

MEASURING SPARSENESS OF NOISY SIGNALS

Juha Karvanen^{1,2} and Andrzej Cichocki²

¹Signal Processing Laboratory
Helsinki University of Technology
P.O. Box 3000, FIN-02015 HUT, Finland
juha.karvanen@hut.fi

²Laboratory for Advanced Brain Signal Processing
Brain Science Institute, Riken
2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan
cia@bsp.brain.riken.go.jp

ABSTRACT

In this paper sparseness measures are reviewed, extended and compared. Special attention is paid on measuring sparseness of noisy data. We review and extend several definitions and measures for sparseness, including the ℓ^0 , ℓ^p and ℓ^ϵ norms. A measure based on order statistics is also proposed. The concept of sparseness is extended to the case where a signal has a dominant value other than zero. The sparseness measures can be easily modified to correspond to this new definition. Eight different measures are compared in three examples. It turns out that different measures may give complete opposite results if the distribution does not have a unique mode at zero. As conclusion, we suggest that the kurtosis should be avoided as a sparseness measure and recommend tanh-functions for measuring noisy sparseness.

1. INTRODUCTION

In image analysis and vision research, sparseness has been demonstrated to be a powerful concept in finding meaningful representations of data [4, 11, 12, 3, 6, 7, 17, 15, 10]. The concept of sparseness or sparsity is also used in speech and music analysis [9, 2], in the statistical modeling of natural languages [16] and in various other applications. Despite the popularity of "the sparse ideology", sparseness is not unambiguously defined. The simplest definition of sparseness states that in a sparse matrix or vector most of elements are zero. This definition is not sufficient in all cases and it leaves open how the sparseness actually should be measured.

In many cases the problem may be formulated as follows: given matrix \mathbf{Y} find a sparse matrix decomposition as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{N}, \quad (1)$$

where \mathbf{N} represents noise and matrices \mathbf{Y} , \mathbf{A} and \mathbf{X} are problem-specific. Typically, the equation $\mathbf{Y} = \mathbf{A}\mathbf{X}$ has infinite many solutions and additional restrictions are needed. For instance, we may want minimize cost function

$$\mathcal{J}(S) = \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F + \lambda J(\mathbf{X}), \quad (2)$$

where λ is a constant, $\|\cdot\|_F$ is a matrix norm and $J(\mathbf{X}) = \sum_i J(\mathbf{x}_i)$ is a cost function for sparseness. In this paper we concentrate to the component-wise cost function $J(\mathbf{x}_i)$ that measures the sparseness of signal \mathbf{x}_i .

The relationship between Independent Component Analysis (ICA) and maximizing sparseness is studied by many authors [3, 7, 9, 17]. The two approaches seem to lead similar results in many cases even if some differences are pointed out [17]. It is not surprising that the similar results are obtained if independent components are found maximizing kurtosis and sparse components are found maximizing kurtosis. As also argued in [3], the choice of the sparseness measure is not a minor detail but may have far-reaching implications on the structure of a solution.

This paper is organized as follows. In Section 2 the ordinary definitions of sparseness for noiseless data are reviewed. In Section 3, measures for noisy data are presented. The ℓ^p and ℓ^ϵ norm based measures are reviewed and some extensions are presented. A measure based on order statistics is also proposed. In Section 4, the concept of sparseness is extended to the case where a signal has a dominant value other than zero. The sparseness measures can be easily modified to correspond to this new definition. In Section 5, examples revealing the differences between the measures are provided. Finally, In Section 6 we give some guidelines for the selection of the sparseness measure.

2. SPARSENESS IN NOISELESS CASE

The definition commonly given for sparseness is based on the ℓ^0 norm defined as the number of non-zero elements

$$\|\mathbf{x}\|_0 = \#\{j, x_j \neq 0\} \quad (3)$$

If $\|\mathbf{x}\|_0 = 0$ the vector x is completely sparse i.e. contains only zeros. If vectors having different length are compared, the ℓ^0 norm should be divided by the length of vector. It is characteristic for the ℓ^0 norm that the magnitude of non-zero elements is ignored. Thus, changing an arbitrary non-zero number to another arbitrary non-zero number does not have any effect on sparseness.

The ℓ^0 norm can be compared with one of basic criteria in ICA, the Shannon entropy defined as follows

$$H(x) = - \int f(x) \log(f(x)) dx, \quad (4)$$

where $f(x)$ is the density function of x . At the first look, it seems that the Shannon entropy and the ℓ^0 norm are very different concepts. However, when dealing with practical data, approximations are used for both the ℓ^0 norm and the entropy. The functional forms of the approximations are often very close to each other. In fact, the kurtosis has been used as a measure of sparseness and as an approximation of the entropy.

3. SPARSENESS AND NOISE

The ℓ^0 norm definition presented in the previous section is not very practical for measuring the sparseness of noisy data. Adding a very small measurement noise makes completely sparse data completely non-sparse. A common solution is to consider the ℓ^p norm instead the ℓ^0 norm

$$\|\mathbf{x}\|_p = \left(\sum_j |x_j|^p \right)^{1/p}, \text{ with } p \leq 1. \quad (5)$$

The connection between the ℓ^p norm and the ℓ^0 norm is $\lim_{p \rightarrow 0} \|\mathbf{x}\|_p^p = \|\mathbf{x}\|_0$. This leads to the use of expectation

$$\nu_p(x) = E\{|x|^p\} \quad (6)$$

as a measure of sparseness. This expectation can be interpreted also as an absolute moment of fractional order p . In order to imitate the ℓ^0 norm small values of p , e.g. 0.1 or 0.01, should be used. The value $p = 1$ is also used commonly. Sometimes the data is whitened prior to measuring sparseness. The connection between (6) and the normalized cumulant based kurtosis κ_4° is the following

$$\kappa_4^\circ(x) = E\{|x_{\text{white}}|^4\} - 3. \quad (7)$$

Instead of the standard normalized kurtosis, we can use the generalized normalized kurtosis (or Gray's variable norm) [5, 8], defined as

$$\kappa_{p,q}(x) = \frac{E\{|x|^p\}}{E^q\{|x|^{p/q}\}} - c_{pq}, \quad (8)$$

where c_{pq} is a positive constant, such that, for the Gaussian distribution $\kappa_{p,q} = 0$ and p, q are chosen suitably positive (typically, $q = 2$ and $p = 1, 3, 4, 6$). In the special case for $p = 4, q = 2$ and $c_{pq} = 3$, the generalized kurtosis reduces to the standard normalized kurtosis.

Another way to take the measurement noise into account is to use the ℓ^ϵ norm defined as follows

$$\|\mathbf{x}\|_{0,\epsilon} = \#\{j, |x_j| \geq \epsilon\}, \quad (9)$$

The parameter ϵ should depend on the noise variance but not on the variance of x . Determining the value of ϵ when the noise variance is not known, is an open problem. Another practical problem is that the ℓ^ϵ norm is non-differentiable and thus cannot be optimized with gradient methods. A solution is to approximate the ℓ^ϵ norm by tanh-functions

$$g(x) = \tanh(|ax|^b), \quad (10)$$

where a and b are positive constants. The main difference between (10) and the ℓ^p norm is that $\tanh(|ax|^b)$ saturates to 1 when $|x| \rightarrow \infty$. In order to imitate the ℓ^ϵ norm, the value of b must be greater than 1, e.g. $b = 2, b = 3$ or $b = 4$.

It is also possible to construct nonparametric sparseness measures. The highest-density intervals are the shortest intervals containing the probability mass θ . We consider a univariate random variable x with the cumulative density function (cdf) $F(x)$. The highest-density interval with the probability mass θ is defined as follows

$$U_\theta^0 = \min_{a,b} (b - a) \text{ on condition} \\ F(b) - F(a) \geq \theta \text{ and } a \leq 0 \leq b. \quad (11)$$

The highest-density interval (11) actually implicitly uses a concept similar to the ℓ^ϵ norm. Instead of defining the parameter ϵ beforehand, the highest-density interval tells what ϵ should be in order to have e.g. 90% of observations in the neighborhood of zero. The only difference to the ℓ^ϵ norm is that the neighborhood is not required to be symmetric around zero.

It is difficult to find *unbiased* nonparametric estimators for the highest-density intervals (11). However, it is relatively easy to construct *biased* estimators for the length of the highest-density intervals. The idea is to use the empirical cdf to estimate $F(x)$ in the equation (11). Calculating the empirical cdf corresponds to ordering the data in ascending order, i.e. using the order statistics. After that, finding the length of the shortest interval with probability mass θ is a linear operation. For the ordered data $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ the measure u_θ^0 may be defined as follows

$$u_\theta^0 = \min_{i,j} (x_{(i)} - x_{(j)}) \text{ on condition} \\ \frac{i - j}{n} \geq \theta \text{ and } x_{(j)} \leq 0 \leq x_{(i)}. \quad (12)$$

Depending on the application, the interesting values of θ might be e.g. 0.5, 0.75, 0.9 or 0.99. Using $\theta = 0.5$ means that 50% of observations are required to be concentrated on a small area, whereas using $\theta = 0.99$ means that 99% of observations are required to be concentrated.

In literature, the ℓ^p norm with $0 < p < 1$ is used in [3, 17, 6]. The $\ell^{1,\epsilon}$ norm is proposed in [14]. The measure $\log(1 + x^2)$ is used in [12, 6]. Tanh-function is used in [6]. Sparse priors (e.g. Laplacian) are utilized in [9, 10, 13, 7]. Note that in many cases the logarithm of the sparse prior density reduces to the ℓ^p norm.

4. SIGNALS WITH NON-ZERO MODE

In Sections 2 and 3, sparseness was defined stating that in sparse data most values are zero or in the neighborhood of zero. In many applications this definition is not always sufficient. The data may be concentrated around a non-zero value that can be called as the mode, the dominant value or the baseline value. An example of this kind of signal is presented in Figure 1. The signal is the number 4 from the dataset ABio7.mat [1] containing typical biological sources.

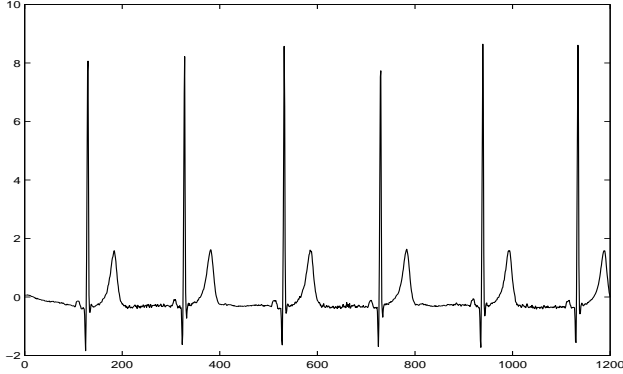


Fig. 1. A typical biological source taken from dataset ABio7.mat [1]. The signal can be considered sparse respect to the baseline value.

A natural solution is to subtract the mode first and then apply sparseness measure designed for zero-mode data. The accurate estimation of the mode is generally a non-trivial task. Nevertheless, in sparse data, the mode can be approximated by the median. This is based on the fact that more 50% of data is in the neighborhood of the mode; otherwise the data cannot be called as sparse. Consequently, the median also belongs to this neighborhood and can be used as an estimate of the mode. The sample mean, on the contrary, is a poor estimator of the mode, because it can be strongly affected by the tails of the data.

The highest-density intervals (11) may be modified replacing the condition $a \leq 0 \leq b$ by the condition $a \leq b$. This leads to the estimator

$$u_{\theta}^1 = \min_{i,j} (x_{(i)} - x_{(j)}) \text{ on condition } \frac{i-j}{n} \geq \theta \text{ and } x_{(j)} \leq x_{(i)}. \quad (13)$$

5. EXAMPLES WITH SYMMETRIC, ASYMMETRIC AND MULTIMODAL DISTRIBUTIONS

To test the measures of sparseness we consider three test cases:

A) Generalized Gaussian distribution (GGD)

$$f(x; \alpha, \beta) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} e^{-(|x|/\beta)^\alpha}, \quad (14)$$

where α is a shape parameter and β is a scaling parameter. The shape parameter is changed while the scaling parameter is fixed to give unit variance.

B) Sparse mixture model

$$x = s + n, \quad (15)$$

where

$$s = 0, \quad \text{with probability } 1 - P \quad (16)$$

$$s = \mu, \quad \text{with probability } P, \quad (17)$$

and n is zero mean Gaussian noise with variance σ^2 . When the probability P and the variance σ^2 are relative small, the majority of the observations from the model (15) is close to zero. When the variance σ^2 and the probability P decrease, the observations become more sparse. Conversely, if σ^2 is relatively small, changing the parameter μ should not have any effect on the sparseness. This follows from the definition of the ℓ^0 norm (3) where only the number of nonzero elements is taken into account, not the deviations from the zero. In the simulations the values $P = 0.01$ and $\sigma^2 = 0.01$ are used.

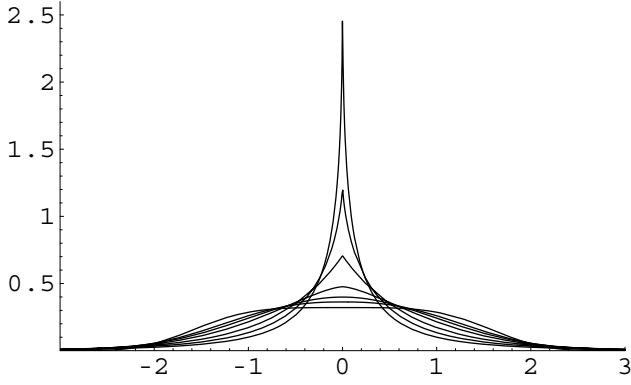
C) Lognormal distribution

$$x = \exp(s) \quad (18)$$

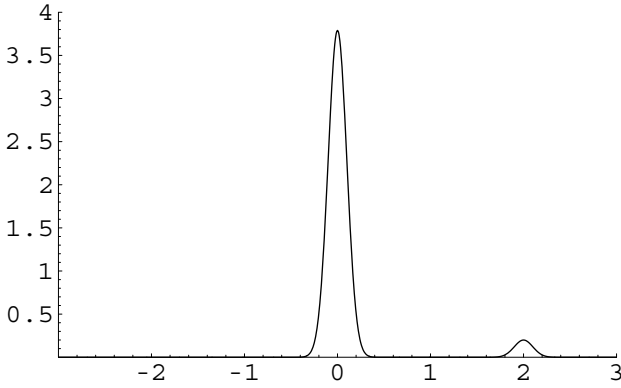
and s has Gaussian distribution with mean μ and variance σ^2 . The sparseness measures are evaluated for different values of σ^2 with $\mu = 0$. When σ^2 is relatively small, the mode of the Lognormal distribution is close to one. The smaller σ^2 is, the higher is the peak.

The density functions of these distributions are plotted with various parameter values in Figure 2. The distributions in the example A are symmetric, unimodal and have a zero mean. In the example B, the probability mass is concentrated around zero but there is another mode in μ . The distributions in the example C are skewed and have non-zero mean, mode and median.

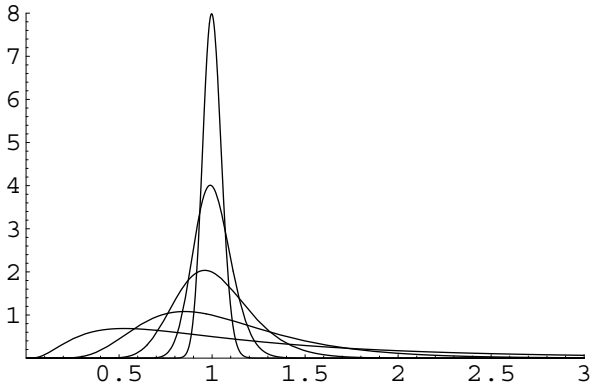
In all three examples, 500 samples were generated from the models with different parameter values. As the measures to be tested we chose the ℓ^p norm with $p = 0.1$ (to be used separately for unmodified data, whitened data and median subtracted data), the kurtosis κ_4° (7), absolute value $|x|$, a logarithm measure $\log(1 + x^2)$, a tanh-measure with the median adjustment $\tanh((x - \text{med}(x))^2)$ and an order statistics measure $u_{0.75}^1$ (13). The results are presented in



(a) GGD with $\alpha \in \{4, 2.5, 2, 1.5, 1, 0.7, 0.5\}$. The pdf with the highest peak is $\alpha = 0.5$



(b) Sparse mixture model with sub-mode $\mu = 2$



(c) Lognormal distribution with $\sigma \in \{0.05, 0.1, 0.2, 0.4, 0.8\}$. The pdf with the highest peak is $\sigma = 0.05$

Fig. 2. Illustration of the distributions used in the examples.

Table 1. The reported values are medians from 101 realizations. In the example A with the GGD, all the measures give analogous results. Except for the kurtosis, the values of the measures decrease when α decreases. Thus, all the measures lead to the same conclusion: the smaller α is, the more sparse the distribution is. In the example B, the consensus is lost. According to the measures $|x_{\text{white}}|^{0.1}$ and κ_4° the sparseness increases when the value of the sub-mode μ

increases. According to the measures $|x|$ and $\log(1+x^2)$ the sparseness decreases when the value of the sub-mode μ increases. According to the measures $|x|^{0.1}$, $|x - \text{med}(x)|^{0.1}$, and $u_{0.75}^1$ the change in the sub-mode does not affect sparseness. Measure $\tanh((x - \text{med}(x))^2)$ tells that the case $\mu = 1$ is the most sparse but there is no difference between the other cases. In the example C, the conclusions are again conflicting. According to the measures $|x|^{0.1}$ and $|x_{\text{white}}|^{0.1}$ the distribution is not sparse and the changing of the parameter has no effect. According to measure κ_4° , the most sparse distribution has $\sigma = 0.8$ According to all other measures the most sparse distribution has $\sigma = 0.05$.

Besides finding when sparseness is maximized it is interesting to compare the values for the sparse and non-sparse cases. For an ideal measure there is a clear difference between the values indicating sparse and non-sparse. The $|\cdot|^{0.1}$ measures seem to be especially poor in that sense. In the example A, the measures for Gaussian ($\alpha = 2$) are around 0.95 whereas the measures for a peaked distribution ($\alpha = 0.5$) are around 0.87. Such a small differences do not correspond to intuition on the sparseness.

The kurtosis has the advantage that it is commonly used to characterize distributions and thus the values of the kurtosis are easy to interpret. This interpretation, however, can be complete misleading as seen from the examples B and C. The measure $u_{0.75}^1$ has a natural interpretation in sense that it is directly the length of an interval. For the other measures there exists no commonly used interpretation. Nevertheless, if the goal is to optimize a sparseness measure by a gradient method, the interpretation of the actual value of the measure has only secondary interest.

None of the sparseness measures tested is fully satisfactory. $\tanh(|(x - \text{med}(x))^b|)$ with $b > 1$ seems to be a good practical choice. The order statistics based $u_{0.75}^1$ also gives good results but has an apparent weakness of being inappropriate for the gradient optimization.

6. CONCLUSION

In this paper sparseness measures are studied systematically. Despite the growing interest towards sparseness, the basic concepts are not clearly defined. Before measuring sparseness it is important to decide what is actually wanted to be measured. Especially, it should be decided whether the extended definition with a non-zero mode is used or not.

This paper does not deal with the problem of maximizing sparseness in practical applications. The measures are tested directly not comparing the resulting sparse representations. This allows us to compare the measures themselves without confusing details. The examples reveal that different measures may give complete opposite results if the distribution does not have a unique mode at zero. Consequently, the kurtosis should be avoided as a sparseness

A) GGD with the unit variance and the parameter α									
α	$ x ^{0.1}$	$ x_{\text{white}} ^{0.1}$	$ x - \text{med}(x) ^{0.1}$	κ_4°	$ x $	$\log(1 + x^2)$	$\tanh(x - \text{med}(x) ^2)$	$u_{0.75}^1$	
4	0.956	0.955	0.954	-0.798	0.843	0.571	0.525	2.395	
2.5	0.949	0.949	0.947	-0.362	0.816	0.549	0.494	2.308	
2	0.943	0.943	0.942	0.017	0.794	0.530	0.470	2.228	
1.5	0.936	0.936	0.935	0.743	0.766	0.506	0.441	2.137	
1	0.919	0.920	0.918	2.522	0.706	0.456	0.381	1.905	
0.7	0.896	0.900	0.896	5.796	0.631	0.399	0.320	1.615	
0.5	0.867	0.872	0.868	12.940	0.551	0.335	0.260	1.299	

B) Sparse mixture model with the sub-mode μ									
μ	$ x ^{0.1}$	$ x_{\text{white}} ^{0.1}$	$ x - \text{med}(x) ^{0.1}$	κ_4°	$ x $	$\log(1 + x^2)$	$\tanh(x - \text{med}(x) ^2)$	$u_{0.75}^1$	
1	0.752	0.916	0.751	23.5	0.0885	0.0165	0.0173	0.226	
2	0.753	0.876	0.752	60.3	0.0989	0.0259	0.0199	0.229	
3	0.754	0.849	0.753	76.8	0.109	0.0328	0.0199	0.228	
4	0.754	0.830	0.753	84.1	0.119	0.0381	0.0199	0.228	
5	0.754	0.817	0.754	87.8	0.129	0.0422	0.0198	0.228	
10	0.755	0.786	0.755	93.1	0.179	0.0559	0.0199	0.228	
100	0.758	0.799	0.758	95.0	1.08	0.1020	0.0200	0.231	
1000	0.762	0.799	0.762	95.0	10.1	0.1480	0.0200	0.228	

D) Lognormal distribution with the parameter σ									
σ	$ x ^{0.1}$	$ x_{\text{white}} ^{0.1}$	$ x - \text{med}(x) ^{0.1}$	κ_4°	$ x $	$\log(1 + x^2)$	$\tanh(x - \text{med}(x) ^2)$	$u_{0.75}^1$	
0.05	1.000	0.944	0.700	0.004	1.001	0.695	0.003	0.112	
0.1	1.000	0.944	0.749	0.123	1.005	0.698	0.010	0.223	
0.2	1.000	0.943	0.803	0.468	1.018	0.711	0.042	0.443	
0.4	1.001	0.939	0.860	2.357	1.088	0.772	0.155	0.860	
0.8	1.003	0.924	0.927	12.23	1.382	0.956	0.364	1.580	

Table 1. Comparison of the measures in the three test cases. The values indicating the highest sparseness are in bold. A column with no value in italics means that there is no significant change in the sparseness measure when the parameter of interest in changed. The presented numbers are medians from 101 realizations with sample size 500.

measure if it is not completely sure that the distribution is unimodal and symmetric. Based on our examples and the idea of noisy sparseness we recommend tanh-functions as a practical choice for measuring sparseness.

7. REFERENCES

- [1] A. Cichocki, S. Amari, K. Siwek et al. ICALAB Toolboxes, <http://www.bsp.brain.riken.go.jp/ICALAB>.
- [2] S. Abdallah and M. Plumbley. Sparse coding of music signals, submitted for publication, 2001.
- [3] D. L. Donoho. Sparse components analysis and optimal atomic decompositions. *Constructive Approximation*, 17:353–382, 2001.
- [4] D. J. Field. What is the goal of sensory coding. *Neural Computation*, 6(4):559–601, 1994.
- [5] W. C. Gray. *Variable Norm Deconvolution*. Ph.D. dissertation, Stanford, 1979.
- [6] A. Hyvärinen and P. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.
- [7] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, to appear.
- [8] R. H. Lambert. *Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures*. Ph.D. dissertation, University of Southern California, 1996.
- [9] M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

- [10] B. A. Olshausen. Principles of image representation in visual cortex. *The Visual Neurosciences*, to appear.
- [11] 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.
- [12] 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.
- [13] B. A. Olshausen and K. J. Millman. Learning sparse codes with a mixture-of-gaussians prior. In *Advances in Neural Information Processing Systems*, volume 12, pages 841–847, 2000.
- [14] C. P. Papageorgiou, F. Girosi, and T. Poggio. Sparse correlation kernel analysis and reconstruction. A.I. Memo 1635, Massachusetts Institute of Technology, 1998. C.B.C.L. Paper No. 162.
- [15] A. Pece. The problem of sparse image coding. *Journal of Mathematical Imaging and Vision*, 17:89–108, 2002.
- [16] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8), 2000.
- [17] N. Saito, B. M. Larson, and B. Benichou. Sparsity vs. statistical independence from a best-basis viewpoint. In *Wavelet Applications in Signal and Image Processing VIII*, volume Proc.SPIE 4119, pages 474–486, 2000.