A SIMPLE TEST TO CHECK THE OPTIMALITY OF SPARSE SIGNAL APPROXIMATIONS

R. Gribonval

IRISA-INRIA
Campus de Beaulieu
F-35042 Rennes Cedex, France

R. M. Figueras i Ventura [*], *P. Vandergheynst*

Signal Processing Institute
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

ABSTRACT

Approximating a signal or an image with a sparse linear expansion from an over-complete dictionary of atoms is an extremely useful tool to solve many signal processing problems. Finding the sparsest approximation of a signal from an arbitrary dictionary is an NP-hard problem. Despite of this, several algorithms have been proposed that provide sub-optimal solutions. However, it is generally difficult to know how close the computed solution is to being "optimal", and whether another algorithm could provide a better result. In this paper we provide a simple test to check whether the output of a sparse approximation algorithm is nearly optimal, in the sense that no significantly different linear expansion from the dictionary can provide both a smaller approximation error and a better sparsity. As a by-product of our theorems, we obtain results on the identifiability of sparse over-complete models in the presence of noise, for a fairly large class of sparse priors.

1. INTRODUCTION

Recovering a sparse approximation of a signal is of great interest in many applications, such as coding [1], source separation [2] or denoising [3]. Several algorithms exist (Matching Pursuits [4, 5], Basis Pursuit [6], FOCUSS [7] ...) that try to decompose a signal in a dictionary in a sparse way, but once the decomposition has been found, it is generally difficult to prove that the computed solution is the sparsest approximation we could obtain given a certain sparsity measure (which can be the number of terms or ℓ^0 "norm", the ℓ^1 norm, or any other metric that may lie "in between", which may be related to the bit-rate needed to represent the coefficients). In this paper, we provide a general tool for checking that the solution computed by some algorithm is nearly optimal, in the sense that no significantly different sparse linear expansion from the dictionary can provide both a smaller approximation error and a better sparsity.

In several aspects, our results extend previous contributions on the topic of recoverability of sparse overcomplete representations :

- previous results on recovery of sparse expansions in the noisy setting [8, 9, 10, 11] make assumptions on the *ideal* sparse approximation which do not seem easy to check in practice. We provide a test that can be implemented in practice since it only depends on the *observed* sparse approximation to determine its optimality. When the test is satisfied we provide a way to recover the ideal sparse approximation (best M-term approximation).
- the test is independent of the particular algorithm used to get the sparse approximation: there is no need to make a new proof or find new optimality conditions when one introduces a new algorithm.
- in the case where the error is measured with the mean square error (MSE), our test is close to being sharp. Moreover, the test is satisfied in some cases where the residual seems "too large" for the previous contributions [8, 9, 10, 11] to provide conclusive results.
- besides the MSE, we can deal with non-quadratic distortion measures, so one could imagine to insert visual criteria if one is dealing with images, or auditive criteria if one is dealing with sounds, or any other criteria more appropriate to the data than the MSE.
- not only do we deal with the ℓ^0 and ℓ^1 sparsity measures but also with all the ℓ^τ sparseness measures[1] $\|\cdot\|_\tau^\tau$, $0 \leq \tau \leq 1$, as well as a much larger class of "admissible" measures, as discussed in Section 2.

This paper is structured as follows: In Section 2 we state the sparse approximation problem and introduce the main concepts and results of the paper. In Section 3 we give the flavour of the proofs briefly discuss the connections between our results and other related work. Section 4 concludes the paper.

[*]This work, which was completed while the second author was visiting IRISA, is supported in part by the European Union's Human Potential Programme, under contract HPRN-CT-2002-00285 (HASSIP).

[1]Throughout this paper we use the notation $\|x\|_0^0$ to denote the ℓ^0 "norm" which counts the number of nonzero coefficients in x.

2. MAIN CONCEPTS AND RESULTS

In a finite or infinite dimensional vector space $\mathcal{H}$ (which may be a Hilbert space or more generally a Banach space) we consider $\mathbf{D}$ a dictionary of atoms $\{\mathbf{g}_k\}$. Using various sparse approximation algorithms (Matching Pursuits [4, 5], Basis Pursuit [6], FOCUSS [7] ...) one can decompose a signal $\mathbf{y} \in \mathcal{H}$ as

$$\mathbf{y} = \sum_k x_k \mathbf{g}_k + \mathbf{e} \tag{1}$$

where the sequence $x = (x_k)$ is "sparse" and the residual $\mathbf{e}$ is "small". Throughout this paper, Eq. (1) will be written $\mathbf{y} = \mathbf{D}x + \mathbf{e}$ where we use the same notation for the dictionary $\mathbf{D}$ and the corresponding synthesis operator which maps representation coefficients to signals. In other words, we will consider the representation coefficients x and the signal $\mathbf{y}$ as column vectors and the dictionary $\mathbf{D}$ as a matrix. We will use bold characters to denote signals (vectors in the space $\mathcal{H}$) and plain characters to denote coefficient sequences.

The goodness of the approximation (1) can be measured by some distortion measure $d(\mathbf{e})$ (such as a norm on $\mathcal{H}$) which only depends on the residual $\mathbf{e}$. The sparseness of a representation x can be measured by an ℓ^τ norm ($0 \leq \tau \leq 1$) or more generally by an f-norm

$$\|x\|_f := \sum_k f(|x_k|), \tag{2}$$

where $f : [0, \infty) \to [0, \infty)$ is non-decreasing, not identically zero, and $f(0) = 0$. The smaller $\|x\|_f$, the sparser the representation x. The most popular sparseness measures are the ℓ^τ "norms" $\|\cdot\|_\tau^\tau = \|\cdot\|_{f_\tau}$ where $f_\tau(t) := t^\tau$ for $0 \leq \tau \leq 1$ (with the convention $0^0 := 0$ and $t^0 = 1$, $t > 0$) but one can imagine many other more exotic sparseness measures, see [12]. Of particular interest will be the class $\mathcal{S}$ of *sub-additive* sparseness measures which, in addition to the above properties, satisfy

$$f(t + u) \leq f(t) + f(u) \text{ for all } t, u \geq 0,$$

and the class $\mathcal{M}$ of *admissible* sparseness measures where

$$t \mapsto f(t)/t \text{ is non-increasing.}$$

It is easy to check that $\mathcal{M} \subset \mathcal{S}$, (see [12]). One can define a partial order on $\mathcal{S}$ by letting $f \ll g$ iff there is some $h \in \mathcal{M}$ such that $f = h \circ g$ ($\mathcal{S}$ is stable by composition). With respect to this partial order, the ℓ^0 and ℓ^1 "norms" are respectively the smallest and the largest admissible sparseness measures, in that $f_0 \ll f \ll f_1$ for each $f \in \mathcal{M}$.

Since different sparse approximation algorithms may optimize different sparseness criteria (ℓ^1 norm for Basis Pursuits, various ℓ^τ norms for FOCUSS, ...), rely on various distortion measures, make a different compromise between sparseness and distortion, or even simply use a heuristic approach such as the greedy approach of Matching Pursuits, it is *a priori* hard to predict how solutions computed through different algorithms are related to one another. Our main theorems provide a simple test to check *a posteriori* if a computed decomposition $\mathbf{y} = \mathbf{D}x + \mathbf{e}$ is nearly optimal, in the sense that x is close to any representation x' which is both sparser and leads to a smaller distortion.

To state the theorems we need to introduce a few notations first. Let $\mathcal{H}$ be a Hilbert space equipped with the norm $\|\mathbf{y}\|_{\mathcal{H}}^2 = \langle \mathbf{y}, \mathbf{y} \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the inner product. For each integer K we denote

$$\sigma_{\min,K}^2(\mathbf{D}) := \inf_{\|\delta\|_0^0 \leq K} \frac{\|\mathbf{D}\delta\|_{\mathcal{H}}^2}{\|\delta\|_2^2} \leq 1 \tag{3}$$

and we consider the norm

$$|\mathbf{e}|_K := \sqrt{\sum_{k \in I_K(\mathbf{e})} |\langle \mathbf{e}, \mathbf{g}_k \rangle|^2} \tag{4}$$

where $I_K(\mathbf{e})$ indexes the K largest inner products $|\langle \mathbf{e}, \mathbf{g}_k \rangle|$. Notice that even though the notation does not make it explicit, $|\mathbf{e}|_K$ also depends on the dictionary $\mathbf{D}$. In infinite dimension, $|\cdot|_K$ is generally not equivalent to the native norm $\|\cdot\|_{\mathcal{H}}$. However, for any integer K we have $\sup_k |\langle \mathbf{e}, \mathbf{g}_k \rangle| = |\mathbf{e}|_1 \leq |\mathbf{e}|_K \leq \sqrt{K} \cdot |\mathbf{e}|_1 \leq \sqrt{K} \cdot \|\mathbf{e}\|_{\mathcal{H}}$, so the norms $|\cdot|_K$ for different K are equivalent. Based on these definitions we can state our first result.

Theorem 1 *Assume the atoms from the dictionary are normalized,* i.e. $\|\mathbf{g}_k\|_{\mathcal{H}} = 1$. *Let* $\mathbf{y} = \mathbf{D}x + \mathbf{e}$ *be a sparse approximation of a signal* $\mathbf{y}$*, which may have been computed with* any *algorithm, let* $M := \|x\|_0^0$ *and let* x' *be any other representation. If* $\|\mathbf{y} - \mathbf{D}x'\|_{\mathcal{H}} \leq \|\mathbf{y} - \mathbf{D}x\|_{\mathcal{H}}$ *and* $\|x'\|_0^0 \leq \|x\|_0^0$*, then*

$$\|x' - x\|_\infty \leq \frac{|\mathbf{e}|_1 + |\mathbf{e}|_{2M}}{\sigma_{\min,2M}^2(\mathbf{D})}. \tag{5}$$

A few additional definitions will be needed to state our second result, which is much stronger since it is valid for *any* admissible sparseness measure. We let $\mathbf{D}_I : \ell^2(I) \to \mathcal{H}$ denote the synthesis matrix associated to the subdictionary $\{\mathbf{g}_k, k \in I\}$ and $\mathbf{D}_I^+ = (\mathbf{D}_I^H \mathbf{D}_I)^{-1} \mathbf{D}_I^H$ be its Moore-Penrose pseudo-inverse. Then, much inspired by the Exact Recovery Coefficient introduced in [10] we consider

$$\lambda_M(\mathbf{D}) := 1 - \sqrt{M} \cdot \sup_{\text{card}(I) \leq M} \sup_{k \notin I} \|\mathbf{D}_I^+ \mathbf{g}_k\|_2. \tag{6}$$

Theorem 2 *Assume the atoms from the dictionary are normalized,* i.e. $\|\mathbf{g}_k\|_{\mathcal{H}} = 1$. *Let* $\mathbf{y} = \mathbf{D}x + \mathbf{e}$ *be a sparse approximation of a signal* $\mathbf{y}$*, which may have been computed with* any *algorithm, let* $M := \|x\|_0^0$ *and let* x' *be any*

other representation. If $\|\mathbf{y}-\mathbf{D}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{D}x\|_{\mathcal{H}}$ *and* $\|x'\|_f \leq \|x\|_f$ *for* some *admissible sparseness measure* f*, and if* $\lambda_M(\mathbf{D}) > 0$*, then*

$$\|x'-x\|_\infty \leq \frac{2}{\lambda_M^2(\mathbf{D})} \cdot \frac{|\mathbf{e}|_1 + |\mathbf{e}|_M}{\sigma^2_{\min,M}(\mathbf{D})}. \tag{7}$$

Note that, in Eq. (7), $2M$ has been replaced with M in the subscripts for $|\mathbf{e}|_\cdot$ and $\sigma^2_{\min,\cdot}(\mathbf{D})$ compared to Eq. (5).

Corollary 1 (Test of ℓ^0 optimality) *Under the hypotheses of Theorem 1, assume that*

$$|\mathbf{e}|_1 + |\mathbf{e}|_{2M} < \frac{\sigma^2_{\min,2M}(\mathbf{D})}{2} \cdot \min_{\{k,|x_k|\neq 0\}} |x_k|. \tag{8}$$

If x' *satisfies* $\|\mathbf{y}-\mathbf{D}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{D}x\|_{\mathcal{H}}$ *and* $\|x'\|_0^0 \leq \|x\|_0^0$*, then* x' *and* x *have the same "support":*

$$\textit{support}(x') := \{k, |x'_k| \neq 0\} = \{k, |x_k| \neq 0\} = \textit{support}(x).$$

In particular, if x and $\mathbf{e}$ satisfy the test (8) then the atoms involved in the best M-term approximation $x^\star$ to $\mathbf{y} = \mathbf{D}x+\mathbf{e}$ are *exactly* the atoms $\{\mathbf{g}_k, k \in \text{support}(x)\}$, and we recover the best M-term approximation $\mathbf{D}x^\star$ by projecting orthogonally $\mathbf{y}$ onto their span.

Corollary 2 (Test of strong optimality) *Under the hypotheses and notations of Theorem 2, assume that*

$$|\mathbf{e}|_1 + |\mathbf{e}|_M < \frac{\sigma^2_{\min,M}(\mathbf{D}) \cdot \lambda^2_M(\mathbf{D})}{4} \cdot \min_{\{k,|x_k|\neq 0\}} |x_k|. \tag{9}$$

If x' *satisfies* $\|\mathbf{y}-\mathbf{D}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{D}x'\|_{\mathcal{H}}$ *and* $\|x'\|_f \leq \|x\|_f$ *for* some *admissible sparseness measure, then* x' *and* x *have* essentially *the same support:*

$$\{k, |x'_k| > \theta\} = \textit{support}(x), \textit{ with } \theta := \frac{1}{2} \min_{\{k,|x_k|\neq 0\}} |x_k|.$$

For sufficiently small M, we have lower estimates of $\lambda_M(\mathbf{D})$ and $\sigma^2_{\min,K}(\mathbf{D})$ in *quasi-incoherent* dictionaries. The estimates are based on the Babel function [8, 10] $\mu_1(M,\mathbf{D})$ and its variant $\mu_2(M,\mathbf{D})$ which we define as

$$\mu_1(M) := \sup_{\text{card}(I)\leq M} \sup_{k\notin I} \sum_{i\in I} |\langle \mathbf{g}_k, \mathbf{g}_i\rangle| \tag{10}$$

$$\mu_2(M) := \sup_{\text{card}(I)\leq M} \sup_{k\notin I} \sqrt{\sum_{i\in I} |\langle \mathbf{g}_k, \mathbf{g}_i\rangle|^2}. \tag{11}$$

Proposition 1 *Let* $\mathbf{D}$ *be a normalized dictionary in a Hilbert space* $\mathcal{H}$*. If* $\mu_1(2M-1) < 1$ *then*

$$\sigma^2_{\min,2M} \geq 1 - \mu_1(2M-1) > 0 \tag{12}$$

If $\sqrt{M}\mu_2(M) + \mu_1(M-1) < 1$ *then* $\lambda_M > 0$ *and*

$$\sigma^2_{\min,M} \cdot \lambda^2_M \geq \frac{\left(1-\sqrt{M}\mu_2(M)-\mu_1(M-1)\right)^2}{1-\mu_1(M-1)} \tag{13}$$

When $\mathbf{D}$ is an orthonormal basis, $\mu_1(M) = \mu_2(M) = 0$ for all M, and the test of ℓ^0 optimality takes the simple form $|\mathbf{e}|_1 + |\mathbf{e}|_{2M} < \min_{\{k,|x_k|\neq 0\}} |x_k|/2$ (which turns out to be sharp [13]). The test of strong optimality becomes $|\mathbf{e}|_1 + |\mathbf{e}|_M < \min_{\{k,|x_k|\neq 0\}} |x_k|/4$. When $\mathbf{D}$ is a union of one or more incoherent orthonormal bases in $\mathbb{C}^N$, such as the Dirac, Fourier and Chirp bases, $\mu_1(M) = M/\sqrt{N}$ and $\mu_2(M) = \sqrt{M}/\sqrt{N}$, and as an example when $M \leq (1+\sqrt{N}/3)/2$ we have $\sigma^2_{\min,2M} \geq 2/3 \approx 0.66$ and $\sigma^2_{\min,M} \cdot \lambda^2_M \geq 4/9 \approx 0.44$.

3. FLAVOUR OF THE PROOF.

Even though the detailed proof of the results given in the previous section is too long to fit in this short paper, it would perhaps be frustrating for the reader to have the statements without at least some idea of the flavour of their proof. Note that some of the ideas are similar to the techniques developed in [11] even though these results were developed totally independently. Moreover, it seems that the test proposed in Corollaries 1-2 is reminiscent of some results of Tropp [10, Correlation Condition Lemma, Theorem 5.2], but with $\sup_k |\langle \mathbf{e}, \mathbf{g}_k\rangle| = |\mathbf{e}|_1$ replaced with $|\mathbf{e}|_1 + |\mathbf{e}|_K$ for $K \in \{M, 2M\}$.

Let $\mathbf{y} = \mathbf{D}x + \mathbf{e}$ be a sparse approximation of a signal $\mathbf{y}$, let $M := \|x\|_0^0$ and assume that x' satisfies $d(\mathbf{y}-\mathbf{D}x') \leq d(\mathbf{y}-\mathbf{D}x)$ and $\|x'\|_f \leq \|x\|_f$. Letting $\delta := x'-x$, we see that $\delta \in D_d(\mathbf{e}) \cap C_f(\mathrm{X}_M)$ with

$$D_d(\mathbf{e}) := \left\{\delta : d(\mathbf{e}-\mathbf{D}\delta) \leq d(\mathbf{e})\right\} \tag{14}$$

$$C_f(\mathrm{X}) := \bigcup_{z\in \mathrm{X}} \{\delta : \|z+\delta\|_f \leq \|z\|_f\} \tag{15}$$

and $\mathrm{X}_M := \{x, \|x\|_0^0 \leq M\}$. Thus, we have

$$\|x'-x\|_\infty \leq |\mathbf{e}|_{f,M} := \sup_{\delta\in D_d(\mathbf{e})\cap C_f(\mathrm{X}_M)} \|\delta\|_\infty. \tag{16}$$

Note that $|\mathbf{e}|_{f,M}$ also depends on the dictionary $\mathbf{D}$ and the distortion measure $d(\cdot)$ but we omit them in the notation.

When f is sub-additive and non-decreasing, we prove in [13] that

$$C_f(\mathrm{X}_M) = \left\{\delta : \sum_{k\in I_M(\delta)} f(|\delta_k|) \geq \frac{\|\delta\|_f}{2}\right\} \tag{17}$$

where $I_M(\delta)$ is the set of the M largest components of $|\delta_k|$. By [12, Lemma 7], for any admissible sparseness measure h, any sequence $z = (z_k)$ and any integer M, we have

$$\frac{\sum_{k\in I_M(z)} h(|z_k|)}{\|z\|_h} \leq \frac{\sum_{k\in I_M(z)} |z_k|}{\|z\|_1}. \tag{18}$$

Thus, for any sub-additive sparseness measures $f \ll g$ we have $C_f(\mathrm{X}_M) \subset C_g(\mathrm{X}_M)$ and $|\cdot|_{f,M} \leq |\cdot|_{g,M}$. In particular, for every admissible sparseness measure f, since $f_0 \ll f \ll f_1$ we have $|\cdot|_{f_0,M} \leq |\cdot|_{f,M} \leq |\cdot|_{f_1,M}$. Theorem 1 and Theorem 2 will follow respectively from the upper estimates

$$|\cdot|_{f_0,M} \leq \frac{|\cdot|_1 + |\cdot|_{2M}}{\sigma^2_{\min,2M}(\mathbf{D})} \quad (19)$$

$$|\cdot|_{f_1,M} \leq \frac{2}{\lambda_M(\mathbf{D})} \cdot \frac{|\cdot|_1 + |\cdot|_M}{\sigma^2_{\min,M}(\mathbf{D})} \quad (20)$$

$$(21)$$

when the distortion $d(\cdot)$ is the MSE (assuming that $\lambda_M(\mathbf{D}) > 0$ to get the second inequality). It is also possible to get bounds on $\|x' - x\|_2$ and $d(\mathbf{D}x' - \mathbf{D}x)$, as well as similar estimates for other distortion measures than the MSE, using the fact that for $0 < q < \infty$

$$\|x' - x\|_q \leq \sup_{\delta \in D_d(\mathbf{e}) \cap C_f(\mathrm{X}_M)} \|\delta\|_q,$$

and

$$d(\mathbf{D}x' - \mathbf{D}x) \leq \sup_{\delta \in D_d(\mathbf{e}) \cap C_f(\mathrm{X}_M)} d(\mathbf{D}\delta).$$

This is where we stop the sketch of the proof, because getting into the details of the above estimates would take twice the space available in this paper, so we refer the reader to our preprint [13] for more details and extensions.

4. DISCUSSION AND CONCLUSION

We provided tools to check if a given sparse approximation of an input signal –which may have been computed using any algorithm– is nearly optimal, in the sense that no other significantly different representation can at the same time be as sparse and provide as good an approximation. In particular we proposed a test to check if the atoms used in a sparse approximation are "the good ones" corresponding to the ideal sparse approximation for a fairly large class of *admissible* sparseness measures. The test is easy to implement, it does not depend on which algorithm was used to obtain the decomposition and does not rely on any prior knowledge on the ideal sparse approximation. In our preprint [13] we give extended results of the same flavour including the case of some non quadratic distortion measures, and we discuss some implications of our results in terms of Bayesian estimation and signal denoising with a fairly large class of sparse priors and random noise. We are currently trying to investigate how this work could also be extended to obtain results on the optimality of simultaneous sparse approximation of several signals, in order to apply the results to blind source separation. In addition, we are investigating the use of the optimality tests to build a stopping criterion for Matching Pursuit or to design provably good sparse approximation algorithms.

5. REFERENCES

[1] E. Le Pennec and S. Mallat, "Sparse geometrical image approximation with bandelets," *IEEE Transaction on Image Processing*, 2004.

[2] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.

[3] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 2, no. 6, pp. 670–684, 2002.

[4] S. Mallat and Z. Zhang., "Matching pursuit with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[5] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition," in *Proceedings of the 27^{th} Annual Asilomar Conf. on Signals, Systems and Computers*, Nov. 1993.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 30–61, 1998.

[7] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss : a reweighted norm minimization algorithm," *IEEE Trans. Signal Proc.*, vol. 45, no. 3, pp. 600–616, mar 1997.

[8] Joel A. Tropp, "Greed is good: Algorithmic results for sparse approximation," Tech. Rep., Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, 2003.

[9] R. Gribonval and P. Vandergheynst, "On the exponential convergence of Matching Pursuit in quasi-incoherent dictionaries," Tech. Rep., IRISA, 2004.

[10] Joel A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," Tech. Rep., Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, 2004.

[11] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," Working draft, Feb. 2004.

[12] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," Tech. Rep. R-2003-16, Dept of Math. Sciences, Aalborg University, Oct. 2003.

[13] R. Gribonval, R. M. Figueras i Ventura, and P. Vandergheynst, "A simple test to check the optimality of sparse signal approximations," Tech. Rep., IRISA, 2004, in preparation.