Computer Science Technical Report TR2008-911, April 2008
Courant Institute of Mathematical Sciences, New York University
http://cs.nyu.edu/web/Research/TechReports/reports.html

Nonlinear extraction of 'Independent Components' of elliptically symmetric densities using radial Gaussianization

Siwei Lyu and Eero P. Simoncelli

Howard Hughes Medical Institute,
Center for Neural Science, and
Courant Institute for Mathematical Sciences
New York University

{*lsw, eero*}*@cns.nyu.edu*

April 23, 2008

Abstract

We consider the problem of efficiently encoding a signal by transforming it to a new representation whose components are statistically independent (also known as factorial). A widely studied family of solutions, generally known as independent components analysis (ICA), exists for the case when the signal is generated as a linear transformation of independent non-Gaussian sources. Here, we examine a complementary case, in which the signal density is non-Gaussian but elliptically symmetric. In this case, no linear transform suffices to properly decompose the signal into independent components, and thus, the ICA methodology fails. We show that a simple nonlinear transformation, which we call radial Gaussianization (RG), provides an exact solution for this case. We then examine this methodology in the context of natural image statistics, demonstrating that joint statistics of spatially proximal coefficients in a multi-scale image representation are better described as elliptical than factorial. We quantify this by showing that reduction in dependency achieved by RG is far greater than that achieved by ICA, for local spatial neighborhoods. We also show that the RG transformation may be closely approximated by *divisive normalization* transformations that have been used to model the nonlinear response properties of visual neurons, and that have been shown to reduce dependences between multi-scale image coefficients.

1 Introduction

Processing of signals is often facilitated by first transforming to a representation in which individual components are statistically independent. In such a "natural" coordinate system,

the components of the signal may be manipulated, transmitted or stored more efficiently. It has been proposed that this principle also plays an important role in the formation of biological perceptual systems (Attneave, 1954; Barlow, 1961). The problem of deriving an appropriate transformation for a given source, based on the statistics of observed samples, has been studied for more than a century. The classical solution, principal components analysis (PCA), is a linear transformation that is derived from the second-order signal statistics (i.e., the covariance structure). Although it may be computed for any source with finite variance, it is only guaranteed to fully eliminate dependencies for Gaussian sources.

Over the past twenty years, a more general family of methods known as *independent component analysis* (ICA) has been developed to handle the case when the signal is formed from linear combinations of independent non-Gaussian sources. Again, the solution is a linear transformation that is derived from statistical properties of the source (typically, the second-order statistics, augmented with a higher-order set of marginal measurements). ICA methods have shown success in blind signal separation problems(Comon, 1994), and in deriving bases for natural signals (Olshausen and Field, 1996; van der Schaaf and van Hateren, 1996; Bell and Sejnowski, 1997; Lewicki, 2002).

As with PCA, the ICA transformations may be computed for nearly any source, but they are only guaranteed to eliminate dependencies when the the assumed linear mixture of independent sources model is correct. And even in cases where the methodology seems to produce a sensible solution, the components of the resulting representation may be far from independent. A case in point is that of natural images, for which derived ICA transformations consist of localized oriented basis functions that appear similar to the receptive field descriptions of neurons in mammalian visual cortex (Olshausen and Field, 1996; Bell and Sejnowski, 1997; van der Schaaf and van Hateren, 1996). But the responses of such linear filters exhibit striking dependencies (Wegmann and Zetzsche, 1990; Zetzsche et al., 1993; Simoncelli, 1997; Buccigrossi and Simoncelli, 1999a), and although dependency between these responses is reduced compared to the original pixels (Zetzsche and Schönecker, 1987), such reduction is relatively small (Bethge, 2006). A number of recent attempts to model local image statistics have proposed the use of spherically or elliptically symmetric non-Gaussian densities, whose components exhibit clear dependencies (Zetzsche and Krieger, 1999; Wainwright and Simoncelli, 2000; Huang and Mumford, 1999; Parra et al., 2001; Hyvärinen et al., 2000; Srivastava et al., 2002; Sendur and Selesnick, 2002; Portilla et al., 2003; Teh et al., 2003; Gehler and Welling, 2006).

Here, we consider the factorization problem for the class of elliptically symmetric densities (ESDs). For this source model, we prove that linear transforms have no effect on the dependencies beyond second order, and thus that ICA decompositions offer no advantage over second-order decorrelation methods such as PCA. We introduce an alternative nonlinear procedure, which we call *radial Gaussianization* (RG), whereby the norms of whitened signal vectors are nonlinearly adjusted to ensure that the resulting output density is a spherical Gaussian, and thus factorized into independent components. We demon-

strate this methodology on data from photographic images. Using nonparametric estimates of the transformation, we show that RG produces much more substantial reductions in multi-information of pairs or blocks of nearby bandpass filter coefficients than does ICA.

Finally, we show that divisive normalization, which have previously been shown empirically to reduce higher-order dependencies in multi-scale image representations (Schwartz and Simoncelli, 2001; Wainwright et al., 2002; Malo et al., 2000b; Valerio and Navarro, 2003a; Gluckman, 2006; Lyu and Simoncelli, 2007), can approximate the RG transform. Thus, RG provides a more principled justification of these previous empirical results in terms of a specific source model.

2 Eliminating Dependencies with Linear Transforms

The problem of selecting a transformation that maps a source signal drawn from a known density to a new representation whose individual components are statistically independent is highly under-constrained (Hyvärinen and Pajunen, 1999). Indeed, even when one specifies a particular target density, there are an infinite number of transformations that can map a random variable associated with the input density into one associated with the target density. This is most easily understood in one dimension, where the well-known process of histogram equalization provides a natural choice for mapping any density to a uniform density (and from there, using an inverse equalization, to any desired density). The histogram equalization operation is clearly not unique, since (for example) it can be followed by any transformation that permutes equal-size intervals of source values, without affecting the uniform distribution of the output. The multiplicity of solutions for the density-mapping problem only becomes worse in higher dimensions.

An intuitively sensible means of selecting a transform from amongst the solution set is to require that it minimize the expected distortion of the original data: $min_f \mathbf{E}\left(|\vec{x} - f(\vec{x})|^2\right)$. This is analogous to solving an under-constrained linear inverse problem by selecting the solution with minimal norm. As an example, the histogram-equalization procedure mentioned above satisfies this property.

In practice, it is important to develop methods for selecting factorizing transforms based on observed data (i.e., samples from the source density). One can attempt to infer a (nonparametric) density from the data samples, but this is generally impractical for high-dimensional data. Instead, many well-known solutions may be derived by assuming the data are drawn from a source density that is a member of some parametric family, estimating the parameters, and then applying a transformation matched to the resulting density. In the following sections, we review several solutions to the problem of dependency elimination, emphasizing the underlying source model assumptions.

2.1 Multi-information

We quantify the statistical dependency for multi-variate sources using the *multi-information* (MI) (Studeny and Vejnarova, 1998), which is defined as the Kulback-Leibler divergence (Cover and Thomas, 2006) between the joint distribution and the product of its marginals:

$$\begin{aligned} I(\vec{x}) &= D_{\mathrm{KL}}\left(p(\vec{x}) \,\middle\|\, \prod_k p(x_k)\right) \\ &= \sum_{k=1}^{d} H(x_k) - H(\vec{x}), \end{aligned} \tag{1}$$

where $H(\vec{x})$ is the differential entropy of $\vec{x}$, and $H(x_k)$ denotes the differential entropy of the kth component of $\vec{x}$. In two dimensions, MI is equivalent to the mutual information (Cover and Thomas, 2006) between the two components.[1] From a coding perspective, MI measures the additional cost of encoding the components of $\vec{x}$ independently, as compared with the cost of jointly encoding $\vec{x}$. As a measure of statistical dependency among the elements of $\vec{x}$, MI is non-negative, and is zero if and only if the components of $\vec{x}$ are mutually independent. Furthermore, MI is invariant to any operation that operates on individual components of $\vec{x}$ (e.g., element-wise rescaling) since such operations produce an equal effect on the two terms in Eq. (1).

When $\vec{x}$ has finite second-order statistics, MI may be further decomposed into two parts, representing second-order and the higher-order dependencies, respectively, as:

$$I(\vec{x}) = \underbrace{\sum_{k=1}^{d} \log(\Sigma_{kk}) - \log|\Sigma|}_{second-order\ dependency} + \underbrace{D_{\mathrm{KL}}\left(p(\vec{x}) \,\|\, \mathcal{G}(\vec{x})\right) - \sum_{k=1}^{d} D_{\mathrm{KL}}\left(p(x_k) \,\|\, \mathcal{G}(x_k)\right)}_{higher-order\ dependency}, \tag{2}$$

where Σ is the covariance matrix of $\vec{x}$, defined as $E((\vec{x} - E\vec{x})(\vec{x} - E\vec{x})^T)$, and $\mathcal{G}(\vec{x})$ and $\mathcal{G}(x_k)$ are zero mean Gaussian densities with the same first and second order statistics (i.e., mean and covariance matrix) as $\vec{x}$ and x_k, respectively. The quantity $D_{\mathrm{KL}}\left(p(\vec{x}) \,\|\, \mathcal{G}(\vec{x})\right)$ is also known as the *negentropy*.

2.2 Principal Components Analysis

The most well-known solution to the dependency elimination problem corresponds to the case of a Gaussian source model for $\vec{x}$. In this case, the the higher-order terms in Eq.(2) are zero, and any linear transform that diagonalizes the covariance matrix is sufficient to completely eliminate statistical dependencies in $\vec{x}$. This is easily seen from the first two

[1] As such, multi-information is sometimes casually referred to as mutual information.

terms of Eq. (2): since the determinant of a diagonal matrix is the product of the diagonal elements, $\log|\Sigma|$ is equal to the sum of the log of the diagonal elements, $\sum_{k=1}^{d}\log(\Sigma_{kk})$. But there are an infinite number of linear transforms that can diagonalize the covariance matrix. Assuming that $\vec{x}$ has zero mean, the covariance matrix is written $\Sigma = E\{\vec{x}\vec{x}^T\}$. We may decompose the covariance matrix in terms of an orthogonal matrix of eigenvectors, U, and a diagonal matrix of eigenvalues, Λ, such that $\Sigma = U\Lambda U^T$. The classical solution, generally known as principal components analysis (PCA) (Jolliffe, 2002), transforms the data with the orthogonal eigenvector matrix, $\vec{y} = U^T\vec{x}$, resulting in a Gaussian density whose diagonal covariance matrix containing the eigenvalues.

2.3 Whitening and ZCA

The diagonalizing transform in PCA is often followed by a "whitening" step, in which each component is re-scaled by its standard deviation, $\vec{x}_{\text{wht}} = \Lambda^{-1/2}U^T\vec{x}$, ensuring that the components of the output signal have unit variance. A two-dimensional illustration of this two-step whitening procedure is illustrated in the left column of Fig. 1.

This whitening transform is not unique: Any matrix of the form $V\Lambda^{-1/2}U^T$ is a whitening transform, where V can be any orthogonal matrix. A common choice, known as zero-phase component analysis (ZCA) (Bell and Sejnowski, 1997), is to select $V = U$, which results in a symmetric transformation matrix. We show in Appendix A that ZCA is the whitening solution that satisfies the minimal distortion principal.

2.4 Independent Component Analysis

A PCA or whitening linear transformation is sufficient to remove all statistical dependencies in Gaussian variables, and it has appealing advantage of efficient computation due to numerical linear algebra. PCA may be applied to any source density with finite covariance, but it is not always guaranteed to generate a factorial output density. A natural question is whether there exists a class of non-Gaussian densities that can also be factorized with linear transforms, and if so, whether those transforms can be easily determined from data. Although PCA is roughly a century old, the answer to this question has only been expressed quite recently. Consider the family of source densities that is generated by linearly transforming a factorial source. That is, $\vec{x} = M\vec{s}$, where $p(\vec{s}) = \prod_k p(s_k)$. Clearly, when the matrix M is invertible, its inverse provides a linear transformation that can factorize the density $p(\vec{x})$ into the original scalar sources. The procedure for recovering the inverse transformation matrix, M^{-1}, and the original factorial source from data $\vec{x}$ is known as *Independent Components Analysis* (ICA) (Comon, 1994; Cardoso, 1999). For our purposes here, we assume M is square and invertible, although the ICA methodology may be generalized to arbitrary matrices.

The ICA computation can be better understood by expanding M in terms of its singular

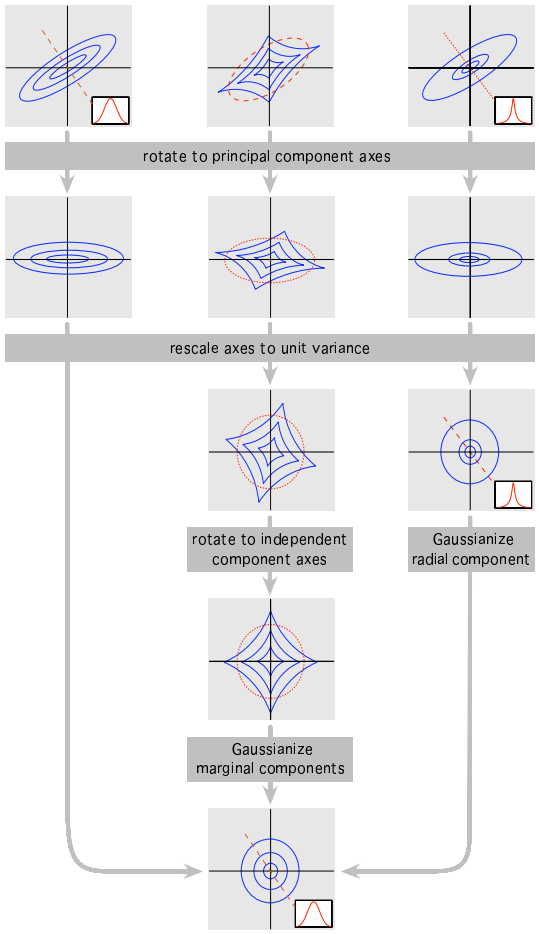

Figure 1: Three methods of dependency removal with their associated source models. Each consists of a sequence of transformations on multi-dimensional probability densities, depicted as two-dimensional contour plots. Red ellipses indicate covariance structure. Inset graphs indicate shape of a slice through the density along the indicated dashed red line. **Left:** Principal components analysis (PCA), applied to a Gaussian source. The first transformation rotates the coordinate system to the principal coordinate axes of the covariance ellipse, and the second rescales each axis according to its standard deviation. The output density is a spherical and unit variance Gaussian. **Middle:** Independent components analysis (ICA), applied to a linearly transformed factorial density. The first two steps are identical to the PCA case, and map the covariance ellipse (red) to the unit circle. The third is an additional rotation that aligns the source components with the axes of the space. A final nonlinear marginal transformation can be used to map the output density to a spherical Gaussian. **Right:** Radial Gaussianization (RG) applied to an elliptically symmetric (but non-Gaussian) density. The first two transformations are again identical to the PCA case. Finally, a nonlinear radial transformation is used to map the density to a spherical Gaussian.

value decomposition: $M = U\Lambda V^T$, where U and V are orthogonal matrices, and Λ is a diagonal matrix containing the singular values of M. Most ICA algorithms assume, without loss of generality, that the components of the initial source $\vec{s}$ are zero-mean with unit variance, and $E\{\vec{s}\vec{s}^T\} = I$. This implies that the transformation $\Lambda^{-1}U^T$ is a whitening operator for $\vec{x}$. The ICA transformation may thus be seen as a concatenation of a traditional whitening operation, followed by orthogonal transform V that eliminates the MI of the whitened variable $\vec{x}_{\text{wht}}$.

$$\begin{aligned} I(V\vec{x}_{\text{wht}}) &= \sum_{k=1}^{d} H((V\vec{x}_{\text{wht}})_k) - H(V\vec{x}_{\text{wht}}) \\ &= \sum_{k=1}^{d} H((V\vec{x}_{\text{wht}})_k) - H(\vec{x}_{\text{wht}}) - \langle \log|V| \rangle_{\vec{x}_{\text{wht}}} \\ &= \sum_{k=1}^{d} H((V\vec{x}_{\text{wht}})_k) - H(\vec{x}_{\text{wht}}). \end{aligned}$$

Since the second term is a constant with regards to V, finding V reduces to minimizing the first term (the sum of the transformed marginal entropies). While some ICA algorithms optimize the sum of marginal entropies directly, most implementations choose to optimize the expected value of a higher-order "contrast function" to avoid the difficulties associated with entropy estimation (Comon, 1994; Bell and Sejnowski, 1997; Cardoso, 1999). Though not usually included, a final nonlinear operation may be applied to map each of the independent components to a unit-variance Gaussian. This *marginal Gaussianization* procedure results in a factorial (and thus spherically symmetric) Gaussian density. This sequence of ICA operations is illustrated in the middle column of Fig. 1.

ICA is a natural generalization of PCA, and can be applied to an arbitrary source (as long as the covariance and the higher-order contrast exist). However, the components of the ICA-transformed source are only guaranteed to be independent when the source is indeed a linearly transformed random variable with a factorial density.

3 Eliminating Dependencies in Elliptical Symmetric Sources

The linear methods of dependency removal described in the previous section have been successfully applied to a wide range of problems across diverse disciplines. However, they are only guaranteed to remove dependencies of sources that are linearly transformed factorial densities (this includes both Gaussian and non-Gaussian cases), and thus it is worth considering dependency elimination for other source models. Here, we focus on a family that may be viewed as an alternative generalization of the Gaussian source model: the *elliptically symmetric density* (ESD) models.

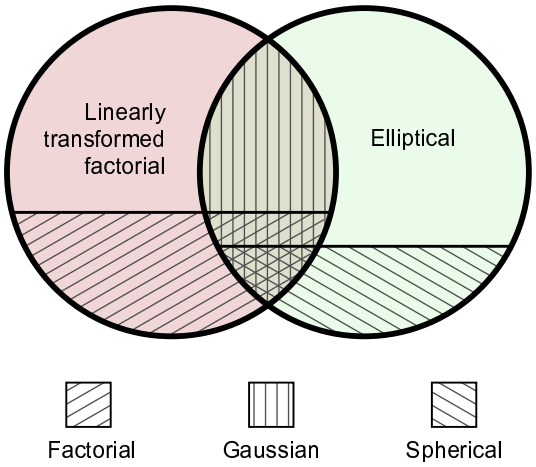

Figure 2: Venn diagram of the relationship between density models. The two circles represent the two density classes considered in this article: the linearly transformed factorial (independent) densities, and elliptically symmetric densities (ESDs). The intersection of these two classes is the set of all Gaussian densities. The factorial densities form a subset of the linearly transformed factorial densities (i.e., those transformed by a diagonal matrix), and the spherically symmetric densities form a subset of the ESDs.

3.1 Elliptically Symmetric Densities

The family of elliptically symmetric random vectors $\vec{x} \in \mathcal{R}^d$ are densities of the form:

$$p(\vec{x}) = \frac{1}{\alpha |\Sigma|^{\frac{1}{2}}} f(-\frac{1}{2}\vec{x}^T \Sigma^{-1} \vec{x}), \tag{3}$$

where Σ is a positive definite matrix (Fang et al., 1990). When $\vec{x}$ has finite second-order statistics, Σ is a multiple of the covariance matrix. With Σ fixed, $p(\vec{x})$ is completely determined by the generating function $f(\cdot) : \mathcal{R}^+ \cup \{0\} \mapsto \mathcal{R}^+ \cup \{0\}$, which has to satisfy $\int_0^\infty f(-r^2/2)\, r^{d-1}\, dr \leq \infty$. The normalizing constant α is then accordingly chosen so that the density integrates to one.

The definitive characteristic of ESDs is that the curves of constant probability are ellipsoids determined by Σ. When $\vec{x}$ is transformed with a whitening matrix as shown in section 2.3, the resulting density of $\vec{x}_{\text{wht}}$ is a special ESD whose Σ is a multiple of the d-dimensional identity matrix, and thus the density is spherically symmetric (also called "isotropic"). The level surfaces of a spherically symmetric density are then hyperspheres in the d-dimensional space (right column in Fig. 1).

When the generating function in the ESD definition is an exponential, the resulting ESD

is a multivariate Gaussian with zero mean and covariance matrix Σ. The same Gaussian variable $\vec{x}$ can also be regarded as a linear combination of independent Gaussian components $\vec{s}$ that each has a unit variance, as $\vec{x} = \Sigma^{-1/2}\vec{s}$. In fact, Gaussian is the only density that is both elliptically symmetric and linearly decomposable into independent components (Nash and Klamkin, 1976). In other words, the Gaussian densities correspond to the intersection of the ESDs and the linearly transformed factorial densities. Restricting this further, an isotropic Gaussian is the only density that is both spherically symmetric and factorial (i.e., with independent components). These relationships between Gaussians, ESDs, and the linearly transformed factorial densities are illustrated in the Venn diagram of Fig. 2.

Besides Gaussians, the ESD family also includes a variety of known densities and density families. Some of these have heavier tails than Gaussians, such the multi-variate Laplacian, the multi-variate Student's t, and the multi-variate Cauchy. More general leptokurtotic ESD families include the α-stable variables (Nolan, 2007) and the Gaussian scale mixtures (GSM) (Kingman, 1963; Yao, 1973; **?**). The ESDs also include densities with lighter tails than a Gaussian, such as the uniform density over the volume of a d-dimensional hyper-ellipsoid.

3.2 Linear Dependency Reduction for ESDs

As described in the section 2, linear transforms can be used to remove statistical dependencies of Gaussians (e.g., PCA, whitening and ZCA), as well as the more general class of linearly transformed factorial densities (ICA). In this section, we show that, apart from the special case of the Gaussian, linear transforms cannot eliminate the dependencies found in ESDs.

A linear whitening operation can be used to transform an elliptically symmetric variable to one that is spherically symmetric, thus eliminating the *second-order* dependencies of Eq.(2). But unlike the ICA case, there is no orthogonal matrix V that can affect the MI of the spherically symmetric density of $\vec{x}_{\text{wht}}$ (again, apart from the case of Gaussians, where $I(V\vec{x}_{\text{wht}})$ is always zero regardless the choice of V). The reason is simple: $p(\vec{x}_{\text{wht}})$ is isotropic (it is a function only of the vector length $|\vec{x}|$), and thus it is invariant under orthogonal linear transformation, as

$$\begin{aligned} p(V\vec{x}_{\text{wht}}) &= \frac{|V|}{\alpha} f(-(V\vec{x}_{\text{wht}})^T(V\vec{x}_{\text{wht}})/2) \\ &= \frac{1}{\alpha} f(-\vec{x}_{\text{wht}}^T V^T V \vec{x}_{\text{wht}}/2) \\ &= \frac{1}{\alpha} f(-\vec{x}_{\text{wht}}^T \vec{x}_{\text{wht}}/2) = p(\vec{x}_{\text{wht}}). \end{aligned}$$

Thus, $I(V\vec{x}_{\text{wht}}) = I(\vec{x})$, since the MI is function of the joint density.

3.3 Radial Gaussianization

Given that linear transforms are ineffective in removing dependencies from a spherically symmetric $\vec{x}_{\text{wht}}$ (and hence the original ESD variable $\vec{x}$), we need to consider non-linear mappings. As described previously, the Gaussian is the only spherically symmetric density that is also factorial. Thus, given a non-Gaussian spherically symmetric variable $\vec{x}_{\text{wht}}$, a natural solution for eliminating dependencies is to map it into a spherical Gaussian. As described in section 2, selecting such a non-linear mapping without any further constraint is a highly ill-posed problem. But in this case, we can select the mapping that satisfies the minimum distortion principle, which is the one that acts *radially*, mapping the radial density function, $f(|\vec{x}_{\text{wht}}|^2)$, to one corresponding to a spherical Gaussian density. We refer to this operation as *radial Gaussianization* (RG). Specifically, the RG transform is defined as:

$$\vec{x}_{\text{rg}} = g(\|\vec{x}_{\text{wht}}\|)\frac{\vec{x}_{\text{wht}}}{\|\vec{x}_{\text{wht}}\|}, \tag{4}$$

where the scalar function $g(\cdot)$ is chosen such that the resulting $\vec{x}_{\text{rg}}$ is an isotropic Gaussian variable. Specifically, note that the generating function for $p(\vec{x}_{\text{wht}})$ is related to the marginal distribution of $r = \|\vec{x}_{\text{wht}}\|$ as:

$$p_r(r) = \frac{r^{d-1}}{\beta} f(-r^2/2), \tag{5}$$

where $\Gamma(\cdot)$ is the standard Gamma function, and β is the normalizing constant that ensures that the density integrates to one. Thus, any $g(\cdot)$ that transforms $p_r(\cdot)$ to the corresponding radial marginal distribution of an isotropic Gaussian density with unit component variance, which is a *chi*-density with d degrees of freedom:

$$p_\chi(r) = \frac{r^{d-1}}{2^{d/2-1}\Gamma(d/2)} \exp(-r^2/2), \tag{6}$$

will transform $\vec{x}_{\text{wht}}$ to an isotropic Gaussian variable $\vec{x}_{\text{rg}}$. This is a one-dimensional density-mapping problem, and the classical solution (which satisfies the minimal distortion criterion) is the continuous monotonic mapping given by composition of the inverse cumulative density function of p_χ with the cumulative density function of p_r: as:

$$g(r) = F_\chi^{-1} F_r(r). \tag{7}$$

An example in the case of transforming a spherically symmetric two dimensional Student's t variable to the corresponding Gaussian variable is illustrated in Fig. 3.

Note that Eq.(4) is not the only operation that transforms $\vec{x}_{\text{wht}}$ to an isotropic Gaussian variable. For example, one could also transform the space with any orthogonal matrix V, since $V\vec{x}_{\text{rg}}$ is also an isotropic Gaussian variable. However, Eq.(4) is the solution that minimizes the expected distortion between $\vec{x}_{\text{wht}}$ and $V\vec{x}_{\text{rg}}$, measured by the mean square error $E(\|\vec{x}_{\text{wht}} - V\vec{x}_{\text{rg}}\|^2)$.

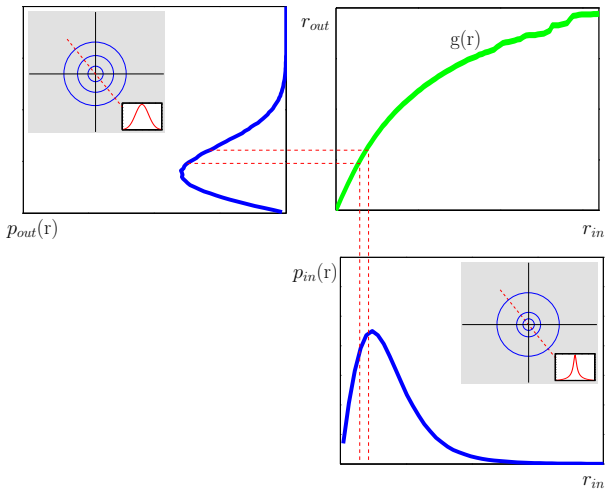

Figure 3: Radial Gaussianization procedure for 2D data. r_{in} are the radii of the input data, and r_{out} are the radii of RG transform output. $p_{out}(r)$ is a *chi*-density with one degree of freedom ($d-1$ for a d-dimensional space).

3.4 RG for General Signal Models

If $\vec{x}$ is not an elliptically symmetric variable, applying RG may not eliminate the higher-order dependencies. In fact, when $\vec{x}$ has a factorial joint density (i.e., the components are already independent), RG will generally *increase* dependency. We can quantify this by re-examining the decomposition of multi-information given in Eq. (2), applied to a whitened source (i.e., where the second-order terms have been eliminated):

$$I(\vec{x}) = D_{\mathrm{KL}}\left(p(\vec{x}) \parallel \mathcal{N}(\vec{x})\right) - \sum_{k=1}^{d} D_{\mathrm{KL}}\left(p(x_k) \parallel \mathcal{N}(x_k)\right). \tag{8}$$

Now rewrite $\vec{x}$ in generalized polar coordinates, $\vec{x} = r \cdot \vec{u}$, where $r = \|\vec{x}\|$, and $\vec{u}$ is a random vector on the surface of the unit hypersphere d-dimensions. For spherically symmetric densities, $p(\vec{x}) = p_\chi(r)\mathcal{U}(\vec{u})$, where $\mathcal{U}(\vec{u})$ denotes a uniform density on the hypersphere. We

can rewrite the first term of the MI expression of Eq. (8) in polar form:

$$\begin{aligned} D_{\mathrm{KL}}\left(p(\vec{x}) \parallel \mathcal{N}(\vec{x})\right) &= \int_{\vec{x}} p(\vec{x}) \log \frac{p(\vec{x})}{\mathcal{N}(\vec{x})} d\vec{x} \\ &= \int_{r,\vec{u}} p(r,\vec{u}) \log \frac{p(r,\vec{u})}{\mathcal{N}(r,\vec{u})} dr d\vec{u} \\ &= \int_{r} p_r(r) \log \frac{p_r(r)}{p_\chi(r)} dr + \int_{r,\vec{u}} p(r,\vec{u}) \log \frac{p(\vec{u}|r)}{\mathcal{N}(\vec{u}|r)} dr d\vec{u} \end{aligned}$$

Therefore the multi-information can be expanded as

$$I(\vec{x}) = D_{\mathrm{KL}}\left(p_r(r) \parallel p_\chi(r)\right) + \left\langle \frac{p(\vec{u}|r)}{\mathcal{U}(\vec{u})} \right\rangle_{\vec{x}} - \sum_{k=1}^{d} D_{\mathrm{KL}}\left(p(x_k) \parallel \mathcal{N}(x_k)\right). \tag{9}$$

The RG operation always eliminates the first term in Eq. (9). When $p(\vec{x})$ is elliptically symmetric, the second term is also zero, since the density of $\vec{u}$ is uniform and independent of r. Finally, for elliptically symmetric sources, the last term will also be zero, since RG ensures that the joint density is spherically Gaussian, and thus that the marginals will be Gaussian with unit variance.

For general sources, the second term of Eq. (9) is typically non-zero, but is not affected by a radial transform such as RG. On the other hand, the RG operation may actually increase the last term. When the density is close to elliptically symmetric, the increase in the last term may be relatively smaller than the reduction caused by the elimination of the first term, and thus RG may still achieve reduction in multi-information. But for densities that are close to factorial, it is possible that RG will result in a net increase in MI. Therefore, the effectiveness of RG in reducing higher-order dependency is determined by the underlying data model.

Summarizing, ICA and RG are procedures for dependency elimination, each producing an optimal result for a complementary generalization of the Gaussian source model, as illustrated to Fig. 2. Each can be optimized for, and applied to, data drawn from an appropriate source model. A natural question then arises: how relevant is the elliptically symmetric family (and the RG transformation) for real-world signals? In the next section, we examine this question in the context of photographic images.

4 Local Image Statistics

The characterization of statistical properties of images is of central importance in solving problems in image processing, and in understanding the design and functionality of biological visual systems. The problem has been studied for more than fifty years (see (Ruderman, 1996) or (Simoncelli and Olshausen, 2001) for reviews). Early analysis, developed in the

television engineering community, concentrated on second-order characterization of local pixel statistics. If one assumes translation-invariance (stationarity), then the Fourier basis should form a suitable Principal Components basis. In practice, when one computes a PCA solution on blocks drawn from photographic images, the basis functions appear as block-global oriented bandpass functions, but are typically not pure sinusoids due to the non-uniqueness of the eigenvalues.

Starting in the 1980's, researchers began to notice striking non-Gaussian behaviors of bandpass filter responses (Burt and Adelson, 1981; Field, 1987), and this led to an influential set of results obtained by using newly developed ICA methodologies to exploit these behaviors (Olshausen and Field, 1996; van der Schaaf and van Hateren, 1996; Bell and Sejnowski, 1997; Lewicki, 2002). These analyses generally produced basis sets containing oriented filters of different sizes with frequency bandwidths of roughly one octave. The nature of these results was widely hailed as a confirmation of central hypotheses that had become standard in both scientific and engineering communities. Specifically, the biological vision community had discovered neurons in the primary visual cortex of mammals whose primary response behaviors could be approximated by local oriented bandpass filters, and these were hypothesized to have been developed under evolutionary pressure as an efficient means of representing the visual environment (Barlow, 1961; Field, 1987). On the other hand, the computer vision and image processing communities (partly motivated by the biological observations, and partly by a desire to capture image features such as object boundaries) had long advocated the use of banks of local oriented filters for representation and analysis of image data (Koenderink, 1984; Granlund, 1978; Adelson et al., 1987; Mallat, 1989).

Despite the success of ICA methods in providing a fundamental motivation for the use of localized oriented filters, there are a number of simple observations that indicate inconsistencies in the interpretation. First, from a biological perspective, it seems odd that the analysis produces a solution that seems to bypass the retina and the lateral geniculate nucleus, two stages of processing that precede visual cortex and exhibit significant nonlinear behaviors in their own responses. Linear approximations of the response properties of these neurons are isotropic (i.e., non-oriented) bandpass filters. If the optimal decomposition for eliminating dependencies are oriented bandpass filters, why do we not see these in retina? Second, the responses of ICA or other bandpass oriented filters exhibit striking dependencies, in which the variance of one filter response can be predicted from the amplitude of another nearby filter response (Simoncelli and Buccigrossi, 1997; Buccigrossi and Simoncelli, 1999b). This suggests that although the histograms of responses of bandpass oriented filters are heavy-tailed, the joint histograms of pairs of responses are not consistent with the factorial source model assumed by ICA. A related observation is that the marginal distributions of a wide variety of bandpass filters (even a "filter" with randomly selected zero-mean weights) are *all* highly kurtotic (Zetzsche et al., 1997). This would not be expected for the ICA source model: projecting the local data onto a random direction should result in a density that

becomes more Gaussian as the neighborhood size increases, in accordance with a generalized version of the Central Limit Theorem (Feller, 1968). A recent quantitative study (Bethge, 2006) showed that the oriented band-pass filters obtained through ICA optimization lead to a surprisingly small improvement in terms of reduction in multi-information relative to second order decorrelation methods such as PCA. Taken together, all of these observations suggest that the filters obtained through ICA optimization are perhaps not as special as initially believed.

In fact, for some time there has been empirical evidence (along with associated modeling efforts) indicating that local joint densities of images are elliptically symmetric. This was first noted with regards to pairwise joint statistics of Gabor filters of differing phase (Wegmann and Zetzsche, 1990), and later extended to filters at nearby positions, orientations and scales (Wainwright and Simoncelli, 2000; Zetzsche and Krieger, 1999). As a result, many recent successful models of local image statistics are based on elliptically symmetric densities (Wainwright, 1999; Huang and Mumford, 1999; Parra et al., 2001; Hyvärinen et al., 2000; Srivastava et al., 2002; Sendur and Selesnick, 2002; Portilla et al., 2003; Teh et al., 2003). As introduced in the previous section of this article, this suggests that Radial Gaussianization may be an appropriate methodology for eliminating local statistical dependencies. In this section, we examine this hypothesis empirically, first by testing the local statistics of bandpass filter responses for ellipticity, and then by comparing the reduction in multi-information (MI) that is obtained using PCA, ICA and RG.

4.1 Ellipticity of Local Image Statistics

We examine the statistics of local blocks of image. We first remove the local mean by convolving with an isotropic band-pass filter[2]. that captures an annulus of frequencies in the Fourier domain ranging from $\pi/4$ to π radians/pixel. We applied this transformation to a set of different photographic images, commonly known as "Barbara", "boats", "cameraman","hill","Lena","baboon","pepper", and "house" [3]. These are 8-bit JPEG-compressed, and thus not directly representative of light intensities, but the results we report here are not significantly different when examined in intensity-calibrated images.

4.1.1 Sphericality of Pixel Pairs

We first examine the statistical properties of pairs of band-pass filter responses with different spatial separations. The two-dimensional densities of such pairs are easy to visualize, and can serve as an intuitive reference when we later extend to the multi-dimensional pixel blocks.

[2]Specifically, we use one subband of a non-oriented steerable pyramid (Simoncelli and Freeman, 1995).

[3]All images are available from Javier Portilla's web page at `http://www.io.csic.es/PagsPers/JPortilla/denoise/`.

The top row of Fig. 4 (labeled "raw") shows example contour plots of the joint histograms obtained from image "boats". The plots were arranged so that a 2D Gaussian density will have equally spaced contour lines. Consistent with previous empirical observations (Wegmann and Zetzsche, 1990; Wainwright and Simoncelli, 2000), the joint densities are non-Gaussian, and have roughly elliptical contours for nearby pairs. For pairs that are distant, both the second order correlation and the higher-order dependency become weaker and the corresponding joint and conditional histograms become more separable, as would be expected for two independent random variables.

The second row in Fig. 4 (labeled "ica") shows the ICA-transformed pairs, denoted as $\vec{x}_{\text{ica}}$, computed using the RADICAL algorithm (Learned-Miller and Fisher, 2000). RADICAL is an implementation of ICA that directly optimizes the MI instead of some surrogate contrast function, using a smoothed grid search over a non-parametric estimate of entropy. Note that for adjacent pairs, the transformed density does not become factorial: it has contours that are approximately circular, yet it is not an isotropic Gaussian, which is is the only case where both spherical symmetry and complete factorization are both satisfied. Thus ICA has not succeeded in removing higher-order dependencies. On the other hand, for pairs that are further apart, the raw density is more factorial, and remains relatively unchanged by the ICA transformation.

Next, we compare the distributions of the ICA-transformed pairs with those of synthesized data with related spherically symmetric or completely factorized distributions. Shown in the third row of Fig. 4 (labeled "ss", for spherically symmetric) are histograms of synthetic 2D samples that preserves the radial component of the ICA-transformed pairs with randomized orientations from samples of a uniform distribution on the unit circle. This implies these samples are from a spherically symmetric density that has the same radial marginal density as the ICA-transformed pairs. Shown in the next row (labeled as "fac", for factorial) are histograms of synthetic 2D samples that preserves the marginal distribution of the ICA-transformed pairs but with no inter-dependency. This results in a factorial density that has the same set of marginal densities as the ICA-transformed pairs. Comparing the histograms in the second, third and fourth rows, we see that the densities of the ICA-transformed adjacent pairs are much more similar to the spherically symmetric density than the factorial density. As the separation increases, the ICA-transformed density becomes less circular and starts to resemble the factorial density.

The isotropy of the above shown 2D joint densities can be further quantified by measuring the sample kurtosis of marginal projections in different directions[4]. The fifth row of Fig. 4 shows the kurtosis of the ICA-transformed pairs (black dashed curve) plotted as a function of marginalization direction. For the spherically symmetric densities of the

[4]We define kurtosis as the ratio between the fourth order centered moments and the squared second order centered moment (i.e., variances): $\kappa(x) = E\left\{(x - E(x))^4\right\} / \left(E\left\{(x - E(x))^2\right\}\right)^2$. With this definition, a Gaussian density has kurtosis of 3.

third row, the marginal kurtosis (blue solid curve) is constant with respect to direction, apart from fluctuations due to estimation errors from finite sampling. On the other hand, the kurtosis of the factorial density (red dashed curve) varies with rotation. For nearby pairs, the density of the ICA-transformed pairs is clearly seen to be better approximated by that of the spherically symmetric distributed samples than that of the factorial-distributed samples. As the distance increases, the marginal kurtosis of the ICA-transformed pairs fluctuates more and begins to resemble that of the factorial-distributed samples.

4.1.2 Sphericality of Pixel Blocks

The analysis of the previous section indicates that pairs of nearby pixels with mean removed have approximately spherical symmetric joint densities after whitening. But this does not necessarily guarantee the densities of $b \times b$ blocks of nearby pixels are also spherically symmetric (after whitening). To examine the isotropy of these densities, we computed the kurtosis for a large set of *random* projections. If the joint density has spherical symmetry, then the kurtosis (and all statistics) should be identical for marginals along any direction. On the other hand, for a non-Gaussian factorial density with identical marginals, such higher-order statistics will vary depending on how close a randomly chosen projection direction is to one of the cardinal axes. We can thus use the distribution of such higher-order statistics over random projections as an indicator of isotropy of the joint density.

Shown in Fig. 5 are distributions of kurtosis over $10^5 \times b^2$ random projections for square blocks of pixels, after their means removed with band-pass filtering and their covariance reduced to identity matrix with whitening. Further more, these blocks are transformed with ICA to align the cardinal axis with the most sparse marginals. In this case the ICA transform is implemented with the FastICA algorithm, which is more efficient and reliable for data of more than a few dimensions. Following (Bethge, 2006), we used contrast function $g(u) = 1 - \exp(-u^2)$ and the optimization was done using the symmetric approach. The factor of b^2 in the number of sampled projections compensates for the expected increase in sampling-induced variability that arises as the block size increases. In each plot, the black curves correspond to the ICA-transformed band-pass filtered pixel blocks. As in the pairwise case, the blue curves correspond to synthetic data sets of the same size and dimensionality that retain the radial density but with randomized orientations, whose joint density is spherically symmetric. Similarly, the red curves correspond to synthetic data sets formed by samples from a factorial density with the same marginal densities as the ICA-transformed pixel blocks.

These distributions of kurtosis can be used as an indicator of the Sphericality of the underlying joint density. Specifically, the mean of these distributions indicates average kurtosis over all marginals, and can be taken as a measure of the Gaussianity of "typical" projections of the data. To understand this, consider the curves (in red) corresponding to the factorized data. For small block sizes, the kurtosis varies substantially, ranging from roughly 3 to over

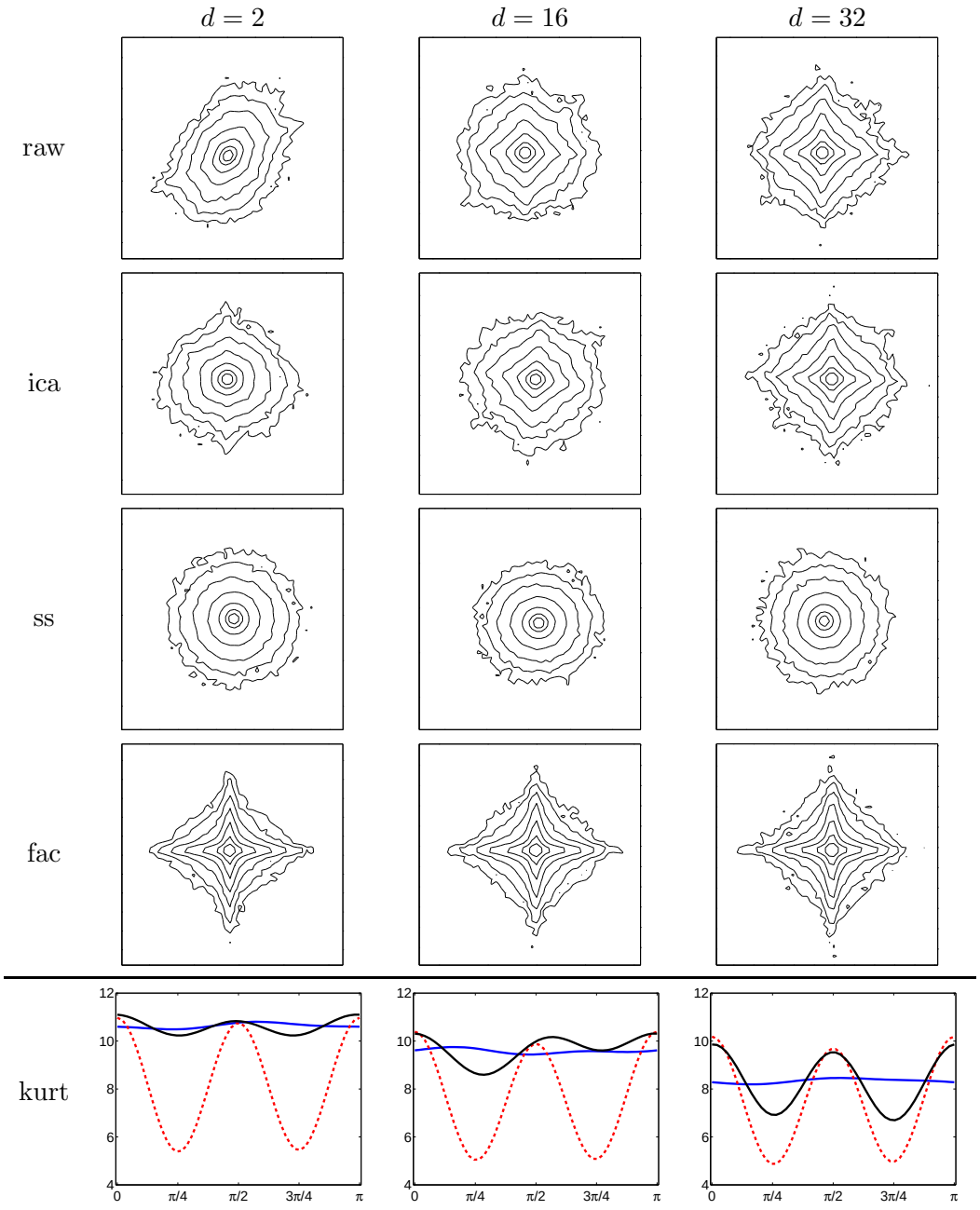

Figure 4: Contour plots of joint histograms of pairs of band-pass filter responses from an isotropic subband of the "boats" image with different spatial separations (given in units of pixels). See text for details.

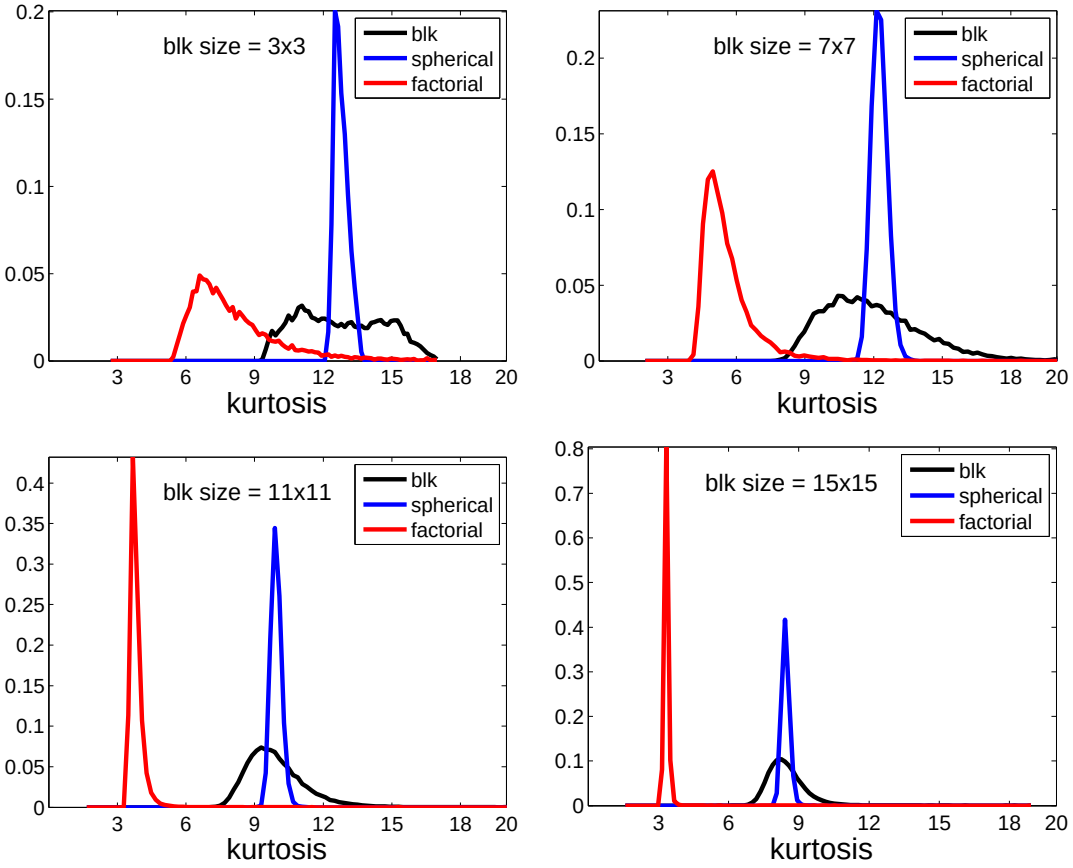

Figure 5: Distribution of kurtosis for whitened and ICA transformed pixel blocks (black), sphericalized samples (blue) and independent components (red). See text for details.

20. The large values correspond to marginal directions that are well-aligned with one of the cardinal axes. The smaller values correspond to marginal directions that are poorly aligned with the cardinal axes (e.g., the marginal along the direction $[1, 1, \cdots, 1]/\sqrt{N}$), and thus are averaging together the independent marginal variables. These averages tend to be significantly more Gaussian than the distributions along the cardinal axes, which is the underlying motivation for most ICA algorithms to find the most non-Gaussian axes. As the block size grows, alignment with the cardinal axes becomes rarer, and the distribution becomes more concentrated toward the kurtosis values expected of "typical" marginal directions. By the Central Limit theorem, these values converge to 3 as the dimensionality goes to infinity.

On the other hand, the width of the kurtosis distribution is determined by two factors: first, the variability of kurtosis due to changes in the shape of the marginal projections along different directions. And second, there is additional variability that arises from sampling. This portion of the variability may be seen directly in the distributions corresponding to the sphericalized data (blue curves). Since these distributions are spherically symmetric, all marginals should have the same kurtosis, and the only source of variability is due to sampling.

Given the above, consider the distributions of kurtoses for the ICA-transformed data (black dashed curves). For all block sizes, these have a mean that is similar to that of the sphericalized data, but consistently and substantially larger than that of the factorized data. We also see that the ICA-transformed data is not as concentrated as the sphericalized data. Thus, there is some real variation in kurtosis that cannot be attributed to sampling artifacts. Although this implies that the ICA-transformed distributions are not perfectly spherical, they are still much closer to spherical than factorial. This suggests that RG is likely to be more effective in reducing statistical dependencies than linear transforms such as PCA and ICA. In the next section, we test this assertion directly.

4.2 Reducing Local Image Dependencies with Radial Gaussianization

We begin by comparing the reduction of statistical dependency in pixel pairs using each of the methods described previously. We estimated the MI for $\vec{x}_{\text{raw}}$, $\vec{x}_{\text{wht}}$, $\vec{x}_{\text{ica}}$ and $\vec{x}_{\text{rg}}$ on pairs of band-pass filtered responses separated by distances ranging from 1 to 32 samples. Here, the MI was computed using a recent non-parametric method based on the order statistics (Kraskov et al., 2004). This approach belongs to the class of "binless" estimator of entropy and mutual information, which alleviates the strong bias and variance intrinsic to the more traditional binning (i.e., "plug-in") estimators. It is especially effective in this particular case, where the data dimension is two.

In order to implement RG, we have to estimate the radial component of the whitened source density from data. rom a set of training data $\{\vec{x}_1, \cdots, \vec{x}_n\}$, a trapezoidal approximation of F_r, $\hat{F}_r$, is obtained as following. First, we re-order the training data into

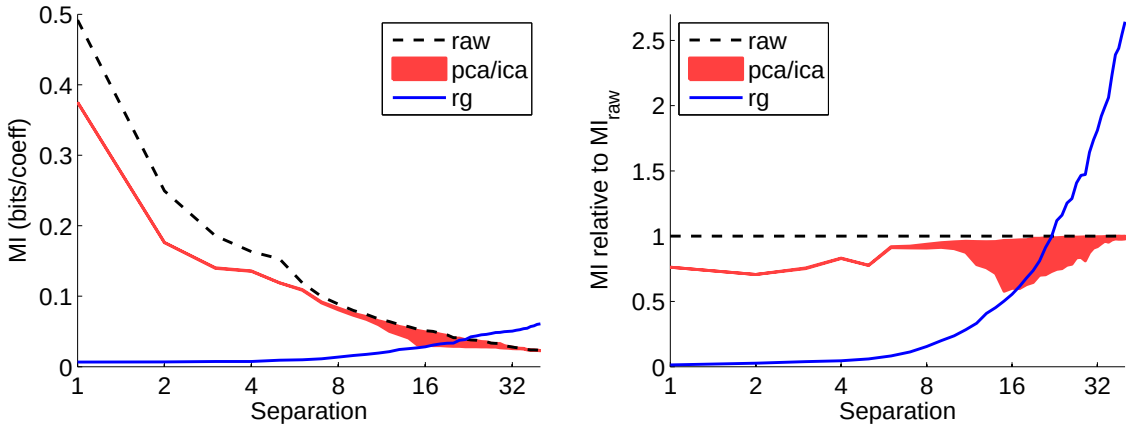

Figure 6: **Left:** MI (bits/pixel) for pairs of band-pass filtered pixels and their transformations with PCA, ICA and RG, as a function of spatial separation. **Right:** same data, re-plotted as a proportion of the MI of the band-pass filtered pixel pairs. The top of the red region corresponds to the results for PCA, and the bottom to those for ICA.

$\{\vec{x}_{i_1}, \cdots, \vec{x}_{i_n}\}$, such that $\|\vec{x}_{i_1}\| \leq \cdots \leq \|\vec{x}_{i_n}\|$. Then $\hat{F}_r$ is computed as

$$\hat{F}_r(r) = \begin{cases} 0 & r \leq \|\vec{x}_{i_1}\| \\ \frac{k}{n} & k = \text{argmax}_j\{j | \|\vec{x}_{i_j}\| \leq r\} \\ 1 & \|\vec{x}_{i_n}\| \leq r \end{cases} \tag{10}$$

In practice, if n is sufficiently large, the obtained $\hat{F}_r(r)$ will be smooth and a good approximation of $F_r(r)$. A non-parametric estimation of $F_\chi(r)$, $\hat{F}_\chi(r)$ can be obtained similarly by generating a set of d-dimensional isotropic Gaussian samples. From $\hat{F}_\chi(r)$ and $\hat{F}_r(r)$, a look-up table can be constructed with proper interpolation, as $\hat{g}(r) = \hat{F}_\chi^{-1}\hat{F}_r(r)$, to approximate the continuous function $g(r)$. It is also possible, though not necessary, to further reduce the complexity in specifying $\hat{g}$ by fitting it with piece-wise smooth functions (e.g., splines).

The averaged results over three images, are plotted in Fig. 6. First, we note that PCA produces a relatively modest reduction in MI: roughly 25% for small separations, decreasing gradually for larger separations. More surprisingly, ICA offers no additional reduction for small separations, and a relatively modest improvement for separations of between 12 and 32 samples. This is consistent with the histograms and kurtosis analysis shown in Fig. 4, which suggest that the joint density of adjacent pairs have roughly elliptical contours. As such, we should not expect ICA to provide much improvement beyond what is obtained with a whitening step.

The behavior for distant pairs is also consistent with the results shown in Fig. 4. These densities are roughly factorial, and thus require no further transformation to reduce MI. So ICA again provides no improvement, as is seen in the plots of Fig. 6 for separations beyond 32 samples. The behavior for intermediate separations is likely due to the fact that during the transition from spherical to factorial symmetry, there is a range where the final rotation transformation of ICA can result in a reduction in MI (e.g., middle columns Fig. 4).

In comparison to PCA and ICA, the non-linear RG transformation achieves an impressive reduction (nearly 100%) in MI for pairs separated by less than 16 samples. Beyond that distance, the joint densities are closer to factorial, and RG can actually make the pairs *more* dependent, as indicated by an increase in MI above that of the original pairs.

In the next set of experiments, we generalize our analysis to examine the effects of RG in reducing dependencies within pixel blocks. As with the kurtosis analyses of the previous section, the generalization from pairs to blocks is more difficult computationally, and more difficult to visualize. Specifically, direct estimation of the MI of pixel blocks becomes increasingly difficult (and less accurate) as the block size grows. This problem may be partially alleviated by instead evaluating and comparing *differences* in MI between different transforms. The details of this computation are provided in Appendix C.

For the sake of comparison, we use $\Delta I_{pca} = I(\vec{x}_{raw}) - I(\vec{x}_{pca})$ as a reference value, and compare this with $\Delta I_{ica} = I(\vec{x}_{raw}) - I(\vec{x}_{ica})$ and $\Delta I_{rg} = I(\vec{x}_{raw}) - I(\vec{x}_{rg})$. Shown in Fig.7 are scatter plots of ΔI_{pca} versus ΔI_{ica} (red circle) and ΔI_{rg} (blue cross) for various block sizes. Each point corresponds to MI computation over blocks from one of eight bandpass-filtered test images. As previously, the ICA algorithm was implemented with FastICA.

As shown in Fig. 7, for small block sizes (e.g., 3×3), RG achieved a significant reduction in MI (roughly double that of PCA) , whereas ICA shows only a small improvement over PCA. Since PCA-based whitening is usually used as preprocessing step for ICA, this suggests that ICA algorithm does not offer much advantage over second-order decorrelation algorithms such as PCA. Similar results were also obtained with the means of each block removed in a slight different manner in (Bethge, 2006). These results may be attributed to the fact that the joint density for small pixel blocks tend to be roughly elliptical. It also suggests the amount of higher order dependency in these blocks is significant compared to the second order correlations measured by the MI reduction of PCA. On the other hand, as the block size increases, the advantage of RG in reducing statistical dependency fades, consistent with the fact that the pairwise densities for coefficients become less elliptical with distance, and thus the multi-dimensional joint density of larger blocks will tend to deviate more from elliptically symmetric.

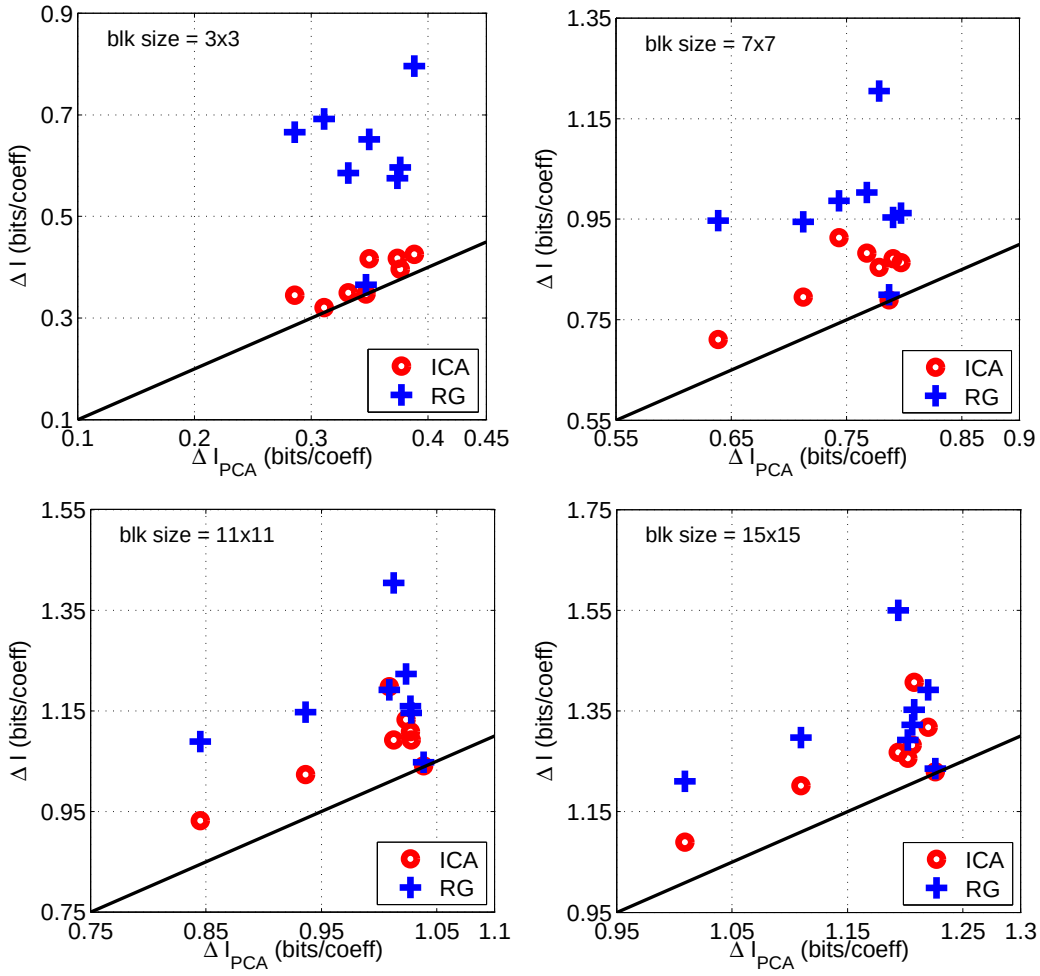

Figure 7: Reduction of MI with PCA, ICA and RG for pixel blocks of various sizes. The x-axis corresponds to ΔI_{pca}, the plus denotes ΔI_{rg}, and circles denotes ΔI_{ica}.

5 Relationship to divisive normalization

In recent years, a local gain control model, sometimes known as *divisive normalization* (DN) has become popular for modeling biological vision. In DN, responses of a band-pass filter are divided by a Minkowski combination of a cluster of neighboring response amplitudes. This type of model has been used to explain nonlinearities in the responses of mammalian cortical neurons (Heeger, 1992; Geisler and Albrecht, 1992), and nonlinear masking phenomenon in human visual perception (Foley, 1994; Watson and Solomon, 1997; Teo and Heeger, 1994). Statistically, it's been shown that locally dividing bandpass-filtered images by local standard deviation can produce approximately Gaussian marginal distributions(Ruderman, 1996), and that a weighted DN nonlinearity can reduce statistical dependencies of oriented bandpass filter responses (Simoncelli, 1997; Buccigrossi and Simoncelli, 1999a; Malo et al., 2000a; Schwartz and Simoncelli, 2001; Valerio and Navarro, 2003b). Recently, several authors have developed invertible image transformations that incorporate DN (Valerio and Navarro, 2003a; Malo et al., 2006; Gluckman, 2006; Lyu and Simoncelli, 2007). Since DN provides a nonlinear means of reducing dependencies in bandpass representations of images, it is natural to ask how it is related to the RG methodology introduced in this article.

Given decorrelated input variable $\vec{x} \in \mathcal{R}^d$, we define the DN transform as (Simoncelli, 1997):

$$r_i = \frac{x_i}{(b + \sum_j c_j x_j^2)^{1/2}}, \quad \text{for} \quad i = 1, \cdots, d. \tag{11}$$

where c_i and b are the transform parameters[5]. When the weights are all identical ($c_i = c, \forall i$), DN becomes a radial transform:

$$\phi_{dn}(\vec{x}) = g_{dn}(\|\vec{x}\|)\frac{\vec{x}}{\|\vec{x}\|}, \tag{12}$$

where

$$g_{dn}(r) = \frac{r}{\sqrt{b + cr^2}}, \tag{13}$$

with scalars b and c as transform parameters.

In practice, the transform parameters in the DN transform are learned from a set of data samples. Previously, the DN parameter learning problem was formulated to maximize likelihood, where specific marginals were assumed for r_i (Schwartz and Simoncelli, 2001;

[5]For biological modeling, the DN transform is commonly defined using a rectified numerator:

$$R_i = \frac{\text{sign}(x_i)x_i^2}{b + \sum_j c_j x_j^2}$$

(Schwartz and Simoncelli, 2001). Note that R_i can be mapped to r_i using a point operation: $r_i = \text{sign}(R_i) \cdot \sqrt{|R_i|}$. As MI is not affected by point-wise operations, we may choose either $(r_1, \cdots, r_d)$ or $(R_1, \cdots, R_d)$ for the analysis of dependency reduction.

Wainwright et al., 2002). In this work, we employ an alternative learning scheme that explicitly optimize the DN transform parameters with regards to the reduction of MI. Specifically, we optimize the difference in MI from input data $\vec{x}$ to the DN transformed data $\vec{y} = \phi_{dn}(\vec{x})$:

$$\Delta I = I(\vec{x}) - I(\vec{y}) = \sum_{i=1}^{d} H(x_i) - \sum_{i=1}^{d} H(y_i) + \left\langle \log \left| \det \left(\frac{\partial \vec{y}}{\partial \vec{x}} \right) \right| \right\rangle_{\vec{x}}. \tag{14}$$

Note that $\sum_{i=1}^{d} H(x_i)$ is a constant with regards to the DN transform parameters, and the Jacobian of DN transform is given as (see Appendix):

$$\det \left(\frac{\partial \vec{y}}{\partial \vec{x}} \right) = \frac{b}{(b + cr^2)^{d/2+1}},$$

the optimization is reduced to

$$\underset{b,c}{\operatorname{argmax}} - \sum_{i=1}^{d} H(y_i) + \log b - (d/2 + 1)\langle \log(b + cr^2) \rangle_r. \tag{15}$$

We then apply a grid search for the (b, c) values that maximizes Eq.(15), where the expectation over r is replaced by averaging oner training data, and the entropy $H(y_i)$ is computed using a non-parametric m-spacing estimator.

Fig.8 shows two comparisons of optimal RG and DN transformations. The first shows results obtained by optimizing over 10^5 25-D multivariate Student's t samples. The multivariate Student's t density is a member of the elliptically symmetric family, and its MI can be computed in closed form (see Appendix). Note that for relative small of r, the DN radial map closely approximates the RG radial transform. But we also see that the DN radial transform saturates at large values while the RG radial transform continues to increase. Finally, note that DN eliminates only about half of the MI, whereas RG eliminates nearly all of it.

The right side of Fig. 8 shows a comparison of RG and DN applied to image "boats". Similar to the case of multi-variate Student's t, the DN radial transform is approximates the RG radial transform, and reduces a substantial fraction of the MI. Nevertheless, it falls significantly short of the performance of the RG transform.

More generally, we can show that the functional form of the DN transform suggests that it cannot remove the dependencies of spherical densities. Specifically, the radial transform in RG, $g(\cdot)$, is a monotonic bijection (one-to-one map) from $[0, \infty)$ to $[0, \infty)$, for the simple reason that the support of the target χ-distribution takes $[0, \infty)$ as its domain. On the other hand, the radial transform in DN, expressed in Eq.(13), saturates for large values of r, i.e., there exist a constant C such that for any $r \in [0, \infty)$, $g_{dn}(\cdot) \leq C$. Therefore, DN transform cannot be used to gaussianize elliptically symmetric variables, and thus is not able to completely remove their higher-order statistical dependencies.

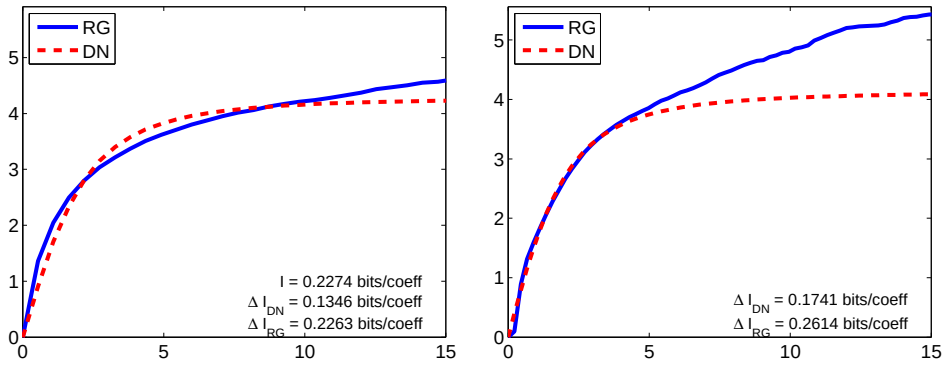

Figure 8: Comparison of radial transforms corresponding to RG (blue solid) and DN (red dashed) optimized to minimize MI for 10^5 samples of a 25-dimensional spherical Student's t density (left), and 5×5 pixel blocks from the the bandpass filtered "boats" image (right). Inset boxes indicate information reduction.

6 Discussion

We have introduced a new form of statistically-adaptive signal transformation, which is able to remove dependencies of sources with elliptically symmetric densities. The methodology is complementary to the ICA approach, which is effective for linearly transformed factorial sources, but ineffective for ESDs. The optimal transformation, which corresponds to a nonlinear adjustment of the signal amplitude, may be estimated non-parametrically from sample data. Although we have not explored it here, it may be possible to express this nonlinear transformation in closed form for specific sub-families of ESD.

An important aspect of our development of this topic is the emphasis on source models. The RG transformation may be applied to data from any source, but it is only guaranteed to produce independent responses when the source is elliptically symmetric, and it may actually increase dependencies of certain class of source models. Thus, RG cannot be applied blindly to data, but requires diagnostic tests to verify that the data are sufficiently close to elliptical.

We have shown that this transformation is highly effective at removing dependencies within local blocks of bandpass filtered images, much more so than ICA or sparse coding methods, which are, in turn, only slightly better than PCA. This may be seen as clear evidence that images are not well-modeled as linear combinations of sparse and/or independent sources. Furthermore, the resulting transformation on the signal amplitude is similar to the divisive normalization operation that has been used to model the responses properties of

visual neurons. But unlike divisive normalization, RG is derived as an optimal procedure for a specific family of density models.

There are several nonlinear methods for dependency removal in the literature. Kernel PCA methods (Mika et al., 1999) operate by nonlinearly transforming the data to a space where PCA is used to remove any remaining dependencies. The concept is quite general, but success relies on choosing nonlinear kernel functions that are capable of Gaussianizing the data. Chen and Gopinath proposed an iterative scheme that alternate between ICA transformations and marginal Gaussianization (Chen and Gopinath, 2000). Although this method is guaranteed to converge for any data source, the overall transformation (which is a composition of the iterated alternating sequence of linear transforms and marginal nonlinearities) is difficult to interpret and will generally result in a substantial distortion of the original source space. This is especially true for elliptically symmetric sources. RG, while only guaranteed to work correctly for elliptically symmetric densities, is a one-step procedure.

Some authors have proposed the use of spherically symmetric densities for representation of image features that are invariant to phases or orientations (Zetzsche and Barth, 1990; Kohonen, 1996; Zetzsche and Krieger, 1999; Hyvärinen et al., 2000, e.g.,). In addition, several recent approaches for unsupervised learning of image structures arrive at related local descriptions. Specifically, independent subspace analysis(Hyvärinen and Hoyer, 2000), topographical ICA (Hyvärinen et al., 2000), and hierarchical scale mixture models (Karklin and Lewicki, 2005) each assume that image data are generated from linearly transformed densities which are formed by combining clusters of variables whose dependency cannot be reduced by linear transform. In all three cases, we believe the densities of these local clusters are approximately elliptical, and thus the RG methodology may be relevant for eliminating the dependencies captured by these generative models.

There are a number of extensions of RG that are worth considering, in the context of image representation. First, we see that RG substantially reduces the multi-information for small blocks, but that performance worsens as the block size increases. The RG solution cannot provide a global solution for removing independence from images. A natural means of extending a local statistical model is through Markov Random Fields, and we have begun exploring such extensions, based on our previous work (Lyu and Simoncelli, 2008). Second, since the RG methodology generates factorial responses, it provides a solution to the Efficient Coding problem for elliptical signals, in the noise-free case (Barlow, 1961; Simoncelli and Olshausen, 2001). It is important to examine how this solution would be affected by the incorporation of sensor noise and/or channel noise. And finally, we are currently examining the statistics of images after local RG transformations, with the expectation that remaining statistical regularities (e.g., orientation and phase dependencies) can be studied, modeled and removed with additional transformations.

A ZCA

In this appendix, we show that ZCA is the whitening transform that minimizes expected square distortion of the transformed data. We formulate this with an optimization:

$$\min_W E(\|\vec{x} - W\vec{x}\|^2) \;\; \text{s.t.} \;\; E((W\vec{x})(W\vec{x})^T) = I. \tag{16}$$

First, note that

$$E((W\vec{x})(W\vec{x})^T) = I \Rightarrow WE(\vec{x}\vec{x}^T)W^T = I \Rightarrow E(\vec{x}\vec{x}^T) = W^{-1}W^{-T}.$$

Denote $C = E(\vec{x}\vec{x}^T)$, we have

$$C = W^{-1}W^{-T}. \tag{17}$$

Next, we rewrite the objective function in Eq.(16) as:

$$\begin{aligned} E(\|\vec{x} - W\vec{x}\|^2) &= E((\vec{x} - W\vec{x})^T(\vec{x} - W\vec{x})) \\ &= E(\text{tr}[(\vec{x} - W\vec{x})(\vec{x} - W\vec{x})^T]) \\ &= tr[E((I - W)\vec{x}\vec{x}^T(I - W)^T))] \\ &= tr[(I - W)E(\vec{x}\vec{x}^T)(I - W)^T)] \\ &= tr[(I - W)C(I - W)^T)] \end{aligned}$$

Using Eq.(17), we further transform

$$\begin{aligned} E(\|\vec{x} - W\vec{x}\|^2) &= tr[(I - W)W^{-1}W^{-T}(I - W)^T)] \\ &= tr[W^{-1}W^{-T} + I - W^{-1} - W^{-T}] \\ &= tr[W^{-1}W^{-T}] + d - 2\text{tr}[W^{-1}]. \end{aligned}$$

Now using Eq.(17) again, we have

$$E(\|\vec{x} - W\vec{x}\|^2) = \text{tr}[C] + d - 2tr[W^{-1}]. \tag{18}$$

Therefore, minimizing $E(\|\vec{x} - W\vec{x}\|^2)$ is equivalent to maximizing $\text{tr}[W^{-1}]$ as the first two terms are constants. Now denote $A = W^{-1}$, the original optimization becomes

$$\max_A \text{tr}[A] \;\; \text{s.t.} \;\; AA^T = C. \tag{19}$$

We solve Eq.(19) by introducing Lagrangian multipliers M_{ij} which forms a matrix M. Note that M is a symmetric matrix as the constraints are symmetric. The Lagrangian of Eq.(19) is then

$$L(A) = \text{tr}[A] - tr[M^T(AA^T - C)]. \tag{20}$$

Taking derivative of Eq.(20) with regards to A and setting the result to zero yeild

$$\frac{\partial L(A)}{\partial A} = 0 \Rightarrow I - (M^T A + MA) = 0. \tag{21}$$

As M is symmetric, this implies $A = -\frac{1}{2}M^{-1}$, and that A as well as W should also be symmetric. We write the eigen-decomposition of $W = U\Gamma U^T$, the optimal solution to Eq.(16) is obtained as

$$WCW^T = I \Rightarrow U\Gamma U^T C U \Gamma U^T = I \Rightarrow U^T C U = \Gamma^{-2}. \tag{22}$$

Therefore, U are the eigen-vectors of C and Γ contains the inverse of eigen-values of C. This shows that W is the square root of C and thus the ZCA solution.

B Jacobian of Radial Transform

In this appendix, we show the Jacobian of a general radial transform as given in Eq.(4). Denote $\vec{y} = \phi(\vec{x})$, we aim to compute $\det\left(\frac{\partial \vec{y}}{\partial \vec{x}}\right)$. Note that $\left(\frac{\partial \vec{y}}{\partial \vec{x}}\right)_{ij} = \frac{\partial y_i}{\partial x_j}$. We will thus compute $\frac{\partial y_i}{\partial x_j}$ for $i \neq j$ and $\frac{\partial y_i}{\partial x_i}$.

First, we have

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j}\left[g(\|\vec{x}\|)\frac{\vec{x}}{\|\vec{x}\|}\right] = x_i \frac{\partial \|\vec{x}\|}{\partial x_j}\left[\frac{g'(\|\vec{x}\|)}{\|\vec{x}\|} - \frac{g(\|\vec{x}\|)}{\|\vec{x}\|^2}\right].$$

As $\frac{\partial \|\vec{x}\|}{\partial x_j} = \frac{x_j}{\|\vec{x}\|}$,

$$\frac{\partial y_i}{\partial x_j} = \frac{x_i x_j}{\|\vec{x}\|}\left[\frac{g'(\|\vec{x}\|)}{\|\vec{x}\|} - \frac{g(\|\vec{x}\|)}{\|\vec{x}\|^2}\right].$$

Similarly,

$$\frac{\partial y_i}{\partial x_i} = \frac{\partial}{\partial x_i}\left[g(\|\vec{x}\|)\frac{\vec{x}}{\|\vec{x}\|}\right] = \frac{g(\|\vec{x}\|)}{\|\vec{x}\|} + \frac{x_i^2}{\|\vec{x}\|}\left[\frac{g'(\|\vec{x}\|)}{\|\vec{x}\|} - \frac{g(\|\vec{x}\|)}{\|\vec{x}\|^2}\right].$$

Writing in matrix form, and denote $r = \|\vec{x}\|$, we have

$$\frac{\partial \vec{y}}{\partial r} = \frac{g(r)}{r} I_d + \frac{\vec{x}\vec{x}^T}{r}\left[\frac{g'(r)}{r} - \frac{g(r)}{r^2}\right], \tag{23}$$

where I_d is d-dimensional identity matrix.

Making use of identity $\det(aI_d + b\vec{x}\vec{x}^T) = a^{d-1}(a + b\vec{x}^T\vec{x})$ (Abadir and Magnus, 2005), we can compute

$$\det\left(\frac{\partial \vec{y}}{\partial r}\right) = \left(\frac{g(r)}{r}\right)^{d-1}\left[\frac{g(r)}{r} + r\left[\frac{g'(r)}{r} - \frac{g(r)}{r^2}\right]\right] = g'(r)\left(\frac{g(r)}{r}\right)^{d-1}.$$

When the radial transform in question is in the form of divisive normalization, Eq.(12), where g has form in Eq.(13), is

$$\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right) = \frac{b}{(b+cr^2)^{d/2+1}},$$

C Computing Differences in Multi-information

Direct estimation or optimization of multi-information is challenging, especially for high-dimensional data. On the other hand, under our current setting, we are not required to compute the multi-information directly, but rather more interested in the reduction of multi-information using different methods. Therefore, it suffices to compute the *difference* in multi-information between raw data and transformed data.

For an invertible transform $\phi : \mathcal{R}^d \mapsto \mathcal{R}^d$, the change in multi-information from $\vec{x}$ to its transformation $\vec{y} = \phi(\vec{x})$ is given as:

$$\begin{aligned}
\Delta I &= I(\vec{x}) - I(\vec{y}) \\
&= \sum_{i=1}^{d} H(x_i) - H(\vec{x}) - \left[\sum_{i=1}^{d} H(y_i) - H(\vec{y})\right] \\
&= \sum_{i=1}^{d} H(x_i) - \sum_{i=1}^{d} H(y_i) - \int_{\vec{y}} p(\vec{y}) \log p(\vec{y}) d\vec{y} - H(\vec{x}) \\
&= \sum_{i=1}^{d} H(x_i) - \sum_{i=1}^{d} H(y_i) - \int_{\vec{x}} p(\vec{x}) \log \frac{p(\vec{x})}{\left|\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right)\right|} d\vec{x} - H(\vec{x}) \\
&= \sum_{i=1}^{d} H(x_i) - \sum_{i=1}^{d} H(y_i) + \int_{\vec{x}} p(\vec{x}) \log \left|\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right)\right| d\vec{x} \\
&= \sum_{i=1}^{d} H(x_i) - \sum_{i=1}^{d} H(y_i) + \left\langle \log \left|\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right)\right| \right\rangle_{\vec{x}}.
\end{aligned}$$

Therefore, the computation of ΔI can be split into two parts: (1) estimating marginal entropies for the input and transformed variables, $H(x_i)$ and $H(y_i)$, and (2) computing the expected log Jacobian $\left\langle \log \left|\frac{\partial\vec{y}}{\partial\vec{x}}\right| \right\rangle_{\vec{x}}$.

C.1 Entropy Estimation

To estimate the entropy for the 1D marginal densities $p(x_i)$ and $p(y_i)$, we employed the non-parametric m-spacing entropy estimator (Vasicek, 1976). We briefly describe this algorithm

density (ent = 1)	bias	$\sqrt{\text{var}}$	mse
Gaussian	2.43×10^{-3}	3.17×10^{-3}	3.99×10^{-3}
Laplacian	5.72×10^{-3}	7.49×10^{-3}	9.42×10^{-3}
Student t	4.36×10^{-3}	6.38×10^{-3}	7.73×10^{-3}

Table 1: Bias and variance of the non-parametric entropy estimator on several example densities. The unit for entropy is nat. Note that $\text{mse}^2 = \text{bias}^2 + \text{var}$.

here: a more comprehensive tutorial can be found at (Learned-Miller and Fisher, 2000). In the m-spacing entropy estimation, one is given a set of *i.i.d.* data samples $(x_1, \cdots, x_N)$. The first step is to sort data into $z_1 \leq \cdots \leq z_N$. Next, we choose an integer m to form the m-spacing entropy estimator as:

$$\hat{H}(z_1, \cdots, z_N) = \frac{1}{N} \sum_{i=1}^{N-m} \log \left(\frac{N}{m} [z_{i+m} - z_i] \right) - \psi(m) + log(m) \tag{24}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ is the digamma function. The m-spacing estimator is strongly consistent, i.e. as $m \to \infty$, and $m/N \to 0$, $\hat{H}(z_1, \cdots, z_N) \to H(z)$ with probability 1. We set $m = \sqrt{N}$ in our implementation.

To evaluate the bias and variance of this estimator we tested it on several densities whose entropy can be computed in closed-form. Specifically, we generated 10^5 random trials of 2.5×10^5 samples, which is in the same order as the amount of data used in subsequent experiments, from a Gaussian, a Laplacian and a Student's t density, respectively. Differential entropies of these densities afford closed-forms and can be found in (Cover and Thomas, 2006). The parameters of each density is chosen so that they all have entropy 1. As summarized in Table 1, the m-spacing estimator leads to relatively small bias and variances in the estimation.

For $\vec{x}_{raw}$ and its linear transformations $\vec{x}_{pca}$ and $\vec{x}_{ica}$, it has been observed that the marginals can be well fitted by the generalized Laplacian family (also known as the generalized Gaussians, or stretched exponential densities) (Mallat, 1989; Simoncelli and Adelson, 1996; Huang and Mumford, 1999):

$$p(x; p, s) = \frac{p}{2s\Gamma(1/p)} \exp\left(- \left(\frac{|x|}{s} \right)^p \right), \tag{25}$$

which is determined by the shape parameter p and scale s. This suggests that we can estimate the marginal entropy with a parametric approach, where we can first fit the generalized Laplacian to the marginals by maximizing likelihood, and then compute the differential en-

tropy as (Farvardin and Modestino, 1984):

$$H(x) = \frac{1}{p} - \log\left(\frac{p}{2s\Gamma(1/p)}\right). \tag{26}$$

In practice, this parametric entropy estimator yields results that are close to that of the non-parametric estimator when the marginal density of the source is well-approximated by the generalized Gaussian family. However, for non-linear transformed $\vec{x}_{rg}$, especially when the coefficient pairs are far apart or the size of coefficient blocks is large, the generalized Laplacian is a poor fit to the marginals, and thus the parametric estimator may introduce a large bias in estimation. For this reason, in subsequent numerical experiments, we report only the results with the non-parametric estimator, as it is more flexible in dealing with data for which no prior knowledge of marginal densities is available.

C.2 Computing Expected Log Jacobian

For linear transforms, the log Jacobian, $\log\left|\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right)\right|$, is a constant equal to the log determinant of the transform matrix. Note that when the linear transform is orthonormal the log Jacobian is zero.

For the nonlinear RG transform, the log Jacobian can be directly computed from the radial transform, $g(r)$, as:

$$\log\left|\det\left(\frac{\partial\vec{y}}{\partial\vec{x}}\right)\right| = \log g'(r) + (d-1)\log\frac{g(r)}{r}. \tag{27}$$

where $r = \|\vec{x}\|$. Then the expectation over $\vec{x}$ of the log Jacobian in this case is computed as

$$\left\langle \log\left|\frac{\partial\vec{y}}{\partial\vec{x}}\right|\right\rangle_{\vec{x}} = \langle \log g'(r)\rangle_r + (d-1)\left\langle \log\frac{g(r)}{r}\right\rangle_r.$$

In practical implementation, the differentiation is computed numerically. The expectation is implemented by averaging over radii of training data.

References

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*. Cambridge.

Adelson, E. H., Simoncelli, E. P., and Hingorani, R. (1987). Orthogonal pyramid transforms for image coding. In *Proc SPIE Visual Communications and Image Processing II*, volume 845, pages 50–58, Cambridge, MA.

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

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

Bell, A. J. and 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.

Buccigrossi, R. W. and Simoncelli, E. P. (1999a). Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12):1688–1701.

Buccigrossi, R. W. and Simoncelli, E. P. (1999b). Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans Image Proc*, 8(12):1688–1701.

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

Cardoso, J.-F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192.

Chen, S. S. and Gopinath, R. A. (2000). Gaussianization. In *Advances in Neural Computation Systems (NIPS)*, pages 423–429.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Process.*, 36:387–314.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory*. Wiley-Interscience, 2nd edition.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

Farvardin, N. and Modestino, J. W. (1984). Optimum quantizer performance for a class of non-gaussian memoryless sources. *IEEE Transactions on Information Theory*, 30(3):485–496.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394.

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

Gehler, P. and Welling, M. (2006). Products of "edge-perts". In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 419–426. MIT Press, Cambridge, MA.

Geisler, W. S. and Albrecht, D. G. (1992). Cortical neurons: Isolation of contrast gain control. *Vision Research*, 8:1409–1410.

Gluckman, J. M. (2006). Higher order pyramids: an early vision representation. In *European Conference on Computer Vision (ECCV)*.

Granlund, G. H. (1978). In search of a general picture processing operator. *Computer Graphics and Image Processing*, 8(2):155–173.

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

Huang, J. and Mumford, D. (1999). Statistics of natural images and models. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(2):1705–1720.

Hyvärinen, A., Hoyer, P. O., and Inki, M. (2000). Topographic ICA as a model of natural image statistics. In *the First IEEE Int'l. Workshop on Bio. Motivated Comp. Vis.*, London, UK.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, 2nd edition.

Karklin, Y. and Lewicki, M. S. (2005). A hierarchical bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423.

Kingman, J. F. C. (1963). Random walks with spherical symmetry. *Acta Math.*, 109(9):11–53.

Koenderink, J. J. (1984). The structure of images. *Biological Cybernetics*, 50:363–370.

Kohonen, T. (1996). Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map. *Biological Cybernetics*, 75(5):281–291.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, 69(6):66–82.

Learned-Miller, E. G. and Fisher, J. W. (2000). ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4(1):1271–1295.

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

Lyu, S. and Simoncelli, E. P. (2007). Statistically and perceptually motivated nonlinear image representation. In Rogowitz, B., Pappas, T. N., and Daly, S. J., editors, *Proc. SPIE, Conf. on Human Vision and Electronic Imaging XII*, volume 6492, San Jose, CA.

Lyu, S. and Simoncelli, E. P. (2008). Modeling wavelet subbands of photographic images with fields of Gaussian scale mixtures. *IEEE Trans. Patt. Analysis and Machine Intelligence*. Accepted for publication, 4/08.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–697.

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

Malo, J., Navarro, R., Epifanio, I., Ferri, F., and Artigas, J. (2000a). Non-linear invertible representation for joint statistical and perceptual fea ture representation. *Lect. Not. Comp. Sci.*, 1876:658–667.

Malo, J., Navarro, R., Epifanio, I., Ferri, F., and Artigas, J. (2000b). Non-linear invertible representation for joint statistical and perceptual feature decorrelation. In *Proc. SPR+SSPR*, pages 658–667. Springer Verlag. Lect. Not. Comp. Sci. 1876.

Mika, S., Schölkopf, B., Smola, A. J., Müller, K.-R., Scholz, M., and Rätsch, G. (1999). Kernel PCA and de–noising in feature spaces. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*. MIT Press.

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

Nolan, J. P. (2007). *Stable Distributions - Models for Heavy Tailed Data*. Birkhäuser, Boston. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

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

Parra, L., Spence, C., and Sajda, P. (2001). Higher-order statistical properties arising from the non-stationarity of natural signals. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Adv. Neural Information Processing Systems (NIPS*00)*, volume 13, pages 786–792, Cambridge, MA. MIT Press.

Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P. (2003). Image denoising using a scale mixture of Gaussians in the wavelet domain. *IEEE Trans Image Processing*, 12(11):1338–1351.

Ruderman, D. L. (1996). The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548.

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

Sendur, L. and Selesnick, I. W. (2002). Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans. on Signal Processing*, 50(11):2744–2756.

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *Proc 31st Asilomar Conf on Signals, Systems and Computers*, volume 1, pages 673–678, Pacific Grove, CA. IEEE Computer Society.

Simoncelli, E. P. and Adelson, E. H. (1996). Noise removal via Bayesian wavelet coring. In *Proc 3rd IEEE Int'l Conf on Image Proc*, volume I, pages 379–382, Lausanne. IEEE Sig Proc Society.

Simoncelli, E. P. and Buccigrossi, R. W. (1997). Embedded wavelet image compression based on a joint probability model. In *Proc 4th IEEE Int'l Conf on Image Proc*, volume I, pages 640–643, Santa Barbara. IEEE Sig Proc Society.

Simoncelli, E. P. and Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc 2nd IEEE Int'l Conf on Image Proc*, volume III, pages 444–447, Washington, DC. IEEE Sig Proc Society.

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

Srivastava, A., Liu, X., and Grenander, U. (2002). Universal analytical forms for modeling image probability. *IEEE Pat. Anal. Mach. Intell.*, 24(9):1200–1214.

Studeny, M. and Vejnarova, J. (1998). The multiinformation function as a tool for measuring stochastic dependence. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 261–297. Dordrecht: Kluwer.

Teh, Y., Welling, M., and Osindero, S. (2003). Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:1235–1260.

Teo, P. C. and Heeger, D. J. (1994). Perceptual image distortion. In *IEEE Int'l. Conf. on Image Proc.*, pages 982–986.

Valerio, R. and Navarro, R. (2003a). Optimal coding through divisive normalization models of V1 neurons. *Network: Computation in Neural Systems*, 14:579–593.

Valerio, R. and Navarro, R. (2003b). Optimal coding through divisive normalization models of v1 neurons. *Network: Computation with neural systems*, pages 579–593.

van der Schaaf, A. and 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. (1999). Visual adaptation as optimal information transmission. *Vision Research*, 39:3960–3974.

Wainwright, M. J., Schwartz, O., and Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In Rao, R., Olshausen, B., and Lewicki, M., editors, *Probabilistic Models of the Brain: Perception and Neural Function*, chapter 10, pages 203–222. MIT Press.

Wainwright, M. J. and Simoncelli, E. P. (2000). Scale mixtures of Gaussians and the statistics of natural images. In Solla, S. A., Leen, T. K., and Müller, K.-R., editors, *Adv. Neural Information Processing Systems (NIPS*99)*, volume 12, pages 855–861, Cambridge, MA. MIT Press.

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

Wegmann, B. and Zetzsche, C. (1990). Statistical dependence between orientation filter outputs used in an human vision based image code. In *Proc Visual Comm. and Image Processing*, volume 1360, pages 909–922, Lausanne, Switzerland.

Yao, K. (1973). A representation theorem and its applications to spherically-invariant random processes. *IEEE Trans. on Information Theory*, 19(9):600–608.

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

Zetzsche, C., Barth, E., Krieger, G., and Wegmann, B. (1997). Neural network models and the visual cortex: the missing link between orientation selectivity and the natural environment. *Neuroscience Letters*, 228:155–158.

Zetzsche, C. and Krieger, G. (1999). The atoms of vision: Cartesian or polar? *J. Opt. Soc. Am. A*, 16(7).

Zetzsche, C. and Schönecker, W. (1987). Orientation selective filters lead to entropy reduction in the processing of natural images. *Perception*, 16:229.

Zetzsche, C., Wegmann, B., and Barth, E. (1993). Nonlinear aspects of primary vision: Entropy reduction beyond decorrelation. In *Int'l Symposium, Society for Information Display*, volume XXIV, pages 933–936.