta = 1$, which corresponds to a Cauchy distribution). The multivariate t model is a generalization of the gaussian model, and when $\alpha, \beta \to \infty$ and $\alpha/(\beta - 1) = \text{const}$, $\mathbf{x}$ converges in distribution to a gaussian random vector with zero mean and covariance matrix $\frac{\alpha\Sigma}{2(\beta-1)}$. Parameters α, β, and Σ can be estimated from data using maximum likelihood (see appendix B).

The marginal distributions of a multivariate t model are one dimensional t densities (also known as the Student-t model; see Figure 3, left), which are symmetric and nongaussian with high kurtosis; the pairwise marginal distributions of a multivariate t model are two-dimensional t models, which are elliptical and nongaussian (see Figure 3, middle panel) (Kotz & Nadarajah, 2004). The following result states that the dependencies shown in the conditional densities of natural sensory signals are the result of an intrinsic property of the multivariate t model (the lemma is proved in appendix A):

Lemma 1 *(Zellner, 1971.) For a d-dimensional multivariate t vector $\boldsymbol{x}$ with zero mean and parameters α, β, and Σ, denote $\boldsymbol{x}_{\setminus i}$ as the vector formed by excluding the ith element from $\boldsymbol{x}$ and $\Sigma_{\setminus i,\setminus i}$ as the submatrix of Σ corresponding with rows and columns of indices $\{1, \ldots, d\} \setminus i$, $\Sigma_{\setminus i,i}$ as the vector formed by the ith column*

of Σ *without its ith element, and* $\Sigma_{i,i}$ *as the ith diagonal of* Σ*. Then we have*

$$\mathrm{E}(x_i|\boldsymbol{x}_{\setminus i}) = \Sigma'_{\setminus i,i}\,\Sigma^{-1}_{\setminus i,\setminus i}\boldsymbol{x}_{\setminus i}, \tag{3.2}$$

$$\mathrm{var}(x_i|\boldsymbol{x}_{\setminus i}) = \frac{\Sigma_{i,i} - \Sigma'_{\setminus i,i}\,\Sigma^{-1}_{\setminus i,\setminus i}\,\Sigma_{\setminus i,i}}{2\beta + d - 3}\left(\alpha + \boldsymbol{x}'_{\setminus i}\,\Sigma^{-1}_{\setminus i,\setminus i}\boldsymbol{x}_{\setminus i}\right). \tag{3.3}$$

Two special cases of this result are of particular interest. First, for $d = 2$, equations 3.2 and 3.3 reduce to equation 3.1, which leads to the bow-tie shapes of the conditional distributions. In addition, if second-order dependencies in the input signal are removed with a whitening operation, where Σ becomes an identity matrix, the resulting model is the isotropic multivariate t density,

$$p_t(\mathbf{x};\alpha,\beta) = \frac{\alpha^\beta \Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)}\left(\alpha + \mathbf{x}'\mathbf{x}\right)^{-\beta-d/2}, \tag{3.4}$$

for which equations 3.2 and 3.3 are simplified to $\mathrm{E}(x_i|\mathbf{x}_{\setminus i}) = \mathbf{0}$ and $\mathrm{var}(x_i|\mathbf{x}_{\setminus i}) \propto \alpha + \mathbf{x}'_{\setminus i}\mathbf{x}_{\setminus i}$, respectively. This result is used in section 5.1

4 Justification

With the multivariate t model capturing important statistical dependencies exhibited in natural sensory signals in the bandpass-filtered domains, according to the efficient coding principle, we seek a transform that can effectively reduce such statistical dependencies. Since second-order dependencies can be trivially removed with a whitening transform, we focus on the isotropic multivariate t model, equation 3.4. However, the residual statistical dependencies in the isotropic multivariate t model cannot be further reduced with any linear transform (Lyu & Simoncelli, 2009b; Eichhorn et al., 2009).

To find a simple nonlinear transform that removes statistical dependency in the isotropic multivariate t model, we note that the isotropic gaussian distribution is the only isotropic model with mutually independent components (Kac, 1939; Nash & Klamkin, 1976). Naturally, if we can obtain a transform that can map an isotropic multivariate t vector $\mathbf{x}$ to an isotropic gaussian vector $\mathbf{u}$, then all statistical dependencies embodied in $p(\mathbf{x})$ are eliminated. In the following, we describe two different approaches that "gaussianize" an isotropic multivariate t vector; the former is based on the equivalency of the multivariate t model as a gaussian scale mixture (GSM), and the latter is based on radial gaussianization (RG). As we will show, both transforms can be closely approximated with the DN transform, which justifies its role in dependency reduction.

4.1 Via the GSM Equivalency of the *t* Model. It is well known that the multivariate t model is a gaussian scale mixture (GSM) (Andrews & Mallows, 1974). Specifically, a d-dimensional isotropic multivariate t vector $\mathbf{x}$ with parameters α and β can be decomposed into the product of two independent random variable, as

$$\mathbf{x} = \mathbf{u} \cdot \sqrt{z}, \tag{4.1}$$

where $\mathbf{u}$ is a d-dimensional isotropic gaussian vector with zero mean and unit component variance and $z > 0$ is an inverse gamma random variable with density $p(z) = \frac{\alpha^\beta}{2^\beta \Gamma(\beta)} z^{-\beta-1} \exp(-\frac{\alpha}{2z})$.

With equation 4.1, one approach to map $\mathbf{x}$ to a gaussian random vector is simply $\mathbf{u} = \mathbf{x}/\sqrt{z}$. However, as z is a latent variable to which we do not have direct access, this transform is not realizable. But it can be approximated by substituting the actual value of z with its estimator based on $\mathbf{x}$. Common choices of estimator include the maximum a posteriori (MAP) estimator $\hat{z}_{\text{MAP}} = \arg\max_z p(z|\mathbf{x})$ and the Bayesian least squares (BLS) estimator $\hat{z}_{\text{BLS}} = \operatorname{argmin}_{\hat{z}} \mathrm{E}_{z|\mathbf{x}}((\hat{z} - z)^2) = \mathrm{E}_{z|\mathbf{x}}(z|\mathbf{x})$. The following lemma, proved in appendix A, shows that both the MAP and BLS estimators for the latent z in an isotropic multivariate t model have similar analytical forms:

Lemma 2. *For a d-dimensional isotropic multivariate t vector $\boldsymbol{x}$ with zero mean and parameters (α, β), the three estimators of the latent variable z in its equivalent GSM definition are:*

1. $\hat{z}_{MAP} = \frac{1}{2\beta+d+2}(\alpha + \boldsymbol{x}'\boldsymbol{x})$
2. $\hat{z}_{BLS} = \frac{1}{2\beta+d-2}(\alpha + \boldsymbol{x}'\boldsymbol{x})$
3. $\hat{z}_{ALT} = \left(E_{z|\boldsymbol{x}}(1/z|\boldsymbol{x})\right)^{-1} = \frac{1}{2\beta+d}(\alpha + \boldsymbol{x}'\boldsymbol{x})$

If we ignore the scaling factors (which has no effect on the statistical dependencies measured by multi-information; see section 5.1), all three estimators have the form of $\alpha + \mathbf{x}'\mathbf{x}$. If we then replace $\sqrt{z}$ with $\sqrt{\alpha + \mathbf{x}'\mathbf{x}}$ in the optimal gaussianization transform, we obtain a nonlinear transform of $\mathbf{x}$ as

$$\phi(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\alpha + \mathbf{x}'\mathbf{x}}} = \frac{\|\mathbf{x}\|}{\sqrt{\alpha + \|\mathbf{x}\|^2}} \frac{\mathbf{x}}{\|\mathbf{x}\|}, \tag{4.2}$$

which is the standard form of the DN transform as shown in Figure 1. (We discuss the relation of this standard form of the DN transform with other alternative definitions in section 5.1.)

It should be mentioned that similar connections between GSM and DN have been noted previously (Wainwright & Simoncelli, 2000; Schwartz, Sejnowski, & Dayan, 2005). On the other hand, the use of the multivariate t model (a special case of GSM) has the advantage that we can

exactly compute the dependency reduction achieved with DN, as we show in section 5.2.

4.2 Via Radial Gaussianization. Radial gaussianization (RG) (Lyu & Simoncelli, 2009b; Sinz & Bethge, 2009) is a deterministic nonlinear transform that maps random vectors with isotropic nongaussian density models to isotropic gaussian vectors. The key component in RG is a nonlinear function $\psi(r)$ of the L_2 norms of input vector, which transform the radial marginal distribution, $p(r)$, of the source isotropic model to that of an isotropic gaussian model. The overall gaussianization transform is then constructed by modifying the radial component of vector $\mathbf{x}$ as $\phi(\mathbf{x}) = \psi(\|\mathbf{x}\|)\mathbf{x}/\|\mathbf{x}\|$.

In particular, the radial marginal distribution of an isotropic gaussian model with variance $1/\alpha$ is a χ density with d degrees of freedom, as

$$p_\chi(r) = \frac{\alpha^{d/2} r^{d-1}}{(2\pi)^{d/2}} \exp\left(-\frac{\alpha r^2}{2}\right),$$

and the radial marginal distribution of an isotropic multivariate t model is

$$p_t(r) = \frac{\alpha^\beta \Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \frac{r^{d-1}}{(\alpha + r^2)^{\beta+d/2}}.$$

To simplify the discussion, we assume that the isotropic multivariate t model has unit variance, which further implies that $\alpha = 2(\beta - 1)$. Under RG, we seek nonlinear map $\psi(\cdot)$ so that $r \sim p_t$ and $\psi(r) \sim p_\chi$. when the rules of changing variables are used, $\psi(r)$ is determined with equation $p_\chi(\psi(r))\left|\psi'(r)\right| = p_t(r)$, which, after expanding all the terms, becomes

$$\begin{aligned} &\frac{\alpha^{d/2}\psi^{d-1}(r)}{(2\pi)^{d/2}} \exp\left(-\frac{\alpha\psi^2(r)}{2}\right)\left|\psi'(r)\right| \\ &\quad = \frac{\alpha^{\alpha/2+1}\Gamma(\alpha/2+1+d/2)}{\pi^{d/2}\Gamma(\alpha/2+1)} \frac{r^{d-1}}{(\alpha + r^2)^{\alpha/2+1+d/2}}. \end{aligned} \tag{4.3}$$

Although equation 4.3 does not have a closed-form solution, the following lemma, proved in the appendix, shows that its solution can be approximated with the radial nonlinear transform in DN.

Lemma 3. *For small r, the dominant term on the right-hand side of equation 4.3, has an approximation, as*

$$\frac{r^{d-1}}{(\alpha + r^2)^{\alpha/2+1+d/2}} \approx \exp\left(-\frac{\alpha}{2}\frac{r^2}{\alpha + r^2}\right) \frac{\alpha^{-\frac{\alpha}{2}} r^{d-1}}{(\alpha + r^2)^{d/2+1}}. \tag{4.4}$$

With this approximation, equation 4.3 has the solution $\hat{\psi}(r) = r/\sqrt{\alpha + r^2}$.

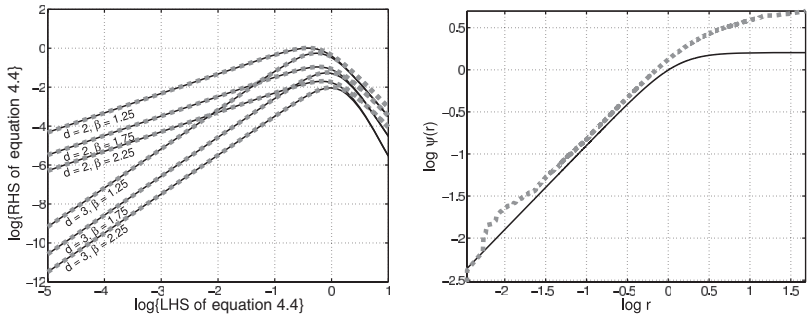

Figure 4: DN approximation to the RG transform. (Left) Illustration of the approximation in equation 4.4 for different d and β values. Dashed curves are the left hand side of equation 4.4, and solid curves are the right hand side of equation 4.4. (Right) Radial maps $\psi(r)$ of DN (solid) and RG (dashed) for an isotropic multivariate t vector. The RG transform is obtained nonparametrically with density transform (Lyu & Simoncelli, 2009b).

With the approximated solution $\hat{\psi}(r)$, the overall signal transform is constructed as

$$\phi(\mathbf{x}) = \psi(\|\mathbf{x}\|)\frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\|\mathbf{x}\|}{\sqrt{\alpha + \|\mathbf{x}\|^2}}\frac{\mathbf{x}}{\|\mathbf{x}\|},$$

which again leads to the standard form of the DN transform, equation 4.2.

The left panel of Figure 4 demonstrates the approximation of equation 4.4 with different values of d and α. The right panel of the figure compares the optimal RG solution obtained by nonparametric estimation (for details, see Lyu & Simoncelli, 2009b), and the approximated solution $\hat{\psi}(r)$. As this plot shows, for small r values, the two transforms agree with each other, and the difference between the two solutions is quite small. On the other hand, the two transforms are fundamentally different for inputs with large magnitudes, as DN saturates at 1, whereas RG transform increases unboundedly. This shows the the suboptimality of DN as a gaussianization (and hence efficient coding) transform for the multivariate t models.

5 Evaluation on Multivariate t Models

We have established DN as an approximation to the optimal efficient coding transforms of the multivariate t model. However, for two important reasons, we still need to precisely quantify the effectiveness of DN in reducing statistical dependencies. First, as seen in section 4, the DN transform is an approximation to the optimal transform that eliminates statistical dependencies in a multivariate t model. Furthermore, the multivariate t model

itself is a proxy model of natural sensory signals. Therefore, we need a quantitative evaluation of the effectiveness of the DN transform in reducing statistical dependencies of natural sensory signals to confirm its usefulness as an efficient coding transform.

We start with this section by applying DN to the isotropic multivariate t models, whose closed-form density allows a precise computation of the statistical dependencies reduced by the DN transform. In the next section, we apply the DN transform to natural sensory signal data, estimate the dependency reduction, and compare the results with those predicted from the optimally fitted multivariate t model.

Subsequently, we employ the multi-information (MI) (Studeny & Vejnarova, 1998), also known as total variation (Watanabe, 1960) or *multivariate constraint* (Garner, 1962), as a quantitative measure of statistical dependencies in multivariate random variables. MI is a multivariate generalization of the mutual information (Cover & Thomas, 2006) between a pair of variables. For a d-dimensional random vector $\mathbf{x}$ with joint density $p(\mathbf{x})$, its MI is the Kullback-Leibler (KL) divergence (Cover & Thomas, 2006) between the joint model and the product of marginals of all its components:

$$I(\mathbf{x}) = \mathrm{KL}\left(p(\mathbf{x}) \,\middle\|\, \prod_{k=1}^{d} p(x_k)\right) = \int_{\mathbf{x}} p(\mathbf{x}) \log\left(p(\mathbf{x}) \Big/ \prod_{k=1}^{d} p(x_k)\right) d\mathbf{x}. \quad (5.1)$$

MI is always nonnegative, and $I(\mathbf{x}) = 0$ if and only if the components of $\mathbf{x}$ are mutually independent. An important property of MI is that for a transform of $\mathbf{x}$ defined as $\phi(\mathbf{x}) = (\phi_1(x_1), \ldots, \phi_d(x_d))'$, where $\{\phi_k(\cdot)\}_{k=1}^{d}$ are univariate and continuously differentiable functions, we have $I(\mathbf{x}) = I(\phi(\mathbf{x}))$. This is a direct result of the change of variable procedure on the probability distributions of continuous random variables.

5.1 Equivalent Forms of DN. We have focused on the standard form of the DN transform in equation 4.2, but there are several alternative forms of the DN transform that are frequently used in the literature. In particular, we list three other forms of DN in terms of their effects on the individual elements of the output vector, as

- $s_i = \dfrac{x_i^2}{\alpha + \mathbf{x}'\mathbf{x}}$ (Heeger, 1992)
- $v_i = \dfrac{x_i}{\sqrt{\alpha + \mathbf{x}'_{\setminus i}\mathbf{x}_{\setminus i}}}$ (Schwartz & Simoncelli, 2001a)
- $t_i = \dfrac{x_i^2}{\alpha + \mathbf{x}'_{\setminus i}\mathbf{x}_{\setminus i}}$ (Wainwright et al., 2002)

However, the outputs of these DN transforms have the same MI as that of equation 4.2. Hence, they are all equivalent as efficient coding transforms.

To better see this, first recall the property of the MI that it is invariant to any continuously differentiable element-wise transform. The three alternative DN transforms are related to the standard DN form by continuously differentiable transforms that map between corresponding elements. Specifically, denoting the output of the standard DN form by $\mathbf{u}$, the output of the first DN transform can be expressed as an element-wise square of $\mathbf{u}$ as $s_i = u_i^2$. The origin of the second form of DN is a division of x_i with the conditional standard deviation of x_i given the remaining components in $\mathbf{x}$ (see section 3). In this case, we have

$$v_i = \frac{x_i}{\sqrt{\alpha + \mathbf{x}'_{-i}\mathbf{x}_{-i}}} = \frac{x_i}{\sqrt{\alpha + \mathbf{x}'\mathbf{x} - x_i^2}} = \frac{x_i/\sqrt{\alpha + \mathbf{x}'\mathbf{x}}}{\sqrt{1 - x_i^2/(\alpha + \mathbf{x}'\mathbf{x})}} = \frac{u_i}{\sqrt{1 - u_i^2}},$$

which is an element-wise nonlinear transformation of $\mathbf{u}$. Finally, the same is true for the third form of the DN transform, as $t_i = v_i^2 = u_i^2/(1 - u_i^2)$.

5.2 MI of *t* Model and Its DN Transform. We next evaluate the effectiveness of DN in reducing statistical dependencies in the isotropic multivariate t model by a direct comparison of their MIs. In doing so, we need the closed-form density of the DN transformed multivariate t vector, as the following result shows (see appendix A for the proof):

Lemma 4 *(Costa, Hero, & Vignat, 2003). If $\mathbf{x} \in R^d$ has an isotropic multivariate t density with parameter (α, β), then its DN transform, $\boldsymbol{u} = \phi(\boldsymbol{x})$, is in the d-dimensional unit hypersphere ($\|\boldsymbol{u}\| \leq 1$), and has density as*

$$p(\boldsymbol{u}) = \frac{\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \left(1 - \boldsymbol{u}'\boldsymbol{u}\right)^{\beta - 1}. \tag{5.2}$$

The density of equation 5.2 is known as the isotropic multivariate r model (Costa et al., 2003). Similar to the multivariate t models, the multivariate r models approaches to gaussians with $\beta \to \infty$. One particular property that distinguishes the multivariate r model from the multivariate t or gaussian models is that it has a finite support, which is the inside of a hypersphere corresponding to $\|\mathbf{u}\| \leq 1$.

One particular important property of the multivariate t and r models is that their entropy are in closed form. We summarize these results in the following lemma (proved in appendix A):

Lemma 5 *(Costa et al., 2003; Guerrero-Cusumano, 1996). The differential entropy of a d-dimensional isotropic multivariate t vector* $\boldsymbol{x}$ *with parameters* (α, β) *is*

$$H(\boldsymbol{x}) = \frac{d}{2}\log\alpha\pi + \log\Gamma(\beta) - \log\Gamma\left(\beta + \frac{d}{2}\right) + \left(\beta + \frac{d}{2}\right)\left[\Psi\left(\beta + \frac{d}{2}\right) - \Psi(\beta)\right], \tag{5.3}$$

where $\Psi(\beta)$ *is the digamma function defined as* $\Psi(\beta) = \frac{d}{d\beta}\log\Gamma(\beta)$. *The differential entropy of its DN transform,* $\boldsymbol{u} = \phi(\boldsymbol{x})$, *which is a d-dimensional r vector, is*

$$H(\boldsymbol{u}) = \frac{d}{2}\ \log\ \pi + \ \log\ \Gamma(\beta) - \ \log\ \Gamma\left(\beta + \frac{d}{2}\right) + (\beta - 1)\left[\Psi\left(\beta + \frac{d}{2}\right) - \Psi(\beta)\right]. \tag{5.4}$$

A direct result of the closed-form differential entropy of the multivariate t and r models is that their MIs also have closed forms (the corollary is proved in appendix A):

Corollary 1. *The MI of a d-dimensional isotropic multivariate t vector* $\boldsymbol{x}$ *with parameters* (α, β) *is*

$$I(\boldsymbol{x}) = (d-1)\log\Gamma(\beta) - d\log\Gamma(\beta + 1/2) + \log\Gamma(\beta + d/2) - (d-1)\beta\Psi(\beta) + d(\beta + 1/2)\Psi(\beta + 1/2) - (\beta + d/2)\Psi(\beta + d/2),$$

where $\Psi(\beta)$ *is the digamma function defined as* $\Psi(\beta) = \frac{d}{d\beta}\log\Gamma(\beta)$. *The MI of its DN transform,* $\boldsymbol{u} = \phi(\boldsymbol{x})$, *which is a d-dimensional r vector, is*

$$I(\mathbf{u}) = d\log\Gamma(\beta + (d-1)/2) - \log\Gamma(\beta) - (d-1)\log\Gamma(\beta + d/2) + (\beta - 1)\Psi(\beta) + (d-1)(\beta + d/2 - 1)\Psi(\beta + d/2) - d(\beta + (d-3)/2)\Psi(\beta + (d-1)/2).$$

As the gamma and the digamma functions can be evaluated to high precision, we can compute $I(\mathbf{x})$ and $I(\mathbf{u})$ directly. Figure 5 shows the surface plot of $I(\mathbf{x})$ and $I(\mathbf{u})$ after normalization by the data dimension as functions of model parameter β and data dimension d. We observe that when β increases, dependencies in both models decrease, as both the multivariate t model and the multivariate r model approach to gaussian has zero MI. On the other hand, while the MI of the multivariate t model tends to increase with data dimensions, it is not always so for the multivariate r model.

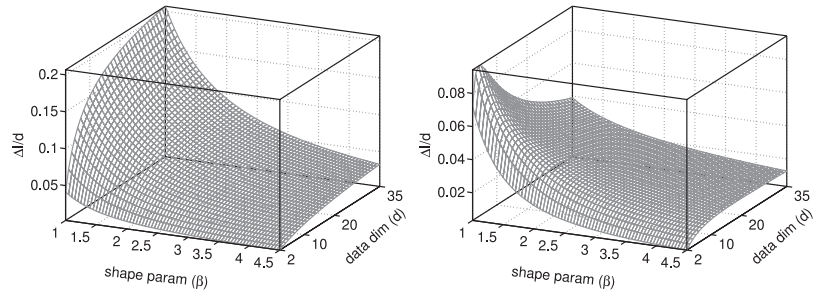

Figure 5: Surface plot of the unit MI normalized by the data dimension for the isotropic multivariate t model (left) and the isotropic multivariate r model (right) with different β and d values.

5.3 Dependency Reduction with DN. With these results, we can compute the change of MI per dimension of the input, $\Delta I/d = (I(\mathbf{x}) - I(\mathbf{u}))/d$, as a function of β and d. The left column of Figure 6 shows $\Delta I/d$, and the right column shows $\Delta I/I(\mathbf{x})$, which is the MI change relative to the raw statistical dependencies in $\mathbf{x}$. For $d > 4$, MI are reduced after DN is applied to the multivariate t models. The reduction increases with data dimensionality for fixed β values (middle row, Figure 6) and decreases with increased shape parameter β for the fixed data dimension (bottom row, Figure 6). The dependencies reduced with the DN transform are also reflected by the gaussianization effect of its outputs. Shown in the top row of Figure 7 are 1D marginal densities of the DN transformed multivariate t vectors with $\beta = 1.1$ and different dimensions. As it shows, for higher data dimension, (e.g., $d = 10$), the marginal distribution becomes quite close to the gaussian model.

However, when $d \leq 4$, the changes in MI are consistently negative for all β values, indicating that the outputs of the DN transform have increased statistical dependencies compared to the inputs. One intuitive explanation is that the small number of components in low-dimensional $\mathbf{x}$ leads to inferior estimations of the latent variable z, and thus a weaker gaussianization effect (Schwartz, Sejnowski, & Dayan, 2006). This can be further confirmed with the marginal distributions of the DN-transformed multivariate t vectors of different dimensions (top row, Figure 7). The marginal distributions for low-dimensional inputs (e.g., $d = 2$) are quite different from a gaussian.

Another interpretation may be obtained by observing the two-dimensional projections of the DN-transformed multivariate t vectors of different dimensions (bottom row, Figure 7). As the plots show, for low-dimensional inputs (e.g., $d = 2$), a significant fraction of the 2D projections are around the unit circle, which is the boundary of support of the corresponding 2D r distribution. Samples near this boundary have strong statistical dependencies (e.g., knowing one has a coordinate near ± 1, we can

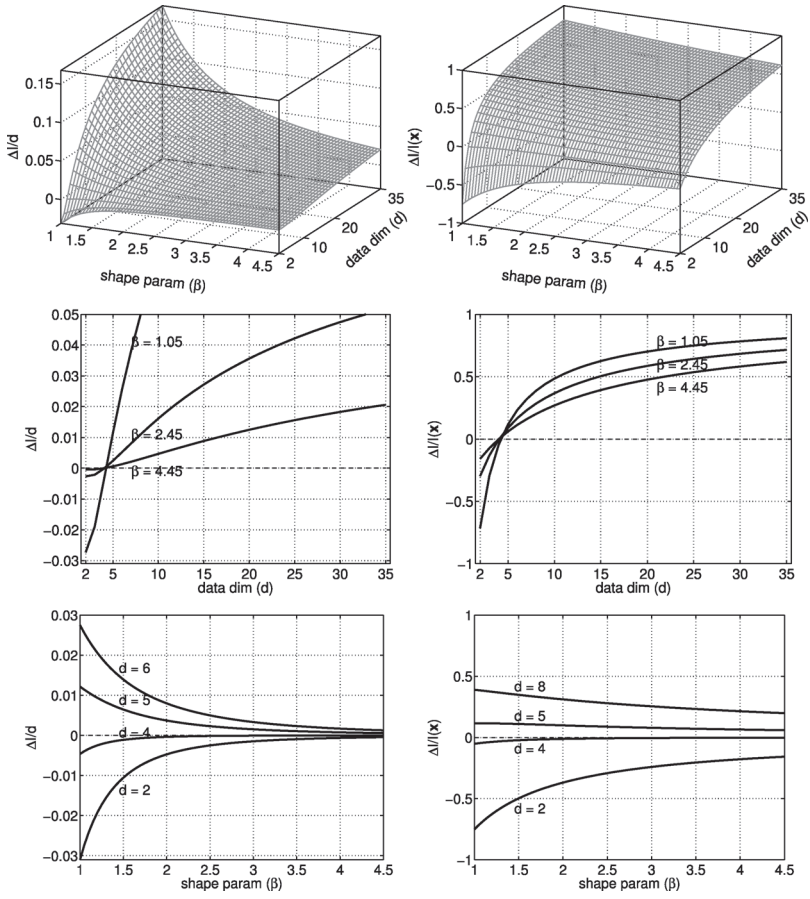

Figure 6: (Left column). Absolute change of MI normalized by the data dimensionality, $\Delta I/d$. (Right column). Same plots for relative changes of MI, $\Delta I/I(\mathbf{x})$. The top row shows surface plots with the full range of data dimensions (d) and the shape parameter of (β). The middle and bottom rows show slices of the corresponding surface plots in the top row for fixed β and d, respectively.

predict that the other component has a coordinate close to 0.0), while samples in the central region are closer to being gaussian distributed and less dependent. For low-dimensional data, the increased dependencies near the boundary may counteract the reduced dependencies of the central region; hence, the net effect is an increased MI. On the other hand, for a higher data dimension (e.g., $d = 10$), the majority of the projected samples are farther away from the unit circle, with weakened dependencies caused by the boundary constraint; hence the overall dependencies are reduced.

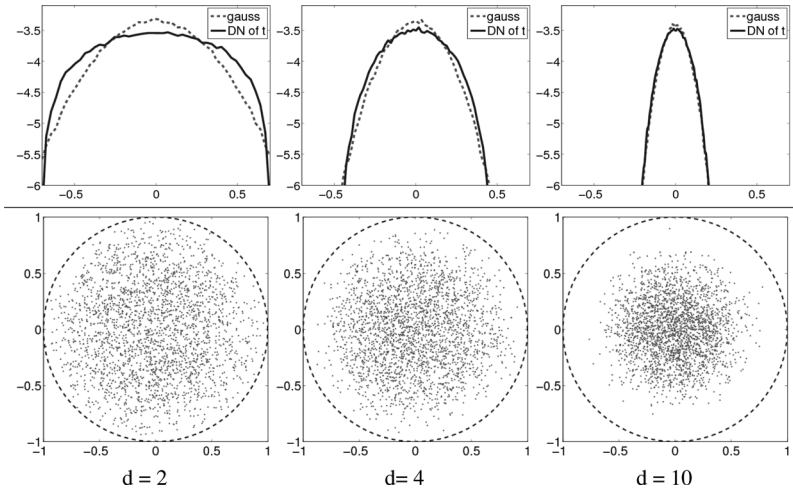

Figure 7: (Top) Marginal densities in the log domain of the components of the DN transformation of an isotropic multivariate t vector (solid curves), compared with gaussian densities (dahsed curves) of the same mean and standard deviation. (Bottom) Two-dimensional projections of the DN transformation of 1000 isotropic multivariate t samples. The three columns correspond to different data dimensions.

6 Evaluation on Natural Sensory Signals

We next evaluate the effectiveness of dependency reduction of the DN transform on natural sensory signals. Not relying on a parametric probabilistic model, we achieve this directly with data in a nonparametric manner. However, nonparametric estimation of entropy and MI from data may be difficult, as straightforward estimation using histograms are prone to strong biases (Paninski, 2003), and this may be further exacerbated by the curse of dimensionality for high-dimensional data. Nevertheless, direct estimation of MI in our case is not necessary, as we only need to compute the difference of MI between data $\mathbf{x}$ and its DN transform $\mathbf{u}$:

$$\Delta I = \sum_{k=1}^{d} (H(x_k) - H(u_k)) + H(\mathbf{u}) - H(\mathbf{x}). \tag{6.1}$$

Here we use the equivalent definition of MI in terms of the differential entropy as $I(\mathbf{x}) = \sum_{k=1}^{d} H(x_k) - H(\mathbf{x})$. $H(\mathbf{u})$ and $H(\mathbf{x})$ are related by

$$H(\mathbf{u}) = H(\mathbf{x}) - \int_{\mathbf{x}} p(\mathbf{x}) \log \left| \det \left(\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}} \right) \right| d\mathbf{x},$$

where the Jacobian determinant of the standard form of DN transform is

$$\det\left(\frac{\partial\phi(\mathbf{x})}{\partial\mathbf{x}}\right) = \frac{\alpha}{(\alpha + \mathbf{x}'\mathbf{x})^{d/2+1}}.$$

Replacing these results back to equation 6.1, we have

$$\Delta I = \sum_{k=1}^{d} (H(x_k) - H(u_k)) + \log\alpha - \left(\frac{d}{2} + 1\right) \int_{\mathbf{x}} p(\mathbf{x}) \log(\alpha + \mathbf{x}'\mathbf{x})\, d\mathbf{x}. \tag{6.2}$$

The first term computes the total difference between differential entropy of corresponding components of $\mathbf{x}$ and $\mathbf{u}$. The entropy of each component is estimated with the nonparametric m-spacing entropy estimator (Vasicek, 1976) (see appendix C). The last term in equation 6.2 is the expectation of $\log(\alpha + \mathbf{x}'\mathbf{x})$ with regard to $p(\mathbf{x})$, which can be well approximated with averages over a sufficient number of samples from $p(\mathbf{x})$. The only free parameter is α, which we determine by a direct search for a value that maximizes the resulting ΔI over a range of values. We test this nonparametric estimation of ΔI with samples from the isotropic multivariate t model and compare the results with the theoretical values computed using lemma 1. Figure 8 shows two cases of this comparison—one with fixed β in the model and varying d values and the other with fixed d and varying β values. As these results show, the nonparametric estimations are very close to the the theoretical ground-truth values, justifying their uses in the subsequent experiments.

6.1 Experiments with Natural Audio and Image Data. We next perform experiments with natural audio and image data. For audio data, we use 10 sound clips of animal vocalization and recordings in natural environments, which have a sampling frequency of 44.1 kHz and a typical length of 15 to 20 seconds. These sound clips are preprocessed with a bandpass gamma-tone filter of 3 kHz center frequency (Johannesma, 1972). For image data, we use the central 1024×1024 cropped regions of 20 images of linearized intensities from the van Hateren database (van der Schaaf & van Hateren, 1996), which are taken from natural scenes such as woods and parks. As a preprocessing step, the intensities of the image data are first subject to a global logarithm nonlinearity, $\log I(x) - C_0$, as in Bethge (2006), where C_0 is a constant so that the adjusted log intensities of an image have mean 0. The logarithm transform loosely simulates the nonlinear intensity transforms found in the cone photoreceptors in vertebrates (McCann, 2005). The images are then convolved with an isotropic bandpass filter obtained from an unoriented steerable pyramid (Simoncelli & Freeman, 1995) that

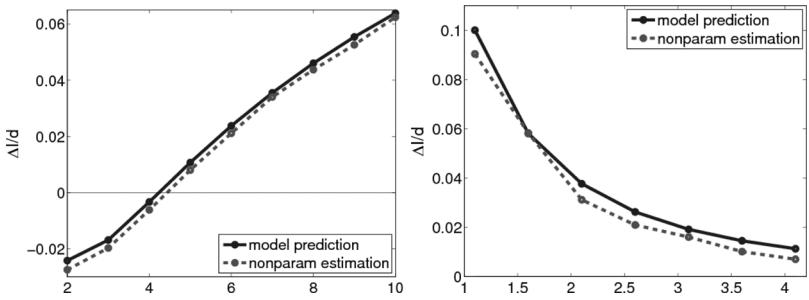

Figure 8: Comparison of theoretical prediction of MI reduction for an isotropic multivariate t model (solid curves) with the nonparametric estimation using equation 6.2, with random samples drawn from the multivariate t models (dashed curves). The left plot corresponds to samples drawn from fixed $\beta = 1.10$ and different data dimensions, and the right plot corresponds to those drawn from a model of $d = 10$ and varying β.

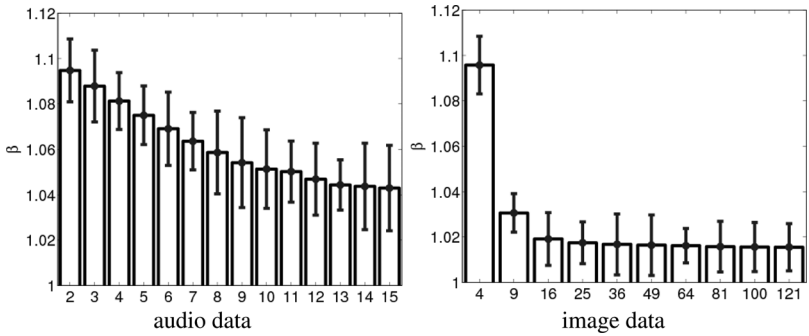

Figure 9: Mean and standard deviation of the estimated shape parameter β on natural sound data sets and natural image data sets with different dimensions.

captures an annulus of frequencies in the Fourier domain ranging from $\pi/4$ to π radians per pixel, followed by a proper downsampling. We then extract adjacent samples using localized 1D temporal (for audios) or 2D spatial (for images) windows of different sizes. These data are vectorized and whitened to have second-order dependencies removed.

We fit isotropic multivariate t models to data using the maximum likelihood estimation (described in appendix B). Shown in Figure 9 are the means and standard deviations of the estimated shape parameter β of different sizes of local windows for audio and image data, respectively (α is determined as $\alpha = 2(\beta - 1)$ for isotropic multivariate t model fitted to whitened data with identity covariance matrix). As these plots show, the estimated β values are typically close to 1, reflecting their high kurtosis.

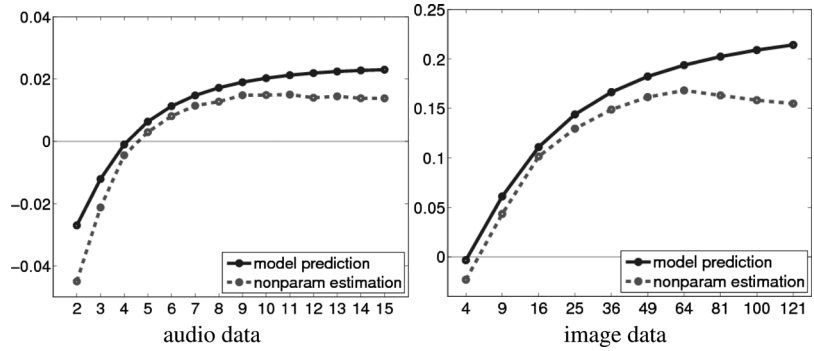

Figure 10: Unit MI changes ($\Delta I/d$) on natural sound data sets and natural image data sets with different dimensions. The solid curve corresponds to the theoretical value obtained with lemma 1; the dashed curve is the nonparametric estimation with equation 4.2.

Furthermore, β decreases as the dimensionality increases, indicating an increasing trend of nongaussianity.

Based on the fitted isotropic multivariate t model, we compute the MI difference predicted by the fitted multivariate t model (see lemma 1) and compared the results with the nonparametric estimation of MI difference in Figure 10. The solid curves correspond to theoretical predictions, and dashed curves are results from nonparametric estimations. In both cases, we report average MI difference over all sounds or images to remove biases of individual samples.

The nonparametric estimations of MI differences after applying DN to natural sensory data are in accordance with their predictions from the multivariate t models. Both results suggest strong positive correlation of the effectiveness of dependency reduction of DN with data dimension. Furthermore, as observed in the previous section, the multivariate t model predicts that DN increases statistical dependency for small input dimensions; similar observations hold for natural sensory signal data. There are also several distinct differences between the two sets of results. First, the predictions based on the multivariate t model tend to overestimate the MI difference achieved by the DN transform. More important, as the data dimension increases, the nonparametric estimations of MI difference seem to have a decreasing trend, even though multivariate t model predictions tend to keep increasing with data dimension. One important reason for such behaviors is the fact that the multivariate t model is still insufficient to represent all statistical properties of natural sensory signals, even though it is mathematically convenient and encapsulates some important types of statistical dependencies. A particular characteristic of natural sensory signals is that the statistical dependencies among their bandpass-filtered

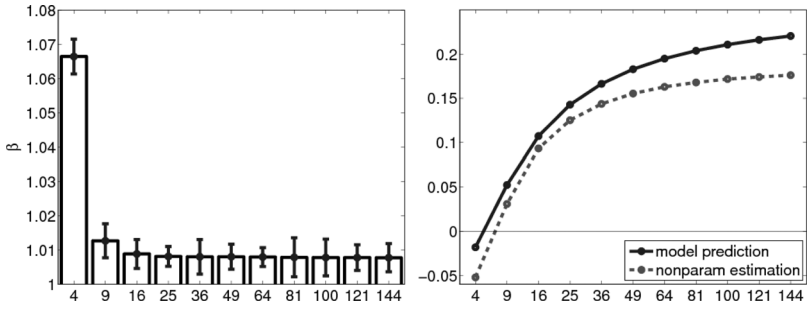

Figure 11: Experimental results with a simple direct cosine transform preprocessing on the image data. (Left) Mean and standard deviation of the estimated shape parameter β. (Right) Unit MI changes ($\Delta I/d$).

responses weaken as their separations increase: responses that are far away from each other tend to be less dependent (Wainwright & Simoncelli, 2000; Lyu & Simoncelli, 2009a). This transition of statistical dependency cannot be effectively captured by the multivariate t model, as any pair of components in an isotropic multivariate t vector, regardless of their relative positions, has the same dependencies. The lack of ability to represent less dependent or independent components using multivariate t models may explain the difference of the model predictions and nonparametric estimations of DN's effect in dependency reduction, particularly for large local patches.

The prefiltering of audio and image data is to gently restrict them to the bandpass-filtered domain, where the statistical properties described in section 3 are present. We observe that the results are robust to different choices of bandpass filters. In Figure 11, we show results for image data with different preprocessing. Instead of using filters from unoriented steerable pyramid, we apply a direct cosine transform to these blocks and separate the DC component. The AC components are further whitened and transformed with DN. The left panel of Figure 11 shows the estimated shape parameters on this data set, and the right panel shows the corresponding dependency reduction estimated by change of MI per dimension as in Figure 10. Note that these results are qualitatively consistent with those in Figure 10.

7 Discussion

In this work, we have presented an analysis justifying divisive normalization as an efficient coding transform for natural sensory signals. We use the multivariate t model to capture several important statistical properties

of natural sensory signals in the bandpass-filtered domains. DN then emerges as an approximation to two different optimal nonlinear transforms that eliminate statistical dependencies in the multivariate t model. Though focusing on one specific form of the DN transform, we show that several alternative forms of DN are equivalent in terms of their effects in dependency reduction. In addition, we use the analytical form of the multivariate t model to provide a precise quantification of the statistical dependencies reduced by the DN transform, which are used as a theoretical prediction of the actual performance of DN on natural sensory signal data. Moving to real natural sensory signal data, we provide a simple method to estimate the dependency reduction with DN nonparametrically. Our analyses confirm DN as an effective and efficient coding transform for natural sensory signals in bandpass-filtered domains. In the experiments, we observe a previously unreported phenomenon that when the input has low dimensions, DN increases statistical dependencies for both the multivariate t models and natural sensory signal ensemble.

In this work, we also studied radial gaussianization analytically in the context of the multivariate t model and derived an explicit expression that directly yields DN as an approximation to it. One distinct characteristic of the form of the DN transform is that the output saturates for large inputs (see the right panel of Figure 4), while the RG map is monotonically increasing. On one hand, this elucidates that DN cannot be the optimal efficient coding transform for any isotropic source model. On the other hand, the DN transform seems more plausible for a biological sensory system, as the inputs and outputs to sensory neurons are always bounded. One interesting open question is whether DN will emerge as an optimal solution to the efficient coding objective with ecological constraints.

Finally, our analyses are based on DN of equation 4.2 and its several equivalent forms, which are all based on the L_2 norm of the input vectors. More flexible forms of the DN transform use the general L_p norms and allow the denominator and the numerator to have different degrees. Such general forms of the DN transform are more flexible. We envision and are working on a similar analysis of the dependency reduction effects of such general forms of the DN transform, using a generalized multivariate t model based on L_p norms (Sinz & Bethge, 2009)).

Appendix A: Proof

A.1 Proof of Lemma 1. To prove the lemma, we use the basic fact that the multivariate t model is a gaussian scale mixture with an inverse gamma scaling variable. Specifically, we can express the joint distribution as $p(\mathbf{x}) = \int_z p(\mathbf{x}|z)p(z)dz$, where $p(\mathbf{x}|z)$ is a gaussian distribution with zero mean and covariance matrix $z\Sigma$, while $p(z)$ is an inverse gamma distribution with parameter (α, β).

The conditional mean of x_i given $\mathbf{x}_{\backslash i}$ is then given by

$$\begin{aligned}\mathrm{E}(x_i|\mathbf{x}_{\backslash i}) &= \int_{x_i} x_i\, p(x_i|\mathbf{x}_{\backslash i})\; dx_i = \frac{\int_{x_i} x_i\, p(\mathbf{x})\; dx_i}{p(\mathbf{x}_{\backslash i})} = \frac{\int_{x_i,z} x_i\, p(\mathbf{x}, z)\; dx_i dz}{\int_{x_i,z} p(\mathbf{x}, z)\; dx_i dz} \\ &= \frac{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz \int_{x_i} x_i\, p(x_i|\mathbf{x}_{\backslash i}, z)\; dx_i}{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz}.\end{aligned}$$

With the property of gaussian distributions (Feller, 1968), we note that $p(x_i|\mathbf{x}_{\backslash i}, z)$ is a 1D gaussian density, whose mean and variance are $\Sigma'_{\backslash i,i}\Sigma^{-1}_{\backslash i,\backslash i}\mathbf{x}_{\backslash i}$ and $z(\Sigma_{i,i} - \Sigma'_{\backslash i,i}\Sigma^{-1}_{\backslash i,\backslash i}\Sigma_{\backslash i,i})$, respectively. Therefore, we have

$$\mathrm{E}(x_i|\mathbf{x}_{\backslash i}) = \mathrm{E}(x_i|\mathbf{x}_{\backslash i}, z) = \Sigma'_{\backslash i,i}\Sigma^{-1}_{\backslash i,\backslash i}\mathbf{x}_{\backslash i}.$$

Next, we compute the conditional variance:

$$\begin{aligned}\mathrm{var}(x_i|\mathbf{x}_{\backslash i}) &= \int_{x_i} (x_i - \mathrm{E}(x_i|\mathbf{x}_{\backslash i}))^2 p(x_i|\mathbf{x}_{\backslash i})\; dx_i \\ &= \frac{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz \int_{x_i} (x_i - \mathrm{E}(x_i|\mathbf{x}_{\backslash i}, z))^2 p(x_i|\mathbf{x}_{\backslash i}, z)\; dx_i}{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz} \\ &= \left(\Sigma_{i,i} - \Sigma'_{\backslash i,i}\Sigma^{-1}_{\backslash i,\backslash i}\Sigma_{\backslash i,i}\right)\frac{\int_z z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz}{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz}.\end{aligned}$$

The ratio in the last step can be further simplified if we notice that

$$\begin{aligned}\frac{\int_z z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz}{\int_z p(z) p(\mathbf{x}_{\backslash i}|z)\; dz} &= \frac{\int_z z p_{\gamma^{-1}}(z;\alpha,\beta)\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz}{\int_z p_{\gamma^{-1}}(z;\alpha,\beta)\; dz\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz} \\ &= \frac{\alpha\Gamma(\beta-1)}{2\Gamma(\beta)}\frac{\int_z p_{\gamma^{-1}}(z;\alpha,\beta-1)\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz}{\int_z p_{\gamma^{-1}}(z;\alpha,\beta)\; dz\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz} \\ &= \frac{\alpha}{2(\beta-1)}\frac{\int_z p_{\gamma^{-1}}(z;\alpha,\beta-1)\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz}{\int_z p_{\gamma^{-1}}(z;\alpha,\beta)\; dz\mathcal{N}(\mathbf{x}_{\backslash i}/\sqrt{z})\; dz}.\end{aligned}$$

Notice that the numerator is the GSM form of a multivariate t model of the $d-1$ dimension with parameter α and $\beta-1$, while the denominator is a multivariate t model of the $d-1$ dimension with parameters α and β, so the last step can be further simplified as

$$\frac{\alpha}{2(\beta-1)}\frac{\frac{\alpha^{\beta-1}\Gamma(\beta-1+(d-1)/2)}{\pi^{(d-1)/2}\Gamma(\beta-1)}\left(\alpha+\mathbf{x}'_{\backslash i}\Sigma^{-1}_{\backslash i,\backslash i}\mathbf{x}_{\backslash i}\right)^{-(\beta-1)-(d-1)/2}}{\frac{\alpha^{\beta}\Gamma(\beta+(d-1)/2)}{\pi^{(d-1)/2}\Gamma(\beta)}(\alpha+\mathbf{x}'_{\backslash i}\Sigma^{-1}_{\backslash i,\backslash i}\mathbf{x}_{\backslash i})^{-\beta-(d-1)/2}},$$

which, after simplifying the terms, becomes $\mathrm{E}(x_i^2|\mathbf{x}_{\setminus i}) = \frac{1}{2\beta+d-3}(\alpha + \mathbf{x}'_{\setminus i}\Sigma^{-1}_{\setminus i,\setminus i}\mathbf{x}_{\setminus i})$. Collecting all terms together, we have

$$\mathrm{var}(x_i|\mathbf{x}_{\setminus i}) = \frac{\Sigma_{i,i} - \Sigma'_{\setminus i,i}\Sigma^{-1}_{\setminus i,\setminus i}\Sigma_{\setminus i,i}}{2\beta + d - 3}(\alpha + \mathbf{x}'_{\setminus i}\Sigma^{-1}_{\setminus i,\setminus i}\mathbf{x}_{\setminus i}).$$

A.2 Proof of Lemma 2.

Corollary 2. *The mean and mode of the inverse gamma density,* $p_{\gamma^{-1}}(z;\alpha,\beta) = \frac{\alpha^\beta}{2^\beta\Gamma(\beta)}z^{-\beta-1}\exp(-\frac{\alpha}{2z})$, *is* $\frac{\alpha}{2(\beta-1)}$ *and* $\frac{\alpha}{2(\beta+1)}$, *respectively. Furthermore, the mean of* $1/z$ *with regard to the inverse gamma density is* $2\beta/\alpha$.

Proof. For the mean of the inverse gamma density, we have

$$\begin{aligned}\int_0^\infty z p_{\gamma^{-1}}(z;\alpha,\beta)dz &= \int_0^\infty z\cdot\frac{\alpha^\beta}{2^\beta\Gamma(\beta)}z^{-\beta-1}\exp\left(-\frac{\alpha}{2z}\right)dz\\ &= \frac{\alpha^\beta}{2^\beta\Gamma(\beta)}\int_0^\infty z^{-\beta}\exp\left(-\frac{\alpha}{2z}\right)dz.\end{aligned}$$

Using the property of the gamma function, the last integral equals $(a/2)^{-(\beta-1)}\Gamma(\beta-1)$. As $\Gamma(\beta)/\Gamma(\beta-1) = \beta-1$, we show that the mean is $\alpha/[2(\beta-1)]$.

Next, the mode of the inverse gamma density corresponds to the z value with the maximum log density, which is given by

$$\log p_{\gamma^{-1}}(z;\alpha,\beta) = \log\alpha^\beta - \log 2^\beta\Gamma(\beta) - (\beta+1)\log z - \frac{\alpha}{2z},$$

whose derivative with regard to z is $-\frac{\beta+1}{z} + \frac{\alpha}{2z^2}$. Setting the derivative to 0 and solving for z, we obtain the mode of the inverse gamma density as $\alpha/[2(\beta+1)]$.

Finally, the mean of $1/z$ in the inverse gamma density is defined as

$$\int_0^\infty z^{-1}p_{\gamma^{-1}}(z;\alpha,\beta)dz = \frac{\alpha^\beta}{2^\beta\Gamma(\beta)}\int_0^\infty z^{-\beta-2}\exp\left(-\frac{\alpha}{2z}\right)dz.$$

Using the property of the gamma function, we find that the last integral equals $(a/2)^{-(\beta+1)}\Gamma(\beta+1)$. Putting this result back and using the property of the gamma function that $\Gamma(\beta+1)/\Gamma(\beta) = \beta$, we simplify the result to $2\beta/\alpha$.

Now we turn to prove lemma 2. We first show that the posterior density of z in the multivariate t model is also an inverse gamma density, as we

have

$$p(z|\mathbf{x};\alpha,\beta) = \frac{p(\mathbf{x},z;\alpha,\beta)}{p_t(\mathbf{x};\alpha,\beta)} = \frac{\frac{1}{(2\pi)^{d/2}z^{d/2}}\exp\left(-\frac{1}{2z}\mathbf{x}'\mathbf{x}\right)\frac{\alpha^\beta}{2^\beta\Gamma(\beta)}z^{-\beta-1}\exp\left(-\frac{\alpha}{2z}\right)}{\frac{\alpha^\beta\Gamma(\beta+d/2)}{\pi^{d/2}\Gamma(\beta)}\frac{1}{(\alpha+\mathbf{x}'\mathbf{x})^{\beta+d/2}}}.$$

Rearranging terms, we have

$$p(z|\mathbf{x};\alpha,\beta) = \frac{(\alpha+\mathbf{x}'\mathbf{x})^{\beta+d/2}}{2^{\beta+d/2}\Gamma(\beta+d/2)}z^{-\beta-d/2-1}\exp\left(-\frac{1}{2z}(\alpha+\mathbf{x}'\mathbf{x})\right),$$

which is an inverse gamma density with $(\alpha+\mathbf{x}'\mathbf{x}, \beta+d/2)$. The result is immediate with corollary 2.

A.3 Proofs of Lemma 3. We first prove equation 6.4. The first-order Taylor series approximation of function $x\log x$ for $x \geq 1$ is $x\log x \approx x-1$. Replacing with $x = 1+u$, we have $\log(1+u) \approx 1 - \frac{1}{1+u} = \frac{u}{1+u}$. Next, set $u = r^2/\alpha$. We have $\log(\alpha+r^2) \approx \log\alpha + \frac{r^2}{\alpha+r^2}$, so $-\frac{\alpha}{2}\log(\alpha+r^2) \approx -\frac{\alpha}{2}\log\alpha - \frac{\alpha}{2}\frac{r^2}{\alpha+r^2}$, or $(\alpha+r^2)^{-\frac{\alpha}{2}} \approx \alpha^{-\frac{\alpha}{2}}\exp(-\frac{\alpha}{2}\frac{r^2}{\alpha+r^2})$.

Next, replace the right-hand side of equation 6.3 with 6.4. Dropping scaling factors, we now show that $\psi(r) = \frac{r}{\sqrt{\alpha+r^2}}$ provides a solution to the resulting differential equation. This is achieved by replacing $\psi'(r) = \frac{\alpha}{(\alpha+r^2)^{3/2}}$ on the left-hand side of equation 6.3. Merging similar terms, we have

$$\psi(r)^{d-1}\exp\left(-\frac{\alpha\psi^2(r)}{2}\right)|\psi'(r)| = \frac{r^{d-1}}{(\alpha+r^2)^{(d-1)/2}}\exp\left(-\frac{\alpha r^2}{2(\alpha+r^2)}\right)\frac{\alpha}{(\alpha+r^2)^{3/2}},$$

which equates the left-hand side of equation 6.3 with the approximation of its right-hand side with equation 6.4.

A.4 Proofs of Lemma 4. The relation between densities of transformed random vector gives that $p(\mathbf{x}) = p(\phi(x))\det(J_\phi(\mathbf{x}))$ or $p(\mathbf{u}) = p(\mathbf{x})\frac{1}{\det(J_\phi(\mathbf{x}))}$. Using the Jacobian determinant of the DN transform, we have

$$p(\mathbf{u}) = \frac{\alpha^\beta\Gamma(\beta+d/2)}{\pi^{d/2}\Gamma(\beta)}\frac{(\alpha+\mathbf{x}'\mathbf{x})^{d/2+1}}{\alpha(\alpha+\mathbf{x}'\mathbf{x})^{\beta+d/2}} = \frac{\alpha^{\beta-1}\Gamma(\beta+d/2)}{\pi^{d/2}\Gamma(\beta)}\frac{1}{-(\alpha+\mathbf{x}'\mathbf{x})^{\beta-1}}. \tag{A.1}$$

Next, note that

$$\mathbf{u} = \frac{\mathbf{x}}{\sqrt{\alpha + \mathbf{x}'\mathbf{x}}} \Rightarrow \mathbf{u}'\mathbf{u} = \frac{\mathbf{x}'\mathbf{x}}{\alpha + \mathbf{x}'\mathbf{x}} \Rightarrow \|\mathbf{x}\| = \frac{\sqrt{\alpha}\|\mathbf{u}\|}{\sqrt{1 - \mathbf{u}'\mathbf{u}}}.$$

Therefore, as $\mathbf{x}/\|\mathbf{x}\| = \mathbf{u}/\|\mathbf{u}\|$, we have $\mathbf{x} = \frac{\sqrt{\alpha}\mathbf{u}}{\sqrt{1-\|\mathbf{u}\|^2}}$. Replacing this with equation A.1, we have

$$p(\mathbf{u}) = \frac{\alpha^{\beta-1}\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \frac{1}{\left(\alpha + \frac{\alpha\mathbf{u}'\mathbf{u}}{1-\mathbf{u}'\mathbf{u}}\right)^{\beta-1}} = \frac{\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)}(1 - \mathbf{u}'\mathbf{u})^{\beta-1}.$$

A.5 Proof of Lemma 5. The entropy of a multivariate t and r models is in closed form (Costa et al., 2003), and we provide the derivation here for completeness of this work.

Evaluating $H(\mathbf{x})$. We first expand $H(\mathbf{x})$ based on its definition as

$$H(\mathbf{x}) = \frac{d}{2}\log\pi - \beta\log\alpha + \log\Gamma(\beta) - \log\Gamma(\beta + d/2) + (\beta + d/2)\int_{\mathbf{x}} \log(\alpha + \mathbf{x}'\mathbf{x})p_t(\mathbf{x};\alpha,\beta)\,d\mathbf{x}. \tag{A.2}$$

To compute the last term in equation A.2, we use the normalizing property of the multivariate t model as

$$\int_{\mathbf{x}} \left(\alpha + \mathbf{x}'\mathbf{x}\right)^{-\beta-d/2} d\mathbf{x} = \frac{\pi^{d/2}\Gamma(\beta)}{\alpha^{\beta}\Gamma(\beta + d/2)}.$$

Taking the derivative with regard to β to both sides, we have

$$-\int_{\mathbf{x}} \left(\alpha + \mathbf{x}'\mathbf{x}\right)^{-\beta-d/2} \log(\alpha + \mathbf{x}'\mathbf{x}) d\mathbf{x} = \frac{\partial}{\partial\beta}\left(\frac{\pi^{d/2}\Gamma(\beta)}{\alpha^{\beta}\Gamma(\beta + d/2)}\right).$$

Multiplying $-\alpha^{\beta}\Gamma(\beta + d/2)/\pi^{d/2}\Gamma(\beta)$ on both sides, we obtain

$$\begin{aligned}\int_{\mathbf{x}} \log(\alpha + \mathbf{x}'\mathbf{x})p_t(\mathbf{x};\alpha,\beta)\,d\mathbf{x} &= -\frac{\partial}{\partial\beta}\left(\frac{\pi^{d/2}\Gamma(\beta)}{\alpha^{\beta}\Gamma(\beta + d/2)}\right)\frac{\alpha^{\beta}\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \\ &= \frac{\partial}{\partial\beta}\log\frac{\alpha^{\beta}\Gamma(\beta + d/2)}{\Gamma(\beta)} = \log\alpha + \Psi(\beta + d/2) - \Psi(\beta).\end{aligned}$$

Replacing the integral in the last term in equation A.2 with $\log\alpha + \Psi(\beta + d/2) - \Psi(\beta)$, we have proved equation 5.3.

Evaluating $H(\mathbf{u})$. Similar to the previous case, we first expand $H(\mathbf{u})$ as

$$H(\mathbf{u}) = d/2 \log \pi + \log \Gamma(\beta) - \log \Gamma(\beta + d/2) - (\beta - 1) \int_{\mathbf{u}} p_\tau(\mathbf{u}; \beta) \log \left(1 - \mathbf{u}'\mathbf{u}\right)_+ d\mathbf{u}. \tag{A.3}$$

We compute the integral in equation A.3 by the normalizing property of the multivariate r model. We start with

$$\frac{\pi^{d/2}\Gamma(\beta)}{\Gamma(\beta + d/2)} = \int_{\mathbf{u}} \left(1 - \mathbf{u}'\mathbf{u}\right)_+^{\beta - 1} d\mathbf{u}.$$

Next, taking the derivatives with regard to β and multiplying $\frac{\Gamma(\beta+d/2)}{\pi^{d/2}\Gamma(\beta)}$, we have

$$\begin{aligned}
\frac{\Gamma(\beta + d/2)}{\Gamma(\beta)} \frac{d}{d\beta} \frac{\Gamma(\beta)}{\Gamma(\beta + d/2)} &= \frac{\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \frac{d}{d\beta} \int_{\mathbf{u}} \left(1 - \mathbf{u}'\mathbf{u}\right)_+^{\beta - 1} d\mathbf{u} \\
&= \int_{\mathbf{u}} \frac{\Gamma(\beta + d/2)}{\pi^{d/2}\Gamma(\beta)} \left(1 - \mathbf{u}'\mathbf{u}\right)_+^{\beta - 1} \log \left(1 - \mathbf{u}'\mathbf{u}\right)_+ d\mathbf{u} \\
&= \int_{\mathbf{u}} p_\tau(\mathbf{u}; \beta) \log \left(1 - \mathbf{u}'\mathbf{u}\right)_+ d\mathbf{u}.
\end{aligned}$$

We can further simplify

$$\frac{\Gamma(\beta + d/2)}{\Gamma(\beta)} \frac{d}{d\beta} \frac{\Gamma(\beta)}{\Gamma(\beta + d/2)} = \frac{d}{d\beta} \log \frac{\Gamma(\beta)}{\Gamma(\beta + d/2)} = \Psi(\beta) - \Psi(\beta + d/2).$$

Putting this result in equation A.3, we have proved equation 5.4.

A.6 Proof of Corollary 1. First, using the relation between differential entropy and MI, we have

$$I(\mathbf{x}) = \sum_{k=1}^{d} H(x_k) - H(\mathbf{x}). \tag{A.4}$$

To compute MI, we need to compute the joint differential entropy for multivariate t and r models and the differential entropy for each component of the multivariate t and r vectors. The former are direct results of lemma 5, which can be obtained using the fact that each component of a multivariate t vector $\mathbf{x}$ has a one-dimensional t density and equation 5.3, so we have

$$H(x_i) = \frac{1}{2} \log \alpha\pi + \log \Gamma(\beta) - \log \Gamma(\beta + 1/2) + (\beta + 1/2) \left[\Psi\left(\beta + 1/2\right) - \Psi\left(\beta\right)\right]. \tag{A.5}$$

Furthermore, the marginal density of each element, u_i, in a multivariate r vector is a one-dimensional r model with parameter $\beta + (d-1)/2$ (Elderton, 1953), as

$$p_\tau(u_i;\beta) = \frac{\beta + (d-1)/2}{\sqrt{\pi}}(1-u_i^2)_+^{\beta-1+(d-1)/2}$$

Applying this to equation 5.4, the differential entropy of u_i is given as (Zografos, 1999)

$$H(u_i) = \frac{1}{2}\log\pi + \log\Gamma\left(\beta + \frac{d-1}{2}\right) - \log\Gamma\left(\beta + \frac{d}{2}\right) + \left(\beta + \frac{d-3}{2}\right)\left[\Psi\left(\beta + \frac{d}{2}\right) - \Psi\left(\beta + \frac{d-1}{2}\right)\right]. \tag{A.6}$$

$I(\mathbf{x})$ and $I(\mathbf{u})$ are obtained by combining equations 5.3, 5.4, A.5, and A.6 into equation A.4.

Appendix B: Maximum Likelihood Fitting of Multivariate *t* Model

We briefly describe the maximum likelihood fitting the multivariate t model to using data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, which have been centered and whitened. Enforcing an identical covariance matrix and using the GSM equivalence of the multivariate t model, we have

$$\begin{aligned} I &= \int_{\mathbf{x}} \mathbf{x}\mathbf{x}^T p_t(\mathbf{x};\alpha,\beta)d\mathbf{x} \\ &= \int_z p_{\gamma^{-1}}(z;\alpha,\beta)dz \int_{\mathbf{x}} \mathbf{x}\mathbf{x}^T \mathcal{N}(\mathbf{x}/\sqrt{z})d\mathbf{x} = \frac{\alpha}{2(\beta-1)} I, \end{aligned}$$

the last step of which is based on fact 2. Immediately, we have $\alpha = 2(\beta-1)$, which means that we only need to estimate β. Replacing this result, the average log likelihood of the multivariate t model, $L(\beta) = \frac{1}{N}\sum_{n=1}^N \log p_t(\mathbf{x}_n; 2(\beta-1), \beta)$, becomes

$$L(\beta) = \beta\log 2(\beta-1) + \log\Gamma(\beta + d/2) - \frac{d}{2}\log\pi - \log\Gamma(\beta) - \frac{\beta + d/2}{N}\sum_{n=1}^N \log(2(\beta-1) + \mathbf{x}_n^T\mathbf{x}_n).$$

Optimal β is the root of the nonlinear equation $\frac{d}{d\beta}L(\beta) = 0$, which is obtained numerically by the Newton-Raphson procedure.

Appendix C: *m*-spacing Entropy Estimator

Let $z_1 \leq \cdots \leq z_N$ be the sorted set of N independent and identically distributed samples of scalar random variable z. With an integer $m = O(\sqrt{N})$, the m-spacing entropy estimation of $H(z)$ is defined as

$$\hat{H}(z_1, \ldots, z_N) = \frac{1}{N} \sum_{i=1}^{N-m} \log \left(\frac{N}{m} [z_{i+m} - z_i] \right) - \Psi(m) + \log(m),$$

where $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. The m-spacing estimator is strongly consistent—$P\left(\lim_{m\to\infty, m/N\to 0} \hat{H}(z_1, \ldots, z_N) = H(z)\right) = 1$.

Acknowledgments

Thanks to Eero Simoncelli for helpful discussions and the two anonymous referees for their critical and constructive comments. This material is based on work supported by the National Science Foundation under CAREER Award grant no. 0953373. Any opinions, findings, and conclusions or recommendations expressed in this material are my own and do not necessarily reflect the views of the National Science Foundation.

References

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 99–102.

Atick, J. J. (1992). Network Could information theory provide an ecological theory of sensory processing? *Network, 3*, 213–251.

Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Computation*, *4*, 559–572.

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

Attneave, F. (1954). Some informational aspects of visual perception. *Psych. Rev.*, *61*, 183–193.

Baddeley, R. (1996). Searching for filters with "interesting" output distributions: An uninteresting direction to explore. *Network*, *7*, 409–421.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barlow, H. (2001). Redundancy reduction revisited. *Network, 12*, 241–253.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Bethge, M. (2006). Factorial coding of natural images: How effective are linear models in removing higher-order dependencies? *J. Opt. Soc. Am. A, 23*(6), 1253–1268.

Brady, N., & Field, D. J. (2000). Local contrast in natural images: Normalisation and coding efficiency. *Perception*, *29*, 1041–1055.

Burt, P., & Adelson, E. (1981). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, *31*(4), 532–540.

Carandini, M., & Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*, 1333–1336.

Carandini, M., Heeger, D. J., & Senn, W. (2002). A synaptic explanation of suppression in visual cortex. *Journal of Neuroscience*, *22*(22), 10053–10065.

Chantas, G., Galatsanos, N., Likas, A., & Saunders, M. (2008). Bayesian image restoration based on a product of t-distributions image prior. *IEEE Transactions on Image Processing*, *17*(10), 1795–1820.

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network*, *12*, 199–213.

Costa, J., Hero, A., & Vignat, C. (2003). On solutions to multivariate maximum α-entropy problems. In *Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition.* New York: Springer.

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Deneve, S., Pouget, A., & Latham, P. (1999). Divisive normalization, line attractor networks and ideal observers. In M. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural processing systems 11* (pp. 104–110). Cambridge, MA: MIT Press.

Eichhorn, J., Sinz, F., & Bethge, M. (2009). Natural image coding in V1: How much use is orientation selectivity? *PLoS Computational Biology, 5*(4), 1–16.

Elderton, W. (1953). *Frequency curves and correlations*. General Books LLC.

Feller, W. (1968). *An Introduction to probability theory and its applications*. Hoboken, NJ: Wiley.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Foley, J. (1994). Human luminance pattern mechanisims: Masking experiments require a new model. *J. of Opt. Soc. of Amer. A*, *11*(6), 1710–1719.

Gao, D., & Vasconcelos, N. (2009). Decision-theoretic saliency: Computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, *21*, 239–271.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. Hoboken, NJ: Wiley.

Guerrero-Cusumano, J.-L. (1996). A measure of total variability for the multivariate t distribution with applications to finance. *Information Sciences*, *92*(1–4), 47–63.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neural Science*, *9*, 181–198.

Holt, G. R., & Koch, C. (1997). Shunting inhibition does not have a divisive effect on firing rates. *Neural Computation*, *9*, 1001–1013.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multistage architecture for object recognition? In *Proceedings of the 12th IEEE Conference on Computer Vision*. Piscataway, NJ: IEEE.

Johannesma, P. (1972). The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory* (pp. 58–69). Eindhoven, Holland: IPO.

Kac, M. (1939). On a characterization of the normal distribution. *American Journal of Mathematics*, *61*(3), 726–728.

Kotz, S., & Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge: Cambridge University Press.

Laparra, V., Muñoz-Mari, J., & Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A*, *27*(4), 852–864.

Lee, J., & Maunsell, J. H. R. (2009). A normalization model of attentional modulation of single unit responses. *PLoS One*, *4*(2), 1–13.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Li, Q., & Wang, Z. (2008). General-purpose reduced-reference image quality assessment based on perceptually and statistically motivated image representation. In *IEEE 15th Int'l. Conf. on Image Proc.* (Vol. 15, pp. 1192–1195). Piscataway, NJ: IEEE.

Lyu, S. (2009). An implicit Markov random field model for natural images in multiscale oriented representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Lyu, S. (2010). Divisive normalization: Justification and effectiveness as efficient coding transform. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.). *Advances in neural information processing systems, 21*. Cambridge, MA: MIT Press.

Lyu, S., & Simoncelli, E. P. (2008). Nonlinear image representation using divisive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Piscataway, NJ: IEEE.

Lyu, S., & Simoncelli, E. P. (2009a). Modeling multiscale subbands of photographic images with fields of gaussian scale mixtures. In *IEEE Trans. Patt. Analysis and Machine Intelligence, 31*(4), 693–706.

Lyu, S., & Simoncelli, E. P. (2009b). Nonlinear extraction of "independent components" of natural images using radial gaussianization. *Neural Computation*, *18*(6), 1–35.

Malo, J., Epifanio, I., Navarro, R., & Simoncelli, E. P. (2006). Non-linear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, *15*(1), 68–80.

Malo, J., & Laparra, V. (2010). Psychophysically tuned divisive normalization factorizes the PDF of natural images. *Neural Computation*, *22*, 3179–3206.

Mante, V., Bonin, V., & Carandini, M. (2008). Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron*, *58*, 625–638.

McCann, J. (2005). Rendering high-dynamic range images: Algorithms that mimic human vision. In *Proc. AMOS Technical Conf.* (pp. 19–28) Maui: Maui Economic Development Board Publications.

Nash, D., & Klamkin, M. S. (1976). A spherical characterization of the normal distribution. *Journal of Multi-Variate Analysis*, *55*, 56–158.

Olsen, S. R., Bhandawat, V., & Wilson, R. I. (2010). Divisive normalization in olfactory population codes. *Neuron*, *66*(2), 287–299.

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

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

Pillow, J. W., & Simoncelli, E. P. (2006). Dimensionality reduction in neural models: An information theoretic generalization of spike-triggered average and covariance analysis. *Journal of Vision*, *6*, 414–428.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*, 168–185.

Ringach, D. L. (2010). Population coding under normalization. *Vision Research*, *50*, 2223–2232.

Roth, S., & Black, M. (2005). Fields of experts: A framework for learning image priors. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 860–867). Piscataway, NJ: IEEE.

Ruderman, D. L., Cronin, T. W., & Chiao, C.-C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A*, *15*(8), 2036–2045.

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, *46*(6), 945–956.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2005). Assignment of multiplicative mixtures in natural images. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems*, *17*. Cambridge, MA: MIT Press.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical model of natural scene statistics. *Neural Computation*, *18*, 2680–2718.

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

Schwartz, O., & Simoncelli, E. P. (2001b). Natural sound statistics and divisive normalization in the auditory system. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp, 166–172). Cambridge, MA: MIT Press.

Shapley, R., & Enroth-Cugell, C. (1984). Visual adaptation and retinal gain control. *Progress in Retinal Research*, *3*, 263–346.

Simoncelli, E. P., & Buccigrossi, R. W. (1997). Embedded wavelet image compression based on a joint probability model. In *Proc 4th IEEE Int'l Conf on Image Proc*. (Vol. I, pp. 640–643). Piscataway, NJ: IEEE.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings of the IEEE International Conference on Image Processing* (Vol. 3, pp. 444–447). Piscataway, NJ: IEEE.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743–761.

Sinz, F. H., & Bethge, M. (2009). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*, *21*. Cambridge, MA: MIT Press.

Solomon, S., Lee, B., & Sun, H. (2006). Suppressive surrounds and contrast gain in magnocellular pathway retinal gangalion cells of macaque. *Journal of Neuroscience*, *26*(34), 8715–8726.

Studeny, M., & Vejnarova, J. (1998). The multi-information function as a tool for measuring stochastic dependence. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 261–297). Dordrecht: Kluwer.

Valerio, R., & Navarro, R. (2003a). Input-output statistical independence in divisive normalization models of V1 neurons. *Network*, *14*(4), 733–745.

Valerio, R., & Navarro, R. (2003b). Optimal coding through divisive normalization models of V1 neurons. *Network*, *14*(3), 579–593.

van der Schaaf, A., & van Hateren, J. H. (1996). Modelling the power spectra of natural images: Statistics and information. *Vision Research*, *28*(17), 2759–2770.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, *38*(1), 54–59.

Wainwright, M. J., Schwartz, O., & Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In R.P.N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 203–222). Cambridge, MA: MIT Press.

Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. In S. A., Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 855–861). Cambridge, MA: MIT Press.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, *4*, 66–82.

Watson, A., & Solomon, J. (1997). A model of visual contrast gain control and pattern masking. *J. Opt. Soc. Amer. A*, *14*, 2379–2391.

Wegmann, B., & Zetzsche, C. (1990). Statistical dependence between orientation filter outputs used in an human vision based image code. In *Proc. Visual Comm. and Image Processing* (Vol. 1360, pp. 909–922). Bellingham, WA: SPIE.

Welling, M., Hinton, G. E., & Osindero, S. (2002). Learning sparse topographic representations with products of Student-*t* distributions. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 1111–1117). Cambridge, MA: MIT Press.

Zellner, A. (1971). *An introduction to bayesian inference in econometrics*. Hoboken, NJ: Wiley-Interscience.

Zetzsche, C., & Barth, E. (1990). Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, *30*, 1111–1117.

Zetzsche, C., & Krieger, G. (1999). The atoms of vision: Cartesian or polar? *Journal of the Optical Society of America, A, 16*, 1554–1565.

Zoccolan, D., Cox, D., & DiCarlo, J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, *25*(36), 8150–8164.

Zografos, K. (1999). On maximum entropy characterization of Pearson's type II and VII multivariate distributions. *Journal of Multivariate Analysis*, *71*, 67–75.

Received September 25, 2010; accepted April 26, 2011.