Consistency of l1 recovery from noisy deterministic measurements.

Charles Dossal, Rémi Tesson

Consistency of ℓ_1 recovery from noisy deterministic measurements.

Charles Dossal

IMB, Université Bordeaux 1, 351, Cours de la Libération, 33405 Talence Cedex, France

Remi Tesson

ENS Cachan - Bretagne Campus de Ker Lann Avenue Robert Schuman 35170 Bruz

Abstract

In this paper a new result of recovery of sparse vectors from deterministic and noisy measurements by ℓ_1 minimization is given. The sparse vector is randomly chosen and follows a *generic p-sparse model* introduced by Candès and al. [1]. The main theorem ensures consistency of ℓ_1 minimization with high probability. This first result is secondly extended to compressible vectors.

Key words: sparsity, ℓ_1 minimization, compressibility, consistency, deterministic matrix.

Introduction

Let A be a real matrix with n rows and m columns with $m > n$. Let x^0 be a sparse vector following a *generic p-sparse model* (see [1] and definition 2) and let y be a data vector $y = Ax^0 + b$, where b is a noise vector. The question we want to address is: can we give a bound on the sparsity of x^0 ensuring x^0 can be recovered or estimated from ℓ_1 minimization with high probability ?

Candès and Plan [1] answer partially to this question under assumptions on the coherence of the matrix A, with a random (Gaussian) noise and with hypotheses on the minimum absolute value of non-zero components of x^0. They proved that with high probability the support and the sign of x^0 can be recovered using ℓ_1 minimization if x^0 is sparse enough.

In this paper we show that, under the same assumption on the coherence and sparsity, with a bounded noise, without any assumption on the minimum absolute value of x^0, ℓ_1 minimization provides a vector $x^\star$ such that $\left\|x^0 - x^\star\right\|_2$ can be bounded. Moreover this new result can be extended to compressible vectors that are close to sparse vectors.

In [1] authors use an explicit formulation of the minimizer $x^\star$ which can be exhibited when the signal to noise ratio is bounded. In this article, the goal is to overtake this setting, that's why different tools must be developed.

In a first part, notations and definitions are given. In a second part the contributions of the paper are developed and connected to prior works. In a third part the proof of main results are given. A last part is devoted to discussion.

1. Notations and Definitions

For a given vector $x^0 \in \mathbb{R}^m$ which support is I, A_I denotes the submatrix of A which columns are columns of A indexed by I. For a given vector x, x_I denotes the subvector which components are components of x indexed by I. The vector $\text{sign}\,(x)$ is the vector whose component indexed by i is 1 if $x(i) > 0$, and -1 if $x(i) < 0$ and 0 if $x(i) = 0$. If the columns $(a_i)_{i\in I}$ are linearly indepedent, the matrix $A_I^t A_I$ is invertible and for any x^0 such that $Supp(x^0) = I$, one can define

$$d(x^0) = A_I(A_I^t A_I)^{-1}\text{sign}\left(x_I^0\right) \qquad \text{and } IC(x^0) = \max_{j\notin I}|a_j^t d(x^0)|.$$

This Identification Coefficient (IC) can be seen as signed ERC (Exact Recovery Coefficient introduced by Tropp [2, 3]). The condition $IC(x^0) < 1$, see Fuchs [4], is a sufficient condition for exact recovery by ℓ_1 minimization. Let us recall the definition of the subgradient of the ℓ_1 norm at a point x which support is I :

$$\partial \|x\|_1 = \{\xi \text{ such that } \|\xi\|_\infty \leqslant 1, \forall i \in I, \xi(i) = \text{sign}\,(x(i))\}.$$

The Bregman distance is defined as follows:

Definition 1. *Let x^1, x be two vectors of $\mathbb{R}^m$. For all $\xi \in \partial \left\|x^1\right\|_1$, the Bregman distance between x and x^1 is defined by*

$$D_\xi(x, x^1) = \|x\|_1 - \left\|x^1\right\|_1 - \langle \xi, x - x^1\rangle.$$

The *generic p-sparse model* is defined by Candès and Plan [1] as follows

Definition 2. *A vector x follows the generic p-sparse model if the support of x is randomly chosen with equiprobability from all supports which cardinal is p and if its sign is randomly chosen with equiprobability from all possible sign vectors, the support and the sign beeing independent.*

Definition 3. *For a matrix B the norm $\|B\|_{p\to q}$ is defined as follows :*

$$\|B\|_{p\to q} = \sup_{x\neq 0}\frac{\|Bx\|_q}{\|x\|_p} \quad \textit{and} \quad \|B\|_p = \|B\|_{p\to p}\,.$$

For a matrix B such that B^tB is invertible, $B^+ = (B^tB)^{-1}B^t$ denotes the Moore-Penrose pseudoinverse of B.
Recall that $\|B^+\|_2 = \sqrt{\|(B^tB)^{-1}\|_2}$ and that $\|B\|_{1\to 2} = \max_i \|b_i\|_2$ where $(b_i)_i$ are the columns of B.

Definition 4. *A $n \times m$ matrix A whose columns are normalized is said to satisfy the A_0-coherence criterion if*

$$\mu(A) = \max_{i\neq j}|\langle a_j, a_i\rangle| \leqslant \frac{A_0}{\ln m} \tag{1}$$

where A_0 is non negative real number.

Section 1.2 of [1] gives several examples of matrices satisfying such conditions, as concatenation of Dirac basis and basis of sinusoids.

2. Contributions and relations with prior works

Suppose $y = Ax^0 + b$ with $\|b\|_2 \leqslant \varepsilon$ and define the minimization problem

$$\min_{x\in\mathbb{R}^m} \|x\|_1 \quad \text{under the constraint} \quad \|Ax - y\|_2 \leqslant \varepsilon. \tag{2}$$

Let $x^\star$ be a minimizer of (2).

The following Theorem holds.

Theorem 1. *Let A be a $n\times m$ matrix satisfying the A_0-coherence criterion with A_0 small enough, suppose that x^0 follows the generic p-sparse model, with $p \leqslant \frac{c_0 m}{\|A\|_2^2 \ln m}$, for c_0 small enough depending on A_0. Suppose $\|b\|_2 \leqslant \varepsilon$ and $y = Ax^0 + b$, then any solution $x^\star$ of (2) satisfies*

$$\left\|x^0 - x^\star\right\|_2 \leqslant \left(2\sqrt{2} + \frac{8(2+\sqrt{2})\sqrt{p}}{3}\right)\varepsilon \tag{3}$$

with probability greater than $1 - 4m^{-2\ln 2}$ if m is large enough.

The $\sqrt{p}$ term seems to be suboptimal. It may be, but the answer is not clear. This point is discussed in section 4. The hard part of proofs of many results on ℓ_2 stability of minimization ℓ_1 minimization is to bound the error $x_0 - x^\star$ outside the support of x^0. If $IC(x^0) < 1$, when the noise is small enough the support of $x^\star$ is equal to the support of x^0. When the noise is higher, with a weak asumption on coherence and without any RIP asumptions it is difficult to find on optimal bound on this part of the error vector. The use of the Bregman distance (see the proof) provides a $\sqrt{p}$ term. It turns out that $\left\|x^\star - x^0\right\|_1$ can also be bounded using a similar proof :

$$\left\|x^\star - x^0\right\|_1 \leqslant \frac{\varepsilon}{3}(14\sqrt{2p} + 16p), \tag{4}$$

with the same probability. The proof of this inequality requires a simple modification of Proposition 1. This extension may be interesting since vectors x^0 and $x^\star$ belong the a space which dimension is m much larger than p. This first result can be extended to approximately sparse signals. Applying the theorem to $x^\star = x^1$ and $b_1 = Ar + b$ one obtains the following corollary :

Corollary 1. *Suppose $x^1 = x^s + r$ where $\|r\|_2 \leqslant C_1\varepsilon_1$ and x^s follows the generic p-sparse model, with $p \leqslant \frac{c_0 m}{\|A\|_2^2 \ln m}$, for c_0 small enough and if A satisfies the A_0-coherence criterion with A_0 small enough. Suppose $\|b\|_2 \leqslant \varepsilon_1$ then any solution $x^\star$ of (2) with $y = Ax^1 + b$ and $\varepsilon = \varepsilon_1(1 + C_1\|A\|_2)$ satisfies*

$$\|x^s - x^\star\|_2 \leqslant \left(2\sqrt{2} + \frac{8(2+\sqrt{2})\sqrt{p}}{3}.\right)\varepsilon \tag{5}$$

with probability greater than $1 - 4m^{-2\ln 2}$ if m is large enough.

To prove the corollary, one can apply the Theorem with $x^0 = x^s$ and $b_1 = Ar + b$.

This result sheds a new light on the understanding of the success of ℓ_1 minimization of the recovery of sparse and compressible vectors from noisy deterministic measurements. No Restricted Isometry Properties (RIP) [5, 6] can be used here. A condition on the coherence is a condition on all pairs of column vectors of the matrix A, where RIP conditions are conditions on sets of k column vectors of A, with large k which is much harder to get and to check. Moreover the condition on the coherence is weak since $\frac{1}{\log p}$ decreases quite slowly to 0.

The geometry of polytopes associated to A (see Donoho [7]) seems hard to use and the classical bound derived by the coherence or the ERC [2] are too weak. In [1] authors propose an approach with a random model on the vector x^0. This work lies on concentration lemmas of singular values of submatrices due to Tropp and on a explicit formulation of the solution of ℓ_1 minimization, see Fuchs [8]. This approach ensures the exact recovery of the support and the sign of the solution and needs conditions on the signal to noise ratio, decorrelation between the noise and the matrix and consequently can not be easily extended to compressible vector.
The present article focuses on the ℓ_2 reconstruction error. In this new setting no signal to noise ratio, no independence between matrix and noise are needed and the result can be easily extended to compressible vectors. However this new approach doesn't give any informations on the support of the solution. Unlike [1], the bound holds for any vector x^1 satisfying $\left\|y - Ax^1\right\|_2 \leqslant \varepsilon$ and $\left\|x^1\right\|_1 \leqslant \left\|x^0\right\|_1$.
Following Grasmair et al. [9], our approach uses Bregman distance to bound the part of the ℓ_1 norm of $x^\star$ that is not supported on the support I of x^0.

3. Proof of Theorem 1.

3.1. Proof of Theorem 1.

The proof lies on two properties, the first one bounds the ℓ_2 error $x^0 - x^\star$ under the hypothesis that $IC(x^0) < 1$

Proposition 1. *Let $x^0 \in \mathbb{R}^m$, whose support is I. If $IC(x^0) < 1$, then for any $x^\star$ solution of* (2)*, the following inequality holds*

$$\left\|x^\star - x^0\right\|_2 \leqslant 2\varepsilon \left(\sqrt{\left\|(A_I^t A_I)^{-1}\right\|_2} + \frac{\left\|d(x^0)\right\|_2}{1 - IC(x^0)} (\sqrt{\left\|(A_I^t A_I)^{-1}\right\|_2}\, \|A_{I^c}\|_{1\to 2} + 1) \right) \tag{6}$$

The second proposition ensures that if x^0 follows the $p-$sparse model for p small enough then with high probability $IC(x^0) < \frac{1}{4}$ and $\left\|(A_I^t A_I)^{-1}\right\|_2 \leqslant 2$:

Proposition 2. *Suppose x^0 follows the generic p-sparse model with $p \leqslant \frac{c_0 m}{\|A\|_2^2 \ln m}$ for c_0 small enough*

$$P\left(\left(\left\|(A_I^t A_I)^{-1}\right\|_2 \leqslant 2\right) \bigcap \left(IC(x^0) < t\right) \right) \geqslant 1 - 2m \exp\left(-\frac{t^2 \ln m}{8c_0^2}\right) - 2m^{-2\ln 2}. \tag{7}$$

Choosing c_0 small enough in Proposition 2 yields

$$P\left(\left(IC(x^0) \leqslant \frac{1}{4}\right) \bigcap \left(\left\|(A_I^t A_I)^{-1}\right\|_2 \leqslant 2\right) \right) \geqslant 1 - 4m^{-2\ln 2}. \tag{8}$$

Moreover

$$\left\|d(x^0)\right\|_2^2 = \langle \text{sign}\left(x_I^0\right), (A_I^t A_I)^{-1} \text{sign}\left(x_I^0\right)\rangle \leqslant \left\|(A_I^t A_I)^{-1}\right\|_2 p. \tag{9}$$

It can be noticed that for any support I, if the columns of A are normalized, $\|A_{I^c}\|_{1\to 2} = 1$.

Applying Proposition 1 to x^0, it follows that with probability greater than $1 - 4m^{-2\ln 2}$,

$$\left\|x^\star - x^0\right\|_2 \leqslant \varepsilon \left(2\sqrt{2} + \frac{8(2+\sqrt{2})\sqrt{p}}{3} \right), \tag{10}$$

which concludes the proof of Theorem 1.

3.2. Proof of proposition 1

The proof of Proposition 1 follows the one of Grasmair et al. in [9] using the fact that, under the assumption $IC(x^0) < 1$, $s = A^t d(x^0) \in \partial \left\|x^0\right\|_1$.

$$\begin{aligned}
\left\|x^\star - x^0\right\|_2 &\leqslant \left\|x_I^\star - x_I^0\right\|_2 + \|x_{I^c}^\star\|_2 \\
&\leqslant \left\|A_I^+ A_I (x_I^\star - x_I^0)\right\|_2 + \|x_{I^c}^\star\|_1 \\
&\leqslant \left\|A_I^+\right\|_2 \left\|A_I (x_I^\star - x_I^0)\right\|_2 + \|x_{I^c}^\star\|_1 \\
&\leqslant \left\|A_I^+\right\|_2 (2\varepsilon + \|A_{I^c} x_{I^c}^\star\|_2) + \|x_{I^c}^\star\|_1 \\
&\leqslant \left\|A_I^+\right\|_2 (2\varepsilon + \|A_{I^c}\|_{1\to 2} \|x_{I^c}^\star\|_1) + \|x_{I^c}^\star\|_1 \\
&\leqslant 2\varepsilon \left\|A_I^+\right\|_2 + (\left\|A_I^+\right\|_2 \|A_{I^c}\|_{1\to 2} + 1) \|x_{I^c}^\star\|_1
\end{aligned}$$

Using the Bregman distance, $\|x_{I^c}^\star\|_1$ can be bounded : indeed, from the definition of $s = A^t d(x^0)$ it follows that

$$\begin{aligned}
D_s(x^\star, x^0) &= \|x^\star\|_1 - \left\|x^0\right\|_1 - \langle s, x^\star - x^0\rangle \\
&= \|x^\star\|_1 - \langle s, x^\star\rangle \\
&= \sum_{i\in I} (\text{sign}\left(x_i^\star\right) - \text{sign}\left(x_i^0\right)) x_i^\star + \sum_{j\notin I} (\text{sign}\left(x_j^\star\right) - s_j) x_j^\star \\
&\geqslant \sum_{j\notin I} (\text{sign}\left(x_j^\star\right) - s_j) x_j^\star.
\end{aligned}$$

Since $\forall j \notin I$, $|s_j| \leqslant IC(x^0)$, one gets

$$D_s(x^\star, x^0) \geqslant \sum_{j\notin I} (1 - IC(x^0)) \text{sign}\left(x_j^\star\right) x_j^\star = (1 - IC(x^0)) \|x_{I^c}^\star\|_1,$$

that is,

$$\|x_{I^c}^\star\|_1 \leqslant \frac{D_s(x^\star, x^0)}{1 - IC(x^0)}.$$

Consequently,

$$\left\|x^\star - x^0\right\|_2 \leqslant 2\varepsilon \left\|A_I^+\right\|_2 + (\left\|A_I^+\right\|_2 \|A_{I^c}\|_{1\to 2} + 1) \frac{D_s(x^\star, x^0)}{1 - IC(x^0)}. \tag{11}$$

The Bregman distance can be bounded as follows. Since $IC(x_0) < 1$, $s = A^t d(x^0) \in \partial \left\|x^0\right\|_1$. Consequently

$$\begin{aligned} D_s(x^\star, x^0) &= \|x^\star\|_1 - \left\|x^0\right\|_1 - \langle A^t d(x^0), x^\star - x^0\rangle \\ &\leqslant -\langle A^t d(x^0), x^\star - x^0\rangle \\ &= -\langle d(x^0), A(x^\star - x^0)\rangle \\ &\leqslant \left\|d(x^0)\right\|_2 \left\|A(x^\star - x^0)\right\|_2 \\ &\leqslant \left\|d(x^0)\right\|_2 (\|Ax^\star - y\|_2 + \|b\|_2) \\ &\leqslant 2 \left\|d(x^0)\right\|_2 \varepsilon. \end{aligned}$$

The fact that $\left\|A_I^+\right\|_2 = \sqrt{\left\|(A_I^t A_I)^{-1}\right\|_2}$ concludes the proof of Proposition 1. □

3.3. Proof of Proposition 2

The proof relies on a proposition due to Tropp [10] (see also [1]):

Proposition 3. *Suppose that the set I is randomly and uniformely chosen among sets of cardinal p with $p \leqslant \frac{m}{4\|A\|_2^2}$. Then for $q = 2\ln m$,*

$$E(\left\|A_I^t A_I - Id\right\|_2^q)^{\frac{1}{q}} \leqslant 30\mu(A)\ln m + 13\sqrt{\frac{2p\left\|A\right\|_2^2 \ln m}{m}} \tag{12}$$

and

$$E(\max_{j\notin I} \left\|A_I^t a_j\right\|_2^q)^{\frac{1}{q}} \leqslant 4\mu(A)\sqrt{\ln m} + \sqrt{\frac{p\left\|A\right\|_2^2}{m}}. \tag{13}$$

From this proposition and Markov inequality, Candès and Plan [1] proved the following Corollary :

Corollary 2. *Suppose that A satisfies the A_0-coherence criterion and that x^0 follows the generic p-sparse model with $p \leqslant \frac{c_0 m}{\|A\|_2^2 \ln m}$ and $30A_0 + 13\sqrt{2c_0} \leqslant \frac{1}{4}$. Then $A_I^t A_I$ is invertible with probability greater than $1 - m^{-2\ln 2}$ and*

$$\left\|(A_I^t A_I)^{-1}\right\|_2 \leqslant 2. \tag{14}$$

From the Proposition 3 and the Hoeffding inequality the following Lemma can be deduced (see Candès and Plan [1]):

Lemma 1. *Suppose x follows the generic p-sparse model and let $(W_j)_{j\in J}$ be a collection of deterministic vectors. For $Z_0 = \max_{j\in J} |\langle W_j, \operatorname{sign}(x_I)\rangle|$ one has*

$$P(Z_0 \geqslant t) \leqslant 2|J|e^{-\frac{t^2}{2\kappa^2}}$$

for $\kappa \geqslant \max_{j\in J} \|W_j\|_2$.

Applying Lemma 1 and Proposition 3, the proof of Proposition 1 can be achieved :

For all $j \notin I$, define $W_j = (A_I^t A_I)^{-1} A_I^t a_j$, where I is the support of x^0. Applying Lemma 1, one gets

$$P(IC(x) \geqslant t) \leqslant 2|I^c| e^{-\frac{t^2}{2\kappa^2}} \leqslant 2m e^{-\frac{t^2}{2\kappa^2}}. \tag{15}$$

We need to estimate the maximum of $\|W_j\|_2$. Using Corollary 2 one gets

$$\|W_j\|_2 = \left\|(A_I^t A_I)^{-1} A_I^t a_j\right\|_2 \leqslant \left\|(A_I^t A_I)^{-1}\right\|_2 \left\|A_I^t a_j\right\|_2 \leqslant 2\left\|A_I^t a_j\right\|_2 \tag{16}$$

with a probability greater than $1 - m^{-2\ln 2}$. Proposition 3 and Markov inequality is then used to estimate $\left\|A_I^t a_j\right\|_2$.

$$\begin{aligned} P\left(\max_{j \notin I} \left\|A_I^t a_j\right\|_2 > \frac{c_0}{\sqrt{\ln m}}\right) &\leqslant c_0^q \frac{E(\max_{j \notin I} \left\|A_I^t a_j\right\|_2^q)}{(\ln m)^{\frac{q}{2}}} \\ &\leqslant \left(\frac{c_0}{\sqrt{\ln m}}\right)^q \left(4\mu(A)\sqrt{\ln m} + \sqrt{\frac{p\|A\|_2^2}{m}}\right)^q \\ &\leqslant \left(\frac{c_0}{\sqrt{\ln m}}\right)^q \left(\frac{4A_0 + \sqrt{c_0}}{\sqrt{\ln m}}\right)^q. \end{aligned}$$

If A_0 and c_0 are small enough, $\dfrac{4A_0 + \sqrt{c_0}}{\sqrt{\ln m}} \leqslant \dfrac{\sqrt{\ln m}}{2c_0}$ and using $q = 2\ln m$ it follows

$$P\left(\max_{j \notin I} \left\|A_I^t a_j\right\|_2 > \frac{c_0}{\sqrt{\ln m}}\right) \leqslant m^{-2\ln 2}. \tag{17}$$

From (16) and (17) it follows

$$P\left(\max_{j \notin I} \|W_j\|_2 > \frac{2c_0}{\sqrt{\ln m}}\right) \leqslant 2m^{-2\ln 2}. \tag{18}$$

Combined with inequality (15), this last inequality concludes the proof of the proposition. □

4. Discussions

The two constants $\frac{1}{4}$ and 2 in the proof of Theorem 1 are arbitrarily chosen. If other bounds are chosen, the optimal value of c_0 changes and the value of $\left(2\sqrt{2} + \frac{8(2+\sqrt{2})\sqrt{p}}{3}\right)$ in Theorem 1 may change. It turns out that these values are numerically pessimistic and that their optimization would not be useful. Two relevant questions may be asked about Theorem 1 :

Can we expect better bounds on the sparsity using the criterion $IC < 1$?

Constants may be optimized but it seems that the asymptotic of the sparsity may not be improved. In [11] and [12] authors proved that for gaussian measurements, beyond sparsity $\frac{m}{2\ln m}$, with high probability, $IC(x^0) > 1$. It would be surprising that better results could be achieved by deterministic measurements.

However the Grasmair approach (see Proposition 1) applies also to any vector η in the subgradient of the ℓ_1 norm at the point x^0 not only to

$s = A^t d(x^0)$. It may possible to improve the sparisty bound using another vector η.

The second question is about the $\sqrt{p}$ scaling in the bound (3). Is this scaling optimal or not ? Can we expect a better bound ?

RIP Theory gives similar bounds where the constant in (3) does not depend on the sparsity p but on RIP constants that can be uniformely bounded if the vector is sparse enough. Moreover Fuchs [8] proved that when the noise is small enough and if $IC(x^0) < 1$ the support of $x^\star$ is equal to the support of x^0, that is $\|x^\star_{I^c}\|_1 = 0$. Looking at the proof of Proposition 1, it appears that if $x^\star_{I^c} = 0$, in Theorem 1 the expression $\left(2\sqrt{2} + \frac{8(2+\sqrt{2})\sqrt{p}}{3}\right)$ can be set to $2\sqrt{2}$ which do not depend on p. Unfortunatly if no assumptions are made on ε, there is no guarantee that $x^\star_{I^c} = 0$. RIP theory solves the problem ensuring that all submatrices with a small number of columns have a good behaviour. Such hypothesis can not be done here and for some noise vectors b it may happen that $\|x^\star_{I^c}\|_1 \neq 0$. If $I^\star$ denotes the support of $x^\star$, the solution $x^\star$ satisfies the following implicit equation (see [8])

$$x^\star_{I^\star} = (A^t_{I^\star} A_{I^\star})^{-1} A^t_{I^\star} y - \lambda (A^t_{I^\star} A_{I^\star})^{-1} \mathrm{sign}\,(x_{I^\star}) \tag{19}$$

where λ depends on ε. This expression shows that the stability of the solution depends widely on the matrix $(A^t_{I^\star} A_{I^\star})^{-1}$ wich depends on x^0 and on the noise b. In practice in many situations $I^\star \not\subset I$ and there is no simple way to control $(A^t_{I^\star} A_{I^\star})^{-1}$.

The scaling $\sqrt{p}$ may be the price to pay of the lack of control on this matrix.

5. Conclusion

These results complete the previous one of Candès and Plan [1] and ensures that under the same hypothesis of sparsity ℓ_1 minimization is robust to noise and compressibility even if the exact support and sign can not be recovered. To control the part of the solution that is not supported on the support I of the objective vector x^0, no RIP can be used here but the Bregman distance provides an interesting bound.

References

[1] E. J. Candès, Y. Plan, Near-ideal model selection by ℓ_1 minimization, Annals of Statistics 37 (5A) (2009) 2145–2177.

[2] J. A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise, IEEE Trans. Info. Theory 52 (3) (2006) 1030–1051.

[3] J. A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Trans. Info. Theory 50 (10) (2004) 2231–2242.

[4] J.-J. Fuchs, On sparse representations in arbitrary redundant bases, IEEE Trans. Info. Theory 50 (6) (2004) 1341–1344.

[5] E. J. Candès, The restricted isometry property and its implications for compressed sensing, Compte Rendus de l'Academie des Sciences, Paris, Série I 346 (2008) 589–592.

[6] E. Candès, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Trans. Info. Theory 52 (12) (2006) 5406–5425.

[7] D. L. Donoho, High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension, Discrete & Computational Geometry 35 (4) (2006) 617–652.

[8] J. Fuchs, Recovery of exact sparse representations in the presence of bounded noise, IEEE Trans. Info. Theory 51 (10) (2005) 3601–3608.

[9] M. Grasmair, O. Scherzer, M. Haltmeier, Necessary and sufficient conditions for linear convergence of l1 regularization, Communications on Pure and Applied Mathematics 64 (2011) 161–182.

[10] J. A. Tropp, Norms of random submatrices and sparse approximation, C. R. Math. Acad. Sci. 346 (2008) 1271–1274.

[11] C. Dossal, M. Chabanol, G. Peyré, J. Fadili, Sharp support recovery from noisy random measurements by l1 minimization, Applied and Computational Harmonic Analysis 33 (1) (2012) 24–43. doi:10.1016/j.acha.2011.09.003.
URL `http://hal.archives-ouvertes.fr/hal-00553670/`

[12] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso), IEEE Trans. Info. Theory 55 (5) (2009) 2183–2202.