THE SPIKE PROCESS: A SIMPLE TEST CASE FOR INDEPENDENT OR SPARSE COMPONENT ANALYSIS

Naoki Saito[*] *and Bertrand Bénichou*[†]

Department of Mathematics
University of California
Davis, CA 95616 USA

ABSTRACT

We examine curious behaviors of the Independent Component Analysis (ICA) and Sparse Component Analysis (SCA) when they are applied to some simple stochastic processes called the "simple" and "generalized" spike processes. Both processes put a single spike at a random location in the zero vector of length n for each realization. The simple spike process puts a unit impulse whereas the generalized spike process puts a sample from the standard normal distribution. We obtained interesting set of theorems for these processes. The behavior of SCA to these processes turned out to be much simpler than that of ICA. Our results are useful for validating any ICA/SCA software package because it is very easy to simulate these processes and the desirable answers are known from our theorems.

1. INTRODUCTION

Both *sparsity* and *statistical independence* are important criteria for signal and image representations as demonstrated by the work in computational neuroscience [1], [2], [3], [4] as well as in computational harmonic analysis [5], [6], [7], to name a few. 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 [1], [2] 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 Gabor-like basis functions, which are similar to the receptive field profiles of the neurons in our primary visual cortex. Bell and Sejnowski [3] 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. We wish to deepen our understanding of this intricate relationship. Therefore we chose to study the so-called "spike" process, a simple synthetic stochastic process that generates an impulse at a random location in an n-dimensional vector for each realization. There are two different spike processes we deal with in this paper. One is the so-called "simple" spike process, which puts a unit impulse whose amplitude is constant 1. The other is the so-called "generalized" spike process, which puts an impulse whose amplitude is sampled from the standard normal distribution $\mathcal{N}(0,1)$. 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. Most of the proofs of the theorems below about the simple and generalized spike process can be found in [7] and [8], respectively.

2. NOTATIONS AND TERMINOLOGY

Let us first set our notation and the terminology. Let $\boldsymbol{X} \in \mathbb{R}^n$ be a random vector with some unknown pdf $f_{\boldsymbol{X}}$. Let $B \in \mathcal{D}$, where $\mathcal{D}$ is the so-called *basis dictionary*. For very high dimensional data, we often use the wavelet packets and local Fourier bases as $\mathcal{D}$ (see [9] and references therein for more about such basis dictionaries). In this paper, however, we use much more 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., their determinants are ± 1). We are interested in searching a basis under which the original stochastic process becomes either the sparsest or the least statistically dependent among the bases in $\mathcal{D}$. Let $\mathcal{C}(B \mid \boldsymbol{X})$ be a numerical measure of *deficiency* or *cost* of the basis B given the input

[*] This work was partially supported by NSF DMS-99-73032, DMS-99-78321, and ONR YIP N00014-00-1-046. Email: saito@math.ucdavis.edu

[†] Current address: ENST, 46, rue Barrault, 75634 Paris cedex 13 France

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 $B_\star = \arg\min_{B\in\mathcal{D}} \mathcal{C}(B \mid \boldsymbol{X})$.

3. SPARSITY VS. STATISTICAL INDEPENDENCE

Let us now define the measure of sparsity and that of statistical independence to evaluate a given basis (coordinate system).

3.1. Sparsity

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 which is defined as

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

i.e., the number of nonzero components in $\boldsymbol{x}$. This measure is, however, very unstable for even small perturbation of the components in a vector. Therefore, a 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 for $0 < p < 1$ since this does not satisfy the triangle inequality. It is easy to show that $\lim_{p\downarrow 0} \|\boldsymbol{x}\|_p^p = \|\boldsymbol{x}\|_0$. See [5] for the details of the ℓ^p norm properties.

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

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

We propose to use the minimization of this cost to select the *best sparsifying basis* (BSB):

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

3.2. 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} is a one-dimensional marginal pdf of $f_{\boldsymbol{Y}}$. 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 difficult to find a truly statistically independent coordinate system for a given stochastic process. Such a coordinate system may not even exist for a certain stochastic process. Therefore, we should be satisfied with finding the least-statistically dependent coordinate system within a basis dictionary. Naturally, then, we need to measure the "closeness" of a 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 marginal pdf's:

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

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

Suppose $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ and $B \in \mathrm{GL}(n,\mathbb{R})$ with $\det B = \pm 1$. We denote such a group of matrices by $\mathrm{SL}^{\pm}(n,\mathbb{R})$. Note that the usual $\mathrm{SL}(n,\mathbb{R})$ is a subgroup 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 such an invertible volume-preserving linear transformation, i.e., $H(B^{-1}\boldsymbol{X}) = H(\boldsymbol{X}) + \log|\det B^{-1}| = H(\boldsymbol{X})$, 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 [9]:

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

Now, we can define the LSDB as

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

We were informed that Pham [10] had proposed the minimization of the same cost (2) earlier. We would like to point out the main difference between our work [9] and Pham's. We used 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 more general dictionary $\mathrm{GL}(n,\mathbb{R})$. In practice, however, the numerical optimization (2) clearly becomes more difficult in his general case particularly if one wants to use this for high dimensional datasets.

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 (2) by higher order moments/cumulants using the Edgeworth expansion and

was derived by Comon [11]:

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

where $\mu_k(Y_i)$ is the kth central moment of Y_i, and $\kappa(Y_i) \,/\, \mu_2^2(Y_i)$ is called the *kurtosis* of Y_i. See also Cardoso [12] for a nice exposition of the various approximations to the mutual information. Now, the KMB is defined as follows:[1]

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

where $\mathcal{C}_\kappa(B \mid \boldsymbol{X}) = -\sum_{i=1}^{n} \kappa(Y_i)$. We note that the LSDB and the KMB are tightly related, yet can be different. After all, (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 [13] 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."

4. THE SIMPLE SPIKE PROCESS

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 jth entry and all the other entries are zero. One can view this process as a unit impulse located at a random position between 1 and n.

4.1. The Karhunen-Loève Basis

Let us first consider the Karhunen-Loève basis of this process from which we can learn a few things.

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

This means that the KLB is not useful for this process. This is because the simple spike process is highly non-Gaussian.

4.2. The Best Sparsifying Basis

It is obvious that the standard basis is the BSB among $\mathrm{O}(n)$ by construction; an expansion of a realization of this process into any other basis simply increases the number of nonzero coefficients. More precisely, we have the following proposition.

Theorem 4.2. *The BSB for the spike process is the standard basis if $\mathcal{D} = \mathrm{O}(n)$ or $\mathrm{SL}^{\pm}(n, \mathbb{R})$. If $\mathcal{D} = \mathrm{GL}(n, \mathbb{R})$, then it must be a scalar multiple of the identity matrix, i.e., aI_n where a is a nonzero constant.*

[1] Note that there is a slight abuse of the terminology; We call the kurtosis-maximizing basis in spite of maximizing unnormalized version (without the division by $\mu_2^2(Y_i)$) of the kurtosis.

Remark 4.3. Note that when we say the basis is a matrix such as aI_n, 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.

4.3. Statistical Dependence and Entropy of the Simple Spike Process

Before considering the LSDB of this process, let us 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 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.

4.4. The LSDB among $\mathrm{O}(n)$

For $\mathcal{D} = \mathrm{O}(n)$, we have the following theorem.

Theorem 4.4. *The LSDB among $\mathrm{O}(n)$ is the following:*

- *for $n \geq 5$, either the standard basis or the basis whose matrix representation is*

$$B_{\mathrm{O}(n)} = \frac{1}{n}\begin{bmatrix} n-2 & -2 & \cdots & -2 & -2 \\ -2 & n-2 & \ddots & & -2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -2 & & \ddots & n-2 & -2 \\ -2 & -2 & \cdots & -2 & n-2 \end{bmatrix}; \qquad (5)$$

- *for $n = 4$, the Walsh basis, i.e.,*

$$B_{\mathrm{O}(4)} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix};$$

- *for $n = 3$, $B_{\mathrm{O}(3)} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix}$; and*

- *for* $n = 2$, $B_{\mathrm{O}(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, *and this is the only case where the true independence is achieved.*

Remark 4.5. There is an important geometric interpretation of (5). This matrix can also be written as:

$$I_n - 2\frac{\mathbf{1}_n}{\sqrt{n}}\frac{\mathbf{1}_n^T}{\sqrt{n}}.$$

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 $\mathbf{1}_n/\sqrt{n}$.

4.5. The LSDB among $\mathrm{GL}(n, \mathbb{R})$

As discussed in [7], 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 the following 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):*

$$\begin{bmatrix} a & a & \cdots & \cdots & \cdots & \cdots & a \\ b_2 & c_2 & b_2 & \cdots & \cdots & \cdots & b_2 \\ b_3 & b_3 & c_3 & b_3 & \cdots & \cdots & b_3 \\ \vdots & \vdots & & \ddots & & & \vdots \\ \vdots & \vdots & & & \ddots & & \vdots \\ b_{n-1} & \cdots & \cdots & \cdots & b_{n-1} & c_{n-1} & b_{n-1} \\ b_n & \cdots & \cdots & \cdots & \cdots & b_n & c_n \end{bmatrix} \tag{6}$$

where a, b_k, c_k *are arbitrary real-valued constants satisfying* $a \neq 0$, $b_k \neq c_k$, $k = 2, \ldots, n$, *and*

$$\begin{bmatrix} \left(1 + \sum_{k=2}^n b_k d_k\right)/a & -d_2 & -d_3 & \cdots & -d_n \\ -b_2 d_2/a & d_2 & 0 & \cdots & 0 \\ -b_3 d_3/a & 0 & d_3 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -b_n d_n/a & 0 & \cdots & 0 & d_n \end{bmatrix} \tag{7}$$

where $d_k = 1/(c_k - b_k)$, $k = 2, \ldots, n$.

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 (5) and the LSDB pair (6), (7) provide us with further insight into the difference between sparsity and statistical independence. In the case of (5), this is the LSDB, yet 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, we have the following proposition.

Proposition 4.8.

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{\mathrm{O}(n)} \mid \boldsymbol{X}\right) = \begin{cases} \infty & \text{if } 0 \leq p < 1; \\ 3 & \text{if } p = 1. \end{cases}$$

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

$$\lim_{n\to\infty} \mathcal{C}_p\left(B_{\mathrm{O}(n)} \mid \boldsymbol{X}\right) \neq \mathcal{C}_p\left(\lim_{n\to\infty} B_{\mathrm{O}(n)} \mid \boldsymbol{X}\right).$$

As for the analysis LSDB (6), the ability to sparsify the 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 $b_k \neq c_k$, let us put $a = 1$, $b_k = 0$, $c_k = 1$, for $k = 2, \ldots, n$. Then we get the following specific LSDB pair:

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

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 this analysis LSDB, we have

$$\mathcal{C}_0 = E\left[\|\boldsymbol{Y}\|_0\right] = \frac{1}{n} \times 1 + \frac{n-1}{n} \times 2 = 2 - \frac{1}{n}.$$

Now, let us take $a = 1$, $b_k = 1$, $c_k = 2$ for $k = 2, \ldots, n$ in (6) and (7). Then we get

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} n & -1 & \cdots & -1 \\ -1 & & & \\ \vdots & & I_{n-1} & \\ -1 & & & \end{bmatrix}. \tag{8}$$

The spike process under this analysis basis is completely dense, i.e., $\mathcal{C}_0 = n$. Yet this is still the LSDB.

Finally, from Theorems 4.4 and 4.6, we can prove the following corollary:

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

5. THE GENERALIZED SPIKE PROCESS

In [14], Donoho et al. analyzed the following generalization of the simple spike process in terms of the KLB and the rate distortion function. 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 follows:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j\neq i}\delta(x_j)\right)g(x_i), \tag{9}$$

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. According to Martin Vetterli, this is one of his favorite stochastic processes. Interestingly enough, this generalized spike process shows rather different behavior (particularly in the statistical independence) from the simple spike process in Section 4.

5.1. 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.

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

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.

5.2. The Best Sparsifying Basis

As for the BSB, there is no difference between the generalized spike process and the simple spike process.

Theorem 5.2. *The BSB for the generalized spike process is the standard basis if* $\mathcal{D} = \mathrm{O}(n)$ *or* $\mathrm{SL}^{\pm}(n,\mathbb{R})$. *If* $\mathcal{D} = \mathrm{GL}(n,\mathbb{R})$, *then it must be a scalar multiple of the identity matrix, i.e.,* aI_n *where* a *is a nonzero constant.*

5.3. The LSDB/KMB among $\mathrm{O}(n)$

Here, we can see some difference from the simple spike process. In order to analyze the LSDB, we need some background work. First, let us compute the pdf of the process relative to a transformation $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$, $B \in \mathrm{SL}^{\pm}(n,\mathbb{R})$ rather than $\mathrm{O}(n)$. From (9), and the fact $|\det B| = 1$, we have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j\neq i}\delta(\boldsymbol{r}_j^T\boldsymbol{y})\right)g(\boldsymbol{r}_i^T\boldsymbol{y}), \tag{10}$$

where $\boldsymbol{r}_j^T$ is the jth row vector of B. The marginal pdf can be written as

$$f_{Y_i}(y) = \frac{1}{n}\sum_{j=1}^{n} g(y;|\Delta_{ij}|), \tag{11}$$

where Δ_{ij} is the (i,j)th cofactor of matrix B, and $g(y;\sigma) = g(y/\sigma)/\sigma$ represents the normal distribution $\mathcal{N}(0,\sigma^2)$. In other words, one can interpret this marginal pdf as a *mixture of Gaussians* with the standard deviations $|\Delta_{ij}|$, $j = 1,\ldots,n$. The proof of (11) is an interesting exercise of multivariate calculus and linear algebra to derive the marginal pdf [8].

Let us now consider a more specific case of $\mathcal{D} = \mathrm{O}(n)$. So far, we have been unable to prove the following conjecture.

Conjecture 5.3. *The LSDB among* $\mathrm{O}(n)$ *is the standard basis.*

However, a major simplification occurs if we consider the KMB instead of the LSDB, and we can prove the following:

Theorem 5.4. *The KMB among* $\mathrm{O}(n)$ *is the standard basis.*

The complete proof can be found in [8]. Here, we simply sketch the proof. Because $E[Y_i] = 0$ and $E[Y_i^2] = \frac{1}{n}\sum_{j=1}^{n}\Delta_{ij}^2$ for all i, the fourth order central moment of Y_i can be written as $\mu_4(Y_i) = \frac{3}{n}\sum_{j=1}^{n}\Delta_{ij}^4$, and consequently the cost function in (4) becomes

$$\mathcal{C}_\kappa(B\,|\,\boldsymbol{X}) = \frac{3}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\Delta_{ij}^4 - \frac{1}{n}\left(\sum_{j=1}^{n}\Delta_{ij}^2\right)^2\right). \tag{12}$$

Note that this is true for any $B \in \mathrm{SL}^{\pm}(n,\mathbb{R})$. If we restrict our basis search within $\mathrm{O}(n)$, another major simplification occurs because we have the following 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, we have

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

Inserting this into (12), we get the following 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 follows:

$$B_\kappa = \arg \max_{B \in \mathrm{O}(n)} \sum_{i,j} b_{ij}^4. \tag{13}$$

The proof of the fact that B_κ is the identity matrix or its permuted/sign-flipped versions is based on the observation that a matrix $P = (p_{ij}) = (b_{ij}^2)$ belongs to a set of *doubly stochastic matrices* $\mathcal{S}(n)$ and the optimization in (13) can be written as

$$\max_{P \in \mathcal{S}(n)} \sum_{i,j} p_{ij}^2.$$

5.4. The LSDB/KMB among $\mathrm{SL}^{\pm}(n, \mathbb{R})$

If we extend our search to this more general case, we have the following theorem.

Theorem 5.5. *The KMB among* $\mathrm{SL}^{\pm}(n, \mathbb{R})$ *does not exist.*

Proof. This simply relies on the fact that the set $\mathrm{SL}^{\pm}(n, \mathbb{R})$ is not compact. Therefore, the objective function $\mathcal{C}_\kappa(B \mid \boldsymbol{X})$ cannot have a minimum value in this set. One can also consider a simple example, $B = \mathrm{diag}(a, a^{-1}, 1, \cdots, 1)$, where a is any nonzero real scalar. Then, one can show that $\mathcal{C}_\kappa(B \mid \boldsymbol{X}) = -(a^4 + a^{-4} + n - 2)$, which tends to $-\infty$ as $a \uparrow \infty$. □

This theorem easily generalizes to the $\mathrm{GL}(n, \mathbb{R})$ case.

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

6. DISCUSSION

Because it is very easy to generate many realizations from these spike processes, and because we have concrete answers to which basis best sparsifies them, makes them least statistically dependent, make them maximally non-Gaussian, these realizations can be used to validate any ICA/SCA software package. We hope to report the actual numerical experiments using some popular ICA/SCA software packages at the conference.

7. REFERENCES

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

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

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

[4] 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*, vol. 265, pp. 359–366, 1998.

[5] D. L. Donoho, "Sparse components of images and optimal atomic decompositions," Tech. Rep., Dept. Statistics, Stanford University, 1998.

[6] N. Saito, B. M. Larson, and B. Bénichou, "Sparsity and statistical independence from a best-basis viewpoint," in *Wavelet Applications in Signal and Image Processing VIII*, A. Aldroubi, A. F. Laine, and M. A. Unser, Eds., 2000, vol. Proc. SPIE 4119, pp. 474–486, Invited paper.

[7] B. Bénichou and N. Saito, "Sparsity vs. statistical independence in adaptive signal representations: A case study of the spike process," *Ann. Inst. Statist. Math.*, 2001, submitted.

[8] N. Saito, B. Bénichou, B. M. Larson, O. Leborne, and S. Lucero, "Sparsity and statistical independence in adaptive signal representations through examples," Tech. Rep., Dept. Math., Univ. California, Davis, 2001, In preparation.

[9] N. Saito, "Image approximation and modeling via least statistically-dependent bases," *Pattern Recognition*, vol. 34, no. 9, pp. 159–178, 2001.

[10] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. Signal Process.*, vol. 44, no. 11, pp. 2768–2779, 1996.

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

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

[13] 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*. IEEE, 1996, pp. 1–4, Jun. 18–21, 1996, Paris, France.

[14] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2435–2476, 1998, Invited paper.