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 95616-8633
email: saito@math.ucdavis.edu
http://math.ucdavis.edu/∼saito

Acknowledgment:
Bruno Olshausen, Center for Neuroscience, UCD
NSF DMS-99-73032, NSF DMS-99-78321
ONR YIP N00014-00-1-046

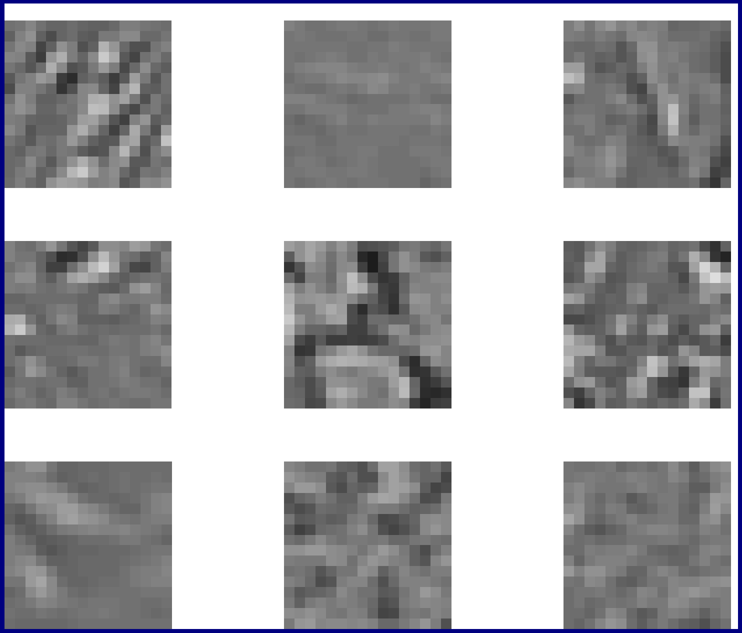
Outline

- Motivation
- Sparsity
- Statistical Independence
- The Simple Spike Process
- The Generalized Spike Process
- Summary

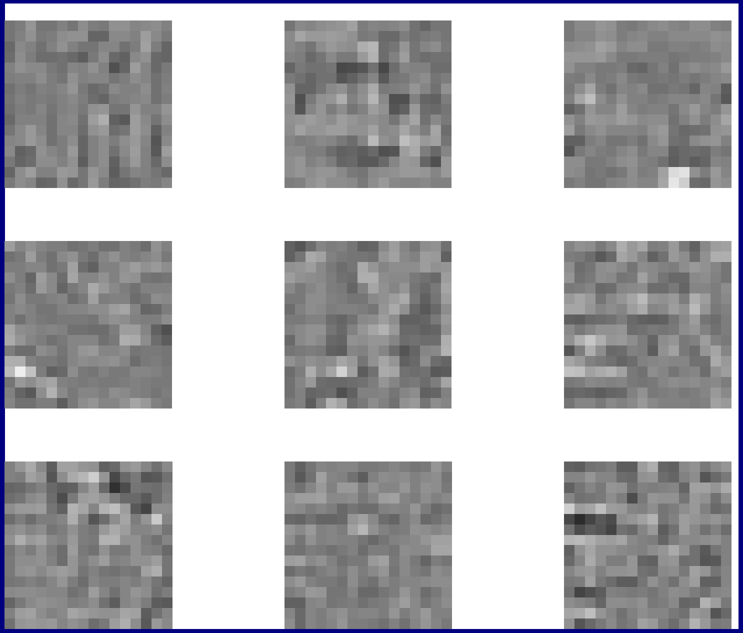
Motivation

- Series of experiments and observations of the basis functions learned from a set of natural scenes:
  - Olshausen & Field: Sparsity
  - Bell & Sejnowski, van Hateren & van der Schaaf: Statistical independence/ICA
- Both approaches produced basis functions that look like edge detectors (i.e., multiscale, oriented DOG functions)
- Why do they have to be the same?
- Natural images are way too complicated to analyze as realizations of a stochastic processes $\Longrightarrow$ Use much simpler stochastic processes to gain deeper understanding about this phenomenon.
- By-product of this research: Our theorems and examples can be used to validate any ICA algorithms/software because it is so simple.

Motivation . . .

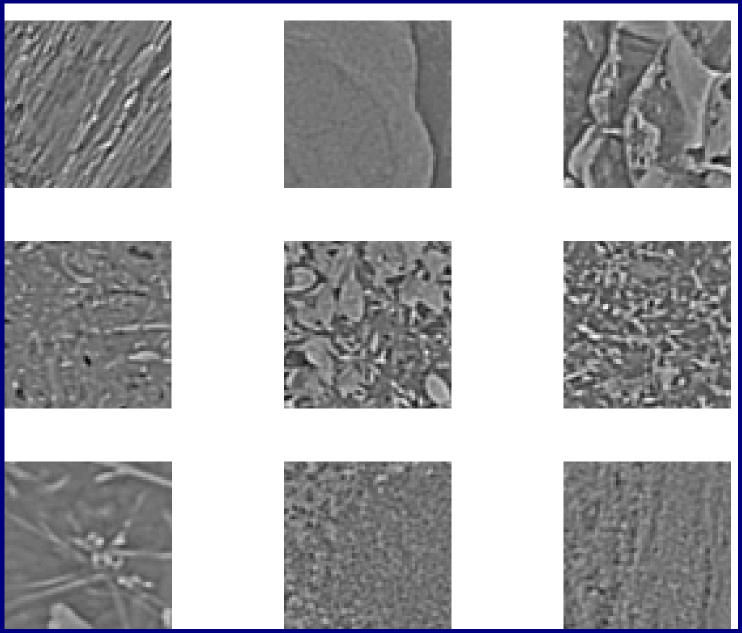


(a) Natural images 16x16

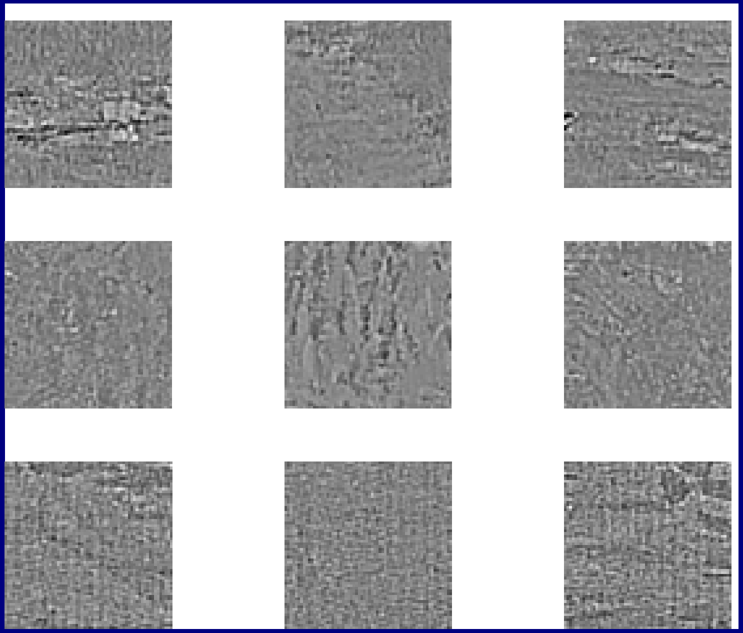


(b) Monet’s paintings 16x16

Motivation . . .

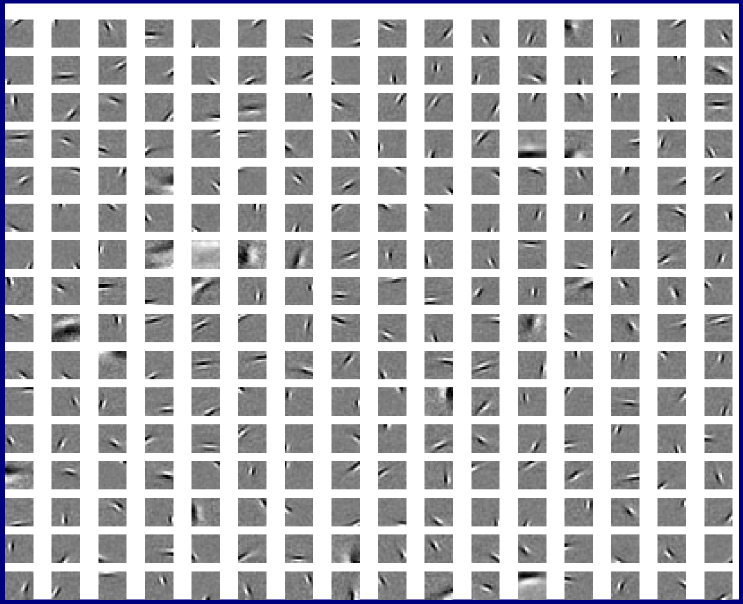


(c) Natural images 64x64

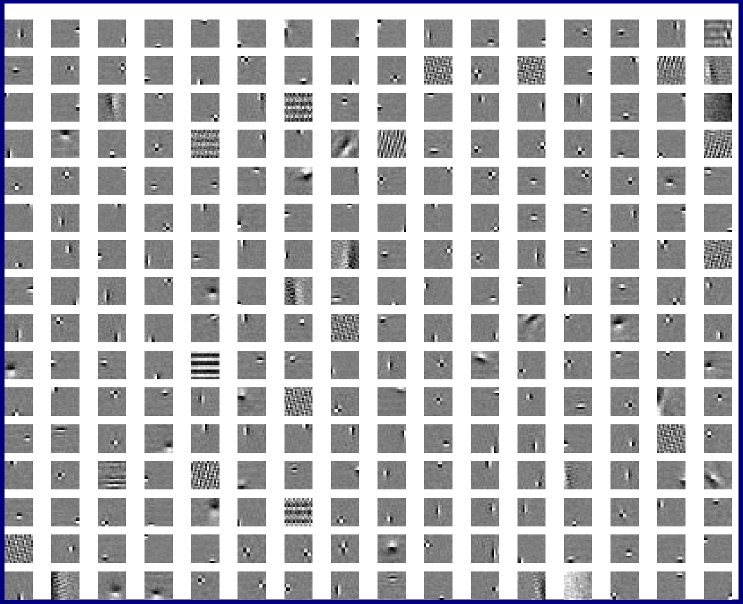


(d) Monet’s paintings 64x64

Motivation . . .



(e) Sparse basis for natural images



(f) Sparse basis for Monet's paintings

Motivation . . .

We want to understand:

1. Why both criteria produced basis functions resembling edge or line detectors?
2. What is the difference between sparsity and independence as a basis selection criterion?
3. What is the effect of the sizes of the image patches used?
4. What is the effect of orthonormality?
5. What is the effect of overcompleteness?
6. What is the effect of orientation selectivities of basis functions?

Methodology: Best-Basis Paradigm

- Let $\boldsymbol{X} \in \mathbb{R}^d$ be a random vector with pdf $f_{\boldsymbol{X}}$.
- Assume that the available data $\mathcal{T} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ were independently generated from this probability model.
- Let $B \in \mathrm{O}(d)$ or $\mathrm{SL}^{\pm}(d, \mathbb{R})$ (i.e., $\mathrm{GL}(d, \mathbb{R})$ with $\det(B) = \pm 1$).
- 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 $\mathcal{C}(B \,|\, \mathcal{T})$ be a numerical measure of deficiency or cost of the basis B given the training dataset $\mathcal{T}$ for the given problem.

Sparsity/SCA

is a key property as a good coordinate system for compression, which can be measured by ℓ^p-norm of the expansion coefficients, where $0 < p \leq 1$.

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

Then, we search the minimizer:

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

- We call B_p the best sparsifying basis (BSB) among $\mathcal{D}$, and this procedure the Sparse Component Analysis (SCA).
- Directly relevant to the compression:

$$\lim_{p \downarrow 0} \|\boldsymbol{Y}\|_p^p = \|\boldsymbol{Y}\|_0 = \#\{i \in [1, d] : Y_i \neq 0\}.$$

- Can compute a best basis for each realization.

Statistical Independence

is a key property as a good coordinate system for compression and modeling.

- Damage of one coordinate does not propagate to the others.
- Easy to model as a set of 1D processes.
- The “closeness” of the random variables $Y_1, \ldots, Y_d$ to the statistical independence can be measured by mutual information among the components of $\boldsymbol{Y}$:

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

- $I(\boldsymbol{Y}) \geq 0$. $I(\boldsymbol{Y}) = 0$ if and only if the components of $\boldsymbol{Y}$ are mutually independent.

Least Statistically-Dependent Basis/ICA

- If $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$ and $B \in \mathrm{SL}^{\pm}(d, \mathbb{R})$, then

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

since the differential entropy is invariant under such a transformation, i.e., $H(B^{-1}\boldsymbol{X}) = H(\boldsymbol{X}) + \log|\det(B^{-1})| = H(\boldsymbol{X})$.

- Define the cost:

$$\mathcal{C}_H(B \,|\, \boldsymbol{X}) = \sum_{i=1}^{d} H(Y_i) \approx -\frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{d} \log \widehat{f}_{Y_i}(y_{i,k}).$$

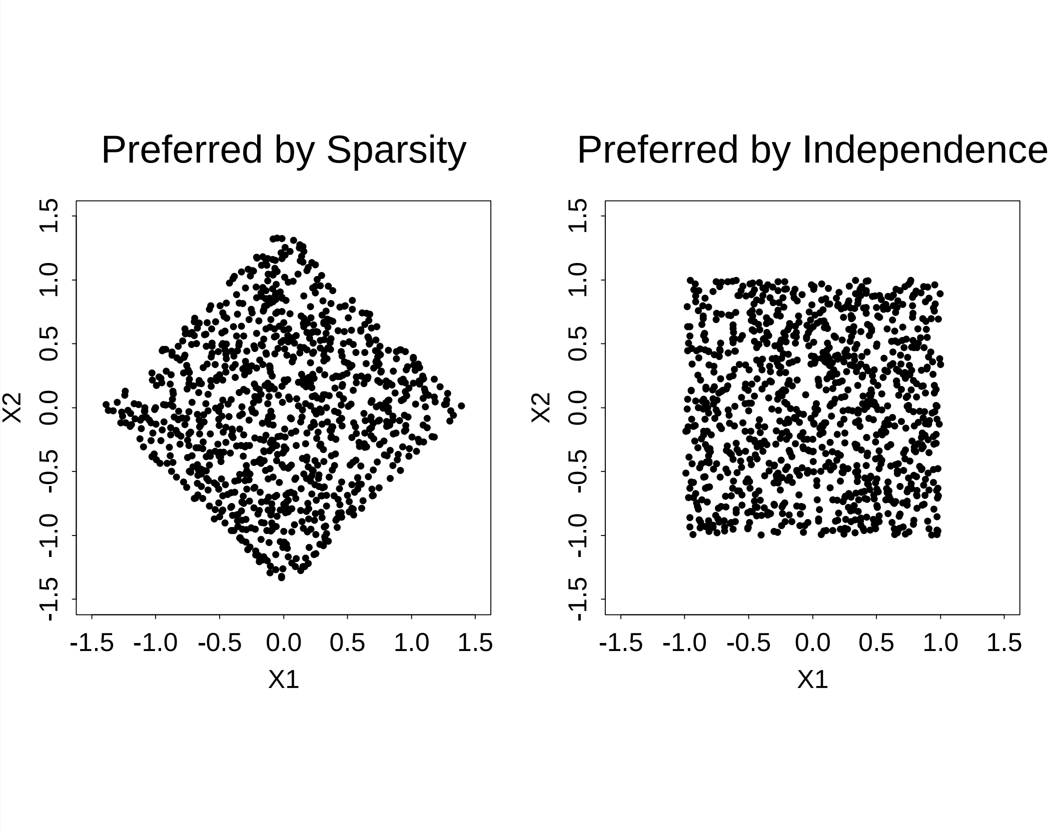
- Then we search the minimizer:

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

We call this basis the least statistically-dependent basis(LSDB) [Saito, 1998, 2001]. This is the same as a certain version of the ICA [Pham, 1996, Cardoso, 1999].

(Counter-)Example: 2D Uniform Distribution

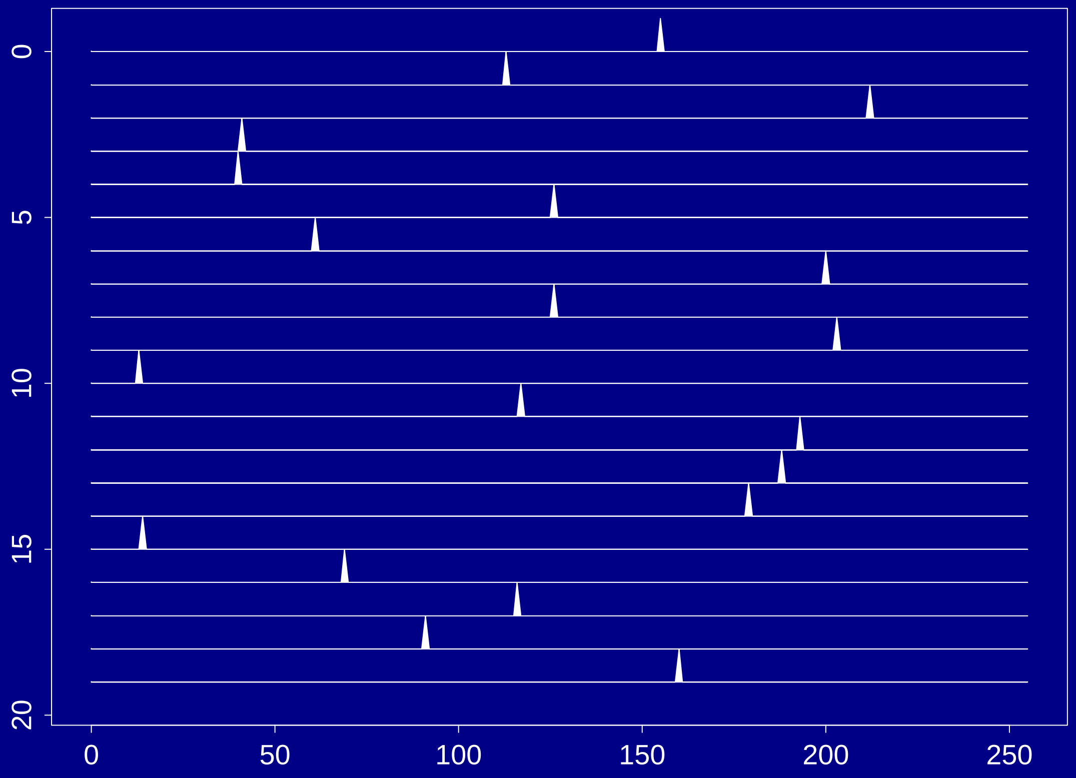
Consider all possible rotations around origin. Then, the rotation giving the sparsest distribution and the independent distribution can be quite different.



The Simple Spike Process

Consider a discrete stochastic process generating a single spike at a random grid location $1, \ldots, d$. This process generates the standard basis vectors $\boldsymbol{e}_j \in \mathbb{R}^d$ randomly.

Some realizations of ‘Spike’ process



The Simple Spike Process . . .

Theorem 1 (BB & NS). *The best sparsifying basis chosen from* $\mathrm{SL}^{\pm}(d, \mathbb{R})$ *for any* $p \in [0, 1]$ *is the standard basis (or its permuted/sign-flipped versions).*

Proposition 2 (BB & NS). *The Karhunen-Loève Basis is any rotation around the "DC" vector,* $\boldsymbol{b} = (1, 1, \ldots, 1)^T / \sqrt{d}$*.*

i.e., the KLB is useless $\Longleftarrow$ the simple spike process is non-Gaussian.

The Simple Spike Process . . .

Theorem 3 (BB & NS). *The LSDB among* $\mathrm{O}(d)$ *is the following:*

$d \geq 5$**:** *either the standard basis or the following basis:*

$$\frac{1}{d}\begin{bmatrix} d-2 & -2 & \cdots & -2 & -2 \\ -2 & d-2 & \ddots & & -2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -2 & & \ddots & d-2 & -2 \\ -2 & -2 & \cdots & -2 & d-2 \end{bmatrix} = I_d - 2\frac{\mathbf{1}_d}{\sqrt{d}}\frac{\mathbf{1}_d^T}{\sqrt{d}};$$

$d = 4$**:** *the Walsh basis,* $\frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix};$

$d = 3$: $\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{bmatrix}$;

$d = 2$: $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, *and this is the only case where the true independence is achieved.*

The Simple Spike Process . . .

Theorem 4 (BB & NS). *The LSDB chosen from* $\mathrm{GL}(d, \mathbb{R})$*,* $d > 2$ *is the following basis pair (analysis and synthesis):*

$$B^{-1}_{\mathrm{GL}(d)} = \begin{bmatrix} a & a & a & \cdots & a \\ b_2 & c_2 & b_2 & \cdots & b_2 \\ b_3 & b_3 & c_3 & & b_3 \\ \vdots & \vdots & & \ddots & \vdots \\ b_d & b_d & b_d & \cdots & c_d \end{bmatrix},$$

where a, b_k, c_k *are arbitrary constants satisfying* $a \neq 0$, $b_k \neq c_k$ *for* $k = 2, \ldots, d$*.*

$$B_{\mathrm{GL}(d)} = \begin{bmatrix} \left(1+\sum_{k=2}^{d} b_k d_k\right)/a & -d_2 & -d_3 & \cdots & -d_d \\ -b_2 d_2/a & d_2 & 0 & \cdots & 0 \\ -b_3 d_3/a & 0 & d_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -b_d d_d/a & 0 & 0 & \cdots & d_d \end{bmatrix},$$

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

Corollary 5 (BB & NS). *There is no invertible linear transformation that provides the truly statistically-independent coordinates for the spike process for* $d > 2$.

Remark: Permuted and sign-flipped versions of these matrices also possess the same quality in sparsity or statistical independence.

The Simple Spike Process . . .

Remark: The LSDB pair chosen from $\mathrm{GL}(d, \mathbb{R})$ shows another contrast between the sparsity and independence as follows.

- Choose $b_k = 0$, $c_k = 1$, for $k = 2, \ldots, d$ to get:

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

This analysis LSDB provides us with a very sparse representation for the spike process. For $\boldsymbol{Y} = B_\star^{-1}\boldsymbol{X}$,

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

- Choose $b_k = 1$, $c_k = 2$ for $k = 2, \ldots, d$ to get:

$$B_\star^{-1} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & & 2 \end{bmatrix}, \quad B_\star = \begin{bmatrix} d & -1 & -1 & \cdots & -1 \\ -1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -1 & 0 & 0 & & 1 \end{bmatrix}.$$

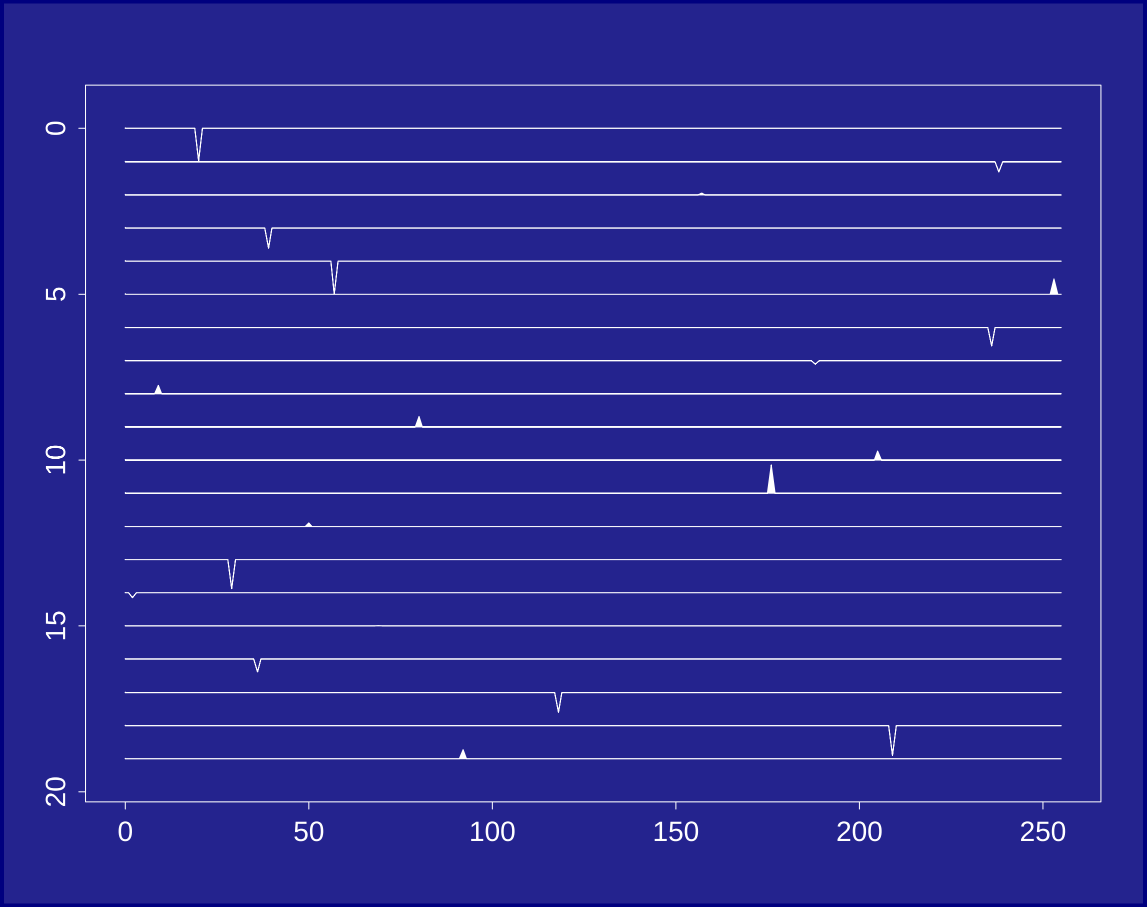
This is the worst (i.e., completely dense) basis in terms of sparsity, i.e.,

$$\mathcal{C}_p = \frac{1}{d} \times d + \frac{d-1}{d} \times \{(d-1) + 2^p\} = d + (2^p - 1)\left(1 - \frac{1}{d}\right),$$

where $0 \le p \le 1$, yet this is still the LSDB.

The 'Generalized' Spike Process

Similar to the simple spike process, but now the amplitude of each spike is sampled from the standard normal distribution.



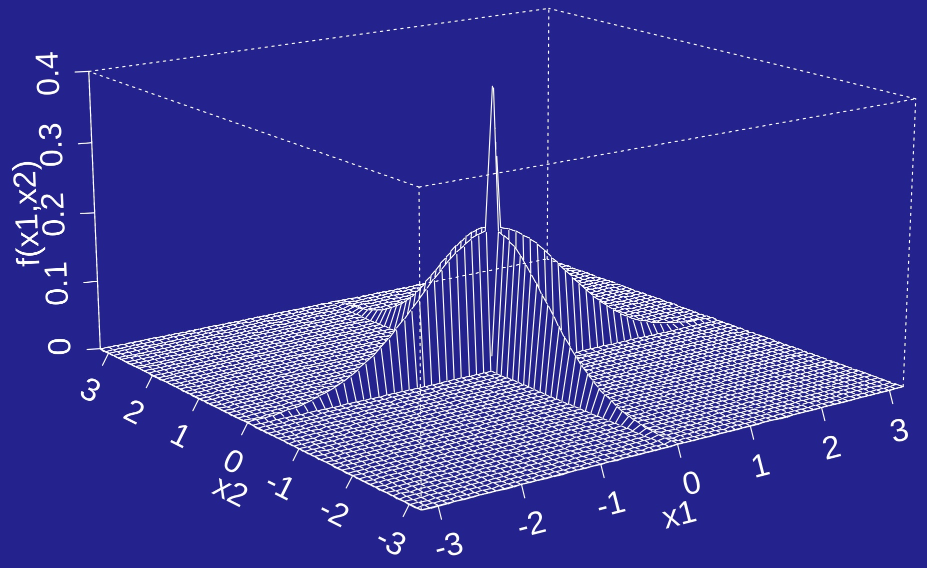
The ‘Generalized’ Spike Process . . .

The pdf of this process can be written as:

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

where $\delta(\cdot)$ is the Dirac delta function, and $g(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$.

The pdf of the ‘Generalized’ Spike Process ($d = 2$)



The Marginal Distributions under $\mathrm{SL}^{\pm}(d, \mathbb{R})$

For $\boldsymbol{Y} = B^{-1}\boldsymbol{X}$, $B \in \mathrm{SL}^{\pm}(d, \mathbb{R})$, the change of variable formula for a pdf generates:

$$f_{\mathbf{Y}}(\boldsymbol{y}) = \frac{1}{d}\sum_{i=1}^{d}\left(\prod_{j\neq i}\delta(\boldsymbol{b}_j^T\boldsymbol{y})\right) g(\boldsymbol{b}_i^T\boldsymbol{y}),$$

where $\boldsymbol{b}_j^T$ is the jth row vector of B. Now, we can compute its marginal pdf as follows:

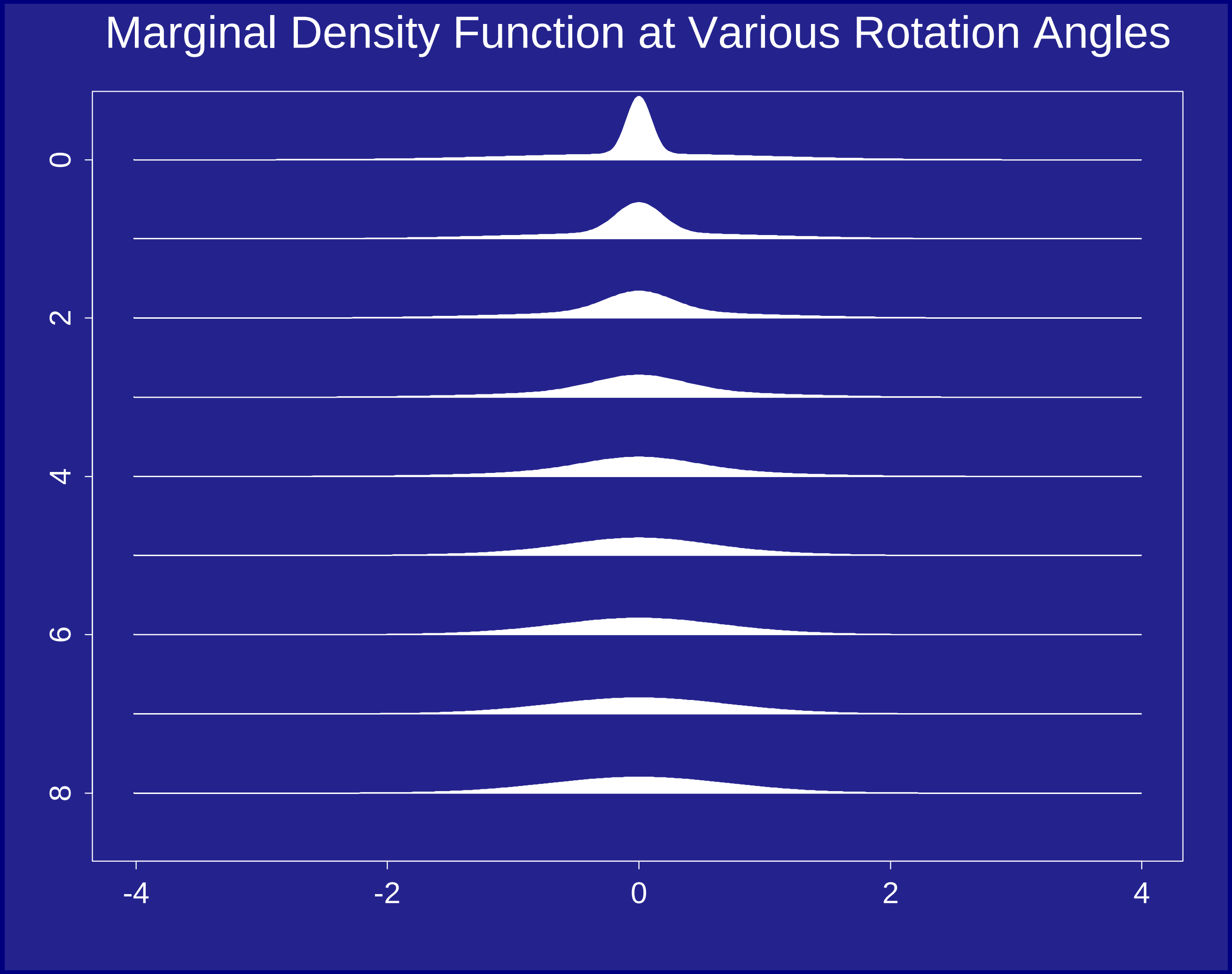
Lemma 6 (NS).

$$f_{Y_j}(y) = \frac{1}{d}\sum_{i=1}^{d} g(y; |\Delta_{ij}|),$$

where Δ_{ij} is the (i, j)th cofactor of matrix B, and $g(y; \sigma) = g(y/\sigma)/\sigma$.

The Marginal Distributions under $\mathrm{SL}^{\pm}(d, \mathbb{R})$ ...

Can interpret $f_{Y_j}(y)$ as a mixture of Gaussians.



The Moments under $\mathrm{SL}^{\pm}(d, \mathbb{R})$

Lemma 7 (NS).

$$E[|Y_j|^p] = \frac{\Gamma(p)}{d2^{p/2-1}\Gamma(p/2)} \sum_{i=1}^{d} |\Delta_{ij}|^p, \quad \textit{for all real } p > 0,\ j = 1, \ldots, d.$$

Remark: Keep Abramowitz & Stegun, Gradshteyn & Ryzhik on your desk!

KLB and BSB

Using all these lemmas, we can prove the following:

Proposition 8 (NS). *The Karhunen-Loève basis for the generalized spike process is any orthonormal basis in* $\mathbb{R}^d$.

Theorem 9 (NS). *The BSB with any* $p \in [0,1]$ *for the generalized spike process is the standard basis if* $\mathcal{D} = \mathrm{O}(d)$ *or* $\mathrm{SL}^{\pm}(d, \mathbb{R})$.

How about LSDB?

Unfortunately, the following is still the conjecture at this point:

Conjecture 10. *The LSDB among* $\mathrm{O}(d)$ *is the standard basis.*

Kurtosis-Maximizing Basis (KMB)

Instead of the LSDB, if we consider the KMB, then we can show much more.

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

where $\kappa(Y_i) = \mu_4(Y_i) - 3\mu_2^2(Y_i)$, and $\mu_k(Y_i)$ is the kth central moment of Y_i. (A slight abuse of notation here: strictly speaking, the kurtosis of Y_i is $\kappa(Y_i)/\mu_2^2(Y_i)$.)

- An approximation to ICA/LSDB
- Based on the approximation of the marginal differential entropy by higher order moments/cumulants using the Edgeworth expansion $H(Y_i) \approx -\kappa(Y_i)/48$. (see Comon (1994), Jones & Sibson (1987)).
- Also proposed independently by Buckheit & Donoho (1996) as a basis exposing maximal non-Gaussianity.

KMB . . .

Then, we have the following theorems:

Theorem 11 (NS). *The KMB among* $\mathrm{O}(d)$ *for the generalized spike process is the standard basis.*

Theorem 12 (NS). *The KMB among* $\mathrm{SL}^{\pm}(d, \mathbb{R})$ *for the generalized spike process does not exist.*

Conclusion

- For the simple spike process,
  - BSB $\neq$ LSDB if $\mathcal{D} = \mathrm{SL}^{\pm}(d, \mathbb{R})$ or $\mathrm{GL}(d, \mathbb{R})$, or if $\mathcal{D} = \mathrm{O}(d)$ with $d \leq 4$.
  - BSB $=$ LSDB if $\mathcal{D} = \mathrm{O}(d)$ with $d \geq 5$, but LSDB is not unique in this case (the Householder reflector).
- For the generalized spike process,
  - BSB $=$ KMB (an alternative to LSDB) if $\mathcal{D} = \mathrm{O}(d)$.
  - $\exists$ BSB whereas $\nexists$ KMB if $\mathcal{D} = \mathrm{SL}^{\pm}(d, \mathbb{R})$.
- The above results can be used to validate any ICA/SCA software.

Conclusion . . .

- Statistical independence and sparsity are completely different notions and criteria in general.
- However, under the best basis setting, both criteria prefer sharply concentrated (i.e., peaky) marginal distributions.
- A fundamental difference: the sensitivity on the location (mean) of the marginal pdf's. The entropy is location invariant whereas the ℓ^p norm is very sensitive to the mean. $\Longrightarrow$ non-uniqueness of the LSDBs for certain cases.
- The LSDB/ICA 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.

Conclusion …

- Numerical issues for general inputs:
  - If $\mathcal{D}$=wavelet packets, local Fourier dictionaries, then BSB/SCA is much simpler and stable than LSDB/ICA.
  - If $\mathcal{D} = \mathrm{O}(d), \mathrm{SL}^{\pm}(d, \mathbb{R})$, then BSB/SCA becomes a very tough nonconvex optimization problem even if $p = 1$. On the other hand, there are several ICA implementations are available.
- Multiple spike processes should be explored.

Papers

http://www.math.ucdavis.edu/∼saito/publications/