Modern Signal Processing
MSRI Publications
Volume **46**, 2003

The Generalized Spike Process, Sparsity, and Statistical Independence

NAOKI SAITO

Abstract. We consider the *best sparsifying basis* (BSB) and the *kurtosis maximizing basis* (KMB) of a particularly simple stochastic process called the "generalized spike process". The BSB is a basis for which a given set of realizations of a stochastic process can be represented most sparsely, whereas the KMB is an approximation to the *least statistically-dependent basis* (LSDB) for which the data representation has minimal statistical dependence. In each realization, the generalized spike process puts a single spike with amplitude sampled from the standard normal distribution at a random location in an otherwise zero vector of length n.

We prove that both the BSB and the KMB select the standard basis, if we restrict our basis search to all possible orthonormal bases in $\mathbb{R}^n$. If we extend our basis search to all possible volume-preserving invertible linear transformations, we prove the BSB exists and is again the standard basis, whereas the KMB does not exist. Thus, the KMB is rather sensitive to the orthonormality of the transformations, while the BSB seems insensitive. Our results provide new additional support for the preference of the BSB over the LSDB/KMB for data compression. We include an explicit computation of the BSB for Meyer's discretized ramp process.

1. Introduction

This paper is a sequel to our previous paper [3], where we considered the *best sparsifying basis* (BSB), and the *least statistically-dependent basis* (LSDB) for input data assumed to be realizations of a very simple stochastic process called the "spike process." This process, which we will refer to as the "simple" spike process for convenience, puts a unit impulse (i.e., constant amplitude of 1) at a random location in a zero vector of length n. Here, the BSB is the basis of $\mathbb{R}^n$ that best sparsifies the given input data, and the LSDB is the basis of $\mathbb{R}^n$ that is the closest to the statistically independent coordinate system (regardless of whether such a coordinate system exists or not). In particular, we considered the BSB and LSDB chosen from all possible orthonormal transformations (i.e., O(n)) or all possible volume-preserving linear transformations (i.e., $\mathrm{SL}^{\pm}(n,\mathbb{R})$, where the determinant of each element is either $+1$ or -1).

In this paper, we consider the BSB and LSDB for a slightly more complicated process, the "generalized" spike process, and compare them with those of the simple spike process. The generalized spike process puts an impulse whose amplitude is sampled from the standard normal distribution $\mathcal{N}(0,1)$.

Our motivation to analyze the BSB and the LSDB for the generalized spike process stems from the work in computational neuroscience [22; 23; 2; 27] as well as in computational harmonic analysis [11; 7; 12]. The concept of sparsity and that of statistical independence are intrinsically different. Sparsity emphasizes the issue of compression directly, whereas statistical independence concerns the relationship among the coordinates. Yet, for certain stochastic processes, these two are intimately related, and often confusing. For example, Olshausen and Field [22; 23] emphasized the sparsity as the basis selection criterion, but they also assumed the statistical independence of the coordinates. For a set of natural scene image patches, their algorithm generated basis functions efficient to capture and represent edges of various scales, orientations, and positions, which are similar to the receptive field profiles of the neurons in our primary visual cortex. (Note the criticism raised by Donoho and Flesia [12] about the trend of referring to these functions as "Gabor"-like functions; therefore, we just call them "edge-detecting" basis functions in this paper.) Bell and Sejnowski [2] used the statistical independence criterion and obtained the basis functions similar to those of Olshausen and Field. They claimed that they did not impose the sparsity explicitly and such sparsity *emerged* by minimizing the statistical dependence among the coordinates. These motivated us to study these two criteria. However, the mathematical relationship between these two criteria in the general case has not been understood completely. Therefore we chose to study these simplified processes, which are much simpler than the natural scene images as a high-dimensional stochastic process. It is important to use simple stochastic processes first since we can gain insights and make precise statements in terms of theorems. By these theorems, we now understand what are the precise conditions for the sparsity and statistical independence criteria to select the same basis for the spike processes, and the difference between the simple and generalized spike processes. Weidmann and Vetterli also used the generalized spike process to make precise analysis of the rate-distortion behavior of sparse memoryless sources that serve as models of sparse signal representations [28].

Additionally, a very important by-product of this paper (as well as our previous paper [3]) is that these simple processes can be used for validating any independent component analysis (ICA) software that uses mutual information or kurtosis as a measure of statistical dependence, and any sparse component analysis (SCA) software that uses ℓ^p-norm ($0 < p \leq 1$) as a measure of sparsity. Actual outputs of the software can be compared with the true solutions obtained by our theorems. For example, the ICA software based on maximization of kurtosis of the inputs should not converge for the generalized spike process unless there is some constraint on the basis search (e.g., each column vector has a unit ℓ^2-norm). Considering the recent popularity of such software ([17; 5; 21]), it is a good thing to have such simple examples that can be generated and tested easily on computers.

The organization of this paper is as follows. The next section specifies notation and terminology. Section 3 defines how to quantitatively measure the sparsity and statistical dependence of a stochastic process relative to a given basis. Section 4 reviews the results on the simple spike process obtained in [3]. Section 5 contains our new results for the generalized spike process. In Section 6, we consider the BSB of Meyer's ramp process [20, p. 19], as an application of the results of Section 5. Finally, we conclude in Section 7 with a discussion.

2. Notation and Terminology

We first set our notation and the terminology. Let $\boldsymbol{X} \in \mathbb{R}^n$ be a random vector with some unknown probability density function (pdf) $f_{\boldsymbol{X}}$. Let $B \in \mathcal{D} \subset \mathbb{R}^{n \times n}$, where $\mathcal{D}$ is the so-called *basis dictionary.* For very high-dimensional data, we often take $\mathcal{D}$ to be the union of the wavelet packets and local Fourier bases (see [25] and references therein for more about such basis dictionaries). In this paper, however, we use much larger dictionaries: $\mathrm{O}(n)$ (the group of orthonormal transformations in $\mathbb{R}^n$) or $\mathrm{SL}^{\pm}(n, \mathbb{R})$ (the group of invertible volume-preserving transformations in $\mathbb{R}^n$, i.e., those with determinants equal to ± 1). We are interested in finding a basis within $\mathcal{D}$ for which the original stochastic process either becomes sparsest or least statistically dependent. Let $\mathcal{C}(B \,|\, \boldsymbol{X})$ be a numerical measure of *deficiency* or *cost* of the basis B given the input stochastic process $\boldsymbol{X}$. Under this setting, the *best basis* for the stochastic process $\boldsymbol{X}$ among $\mathcal{D}$ relative to the cost $\mathcal{C}$ is written as $B_\star = \arg\min_{B \in \mathcal{D}} \mathcal{C}(B \,|\, \boldsymbol{X})$.

We also note that log in this paper implies $\log_2$, unless stated otherwise. The $n \times n$ identity matrix is denoted by I_n, and the $n \times 1$ column vector whose entries are all ones, i.e., $(1, 1, \ldots, 1)^T$, is denoted by $\mathbf{1}_n$.

3. Sparsity vs. Statistical Independence

We now define measures of sparsity and statistical independence for the basis of a given stochastic process.

Sparsity. Sparsity is a key property for compression. The true sparsity measure for a given vector $\boldsymbol{x} \in \mathbb{R}^n$ is the so-called ℓ^0 quasi-norm which is defined as

$$\|\boldsymbol{x}\|_0 \overset{\mathrm{def}}{=} \#\{i \in [1, n] : x_i \neq 0\},$$

i.e., the number of nonzero components in $\boldsymbol{x}$. This measure is, however, very unstable for even small geometric perturbations of the components in a vector.

Therefore, a better measure is the ℓ^p norm:

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

In fact, this is a quasi-norm for $0 < p < 1$ since it does not satisfy the triangle inequality, but only the weaker conditions: $\|\boldsymbol{x}+\boldsymbol{y}\|_p \le 2^{-1/p'}(\|\boldsymbol{x}\|_p+\|\boldsymbol{y}\|_p)$ where $p' = p/(p-1)$ is the conjugate exponent of p; and $\|\boldsymbol{x}+\boldsymbol{y}\|_p^p \le \|\boldsymbol{x}\|_p^p + \|\boldsymbol{y}\|_p^p$. It is easy to show that $\lim_{p \downarrow 0} \|\boldsymbol{x}\|_p^p = \|\boldsymbol{x}\|_0$. See [11] for the details of the ℓ^p norm properties.

Thus, we use the expected ℓ^p-norm minimization as a criterion to find the best basis for a given stochastic process in terms of sparsity:

$$\mathcal{C}_p(B \,|\, \boldsymbol{X}) = E\|B^{-1}\boldsymbol{X}\|_p^p, \tag{3–1}$$

We propose to minimize this cost in order to select the *best sparsifying basis* (BSB):

$$B_p = \arg\min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \boldsymbol{X}).$$

Remark 3.1. It should be noted that *minimization of the ℓ^p norm can also be achieved for each realization.* Without taking the expectation in (3–1), we can select the BSB, $B_p = B_p(\boldsymbol{x}, \mathcal{D})$ for each realization $\boldsymbol{x}$. We can guarantee that

$$\min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \boldsymbol{X} = \boldsymbol{x}) \le \min_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \boldsymbol{X}) \le \max_{B \in \mathcal{D}} \mathcal{C}_p(B \,|\, \boldsymbol{X} = \boldsymbol{x}).$$

For highly variable or erratic stochastic processes, $B_p(\boldsymbol{x}, \mathcal{D})$ may change significantly for each $\boldsymbol{x}$. Thus if we adopt this strategy to compress an entire training dataset consisting of N realizations, we need to store additional information in order to describe a set of N bases.

Whether we should adapt a basis per realization or on the average is still an open issue. See [26] for more details.

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 each f_{Y_k} is a one-dimensional marginal pdf of $f_{\boldsymbol{Y}}$. Statistical independence is a key property for compressing and modeling a stochastic process because: (1) an n-dimensional stochastic process of interest can be modeled as a set of one-dimensional processes; and (2) damage of one coordinate does not propagate to the others. Of course, in general, it is difficult to find a truly statistically independent coordinate system for a given stochastic process. Such a coordinate system may not even exist for a given stochastic process. Therefore, the next best thing is to find the least statistically-dependent coordinate system within a basis dictionary. Naturally, then, we need to measure the "closeness" of a coordinate system (or random variables) $Y_1, \ldots, Y_n$ 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 marginal pdfs:

$$I(\boldsymbol{Y}) \stackrel{\text{def}}{=} \int f_{\boldsymbol{Y}}(\boldsymbol{y}) \log \frac{f_{\boldsymbol{Y}}(\boldsymbol{y})}{\prod_{i=1}^{n} f_{Y_i}(y_i)} \, \mathrm{d}\boldsymbol{y}$$
$$= -H(\boldsymbol{Y}) + \sum_{i=1}^{n} H(Y_i),$$

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

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

We note that $I(\boldsymbol{Y}) \geq 0$, and $I(\boldsymbol{Y}) = 0$ if and only if the components of $\boldsymbol{Y}$ are mutually independent. See [9] for more details of the mutual information.

Suppose $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ and $B \in \mathrm{GL}(n, \mathbb{R})$ with $\det B = \pm 1$. We denote this set of matrices by $\mathrm{SL}^{\pm}(n, \mathbb{R})$. Note that the usual $\mathrm{SL}(n, \mathbb{R})$ is a subset of $\mathrm{SL}^{\pm}(n, \mathbb{R})$. 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 an invertible volume-preserving linear transformation:

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

because $|\det B^{-1}| = 1$. Based on this fact, we proposed the minimization of the following cost function as the criterion to select the so-called *least statistically-dependent basis* (LSDB) in the basis dictionary context [25]:

$$\mathcal{C}_H(B \,|\, \boldsymbol{X}) = \sum_{i=1}^{n} H\left((B^{-1}\boldsymbol{X})_i\right) = \sum_{i=1}^{n} H(Y_i). \tag{3–2}$$

Now we can define the LSDB as

$$B_{LSDB} = \arg\min_{B \in \mathcal{D}} \mathcal{C}_H(B \,|\, \boldsymbol{X}).$$

Closely related to the LSDB is the concept of the *kurtosis-maximizing basis* (KMB). This is based on the approximation of the marginal differential entropy $H(Y_i)$ in (3–2) by higher order moments/cumulants using the Edgeworth expansion and was derived by Comon [8]:

$$H(Y_i) \approx -\frac{1}{48}\kappa(Y_i) = -\frac{1}{48}(\mu_4(Y_i) - 3\mu_2^2(Y_i)) \tag{3–3}$$

where $\mu_k(Y_i)$ is the k-th central moment of Y_i, and $\kappa(Y_i)\ /\ \mu_2^2(Y_i)$ is called the *kurtosis* of Y_i. See also Cardoso [6] for a nice exposition of the various approximations to the mutual information. Now, the KMB is defined as follows:

$$B_\kappa = \arg\min_{B\in\mathcal{D}} \mathcal{C}_\kappa(B \mid \boldsymbol{X}) = \arg\max_{B\in\mathcal{D}} \sum_{i=1}^{n} \kappa(Y_i), \tag{3–4}$$

where $\mathcal{C}_\kappa(B \mid \boldsymbol{X}) = -\sum_{i=1}^{n} \kappa(Y_i)$. (This involves a slight abuse of terminology: the name is "kurtosis-maximizing basis" although what is maximized is the unnormalized κ, without the factor $1/\mu_2^2$.) Note that the LSDB and the KMB are tightly related, yet can be different. After all, (3–3) is simply an approximation to the entropy up to the fourth order cumulant. We also would like to point out that Buckheit and Donoho [4] independently proposed the same measure as a basis selection criterion, whose objective was to find a basis under which an input stochastic process looks maximally "non-Gaussian."

Remark 3.2. Earlier work of Pham [24] also proposes minimization of the cost (3–2). We would like to point out the main difference between our work [25] and Pham's. We use the basis libraries such as wavelet packets and local Fourier bases that allow us to deal with datasets with large dimensions such as face images whereas Pham used the more general dictionary $\mathrm{GL}(n, \mathbb{R})$. In practice, however, the numerical optimization (3–2) clearly becomes more difficult in his general case particularly if we want to use this for high dimensional datasets.

4. Review of Previous Results on the Simple Spike Process

In this section, we briefly summarize the results of the simple spike process. See [3] for the details and proofs.

An n-dimensional *simple spike process* generates the standard basis vectors $\{\boldsymbol{e}_j\}_{j=1}^{n} \subset \mathbb{R}^n$ in a random order, where $\boldsymbol{e}_j$ has one at the j-th entry and all the other entries are zero. We can view this process as a unit impulse located at a random position between 1 and n.

The Karhunen–Loève basis. The Karhunen–Loève basis of this process is not unique and not useful because of the following proposition.

Proposition 4.1. *The Karhunen–Loève basis for the simple spike process is any orthonormal basis in* $\mathbb{R}^n$ *containing the "DC" vector* $\mathbf{1}_n = (1, 1, \dots, 1)^T$.

This proposition reflects the non-Gaussian nature of the simple spike process, i.e., the optimality of the KLB can be claimed only for the Gaussian processes.

The Best Sparsifying Basis. As for the BSB, we have the following result:

Theorem 4.2. *The BSB with any* $p \in [0, 1]$ *for the simple spike process is the standard basis if* $\mathcal{D} = \mathrm{O}(n)$ *or* $\mathrm{SL}^{\pm}(n, \mathbb{R})$.

Statistical dependence and entropy of the simple spike process. Before stating the results on the LSDB of this process, we note a few specifics about the simple spike process. First, although the standard basis is the BSB for this process, it clearly does not provide the statistically independent coordinates. The existence of a single spike at one location prohibits spike generation at other locations. This implies that these coordinates are highly statistically dependent.

Second, we can compute the true entropy $H(\boldsymbol{X})$ for this process unlike other complicated stochastic processes. Since the simple spike process selects one possible vector from the standard basis vectors of $\mathbb{R}^n$ with uniform probability $1/n$, the true entropy $H(\boldsymbol{X})$ is clearly $\log n$. This is one of the rare cases where we know the true high-dimensional entropy of the process.

The LSDB among $\mathrm{O}(n)$**.** For $\mathcal{D} = \mathrm{O}(n)$, we have:

THEOREM 4.3. *The LSDB among* $\mathrm{O}(n)$ *is*:

- *for* $n \geq 5$, *either the standard basis or the basis whose matrix representation is*
$$\frac{1}{n}\begin{bmatrix} n-2 & -2 & \cdots & -2 & -2 \\ -2 & n-2 & \ddots & & -2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -2 & & \ddots & n-2 & -2 \\ -2 & -2 & \cdots & -2 & n-2 \end{bmatrix}; \tag{4–1}$$
- *for* $n = 4$, *the Walsh basis*, *i.e.*,
$$\frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix};$$
- *for* $n = 3$,
$$\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix};$$
- *for* $n = 2$, $\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, *and this is the only case where the true independence is achieved*.

REMARK 4.4. Note that when we say the basis is a matrix as above, we really mean that the column vectors of that matrix form the basis. This also means that any permuted and/or sign-flipped (i.e., multiplied by -1) versions of those column vectors also form the basis. Therefore, when we say the basis is a matrix A, we mean not only A but also its permuted and sign-flipped versions of A. This remark also applies to all the propositions and theorems below, unless stated otherwise.

Remark 4.5. There is an important geometric interpretation of (4–1). This matrix can also be written as:

$$B_{HR(n)} \overset{\text{def}}{=} I_n - 2\frac{\mathbf{1}_n}{\sqrt{n}}\frac{\mathbf{1}_n^T}{\sqrt{n}}.$$

In other words, this matrix represents the *Householder reflection* with respect to the hyperplane $\{\boldsymbol{y} \in \mathbb{R}^n \mid \sum_{i=0}^{n} y_i = 0\}$ whose unit normal vector is $\mathbf{1}_n/\sqrt{n}$.

Below, we use the notation $B_{\mathrm{O}(n)}$ for the LSDB among $\mathrm{O}(n)$ to distinguish it from the LSDB among $\mathrm{GL}(n,\mathbb{R})$, which is denoted by $B_{\mathrm{GL}(n)}$. So, for example, for $n \geq 5$, $B_{\mathrm{O}(n)} = I_n$ or $B_{HR(n)}$.

The LSDB among $\mathrm{GL}(n,\mathbb{R})$**.** As discussed in [3], for the simple spike process, there is no important distinction in the LSDB selection from $\mathrm{GL}(n,\mathbb{R})$ and from $\mathrm{SL}^{\pm}(n,\mathbb{R})$. Therefore, we do not have to treat these two cases separately. On the other hand, the generalized spike process in Section 5 requires us to treat $\mathrm{SL}^{\pm}(n,\mathbb{R})$ and $\mathrm{GL}(n,\mathbb{R})$ differently due to the continuous amplitude of the generated spikes.

We now have a curious theorem:

Theorem 4.6. *The LSDB among* $\mathrm{GL}(n,\mathbb{R})$ *with* $n > 2$ *is the following basis pair* (*for analysis and synthesis respectively*):

$$B_{\mathrm{GL}(n)}^{-1} = \begin{bmatrix} a & a & \cdots & \cdots & \cdots & \cdots & a \\ b_2 & c_2 & b_2 & \cdots & \cdots & \cdots & b_2 \\ b_3 & b_3 & c_3 & b_3 & \cdots & \cdots & b_3 \\ \vdots & \vdots & & \ddots & & & \vdots \\ \vdots & \vdots & & & \ddots & & \vdots \\ b_{n-1} & \cdots & \cdots & \cdots & b_{n-1} & c_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & \cdots & \cdots & b_n & c_n \end{bmatrix}, \qquad (4\text{–}2)$$

$$B_{\mathrm{GL}(n)} = \begin{bmatrix} \left(1+\sum_{k=2}^{n} b_k d_k\right)/a & -d_2 & -d_3 & \cdots & -d_n \\ -b_2 d_2/a & d_2 & 0 & \cdots & 0 \\ -b_3 d_3/a & 0 & d_3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -b_n d_n/a & 0 & \cdots & 0 & d_n \end{bmatrix} \qquad (4\text{–}3)$$

where a, b_k, c_k *are arbitrary real-valued constants satisfying* $a \neq 0$, $b_k \neq c_k$, *and* $d_k = 1/(c_k - b_k)$, $k = 2, \ldots, n$.

If we restrict ourselves to $\mathcal{D} = \mathrm{SL}^{\pm}(n,\mathbb{R})$, *then the parameter* a *must satisfy*:

$$a = \pm \prod_{k=2}^{n} (c_k - b_k)^{-1}.$$

REMARK 4.7. The LSDB such as (4–1) and the LSDB pair (4–2), (4–3) provide us with further insight into the difference between sparsity and statistical independence. In the case of (4–1), this is the LSDB, yet it does not sparsify the simple spike process at all. In fact, these coordinates are completely dense, i.e., $\mathcal{C}_0 = n$. We can also show that the sparsity measure $\mathcal{C}_p$ gets worse as $n \to \infty$. More precisely:

PROPOSITION 4.8.

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{HR(n)} \,|\, \boldsymbol{X}\right) = \begin{cases} \infty & \textit{if } 0 \le p < 1, \\ 3 & \textit{if } p = 1. \end{cases}$$

It is interesting to note that this LSDB approaches the standard basis as $n \to \infty$. This also implies that

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{HR(n)} \,|\, \boldsymbol{X}\right) \neq \mathcal{C}_p\left(\lim_{n\to\infty} B_{HR(n)} \,|\, \boldsymbol{X}\right).$$

As for the analysis LSDB (4–2), the ability to sparsify the simple spike process depends on the values of b_k and c_k. Since the parameters a, b_k and c_k are arbitrary as long as $a \neq 0$ and $b_k \neq c_k$, we put $a = 1$, $b_k = 0$, $c_k = 1$, for $k = 2, \dots, n$. Then we get the following specific LSDB pair:

$$B^{-1}_{\mathrm{GL}(n)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}, \quad B_{\mathrm{GL}(n)} = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \end{bmatrix}.$$

This analysis LSDB provides us with a sparse representation for the simple spike process (though this is clearly not better than the standard basis). For $\boldsymbol{Y} = B^{-1}_{\mathrm{GL}(n)}\boldsymbol{X}$,

$$\mathcal{C}_p = E\left[\|\boldsymbol{Y}\|_p^p\right] = \frac{1}{n} \times 1 + \frac{n-1}{n} \times 2 = 2 - \frac{1}{n}, \quad 0 \le p \le 1.$$

Now take $a = 1$, $b_k = 1$, $c_k = 2$ for $k = 2, \dots, n$ in (4–2) and (4–3). Then

$$B^{-1}_{\mathrm{GL}(n)} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{bmatrix}, \quad B_{\mathrm{GL}(n)} = \begin{bmatrix} n & -1 & \cdots & -1 \\ -1 & & & \\ \vdots & & I_{n-1} & \\ -1 & & & \end{bmatrix}.$$

The sparsity measure of this process is

$$\mathcal{C}_p = \frac{1}{n} \times n + \frac{n-1}{n} \times \{(n-1) + 2^p\} = n + (2^p - 1)\Big(1 - \frac{1}{n}\Big), \quad 0 \le p \le 1.$$

Therefore, the simple spike process under this analysis basis is completely dense, i.e., $\mathcal{C}_p \ge n$ for $0 \le p \le 1$ and the equality holds if and only if $p = 0$. Yet this is still the LSDB.

Finally, from Theorems 4.3 and 4.6, we have:

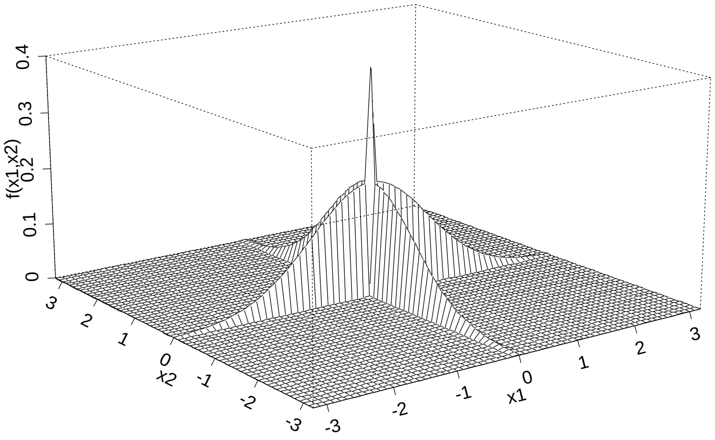

Figure 1. The pdf of the generalized spike process ($n = 2$).

Corollary 4.9. *There is no invertible linear transformation providing the statistically independent coordinates for the simple spike process for $n > 2$.*

5. The Generalized Spike Process

In [13], Donoho et al. analyze the following generalization of the simple spike process in terms of the KLB and the rate distortion function, which was recently followed up in details by Weidmann and Vetterli [28]. This process first picks one coordinate out of n coordinates randomly as before, but then the amplitude of this single spike is picked according to the standard normal distribution $\mathcal{N}(0, 1)$. The pdf of this process can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{j \neq i} \delta(x_j) \right) g(x_i), \tag{5–1}$$

where $\delta(\cdot)$ is the Dirac delta function, and $g(x) = (1/\sqrt{2\pi}) \cdot \exp(-x^2/2)$, i.e., the pdf of the standard normal distribution. Figure 1 shows this pdf for $n = 2$. Interestingly enough, this generalized spike process shows rather different behavior (particularly in the statistical independence) from the simple spike process in Section 4. We also note that our proofs here are rather analytical compared to those for the simple spike process presented in [3], which have a more combinatorial flavor.

The Karhunen–Loève basis. We can easily compute the covariance matrix of this process, which is proportional to the identity matrix. In fact, it is just I_n/n. Therefore, we have the following proposition, which was also stated without proof by Donoho et al. [13]:

Proposition 5.1. *The Karhunen–Loève basis for the generalized spike process is any orthonormal basis in* $\mathbb{R}^n$.

Proof. We first compute the marginal pdf of (5–1). By integrating out all x_i, $i \neq j$, we can easily get:

$$f_{X_j}(x_j) = \frac{1}{n} g(x_j) + \frac{n-1}{n} \delta(x_j).$$

Therefore, we have $E[X_j] = 0$. Since X_i and X_j cannot be simultaneously nonzero, we have

$$E[X_i X_j] = \delta_{ij} E[X_j^2] = \frac{1}{n} \delta_{ij},$$

since the variance of X_j is $1/n$, which is easily computed from the marginal pdf f_{X_j}. Therefore, the covariance matrix of this process is, as announced, I_n/n. Therefore, any orthonormal basis is the KLB. □

In other words, the KLB for this process is less restrictive than that for the simple spike process (Proposition 4.1), and the KLB is again completely useless for this process.

5.1. Marginal distributions and moments under $\mathrm{SL}^{\pm}(n,\mathbb{R})$**.** Before analyzing the BSB and LSDB, we need some background. First, we compute the pdf of the process relative to a transformation $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$, $B \in \mathrm{SL}^{\pm}(n,\mathbb{R})$. In general, if $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$, then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\det B^{-1}|} f_{\boldsymbol{X}}(B\boldsymbol{y}).$$

Therefore, from (5–1), and the fact $|\det B| = 1$, we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{j \neq i} \delta(\boldsymbol{r}_j^T \boldsymbol{y}) \right) g(\boldsymbol{r}_i^T \boldsymbol{y}), \tag{5–2}$$

where $\boldsymbol{r}_j^T$ is the j-th row vector of B. As for its marginal pdf, we have:

Lemma 5.2.

$$f_{Y_j}(y) = \frac{1}{n} \sum_{i=1}^{n} g(y; |\Delta_{ij}|), \quad j = 1, \dots, n, \tag{5–3}$$

where Δ_{ij} *is the* (i,j)*-th cofactor of matrix* B, *and* $g(y;\sigma) = g(y/\sigma)/\sigma$ *represents the pdf of the normal distribution* $\mathcal{N}(0,\sigma^2)$.

In other words, we can interpret the j-th marginal pdf as a *mixture of Gaussians* with the standard deviations $|\Delta_{ij}|$, $i = 1, \dots, n$. Figure 2 shows several marginal pdfs for $n = 2$. As we can see from this figure, it can vary from a very spiky distribution to a usual normal distribution depending on the rotation angle of the coordinate.

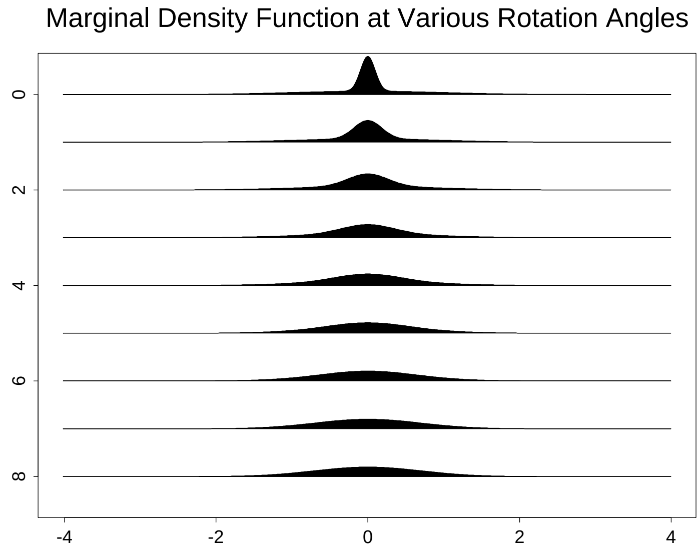

Figure 2. The marginal pdfs of the generalized spike process ($n = 2$). All the pdfs shown here are projections of the 2D pdf in Figure 1 onto the rotated 1D axis. The axis angle in the top row is 0.088 rad., which is close to the the first axis of the standard basis. The axis angle in the bottom row is $\pi/4$ rad., i.e., 45 degree rotation, which gives rise to the exact normal distribution. The other axis angles are equispaced between these two.

Proof. Rewrite (5–2) as

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{r}_1^T\boldsymbol{y})\cdots\delta(\boldsymbol{r}_{i-1}^T\boldsymbol{y})\delta(\boldsymbol{r}_{i+1}^T\boldsymbol{y})\cdots\delta(\boldsymbol{r}_n^T\boldsymbol{y})g(\boldsymbol{r}_i^T\boldsymbol{y}). \tag{5–4}$$

The j-th marginal pdf can be written as

$$f_{Y_j}(y_j) = \int f_{\boldsymbol{Y}}(y_1,\cdots,y_n)\,\mathrm{d}y_1\cdots\mathrm{d}y_{j-1}\,\mathrm{d}y_{j+1}\cdots\mathrm{d}y_n.$$

Consider the i-th term in the summation of (5–4) and integrate it out with respect to $y_1,\ldots,y_{j-1},y_{j+1},\ldots,y_n$:

$$\int \delta(\boldsymbol{r}_1^T\boldsymbol{y})\cdots\delta(\boldsymbol{r}_{i-1}^T\boldsymbol{y})\delta(\boldsymbol{r}_{i+1}^T\boldsymbol{y})\cdots\delta(\boldsymbol{r}_n^T\boldsymbol{y})g(\boldsymbol{r}_i^T\boldsymbol{y})\,\mathrm{d}y_1\cdots\mathrm{d}y_{j-1}\,\mathrm{d}y_{j+1}\cdots\mathrm{d}y_n. \tag{5–5}$$

We use the change of variable formula to integrate this. Let $\boldsymbol{r}_k^T\boldsymbol{y} = x_k$, $k = 1,\ldots,n$, and let $\boldsymbol{b}_\ell$ be the ℓ-th column vector of B. The relationship $B\boldsymbol{y} = \boldsymbol{x}$ can be rewritten as

$$B^{(i,j)}\boldsymbol{y}^{(j)} + y_j\boldsymbol{b}_j^{(i)} = \boldsymbol{x}^{(i)},$$

where $B^{(i,j)}$ is the $(n-1)\times(n-1)$ matrix by removing i-th row and j-th column, and the vectors with superscripts indicate the length $n-1$ column vectors by removing the elements whose indices are specified in the parentheses. The above equation can be rewritten as

$$\boldsymbol{y}^{(j)} = \left(B^{(i,j)}\right)^{-1}\left(\boldsymbol{x}^{(i)} - y_j \boldsymbol{b}_j^{(i)}\right).$$

Thus,

$$\begin{aligned} \mathrm{d}\boldsymbol{y}^{(j)} &= \mathrm{d}y_1 \cdots \mathrm{d}y_{j-1}\,\mathrm{d}y_{j+1} \cdots \mathrm{d}y_n \\ &= \frac{1}{|\det B^{(i,j)}|}\,\mathrm{d}\boldsymbol{x}^{(i)} \\ &= \frac{1}{|\Delta_{ij}|}\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n. \end{aligned}$$

We now express $\boldsymbol{r}_i^T \boldsymbol{y} = x_i$ in terms of y_j and $\boldsymbol{x}$.

$$\begin{aligned} \boldsymbol{r}_i^T \boldsymbol{y} &= \left(\boldsymbol{r}_i^{(j)}\right)^T \boldsymbol{y}^{(j)} + b_{ij} y_j \\ &= \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1}\left(\boldsymbol{x}^{(i)} - y_j \boldsymbol{b}_j^{(i)}\right) + b_{ij} y_j \\ &= \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{x}^{(i)} + y_j\left(b_{ij} - \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{b}_j^{(i)}\right) \\ &\overset{(*)}{=} \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{x}^{(i)} + \frac{y_j}{\Delta_{ij}} \det B \\ &= \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{x}^{(i)} \pm \frac{y_j}{\Delta_{ij}}, \end{aligned} \tag{5–6}$$

where $(*)$ follows from a lemma proved in Appendix A:

Lemma 5.3. *For any* $B = (b_{ij}) \in \mathrm{GL}(n, \mathbb{R})$,

$$b_{ij} - \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{b}_j^{(i)} = \frac{1}{\Delta_{ij}} \det B, \quad 1 \le i, j \le n.$$

Now let's go back to the integration (5–5). Thanks to the property of the delta function with Equation (5–6), we have

$$\begin{aligned} \int \cdots \int \delta(x_1) \cdots \delta(x_{i-1}) \delta(x_{i+1}) \cdots \delta(x_n) g(\boldsymbol{r}_i^T \boldsymbol{y}) \frac{1}{|\Delta_{ij}|}\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n \\ = \frac{1}{|\Delta_{ij}|} g(\pm y_j / \Delta_{ij}) \\ = g(y_j; |\Delta_{ij}|), \end{aligned}$$

where we used the fact that $g(\cdot)$ is an even function. Therefore, we can write the j-th marginal distribution as announced in (5–3). □

We now compute the moments of Y_i, which will be used later. We use the fact that this is a mixture of n Gaussians each of which has mean 0 and variance $|\Delta_{ij}|^2$. The following lemma computes the higher order moments.

Lemma 5.4.

$$E[|Y_j|^p] = \frac{\Gamma(p)}{n\, 2^{p/2-1} \Gamma(p/2)} \sum_{i=1}^{n} |\Delta_{ij}|^p, \quad \textit{for all } p > 0. \tag{5–7}$$

Proof. We have:

$$E[|Y_j|^p] = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} |y|^p g(y; |\Delta_{ij}|)\, \mathrm{d}y$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{2}{\pi}}|\Delta_{ij}|^p\Gamma(1+p)D_{-1-p}(0)$$

by Gradshteyn and Ryzhik [14, Formula 3.462.1], where $D_{-1-p}(\cdot)$ is Whittaker's function as defined by Abramowitz and Stegun [1, pp.687]:

$$D_{-a-1/2}(0) = U(a,0) = \frac{\sqrt{\pi}}{2^{a/2+1/4}\,\Gamma(a/2+3/4)}.$$

Thus, putting $a = p + 1/2$ to the above equation yields:

$$D_{-1-p}(0) = \frac{\sqrt{\pi}}{2^{1/2+p/2}\,\Gamma(1+p/2)}.$$

Therefore,

$$E[|Y_j|^p] = \frac{1}{n}\sum_{i=1}^{n}|\Delta_{ij}|^p\frac{\Gamma(1+p)}{2^{p/2}\,\Gamma(1+p/2)}$$
$$= \frac{1}{n}\sum_{i=1}^{n}|\Delta_{ij}|^p\frac{\Gamma(p)}{2^{p/2-1}\,\Gamma(p/2)}$$
$$= \frac{\Gamma(p)}{n\,2^{p/2-1}\,\Gamma(p/2)}\sum_{i=1}^{n}|\Delta_{ij}|^p,$$

as we desired. □

The Best Sparsifying Basis. As for the BSB, there is no difference after all between the generalized spike process and the simple spike process.

Theorem 5.5. *The BSB with any* $p \in [0,1]$ *for the generalized spike process is the standard basis if* $\mathcal{D} = \mathrm{O}(n)$ *or* $\mathrm{SL}^{\pm}(n,\mathbb{R})$.

Proof. We first consider the case $p \in (0,1]$. Using Lemma 5.4, the cost function (3–1) can be rewritten as

$$\mathcal{C}_p(B \,|\, \boldsymbol{X}) = \sum_{j=1}^{n} E[|Y_j|^p] = \frac{\Gamma(p)}{n\,2^{p/2-1}\,\Gamma(p/2)}\sum_{i=1}^{n}\sum_{j=1}^{n}|\Delta_{ij}|^p.$$

We now define a matrix $\tilde{B} \overset{\text{def}}{=} (\Delta_{ij})$. Then $\tilde{B} \in \mathrm{SL}^{\pm}(n,\mathbb{R})$ since

$$B^{-1} = \frac{1}{\det B}(\Delta_{ji}) = \pm(\Delta_{ji}),$$

and $B^{-1} \in \mathrm{SL}^{\pm}(n,\mathbb{R})$. Therefore, this reduces to

$$\mathcal{C}_p(B \,|\, \boldsymbol{X}) = \frac{\Gamma(p)}{n\,2^{p/2-1}\,\Gamma(p/2)}\sum_{i=1}^{n}\sum_{j=1}^{n}|\tilde{b}_{ij}|^p = K(p,n)\cdot\mathcal{C}_p(\tilde{B} \,|\, \tilde{\boldsymbol{X}}),$$

where $\tilde{\boldsymbol{X}}$ represents the simple spike process, and $K(p,n)$ is the constant before the double summations above, which is dependent only on p and n. This means that for fixed p and n, searching for the B that minimizes the sparsity cost for the generalized spike process is equivalent to searching for the $\tilde{B}$ that minimizes the sparsity cost for the simple spike process. Thus, Theorem 9.5.1 in [3] (or Theorem 4.2 in this paper) asserts that the $\tilde{B}$ must be the identity matrix I_n or its permuted or sign flipped versions. Suppose $\Delta_{ij} = \delta_{ij}$. Then, $B^{-1} = \pm(\Delta_{ji}) = \pm I_n$, which implies that $B = \pm I_n$. If (Δ_{ji}) is any permutation matrix, then B^{-1} is just that permutation matrix or its sign flipped version. Therefore, B is also a permutation matrix or its sign flipped version.

Finally, consider the case $p = 0$. Then, any linear invertible transformation except the identity matrix or its permuted or sign-flipped versions clearly increases the number of nonzero elements after the transformation. Therefore, the BSB with $p = 0$ is also a permutation matrix or its sign flipped version.

This completes the proof of Theorem 5.5. □

The LSDB/KMB among $\mathrm{O}(n)$**.** As for the LSDB/KMB, we can see some differences from the simple spike process.

We first consider the case of $\mathcal{D} = \mathrm{O}(n)$. So far, we have been unable to prove the following conjecture.

Conjecture 5.6. *The LSDB among* $\mathrm{O}(n)$ *is the standard basis.*

The difficulty is the evaluation of the sum of the marginal entropies (3–2) for the pdfs of the form (5–3). However, a major simplification occurs if we consider the KMB instead of the LSDB, and we can prove:

Theorem 5.7. *The KMB among* $\mathrm{O}(n)$ *is the standard basis.*

Proof. Because $E[Y_j] = 0$, $E[Y_j^2] = \frac{1}{n}\sum_{i=1}^n \Delta_{ij}^2$, and $\mu_4(Y_j) = \frac{3}{n}\sum_{i=1}^n \Delta_{ij}^4$ by (5–7), the cost function in (3–4) becomes

$$\mathcal{C}_\kappa(B \mid \boldsymbol{X}) = \frac{3}{n}\sum_{j=1}^n \left(\sum_{i=1}^n \Delta_{ij}^4 - \frac{1}{n}\left(\sum_{i=1}^n \Delta_{ij}^2 \right)^2 \right). \tag{5–8}$$

Note that this is true for any $B \in \mathrm{SL}^\pm(n, \mathbb{R})$. If we restrict our basis search to the set $\mathrm{O}(n)$, another major simplification occurs because we have a special relationship between Δ_{ij} and the matrix element b_{ji} of $B \in \mathrm{O}(n)$:

$$B^{-1} = \frac{1}{\det B}\,(\Delta_{ji}) = B^T.$$

In other words,

$$\Delta_{ij} = (\det B) b_{ij} = \pm b_{ij}.$$

Therefore,

$$\sum_{i=1}^n \Delta_{ij}^2 = \sum_{i=1}^n b_{ij}^2 = 1.$$

Inserting this into (5–8), we get a simplified cost for $\mathcal{D} = \mathrm{O}(n)$:

$$\mathcal{C}_\kappa(B \,|\, \boldsymbol{X}) = -\frac{3}{n}\left(1 - \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij}^4\right).$$

This means that the KMB can be rewritten as

$$B_\kappa = \arg \max_{B\in\mathrm{O}(n)} \sum_{i,j} b_{ij}^4. \tag{5–9}$$

Note that the existence of the maximum is guaranteed because the set $\mathrm{O}(n)$ is *compact* and the cost function $\sum_{i,j} b_{ij}^4$ is continuous.

Now consider a matrix $P = (p_{ij}) = (b_{ij}^2)$. Then, from the orthonormality of columns and rows of B, this matrix P belongs to a set of *doubly stochastic matrices* $\mathcal{S}(n)$. Since doubly stochastic matrices obtained by squaring the elements of $\mathrm{O}(n)$ consist of a proper subset of $\mathcal{S}(n)$, we have

$$\max_{B\in\mathrm{O}(n)} \sum_{i,j} b_{ij}^4 \le \max_{P\in\mathcal{S}(n)} \sum_{i,j} p_{ij}^2.$$

Now we prove that such P must be an identity matrix or its permuted version.

$$\max_{P\in\mathcal{S}(n)} \sum_{j=1}^{n}\sum_{i=1}^{n} p_{ij}^2 \le \sum_{j=1}^{n}\left(\max_{\sum_{i=1}^{n} p_{ij}=1} \sum_{i=1}^{n} p_{ij}^2\right) = \sum_{j=1}^{n} 1 = n,$$

where the first equality follows from the fact that maxima of the radius of the sphere $\sum_i p_{ij}^2$ subject to $\sum_i p_{ij} = 1$, $p_{ij} \ge 0$ occur only at the vertices of that simplex, i.e., $\boldsymbol{p}_j = \boldsymbol{e}_{\sigma(j)}$, $j = 1, \ldots, n$ where $\sigma(\cdot)$ is a permutation of n items. That is, the column vectors of P must be the standard basis vectors. This implies that the matrix B corresponding to $P = I_n$ or its permuted version must be either I_n or its permuted and/or sign-flipped version. □

The LSDB/KMB among $\mathrm{SL}^{\pm}(n,\mathbb{R})$**.** If we extend our search to this more general case, we have:

Theorem 5.8. *The KMB among* $\mathrm{SL}^{\pm}(n,\mathbb{R})$ *does not exist.*

Proof. The set $\mathrm{SL}^{\pm}(n,\mathbb{R})$ is *not* compact. Therefore, there is no guarantee that the cost function $\mathcal{C}_\kappa(B \,|\, \boldsymbol{X})$ has a minimum value on this set. In fact, there is a simple counterexample: let $B = \mathrm{diag}(a, a^{-1}, 1, \cdots, 1)$, where a is any nonzero real scalar. Then $\mathcal{C}_\kappa(B \,|\, \boldsymbol{X}) = -(a^4 + a^{-4} + n - 2)$ tends to $-\infty$ as a increases to ∞. □

As for the LSDB, we do not know whether the LSDB exists among $\mathrm{SL}^{\pm}(n,\mathbb{R})$ at this point, although we believe that the LSDB is the standard basis. The negative result in the KMB does not necessarily imply the negative result in the LSDB.

6. An Application to the Ramp Process

Although the generalized spike process is a simple stochastic process, we have the following important interpretation. Consider a stochastic process generating a basis vector randomly selected from some fixed orthonormal basis and multiplied by a scalar varying as the standard normal distribution at a time. Then, that basis itself is simultaneously the BSB and the KMB among $\mathrm{O}(n)$. Theorems 5.5 and 5.7 claim that once we transform the data to the generalized spike process, we cannot do any better than that, both in terms of sparsity and independence within $\mathrm{O}(n)$.

Along this line of thought, we now consider the following stochastic process as an application of the theorems in this paper:

$$X(t) = \nu \cdot (t - H(t-\tau)), \quad t \in [0,1),\, \nu \sim \mathcal{N}(0,1),\, \tau \sim \mathrm{unif}[0,1), \qquad (6\text{–}1)$$

where $H(\cdot)$ is the Heaviside step function, i.e., $H(t) = 1$ if $t \geq 0$ and 0 otherwise. This is a generalized version of the ramp process of Yves Meyer [20, p. 19]. Some realizations of the simple ramp process are shown in Figure 3.

We now consider the discrete version of (6–1). Let our sampling points be $t_k = \frac{2k+1}{2n}$, $k = 0, \ldots, n-1$. Suppose the discontinuity (at $t = \tau$) does not happen at the exact sampling points. Then all the realizations whose discontinuities are located anywhere in the open interval $(\frac{2k-1}{2n}, \frac{2k+1}{2n})$ have the same discretized

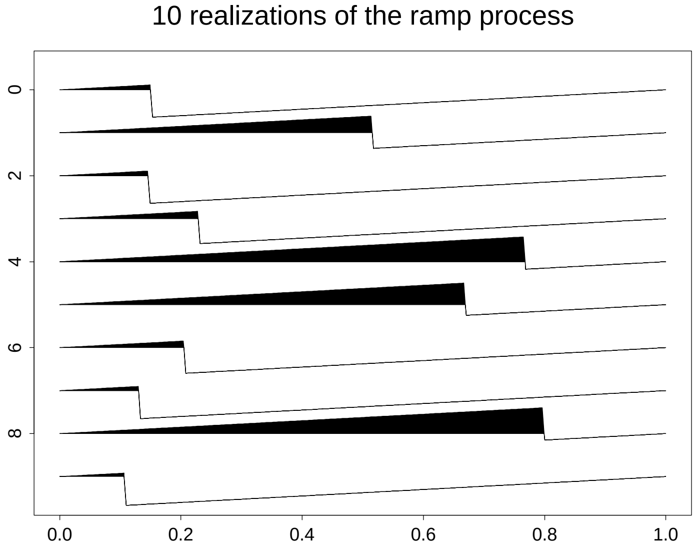

Figure 3. Ten realizations of the simple ramp process. The position of the discontinuity is picked uniformly randomly from the interval $[0,1)$. A realization of the generalized ramp process can be obtained by multiplying a scalar picked from the standard normal distribution to a realization of the simple ramp process.

version. Therefore, any realization now has the form

$$\tilde{\boldsymbol{x}}_j = \nu \boldsymbol{x}_j = \nu(x_{0j}, \ldots, x_{n-1,j})^T, \quad x_{kj} = \begin{cases} \frac{2k+1}{2n} & \text{for } k = 0, \ldots, j-1, \\ \frac{2k+1}{2n} - 1 & \text{for } k = j, \ldots, n-1, \end{cases}$$

where j is picked uniformly randomly from the set $\{0, 1, \cdots, n-1\}$. (Note that the index of the vector components starts with 0 for convenience). Then:

THEOREM 6.1. *The BSB pair of the discretized version of the generalized ramp process* (6–1), *selected from* $\mathrm{SL}^{\pm}(n, \mathbb{R})$, *are*:

$$B_{\text{ramp}}^{-1} = (2n)^{-1/n} \begin{bmatrix} -1 & 0 & \cdots & \cdots & 0 & -1 \\ 1 & -1 & 0 & \cdots & 0 & -2 \\ 0 & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & 1 & -1 & 0 & -2 \\ \vdots & & \ddots & 1 & -1 & -2 \\ 0 & \cdots & \cdots & 0 & 1 & -3 \end{bmatrix}, \tag{6–2}$$

$$B_{\text{ramp}} = (2n)^{1/n} \left[\begin{array}{c|c|c|c} \boldsymbol{x}_0 & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{n-1} \end{array}\right]. \tag{6–3}$$

PROOF. It is straightforward to show that the matrix without the factor $(2n)^{-1/n}$ in (6–2) is the inverse of the matrix $[\boldsymbol{x}_0|\boldsymbol{x}_1|\cdots|\boldsymbol{x}_{n-1}]$. Then, the factors $(2n)^{-1/n}$ and $(2n)^{1/n}$ in (6–2) and (6–3), which are easily obtained, are necessary for these matrices to be in $\mathrm{SL}^{\pm}(n, \mathbb{R})$. It is now clear that the analysis basis B_{ramp}^{-1} transforms the discretized version of the generalized ramp process to the generalized spike process whose amplitudes obey $\mathcal{N}(0, (2n)^{-2/n})$ instead of $\mathcal{N}(0, 1)$. Once converted to the generalized spike process, then from Theorem 5.5, we know that we cannot do any better than the standard basis in terms of the sparsity cost (3–1). This implies that the BSB among $\mathrm{SL}^{\pm}(n, \mathbb{R})$ is the basis pair (6–2) and (6–3). □

In fact, this matrix is a difference operator (with DC measurement) so that it detects the location of the discontinuity in each realization, while the synthesis basis vectors (6–3) are the realizations of this process themselves modulo scalar multiplications. Clearly, this matrix also transforms the discretized version of the simple ramp process (i.e., with $\nu \equiv 1$ in (6–1)) to the simple spike process whose nonzero amplitude is $(2n)^{-1/n}$. Therefore, if the realizations of the simple or generalized ramp process is fed to any software that is supposed to find a sparsifying basis among $\mathrm{SL}^{\pm}(n, \mathbb{R})$, then that software should be able to find (6–2) and (6–3). As a demonstration, we conducted a simple experiment using Cardoso's JADE (Joint Approximate Diagonalization of Eigenmatrices) algorithm [6] applied to the discretized version of the simple ramp process.

The JADE algorithm was designed to find a basis minimizing the sum of the squared fourth order cross-cumulants of the input data (i.e., essentially the

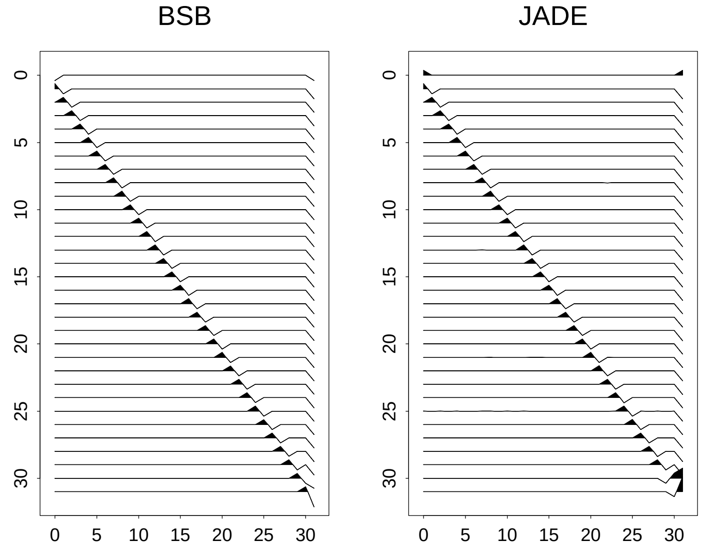

Figure 4. The analysis BSB vs. the analysis basis obtained by JADE algorithm ($n = 32$). A row permutation and a global amplitude normalization were applied to the JADE analysis basis to have a better correspondence with the BSB.

KMB) under the whitening condition, $\boldsymbol{E}\boldsymbol{Y}\boldsymbol{Y}^T = I_n$. In fact, the best basis is searched for within a subset of $\mathrm{GL}(n, \mathbb{R})$, which has a very special structure: every element in this set is of the form $B = W^{-1}U$ where W is the whitening matrix of the inputs $\boldsymbol{X}$ and $U \in \mathrm{O}(n)$. Note that this subset is neither $\mathrm{O}(n)$ nor $\mathrm{SL}^{\pm}(n, \mathbb{R})$. For our numerical experiment with JADE, we modified the code available from [5] so that it does not remove the mean of the input dataset. (Otherwise, we could only extract $n - 1$ basis vectors.) In Figure 4, we compare the theoretical optimum, i.e., the analysis BSB (6–2), and the analysis basis obtained by JADE, which is almost identical to the BSB (modulo permutations and sign flips).

Now, what happens if we restrict the basis search to the set $\mathrm{O}(n)$? The basis pair (6–2) and (6–3) are not orthogonal matrices. Therefore, we will never be able to find the basis pair (6–2), (6–3) within $\mathrm{O}(n)$. Consequently, even if we found the BSB among $\mathrm{O}(n)$, the ramp process would be less sparsified by that orthonormal BSB than by (6–2). Yet, it is of interest to determine the BSB within $\mathrm{O}(n)$ due to the numerical experiments of Cardoso and Donoho [7]. They apply the JADE algorithm without imposing the whitening condition to the discretized version of the simple ramp process. This strategy is essentially equivalent to searching the KMB within $\mathrm{O}(n)$. The resulting KMB, which they call “jadelets” [7], is very similar to Daubechies's almost symmetric wavelet basis called “symmlets” [10, Sec. 6.4]. For the generalized ramp process, the KMB among $\mathrm{SL}^{\pm}(n, \mathbb{R})$ may not exist as Theorem 5.8 shows, because within

$\mathrm{SL}^{\pm}(n,\mathbb{R})$, the generalized ramp process is equivalent to the generalized spike process via (6–2) and (6–3). On the other hand, we cannot convert the generalized ramp process to the generalized spike process within $\mathrm{O}(n)$, although the KMB among $\mathrm{O}(n)$ exists for the generalized spike process. These observations indicate that the orthonormality may be a key to generate the wavelet-like multiscale basis for the generalized ramp process. At this point, however, we do not fully understand why orthonormality has to be a key for generating such a wavelet-like multiscale basis. The mystery of the orthonormality was intensified after we failed to reproduce their results using the modified JADE algorithm. This issue needs to be investigated in the near future.

7. Discussion

Unlike the simple spike process, the BSB and the KMB (an alternative to the LSDB) selects the standard basis if we restrict our basis search to the set $\mathrm{O}(n)$. If we extend our basis search to $\mathrm{SL}^{\pm}(n,\mathbb{R})$, then the BSB exists and is again the standard basis whereas the KMB does not exist. Of course, if we extend the search to nonlinear transformations, then it becomes a different story. We refer the reader to our recent articles [18; 19], for the details of a nonlinear algorithm.

The results of this paper further support the conclusion of the previous work [3]: dealing with the BSB is much simpler than the LSDB. To deal with statistical dependency, we need to consider the probability law of the underlying process (e.g., entropy or the marginal pdfs) explicitly. That is why we need to consider the KMB instead of the LSDB to prove the theorems. Also in practice, given a finite set of training data, it is a nontrivial task to reliably estimate the marginal pdfs. Moreover, the LSDB unfortunately cannot tell how close it is to the true statistical independence; it can only tell that it is the best one (i.e., the closest one to the statistical independence) among the given set of possible bases. In order to quantify the absolute statistical dependence, we need to estimate the true high-dimensional entropy of the original process, $H(\boldsymbol{X})$, which is an extremely difficult task in general. We would like to note, however, a recent attempt to estimate the high-dimensional entropy of the process by Hero and Michel [15], which uses the minimum spanning trees of the input data and does not require us to estimate the pdf of the process. We feel that this type of techniques will help assessing the absolute statistical dependence of the process under the LSDB coordinates. Another interesting observation is that the KMB is rather sensitive to the orthonormality of the basis dictionary whereas the BSB is insensitive to that. Our previous results on the simple spike process (e.g., Theorems 4.3 and 4.6) also suggest the sensitivity of the LSDB to the orthonormality of the basis dictionary.

On the other hand, the sparsity criterion neither requires estimation of the marginal pdfs nor reveals the sensitivity to the orthonormality. Simply computing the expected ℓ^p norms suffices. Moreover, we can even adapt the BSB for each realization rather than for the whole realizations, which is impossible for the LSDB, as we discussed in [3; 26]. These observations, therefore, suggest that the pursuit of sparse representations should be encouraged rather than that of statistically independent representations. This is also the viewpoint indicated by Donoho [11].

Finally, there are a few interesting generalizations of the spike processes, which need to be addressed in the near future. We need to consider a stochastic process that randomly throws in multiple spikes to a single realization. If we throw in more and more spikes to one realization, the standard basis is getting worse in terms of sparsity. Also, we can consider various rules to throw in multiple spikes. For example, for each realization, we can select the locations of the spikes statistically independently. This is the simplest multiple spike process. Alternatively, we can consider a certain dependence in choosing the locations of the spikes. The ramp process of Yves Meyer ((6–1) with $\nu \equiv 1$) represented in the wavelet basis is such an example; each realization of the ramp process generates a small number of nonzero wavelet coefficients around the location of the discontinuity of that realization and across the scales. See [4; 13; 20; 26] for more about the ramp process.

Except in very special circumstances, it would be extremely difficult to find the BSB of a complicated stochastic process (e.g., natural scene images) that truly converts its realizations to the spike process. More likely, a theoretically and computationally feasible basis that sparsifies the realizations of a complicated process well (e.g., curvelets for the natural scene images [12]) may generate expansion coefficients that may be viewed as an amplitude-varying multiple spike process. In order to tackle this scenario, we certainly need to identify interesting, useful, and simple enough specific stochastic processes, develop the BSB adapted to such specific processes, and deepen our understanding of the amplitude-varying multiple spike process.

Acknowledgment

I would like to thank Dr. Jean-François Cardoso of ENST, Paris, and Dr. Motohico Mulase and Dr. Roger Wets, both of UC Davis, for fruitful discussions. This research was partially supported by NSF DMS-99-73032, DMS-99-78321, and ONR YIP N00014-00-1-0469.

Appendix A. Proof of Lemma 5.3

Proof. Consider the system of linear equations

$$B^{(i,j)}\boldsymbol{z}^{(j)} = \boldsymbol{b}_j^{(i)},$$

where $\boldsymbol{z}^{(j)} = (z_1, \cdots, z_{j-1}, z_{j+1}, \cdots, z_n)^T \in \mathbb{R}^{n-1}$, $j = 1, \ldots, n$. Using Cramer's rule (e.g., [16, pp.21]), we have, for $k = 1, \ldots, j-1, j+1, \ldots, n$,

$$\begin{aligned} z_k^{(j)} &= \frac{1}{\det B^{(i,j)}} \det \left[\boldsymbol{b}_1^{(i)} \middle| \cdots \middle| \boldsymbol{b}_{k-1}^{(i)} \middle| \boldsymbol{b}_j^{(i)} \middle| \boldsymbol{b}_{k+1}^{(i)} \middle| \cdots \middle| \boldsymbol{b}_n^{(i)} \right] \\ &\overset{(a)}{=} (-1)^{|k-j|-1} \frac{B^{(i,k)}}{B^{(i,j)}} \\ &\overset{(b)}{=} (-1)^{|k-j|-1} \frac{\Delta_{ik}/(-1)^{i+k}}{\Delta_{ij}/(-1)^{i+j}} = -\frac{\Delta_{ik}}{\Delta_{ij}}, \end{aligned}$$

where (a) follows from the $(|k-j|-1)$ column permutations to move $\boldsymbol{b}_j^{(i)}$ located at the k-th column to the j-th column of $B^{(i,j)}$, and (b) follows from the definition of the cofactor. Hence,

$$\begin{aligned} b_{ij} - \left(\boldsymbol{r}_i^{(j)}\right)^T \left(B^{(i,j)}\right)^{-1} \boldsymbol{b}_j^{(i)} = b_{ij} - \left(\boldsymbol{r}_i^{(j)}\right)^T \boldsymbol{z}^{(j)} &= b_{ij} + \frac{1}{\Delta_{ij}} \sum_{k \neq j} b_{ik} \Delta_{ik} \\ &= \frac{1}{\Delta_{ij}} \sum_{k=1}^{n} b_{ik} \Delta_{ik} = \frac{1}{\Delta_{ij}} \det B. \end{aligned}$$

This completes the proof of Lemma 5.3. □

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions*, 9th printing, Dover, New York, 1972.

[2] A. J. Bell and T. J. Sejnowski. "The 'independent components' of natural scenes are edge filters", *Vision Research*, 37:3327–3338, 1997.

[3] B. Bénichou and N. Saito, "Sparsity vs. statistical independence in adaptive signal representations: A case study of the spike process", pp. 225–257 in *Beyond wavelets*, Studies in Computational Mathematics **10**, edited by G. V. Welland, Academic Press, San Diego, 2003.

[4] J. B. Buckheit and D. L. Donoho. "Time-frequency tilings which best expose the non-Gaussian behavior of a stochastic process", pp. 1–4 in *Proc. International Symposium on Time-Frequency and Time-Scale Analysis* (Jun. 18–21, 1996, Paris), IEEE, 1996.

[5] J. F. Cardoso, "An efficient batch algorithm: JADE", available at http://sig.enst.fr/~cardoso/guidesepsou.html. See http://tsi.enst.fr/~cardoso/icacentral/index.html for collections of contributed ICA software.

[6] J.-F. Cardoso, "High-order contrasts for independent component analysis", *Neural Computation*, 11:157–192, 1999.

[7] J.-F. Cardoso and D. L. Donoho, "Some experiments on independent component analysis of non-Gaussian processes", pp. 74–77 in *Proc. IEEE Signal Processing International Workshop on Higher Order Statistics* (Cesarea, Israel), 1999.

[8] P. Comon, "Independent component analysis, a new concept?", *Signal Processing*, 36:287–314, 1994.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley–Interscience, New York, 1991.

[10] I. Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics **61**, SIAM, Philadelphia, PA, 1992.

[11] D. L. Donoho, "Sparse components analysis and optimal atomic decomposition", *Constructive Approximation*, 17:353–382, 2001.

[12] D. L. Donoho and A. G. Flesia, "Can recent innovations in harmonic analysis 'explain' key findings in natural image statistics?", *Network: Comput. Neural Syst.*, 12:371–393, 2001.

[13] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis", *IEEE Trans. Inform. Theory*, 44(6):2435–2476, 1998.

[14] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, sixth edition, Academic Press, 2000.

[15] A. O. Hero and O. J. J. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs", *IEEE Trans. Inform. Theory*, 45(6):1921–1938, 1999.

[16] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge Univ. Press, 1985.

[17] A. Hyvärinen, The FastICA package for MATLAB, http://www.cis.hut.fi/projects/ica/fastica/.

[18] J.-J. Lin, N. Saito, and R. A. Levine, "An iterative nonlinear Gaussianization algorithm for resampling dependent components", pp. 245–250 in *Proc. 2nd International Workshop on Independent Component Analysis and Blind Signal Separation* (June 19–22, 2000, Helsinki), edited by P. Pajunen and J. Karhunen, IEEE, 2000.

[19] J.-J. Lin, N. Saito, and R. A. Levine, An iterative nonlinear Gaussianization algorithm for image simulation and synthesis, Technical report, Dept. Math., Univ. California, Davis, 2001. Submitted for publication.

[20] Y. Meyer, *Oscillating patterns in image processing and nonlinear evolution equations*, University Lecture Series **22**, AMS, Providence, RI, 2001.

[21] B. A. Olshausen, Sparse coding simulation software, http://redwood.ucdavis.edu/bruno/sparsenet.html.

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

[23] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?", *Vision Research*, 37:3311–3325, 1997.

[24] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis", *IEEE Trans. Signal Process.*, 44(11):2768–2779, 1996.

[25] N. Saito, "Image approximation and modeling via least statistically dependent bases", *Pattern Recognition*, 34:1765–1784, 2001.

[26] N. Saito, B. M. Larson, and B. Bénichou, "Sparsity and statistical independence from a best-basis viewpoint. pp. 474–486 in *Wavelet Applications in Signal and*

Image Processing VIII, edited by A. Aldroubi et al., Proc. SPIE **4119**, 2000. Invited paper.

[27] 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, 265:359–366, 1998.

[28] C. Weidmann and M. Vetterli, "Rate distortion behavior of sparse sources", submitted to IEEE Trans. Info. Theory, Oct. 2001.

Naoki Saito
Department of Mathematics
University of California
Davis, CA 95616
United States
saito@math.ucdavis.edu