Sparsity vs. statistical independence from a best-basis viewpoint

Naoki Saito, Brons M. Larson, and Bertrand Bénichou

Department of Mathematics, University of California, Davis, CA 95616 USA

ABSTRACT

We examine the similarity and difference between sparsity and statistical independence in image representations in a very concrete setting: use the best basis algorithm to select the sparsest basis and the least statistically-dependent basis from basis dictionaries for a given dataset. In order to understand their relationship, we use synthetic stochastic processes (e.g., spike, ramp, and generalized Gaussian processes) as well as the image patches of natural scenes. Our experiments and analysis so far suggest the following: 1) Both sparsity and statistical independence criteria selected similar bases for most of our examples with minor differences; 2) Sparsity is more computationally and conceptually feasible as a basis selection criterion than the statistical independence, particularly for data compression; 3) The sparsity criterion can and should be adapted to individual realization rather than for the whole collection of the realizations to achieve the maximum performance; 4) The importance of orientation selectivity of the local Fourier and brushlet dictionaries was not clearly demonstrated due to the boundary effect caused by the folding and local periodization.

These observations seem to encourage the pursuit of sparse representations rather than that of statistically independent representations.

Keywords: Sparse representation, statistical independence, Independent Component Analysis, basis dictionary, best basis, least statistically-dependent basis

1. INTRODUCTION

Statistical analysis of natural scene images has recently drawn considerable attention particularly in the field of neuroscience such as Field,[1] Olshausen & Field,[2,3] Bell & Sejnowski,[4] van Hateren and van der Schaaf,[5] to name a few. Their main motivation was to understand the receptive field properties of simple cells in the mammalian primary visual cortex (i.e., V1 area) by analyzing the statistics of natural scenes. Barlow[6] suggested that mammals may be using the *statistically independent* coding strategy for visual stimuli, and proposed the *minimum entropy coding* (also known as *factorial coding*). Atick[7] also argued the importance of redundancy reduction and entropy minimization in the visual pathway. Field[1] suggested that neurons with line and edge selectivities in V1 may provide *sparse* representation of natural scenes. This may imply that mammals exploit the sparsity for image representations in their brain. Many neuroscientists have tried to answer the following reverse proposition, which is also very interesting: Immersed in the natural environment, whether the receptive fields of simple cells of mammals autonomously form edge or line detectors. If one can demonstrate this, it may be a convincing argument about why mammals have edge detectors and why mammals are exploiting sparse and efficient representations of natural scenes. Many interesting algorithms were developed and numerical experiments were performed on image patches of natural scenes.[2–5] All of these approaches essentially try to find the basis functions from some overcomplete set of bases by optimizing either *sparsity* or *statistical independence* (among the expansion coefficients) using some form of learning algorithms such as neural networks. In fact, they found that the estimated basis functions all resemble Gabor functions or oriented DOG filters regardless of the basis selection criteria if one uses the image patches of small size (e.g., 16×16 pixels) selected randomly from natural scene images. On one hand, their algorithms are truly self-organizing since they can build such basis functions completely from scratch. On the other hand, their computational cost prevents one from conducting experiments on image patches of large size (e.g., 64×64 or larger).

Inspired by this line of research, Donoho[8] recently proposed the concept of the Sparse Component Analysis (SCA), conducted a detailed mathematical analysis for a specific class of functions, and argued that sparsity and overcompleteness may be more plausible on biological grounds and are more important for practical data compression purposes than statistical independence.

Correspondence: E-mail: saito@math.ucdavis.edu; WWW: http://math.ucdavis.edu/∼saito

In the mean time, from a completely different motivation (stochastic modeling of a class of similar images), we developed an algorithm of computing the *least statistically-dependent basis* (LSDB) from time-frequency dictionaries.[9,10] This can be viewed as a dictionary version of the independent component analysis (ICA).[11,12]

This series of work has motivated us to study the following questions:

- Why both sparsity and statistical independence criteria produced edge or line detectors for natural scene datasets?
- What is the similarity and difference between sparsity and statistical independence criteria?
- What is the effect of the sizes of the image patches used?
- What is the effect of orthonormality?
- What is the effect of overcompleteness?
- What is the effect of orientation selectivities of basis functions?

In this paper, we will examine some of these questions, in particular, the similarity and difference between sparsity and statistical independence under a concrete and simpler setting: extracting best bases from a set of basis dictionaries (such as wavelet packet, brushlet, local cosine, local Fourier dictionaries) with respect to sparsity or statistical independence criteria using simple synthetic examples as well as the natural scene data.

Let us first describe our notation and the terminology of basis dictionaries and best bases. Let $\boldsymbol{X} \in \mathbb{R}^n$ be a random vector with some unknown probability density function (pdf) $f_{\boldsymbol{X}}$. Let us assume that the available data $\mathcal{T} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ were independently generated from this probability model. Let $B = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) \in SO(n, \mathbb{R})$ (a group of orthonormal transformations in $\mathbb{R}^n$) or $SL(n, \mathbb{R})$ (a group of volume-preserving transformations in $\mathbb{R}^n$). The best-basis paradigm is to find a basis B or a subset of basis vectors such that the features (expansion coefficients) $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ are useful for the problem at hand (e.g., compression, modeling, discrimination, regression, segmentation) in a computationally fast manner. Let $\mathcal{C}(B \mid \mathcal{T})$ be a numerical measure of *deficiency* or *cost* of the basis B given the training dataset $\mathcal{T}$ for the given problem. Often we restrict our search within the basis dictionary $\mathcal{D} \subset SL(n, \mathbb{R})$, such as the orthonormal or biorthogonal wavelet packet dictionaries or local cosine or Fourier dictionaries where we never need to compute the full matrix-vector product or the matrix inverse for analysis and synthesis. Under this setting, $B_\star = \arg\min_{B \in \mathcal{D}} \mathcal{C}(B \mid \mathcal{T})$ is called the *best basis* relative to the cost $\mathcal{C}$ and the training dataset $\mathcal{T}$.

2. SPARSITY VS STATISTICAL INDEPENDENCE

The concept of sparsity and that of statistical independence are intrinsically different. The sparsity emphasizes the issue of the compression directly whereas the statistical independence concerns more about the relationship among the coordinates. Yet, for certain stochastic processes, these two are intimately related, and often confusing. For example, Olshausen and Field[2,3] emphasized the sparsity as the basis selection criterion, but they also assumed the statistical independence of the coordinates. Bell and Sejnowski[4] used the ICA and obtained similar results. They claimed that they did not impose the sparsity explicitly and such sparsity emerged by maximizing independence (or minimizing the dependence). These motivated us to study these two criteria.

First let us define the measure of sparsity and that of statistical independence in our context.

2.1. Sparsity

The sparsity is a key property as a good coordinate system for compression. The true sparsity measure for a given vector $\boldsymbol{x} \in \mathbb{R}^n$ is the so-called ℓ^0 quasi-norm,[8,13] which is defined as

$$\|\boldsymbol{x}\|_0 \stackrel{\Delta}{=} \#\{i \in [1, n] : x_i \neq 0\}.$$

This measure is, however, very unstable for even small perturbation of the elements in a vector. Therefore, the better measure is the ℓ^p norm:

$$\|\boldsymbol{x}\|_p \stackrel{\Delta}{=} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 0 < p \leq 1.$$

In fact, this is a quasi-norm since this does not satisfy triangle inequality, but only satisfies the weaker condition, $\|\boldsymbol{x}+\boldsymbol{y}\|_p \le 2^{-1/p'}(\|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p)$ where p' is the conjugate exponent of p,[14] and $\|\boldsymbol{x}+\boldsymbol{y}\|_p^p \le \|\boldsymbol{x}\|_p^p + \|\boldsymbol{y}\|_p^p$. It is easy to show that $\lim_{p\downarrow 0}\|\boldsymbol{x}\|_p^p = \|\boldsymbol{x}\|_0$.

We can use ℓ^p norm minimization as a basis selection criterion for the best basis in terms of sparsity from a dictionary/library. Therefore, we propose to use the expected ℓ^p norm minimization for a given training dataset to select the sparse basis:

$$\mathcal{C}_p(B \mid \mathcal{T}) = E\|\boldsymbol{Y}\|_p^p \approx \frac{1}{N}\sum_{k=1}^{N}\|\boldsymbol{y}_k\|_p^p = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n}|y_{i,k}|^p,$$

$$B_p = \arg\min_{B\in\mathcal{D}} \mathcal{C}_p(B \mid \mathcal{T}).$$

Here, $\boldsymbol{y}_k = (y_{1,k},\ldots,y_{n,k})^T = B^{-1}\boldsymbol{x}_k$, and $\boldsymbol{x}_k$ is the kth sample (or realization) in $\mathcal{T}$. It should be noted that *the minimization of the ℓ^p norm can also be achieved for each realization*. We will discuss more about this in Section 3.

2.2. Statistical Independence

The statistical independence of the coordinates of $\boldsymbol{Y} \in \mathbb{R}^n$ means

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2)\cdots f_{Y_n}(y_n),$$

where $f_{Y_k}(y_k)$ is a one-dimensional marginal pdf defined as

$$f_{Y_k}(y_k) = \int\cdots\int f_{\boldsymbol{Y}}(y_1,\ldots,y_k,\ldots,y_n)\mathrm{d}y_1\cdots\mathrm{d}y_{k-1}\mathrm{d}y_{k+1}\cdots\mathrm{d}y_n.$$

The statistical Independence is a key property as a good coordinate system for compression and particularly modeling because: 1) Damage of one coordinate does not propagate to the others; and 2) it allows us to model the n-dimensional stochastic process of interest as a set of 1D processes. Of course, in general, it is very difficult or almost impossible to find a truly statistically independent coordinate system for a given stochastic process. Therefore, we should be satisfied with finding the least-statistically dependent coordinate system from a dictionary/library. Naturally, then, we need to measure the "closeness" of a given coordinate system $Y_1,\ldots,Y_n$ to the statistical independence. This can be measured by *mutual information* or relative entropy between the true pdf $f_{\boldsymbol{Y}}$ and the product of its marginals:

$$\begin{aligned} I(\boldsymbol{Y}) &\triangleq \int f_{\boldsymbol{Y}}(\boldsymbol{y})\log\frac{f_{\boldsymbol{Y}}(\boldsymbol{y})}{\prod_{i=1}^n f_{Y_i}(y_i)}\mathrm{d}y_1\cdots\mathrm{d}y_n \\ &= -H(\boldsymbol{Y}) + \sum_{i=1}^{n} H(Y_i), \end{aligned}$$

where $H(\boldsymbol{Y})$ and $H(Y_i)$ are the differential entropy of $\boldsymbol{Y}$ and Y_i respectively:

$$H(\boldsymbol{Y}) = -\int f_{\boldsymbol{Y}}(\boldsymbol{y})\,\log f_{\boldsymbol{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \quad H(Y_i) = -\int f_{Y_i}(y_i)\,\log f_{Y_i}(y_i)\mathrm{d}y_i.$$

We note that $I(\boldsymbol{Y}) \ge 0$, and $I(\boldsymbol{Y}) = 0$ if and only if the components of $\boldsymbol{Y}$ are mutually independent.

Suppose we constrain our search of a good basis within the volume-preserving transformations. In other words, suppose $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ and $B \in SL(n,\mathbb{R})$. Then, we have

$$I(\boldsymbol{Y}) = -H(\boldsymbol{Y}) + \sum_{i=1}^{n} H(Y_i) = -H(\boldsymbol{X}) + \sum_{i=1}^{n} H(Y_i),$$

since the differential entropy is *invariant* under such a transformation, i.e.,

$$H(B^{-1}\boldsymbol{X}) = H(\boldsymbol{X}) + \log|\det(B^{-1})| = H(\boldsymbol{X}).$$

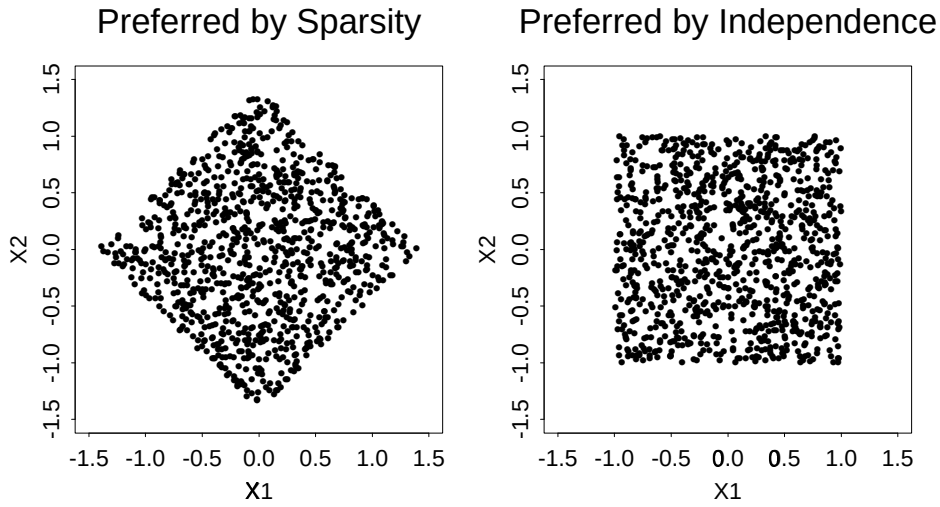


Figure 1. Sparsity and statistical independence prefer the different coordinates.

Based on this fact, we proposed the following cost function to select the so-called *least statistically-dependent basis* (LSDB)[9,10]:

$$\begin{aligned} \mathcal{C}_{LSDB}(B \,|\, \mathcal{T}) &= \sum_{i=1}^{n} H(Y_i) = -\sum_{i=1}^{n} \int f_{Y_i}(y) \log f_{Y_i}(y) \mathrm{d}y \\ &\approx -\frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{n} \log \widehat{f}_{Y_i}(y_{i,k}), \end{aligned}$$

where $\widehat{f}_{Y_i}(y_{i,k})$ is an empirical pdf of the coordinate Y_i, which must be estimated from the training data $\mathcal{T}$ by an algorithm such as the histogram with optimal bin-width search.[15] Now, we can define the LSDB as

$$B_{LSDB} = \arg\min_{B \in \mathcal{D}} \mathcal{C}_{LSDB}(B \,|\, \mathcal{T}).$$

3. SIMPLE EXAMPLES

In this section, we examine the sparsest coordinates and the statistically-independent or least statistically-dependent coordinates for several simple stochastic processes. Although these processes are much simpler than the natural image patches, we gain a good insight into the similarity and difference between sparsity and independence by studying such processes.

3.1. Two-dimensional Counterexample

Let us consider a simple process $\boldsymbol{X} = (X_1, X_2)$ where X_1 and X_2 are independently and identically distributed as the uniform distribution $\mathrm{unif}[-1, 1]$. Thus, the realizations of this process are distributed as the righthand side of Figure 1. Let us consider all possible rotations around origin as a basis dictionary, i.e., $\mathcal{D} = \mathrm{SO}(2, \mathbb{R})$. Then, the sparsity and independence criteria select completely different bases as shown in Figure 1. This example clearly demonstrates that the sparsest coordinates and the statistically independent coordinates are generally different. One can generalize this example to higher dimensions.

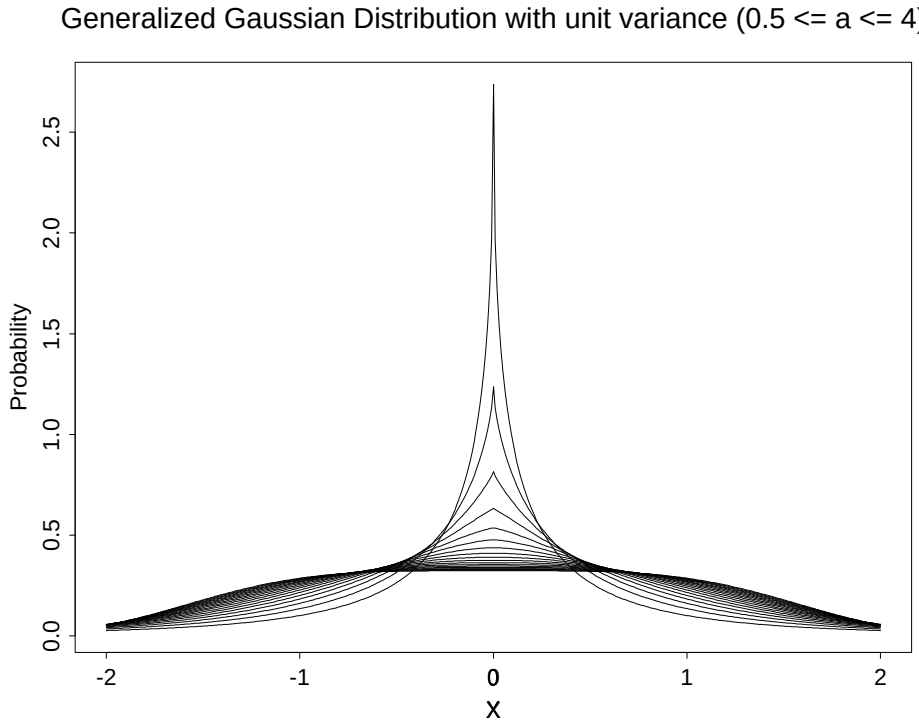


Figure 2. Examples of generalized Gaussian distributions with $0.5 \leq \alpha \leq 4.0$.

3.2. One-dimensional generalized Gaussian distributions

Despite the intrinsic difference between the sparsity and the statistical independence, we often see the similar coordinate system give both good sparsity and small statistical dependence. The experiments of Olshausen-Field, Bell-Sejnowski, and van Hateren and van der Schaaf all suggest this similarity. In this subsection and the following, we will show more concretely this is the case for certain simple stochastic processes.

Let us first consider a particular one-dimensional stochastic process, which obeys the *generalized Gaussian distribution*. This is a widely used probability distribution for modeling various naturally-occurring phenomena (particularly in geophysics[16–18]), and is defined as follows:

$$g(x;\alpha,\beta) \triangleq \frac{\alpha}{2\beta\Gamma(1/\alpha)}\mathrm{e}^{-(|x|/\beta)^{\alpha}},$$

where $\alpha > 0$ prescribes the *shape* of the distribution, and $\beta > 0$ specifies its *scale*. One can easily see that the case of $(\alpha,\beta) = (2,\sqrt{2})$ reduces to the standard Gaussian distribution $N(0,1)$, and $(\alpha,\beta) = (1,1/\sqrt{2})$ reduces to the standard Laplacian distribution. Moreover, as $\alpha \to \infty$, it reaches to the uniform distribution. Figure 2 shows several generalized Gaussian distributions with various values of α.

Since this is a one-dimensional process, the statistical independence among coordinates cannot be defined properly. Thus we only consider the entropy of the one coordinate instead of the statistical independence. A simple calculation[19,18] shows that the differential entropy of the random variable $X \sim g(\cdot;\alpha,\beta)$ is

$$H(X;\alpha,\beta) = E[\ln(1/g(X;\alpha,\beta))] = -\int_{-\infty}^{\infty} g(x;\alpha,\beta)\ln g(x;\alpha,\beta)\,\mathrm{d}x = \frac{1}{\alpha} - \ln\left[\frac{\alpha}{2\beta\Gamma(1/\alpha)}\right]. \tag{1}$$

On the other hand, the sparsity can be computed[19] as

$$S(X;p,\alpha,\beta) = E|X|^{p} = \int_{-\infty}^{\infty} |x|^{p} g(x;\alpha,\beta)\,\mathrm{d}x = \beta^{p}\Gamma\left(\frac{p+1}{\alpha}\right) / \Gamma\left(\frac{1}{\alpha}\right). \tag{2}$$

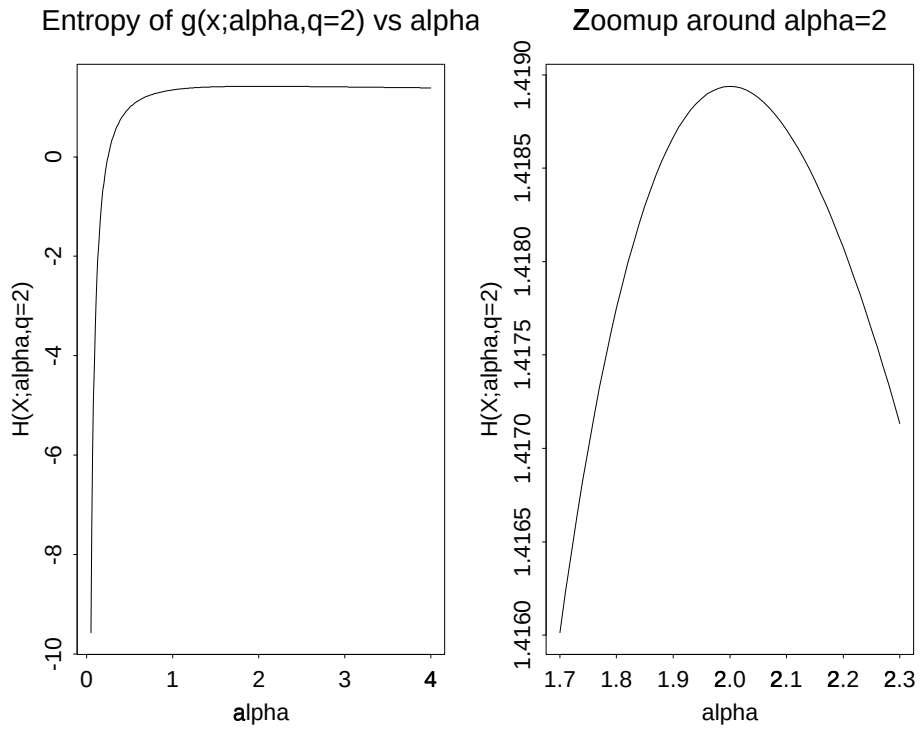


Figure 3. Plots of entropy $H(X;\alpha,q=2)$ vs the shape parameter α. The right figure is zoomed version around $\alpha=2$ to show that the maximum entropy is attained at $\alpha=2$.

At this point, let us assume that the random variable X satisfies some normalization condition, for example, $E|X|^q=1$. After proper normalization (e.g., centering and sphering), it is common to assume that the distribution has a zero mean and a unit variance, i.e., $q=2$. This normalization condition determines the parameter β as

$$\beta_q=\left[\Gamma\left(\frac{1}{\alpha}\right)/\Gamma\left(\frac{q+1}{\alpha}\right)\right]^{1/q}. \tag{3}$$

For this β, (1) becomes

$$H(X;\alpha,q)=H(X;\alpha,\beta_q)=\frac{1}{\alpha}+\ln\frac{2}{\alpha}+\left(1+\frac{1}{q}\right)\ln\Gamma\left(\frac{1}{\alpha}\right)-\frac{1}{q}\ln\Gamma\left(\frac{q+1}{\alpha}\right).$$

The entropy of this normalized generalized Gaussian distribution is shown in Figure 3. As one can see from this figure, we can show the following lemma:

Lemma 3.1.

$$\lim_{\alpha\downarrow 0}H(X;\alpha,q)=-\infty,\quad \lim_{\alpha\to\infty}H(X;\alpha,q)=\frac{1}{q}\ln(1+q)+\ln 2.$$

$$\arg\max_{0\le\alpha<\infty}H(X;\alpha,q)=q.$$

In other words, for $0\le\alpha\le q$, the entropy monotonically increases while for $\alpha\ge q$, it monotonically decreases.

Proof. See[19] for the proof. □

Using the same β in (3), the sparsity (2)becomes

$$S(X;p,\alpha,q)=S(X;p,\alpha,\beta_q)=\Gamma\left(\frac{p+1}{\alpha}\right)\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^{p/q-1}\left[\Gamma\left(\frac{q+1}{\alpha}\right)\right]^{-p/q}.$$

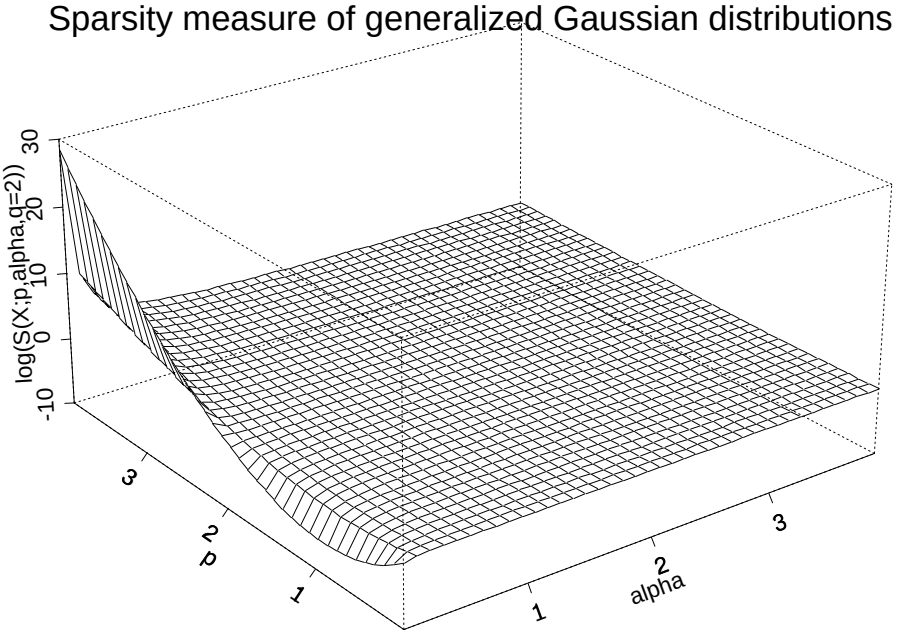


Figure 4. Perspective view of the sparsity measure $S(X; p, \alpha, q = 2)$ in the log scale.

The sparsity of this normalized generalized Gaussian distribution is shown in Figure 4. We note that the sparsity measure changes its behavior according to $p > q$ or $p < q$. In fact we can show the following:

Lemma 3.2. *For each $\alpha > 0$, the sparsity measure $S(X; p, \alpha, q)$ attains minimum (as a function of p) in the interval $(0, q)$.*

Proof. See[19] for the proof. □

From these lemmas and figures, we conclude that the one-dimensional generalized Gaussian distribution with small α is preferred by both the sparsity and entropy criteria.

Remark 3.3. Consider the high dimensional stochastic process $\boldsymbol{X} \in \mathbb{R}^n$ and let us assume that we seek the new coordinates $\boldsymbol{Y}$ where the components are independently and identically distributed as the generalized Gaussian distribution with shape parameter $\alpha = p$, although such a situation is rather specific and restrictive. More precisely, let us assume $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{k=1}^{n} f_{Y_k}(y_k)$ and $f_{Y_k}(y_k) = g(y_k; p, \beta)$. Then,

$$\sum_{k=1}^{n} H(Y_k) = \sum_{k=1}^{n} E[\ln(1/g(Y_k; p, \beta))] = \text{const}(p, \beta) + \frac{1}{\beta^p} \sum_{k=1}^{n} E|Y_k|^p.$$

Therefore, minimizing the ℓ^p norm is equivalent to minimizing the sum of the coordinate-wise entropy in this case. In other words, the independence and sparsity criteria exactly coincide. Furthermore, suppose we are given a collection of the bases $B_1, \ldots, B_K$, under which the stochastic process of our interest is independently and identically distributed with the generalized Gaussian distributions with the shape parameters, $\alpha_1, \ldots, \alpha_K$, respectively. Suppose also the scale parameter for each coordinate in each basis is adjusted to have unit variance. Let $k^\star = \arg\min_{1 \le k \le K} \alpha_k$. Then

$$\arg \min_{1 \le k \le K} \mathcal{C}_{LSDB}(B_k \,|\, \mathcal{T}) = \arg \min_{1 \le k \le K} \mathcal{C}_p(B_k \,|\, \mathcal{T}) = k^\star.$$

In other words, both the sparsest basis and the LSDB is the same basis, $B_{k^\star}$. This remark also suggests that the sparsity and independence criteria optimize different quantities as soon as one varies β for each coordinate.

3.3. The Spike Process

An n-dimensional *spike process* simply generates the standard basis vectors $\{\boldsymbol{e}_j\} \subset \mathbb{R}^n$ in a random fashion, where $\boldsymbol{e}_j$ has one at the jth entry and all the other entries are zeros. One can view this process as a single spike of unit amplitude located at a random position between 1 and n. In this case, it is easy to show[19] that the Karhunen-Loève basis is any orthonormal basis in $\mathbb{R}^n$ containing the constant "DC" vector $\boldsymbol{b} = (1, 1, \ldots, 1)^T$.

In terms of sparsity, it is clear that the standard basis is the sparsest basis. It is clear that this basis is not statistically independent; existence of the single spike constrains the probability of spike generation at other locations. We have the following theorems for this simple process.

Theorem 3.4. *Suppose we restrict our search of the bases within the Haar-Walsh dictionary. Then,*

- *The sparsest basis is the standard basis.*
- *The LSDB is:*
  - *the standard basis if $n > 4$; and*
  - *the Walsh basis if $n = 2, 4$.*
- *The true independence can be achieved only for $n = 2$.*

Note: n is always a dyadic number in this dictionary.

Proof. See[19] for the proof. □

Now, we can generalize this theorem not only to the Haar-Walsh dictionary but also to any basis chosen from $\mathrm{SL}(n, \mathbb{R})$.

Theorem 3.5. *Suppose we extend our search of the bases to $\mathrm{SO}(n, \mathbb{R})$ or $\mathrm{SL}(n, \mathbb{R})$.*

- *The sparsest basis among $\mathrm{SL}(n, \mathbb{R})$ is the standard basis.*
- *The LSDB among $\mathrm{SO}(n, \mathbb{R})$ is:*
  - *the standard basis if $n \geq 5$;*
  - *the Walsh basis if $n = 2, 4$; and*
  - $\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix}$ *or its permutation of rows and columns if $n = 3$.*
- *The LSDB among $\mathrm{SL}(n, \mathbb{R})$ with $n > 2$ is the following basis pair (for synthesis and analysis):*

$$B_\star = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}, \quad B_\star^{-1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}.$$

Proof. See[19] for the proof. □

Finally, we have the following conclusive theorem:

Theorem 3.6. *There is no linear transformation providing the statistically-independent coordinates for the spike process for $n > 2$.*

Proof. See[19] for the proof. □

Remark 3.7. Although this process is very simple, we have the following interpretation. Consider a stochastic process generating a basis vector randomly selected from a specific basis at a time. Then, both the sparsest basis and the LSDB are that particular basis. Theorems 3.5 and 3.6 claims that once we transform the data to spikes, one cannot do any better than that both in sparsity and independence within the linear context. Of course, if one extends the search to nonlinear transformations, then it is a completely different story. We refer the reader to our recent article[20] for a nonlinear algorithm.

Remark 3.8. Next steps beyond this simple spike process are: 1) a spike process with varying amplitude, which was also considered in[21] for rate-distortion calculation; 2) a multi-spike process generating more than one spikes at a time.

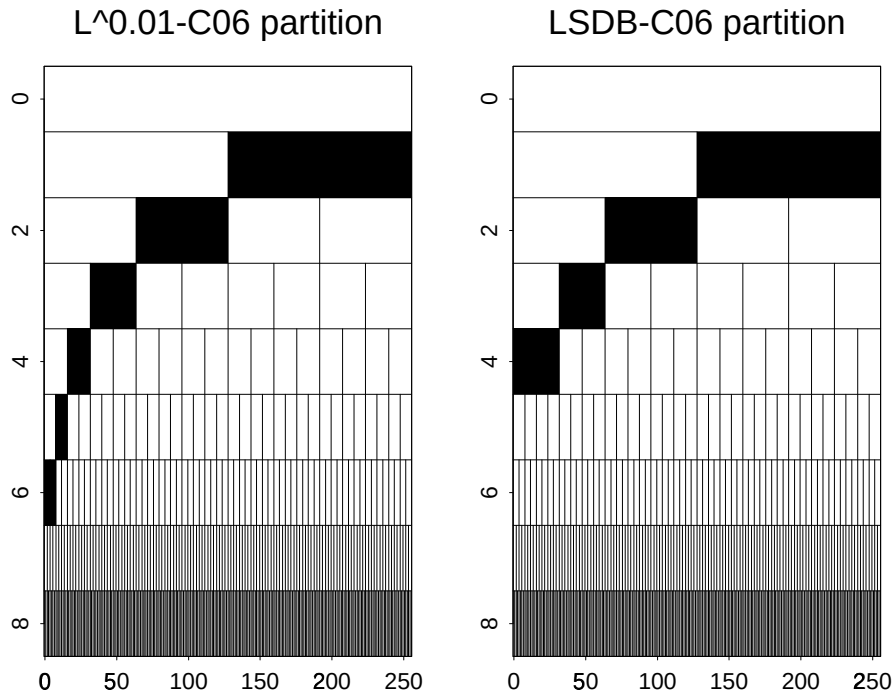


Figure 5. Comparison of basis partition patterns between the sparsest basis with $\ell^{0.01}$ norm and the LSDB.

3.4. The Ramp Process

At the CIRM Luminy meeting of Wavelets and Applications in 1992, Yves Meyer proposed the following simple stochastic process.

$$X(t) = t - H(t - \tau), \quad 0 \le t \le 1,$$

where $H(\cdot)$ is the Heaviside step function, and τ obeys the uniform distribution unif$[0, 1]$. As also described in,[21,22] the covariance of this process is

$$\Gamma(s, t) = \min(s, t) - st,$$

which is the same as that of the Brownian bridge. As a consequence, the KLB of this process is the Fourier sine basis $\phi_k(t) = \sqrt{2}\sin(2\pi kt)$, and the corresponding eigenvalues are $\lambda_k = (2\pi k)^{-2}$. Apparently, this Fourier sine basis is not efficient to compress realizations of this process due to its discontinuity. In fact, we can precisely measure the sparsity and entropy of this KLB coordinates since we can derive the pdf of each KLB coordinate.[19] In particular, the sparsity of this KLB is $\sum_k c_p/k^p$ where c_p is some constant only dependent on p. In other words, the sparsity measure blows up for $0 < p \le 1$. This means that the KLB provides a completely dense coordinate system. This is a clear example of the importance of non-Gaussianity: the KLB is not useful, even harmful, for non-Gaussian processes. Intuitively and clearly, the wavelet basis should perform much better since it is good at capturing such a discontinuity efficiently. So, let us examine this process under the best basis setting.

In our numerical experiments, we discretized each realization by $n = 256$ grid points, and generated $N = 2560$ realizations of this process so that we have exactly 10 realizations for each shift. We also added weak white Gaussian noise ($\sigma = 10^{-7}$) to each realization for numerical stabilization. We used the 6-tap Coiflet as a conjugate mirror filter to generate the wavelet packet coefficients of the input signals. Figure 5 compares the basis patterns of and the $\ell^{0.01}$-sparse basis and the LSDB selected from this set of realizations. As one can see, both bases are essentially the wavelet basis. The only difference is the coarse scale subspaces. We also note that the pattern of the sparse basis are insensitive to the choice of p for $0 < p \le 1$. We are currently working on the proof of the following conjecture:

Conjecture 3.9. *Within any wavelet packet dictionaries, both the ℓ^p-sparse basis ($0 < p \le 1$) and the LSDB for the ramp process are the wavelet basis.*

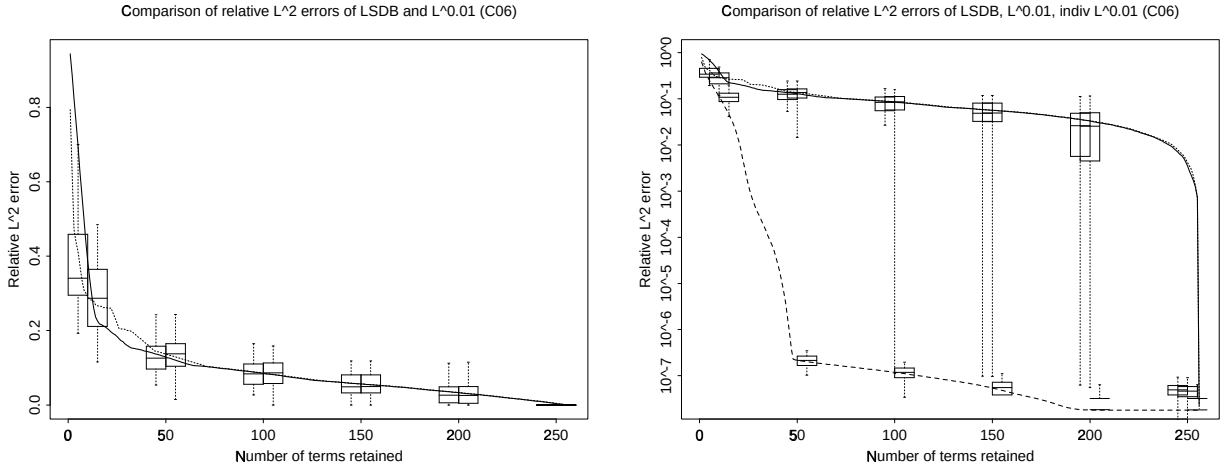


Figure 6. Relative ℓ^2 approximation errors of the $\ell^{0.01}$-sparse basis and the LSDB. The solid curve represents the mean of the error vs. the number of terms retained in the LSDB. The dotted curve represents that of the $\ell^{0.01}$-sparse basis. The associated boxplots are also displayed to show the scatter of individual errors for these bases. The right figure compares these two errors with those by individually-adapted sparse bases.

We hope that we can report the complete proof in the near future.[23] The vanishing moment property of the wavelet packet functions will play an important role in the proof since it kills the polynomial portion of the process and each basis coordinate becomes a stochastic process parameterized by τ, and it is essentially the integration of the wavelet packet functions with the integration interval as a function of τ and scale and location parameter of the wavelet packet basis function. Then, we can show that

- The compact support of the basis function creates the large population of coefficients at zero. This explains the frequent observation of the spiky distribution at the origin with heavy tails such as the generalized Gaussian distributions with $\alpha \leq 1$ when one examines the statistics of the wavelet coefficients.
- The scale of the basis function controls the size of the spike at zero of the pdf of the corresponding coordinate. The larger the scale, the smaller the size of the spike at zero; therefore, the less spiky the distributions.
- The oscillation of the basis function decreases the smoothness of the pdf of the corresponding coefficients. This in turn increases the entropy and sparsity of the distribution.

Now, let us demonstrate that the concept of sparsity is more suitable for individual realizations rather than for the whole collection of the realizations. First of all, as Yves Meyer and Lars Villemoes pointed out to us, the local cosine basis could provide us with the coordinates as sparse as the wavelet basis *if the segmentation of the interval is optimally adapted to capture the discontinuity.* In other words, one can adaptively segment the interval to have large segments away from the discontinuity and progressively smaller segments toward the discontinuity, and then use local cosines in each segment. The time-frequency local cosines proposed by Villemoes[24] should be particularly efficient for this. However this adapted basis is very specific to the location of the discontinuity, and only suitable for the particular realization. Second, let us conduct the approximation experiments using the wavelet packet dictionary with the 6-tap Coiflet. Figure 6 compares the approximation errors of the $\ell^{0.01}$-sparse basis and the LSDB. From the mean curves and boxplots of these errors, we observe that both the $\ell^{0.01}$-sparse basis and the LSDB equally perform with some fluctuations in the first 60 coordinates. However, the strength of the sparsity criterion is that it can also be optimized for individual realization to have one basis per realization. The approximation errors of such strategy is also shown in Figure 6. All the individually-adapted bases are essentially the wavelet basis; the only differences are the coarse scale subspaces ranging from level 6 to 8. Of course, comparing this set of individually-adapted bases with the single LSDB and the overall sparse basis is not fair. It also requires more bits to describe all these individually-adapted bases. Yet, this is clearly one of the good features of the sparsity criterion. Thus we can claim that sparsity may be a concept much more suitable for individual realization rather than a whole collection of realizations. Put it differently, in order to pursue the truly sparsest basis, *it is necessary to adapt our strategy for each realization.*

Remark 3.10. We note that the scatter of the ℓ^p-sparse basis for the whole realizations could be generally much worse than the one shown in Figure 6. In this case, we made sure that the number of realizations for each shift are the same. When we conducted the same experiments with 1000 realizations, i.e., uneven number of realizations for each shift, we observed that the scatter of the sparse basis was far greater than that of the LSDB.

4. EXPERIMENTS ON NATURAL IMAGES

Finally, we report our numerical experiments on the natural scene images. The more detailed analysis can be found in our full report.[19] The setting of these experiments is the following.

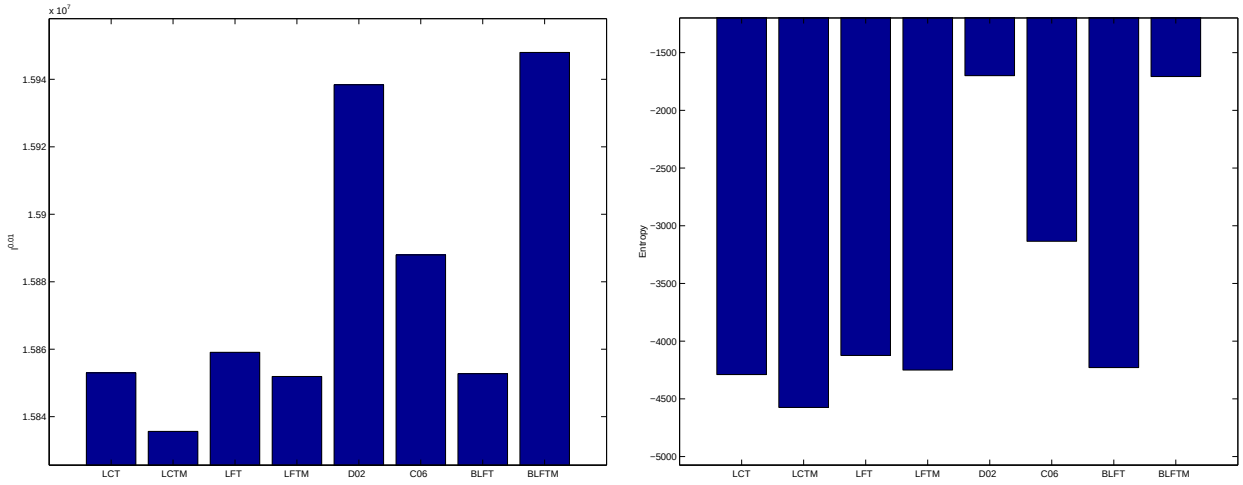
- A database of 10 natural scene images of size 512×512 was received from Bruno Olshausen.
- The bandpass filter was applied to "whiten" the spectrum of each image as discussed in.[3]
- 4000 patches were randomly picked from this database.
- The patch sizes we examined were 16×16 and 64×64.
- The dictionaries used were: Haar-Walsh, Coiflet 6, Local cosines, *real-valued* Local Fourier, and *real-valued* Brushlets.
- Both fixed and multiple folding were examined in local cosines, local Fourier, and brushlet dictionaries.
- Each patch was evenly extended, folded, and cropped at the boundaries; in addition, smooth periodization was applied for wavelet packets, local Fourier, and brushlets, but not for local cosines.
- The sparsity parameters examined were: $p = 1,\ 0.1,\ 0.01$.
- In the LSDB, the Hall-Morton entropy estimator[15] (histogram with optimal bin width search) was used to estimate entropy of each basis coordinate.

We note that the use of the real-valued local Fourier and brushlet dictionaries[25] are important because the ℓ^p norm of the complex coefficients are directly related to the ℓ^2 norm and cannot properly compare the sparsity of the representations with the other dictionaries.

In this paper, we only show the results of performance or costs of the $\ell^{0.01}$-sparse basis and LSDB criteria for this dataset in Figure 7. Our observation and interpretation of the results are the following:

- The local cosine basis with multiple folding performed best among all the dictionaries we examined in terms of both sparsity and statistical independence.
- Although we do not show the basis partition patterns here, we observed that both the sparse bases and the LSDBs tend to select similar partition patterns, and these patterns are insensitive to the change in p as long as $0 < p \leq 1$.
- The brushlets with fixed folding perform relatively well (after LCTM), whereas the brushlets with multiple folding perform very poorly.
- The importance of the orientation selectivity of the dictionaries (via local Fourier or brushlets) was not clearly demonstrated due to the boundary effect caused by folding and periodization.
- Multiple folding cases have always smaller entropy and ℓ^p norm than fixed folding cases both for the local cosine and the local Fourier dictionaries, but not for the brushlets.
- For the smaller patches of 16×16 pixels, we essentially observe the same (not shown here). The only difference from the results of 64×64 patches is that the windowing effect is more severe so that the performance of the brushlets are not so good as the case of 64×64 patches.

We are very curious why the local cosine dictionary with multiple folding performed best in our experiments. We are currently investigating this issue by varying the boundary treatment of the image patches, and comparing its performance with that of the standard block DCT used in JPEG as well as that of the 7/9-taps biorthogonal wavelets which will be incorporated in the JPEG 2000 standard.



(a) $\mathcal{C}_{0.01}$ of the basis dictionaries

(b) $\mathcal{C}_{LSDB}$ of the basis dictionaries

Figure 7. The performance (i.e., cost) of each dictionary in terms of $\ell^{0.01}$ sparsity and the statistical dependence for natural image patches of 64×64 pixels. In both cases, the smaller the cost, the better. The letter M in the end of the abbreviations of the basis dictionaries indicates the use of multiple folding. BLFT means a brushlet dictionary.

5. CONCLUSION

Except for the synthetic example of 2D uniform distributions of Section 3.1, in all the examples we worked on, essentially, the sparsity and the statistical independence criteria produced nearly the same basis patterns. Optimizing the sparsity turned out to be much more feasible both computationally and conceptually. The entropy estimation based on the empirical pdf estimation is more computationally intensive than the ℓ^p sparsity criterion. Proving the theorems in the spike processes was much harder for the statistical independence than for the sparsity. We anticipate the same situation for proving Conjecture 3.9 for the ramp process. Moreover, the sparsity criterion can be adapted to each realization to maximize the sparsity of representations. On the other hand, even though we can compute the LSDB, that does not guarantee the true statistical independence in general. Therefore, as long as we stay in the linear framework and we are interested in the data compression, it seems to us that the pursuit of sparse representations should be encouraged than that of statistically independent representations.

ACKNOWLEDGMENTS

The first author would like to thank Bruno Olshausen for stimulating discussions and for providing the digitized natural scene images. He also would like to thank fruitful discussions with Stéphane Mallat, Yves Meyer, and Lars Villemoes. This research was supported in part by NSF grants DMS-99-73032 and DMS-99-78321.

REFERENCES

1. D. J. Field, "What is the goal of sensory coding?," *Neural Computation* **6**, pp. 559–601, 1994.
2. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, pp. 607–609, 1996.
3. B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research* , pp. 3311–3325, 1997.
4. A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research* **37**, pp. 3327–3338, 1997.
5. J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. Royal Soc. London, Ser. B* **265**, pp. 359–366, 1998.
6. H. B. Barlow, "Unsupervised learning," *Neural Computation* **1**, pp. 295–311, 1989.

7. J. J. Atick, "Could information theory provide an ecological theory of sensory processing?," *Network* **3**, pp. 213–251, 1992.
8. D. L. Donoho, "Sparse components of images and optimal atomic decompositions," tech. rep., Dept. Statistics, Stanford University, 1998.
9. N. Saito, "Least statistically-dependent basis and its application to image modeling," in *Wavelet Applications in Signal and Image Processing VI*, A. F. Laine, M. A. Unser, and A. Aldroubi, eds., vol. Proc. SPIE 3458, pp. 24–37, 1998. Invited paper.
10. N. Saito, "The least statistically-dependent basis and its applications," in *Proc. 32nd Asilomar Conference on Signals, Systems, and Computers*, pp. 732–736, IEEE, 1998.
11. P. Comon, "Independent component analysis, a new concept?," *Signal Processing* **36**, pp. 287–314, 1994.
12. A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation* **7**, pp. 1129–1159, 1995.
13. D. L. Donoho, "On minimum entropy segmentation," Technical Report 450, Dept. Statistics, Stanford University, Apr. 1994.
14. M. M. Day, "The spaces L^p with $0 < p < 1$," *Bull. Amer. Math. Soc.* **46**, pp. 816–823, 1940.
15. P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.* **45**(1), pp. 69–88, 1993.
16. A. T. Walden, "Non-Gaussian reflectivity, entropy, and deconvolution," *Geophysics* **50**(12), pp. 2862–2888, 1985.
17. A. T. Walden and J. W. J. Hosken, "The nature of the non-Gaussianity of primary reflection coefficients and its significance for deconvolution," *Geophys. Prospect.* **34**, pp. 1038–1066, 1986.
18. W. C. Gray, *Variable Norm Deconvolution*. PhD thesis, Stanford University, 1979.
19. N. Saito, B. M. Larson, and B. Benichou, "Sparsity and statistical independence in adaptive signal representations," tech. rep., Dept. Math., Univ. California, Davis, 2000. In preparation.
20. J.-J. Lin, N. Saito, and R. A. Levine, "An iterative nonlinear Gaussianization algorithm for resampling dependent components," in *Proc. 2nd International Workshop on Independent Component Analysis and Blind Signal Separation*, P. Pajunen and J. Karhunen, eds., pp. 245–250, IEEE. June 19–22, 2000, Helsinki, Finland.
21. D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory* **44**(6), pp. 2435–2476, 1998. Invited paper.
22. J. B. Buckheit and D. L. Donoho, "Time-frequency tilings which best expose the non-Gaussian behavior of a stochastic process," in *Proc. International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 1–4, IEEE. Jun. 18–21, 1996, Paris, France.
23. N. Saito, "The wavelet basis is the best basis for the ramp process," tech. rep., Dept. Math., Univ. California, Davis, 2000. In preparation.
24. L. F. Villemoes, "Adapted bases of time-frequency local cosines," *Applied and Computational Harmonic Analysis* , 1999. submitted.
25. N. Saito and B. M. Larson, "The local Fourier and brushlet dictionaries: Theory and applications," in *Wavelets and Their Applications: Case Studies II*, M. Kobayashi, ed., ch. 2, SIAM, Philadelphia, PA, 2001. To appear.