Séminaire BOURBAKI Juin 2016
68ème année, 2015-2016, n$^{\text{o}}$ 1117

# PARCIMONIE ET SYSTÈMES LINÉAIRES SOUS-DETERMINÉS [d'après Emmanuel Candès]

par **Francis BACH**

## INTRODUCTION

Les systèmes linéaires sous-déterminés, avec plus d'inconnues que d'équations, sont très courants dans de nombreux domaines d'applications des mathématiques. Dans ce texte, nous considérons le système suivant :

$$y = Ax, \tag{1}$$

dans la variable $x \in \mathbb{R}^n$, où $A$ est une matrice $m \times n$ et $y \in \mathbb{R}^m$. Nous ferons toujours l'hypothèse que (a) le système a au moins une solution, i.e., $y$ s'écrit $Ax^*$ pour un certain $x^* \in \mathbb{R}^n$ a priori inconnu et non unique, et que (b) le système est sous-déterminé, i.e., $m$ est inférieur à $n$. La difficulté majeure est alors notamment l'absence de solution unique.

Ces systèmes sont très fréquents dans de multiples domaines d'applications de plusieurs branches des mathématiques, avec des tailles $m$ et $n$ pouvant atteindre l'ordre du million. Par exemple :

(a) En traitement du signal, $x$ peut représenter un signal dont on n'observe qu'un sous-ensemble $y$ de sa transformée de Fourier discrète ($A$ est alors une matrice de cosinus et de sinus), une situation courante par exemple en imagerie médicale.

(b) En statistique, $A$ peut représenter l'expression d'un très grand nombre $n$ de gènes chez $m$ individus, et pour chacun de ces individus $i \in \{1, \dots, m\}$, on observe une réponse $y_i \in \mathbb{R}$ (codant par exemple l'apparition d'une maladie), que l'on cherche à prédire comme combinaison linéaire des expressions des gènes, le vecteur $x$ représentant alors les coefficients inconnus de cette combinaison.

(c) Le vecteur $x$ peut aussi être une matrice dont on n'observe que certains éléments et que l'on souhaite compléter, dans des applications comme le "filtrage collaboratif" (certains utilisateurs ont donné une note à certains produits, et on cherche à proposer une note pour tous les couples utilisateur-produit), ou le "criblage virtuel" (certaines molécules ont été expérimentalement testées sur certain agents pathogènes, et on cherche à trouver des molécules actives pour chaque agent).

Pour pallier l'absence de solutions uniques, certaines structures, dites de parcimonie, peuvent être imposées sur les solutions :

— **Parcimonie** : le vecteur $x$ est supposé *creux*, i.e., la plupart de ses composantes sont égales à 0. On appellera vecteur $k$-creux, un vecteur dont le nombre de composantes non nulles est inférieur ou égal à $k$. En pratique, $k$ sera très petit par rapport à $n$. Dans un cadre de traitement du signal où $x$ est un signal et $y$ un vecteur de mesures, l'échantillonnage sera dit *compressé* car $m$ sera très inférieur à la taille du signal $n$.
— **Rang-faible** : dans le cas de signaux matriciels, la matrice $x$ est supposée avoir un rang faible. En pratique le rang $r$ sera très inférieur aux nombres de lignes et de colonnes de la matrice.

Par exemple, dans l'application en génomique présentée ci-dessus, un petit nombre de gènes est supposé impliqué dans la prédiction de la réponse, alors que pour la complétion de matrices, un petit nombre de facteurs est supposé expliquer chaque entrée de la matrice.

Ces structures ne permettent pas par elles-mêmes d'obtenir des solutions uniques sans hypothèse supplémentaire. Par exemple, si on considère un vecteur 1-creux dans $\mathbb{R}^n$, dont seule la dernière composante est non nulle, observer les premières composantes, ce qui correspond à une matrice $A$ elle-même très creuse, ne permet pas de retrouver le vecteur initial. Il est donc nécessaire de faire des hypothèses supplémentaires sur la matrice $A$. En particulier, les lignes de la matrice $A$ (elles-mêmes des signaux de même taille que le signal $x$ à estimer) ne doivent pas être trop corrélées avec le signal $x$ de telle sorte que chaque mesure contienne de l'information sur les composantes non nulles de $x$.

Le but de ce texte est de présenter les travaux récents sur la résolution de tels systèmes avec hypothèse de parcimonie ou de rang faible. Cette simple hypothèse donne lieu à une théorie riche mettant en jeu des concepts de convexité et de matrices aléatoires et nous nous focaliserons principalement sur les contributions d'Emmanuel Candès et de ses co-auteurs, sur l'échantillonnage compressé et la complétion de matrices, qui sont deux instantiations marquantes de ces systèmes sous-déterminés.

Ce texte sera organisé comme suit : en section 1, nous présenterons la méthode de résolution proposée par optimisation convexe et les deux principaux résultats ; la preuve du premier résultat sera présentée en section 2 alors que la preuve du deuxième résultat sera esquissée en section 3. Enfin, en section 4, nous décrirons des aspects de transition de phases quand la matrice $A$ est gaussienne, faisant le lien avec d'autres branches des mathématiques. Pour finir, en section 5, nous présenterons certaines perspectives que ces travaux ouvrent, ainsi qu'une description non exhaustive d'autres problèmes liés à la parcimonie. Ce texte s'inspire de l'article d'Emmanuel Candès publié au Congrès International des Mathématiciens [Can1].

## 1. RÉSOLUTION PAR OPTIMISATION CONVEXE

Étant données les grandes tailles des problèmes à résoudre, il est crucial de considérer des méthodes de résolution dont le temps de calcul ne grandit pas trop vite avec les tailles $m$ et $n$. Ainsi, la méthode naïve consistant à résoudre les systèmes linéaires en essayant tous les $\binom{n}{k}$ supports de taille $k$ n'est pas utilisable en pratique. Nous considérerons des méthodes d'optimisation convexe, dont le temps de calcul est polynomial en $k$, $m$ et $n$. Dans la suite, on notera $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ la norme $\ell_p$ pour $p \in [1, \infty[$, et $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$ la norme $\ell_\infty$ de $x$.

### 1.1. Résolution par pénalisation d'une norme

La méthode naïve présentée ci-dessus correspond à minimiser par rapport à $x \in \mathbb{R}^n$ le nombre de non-zéros de $x$, que nous noterons $\|x\|_0$, sous la contrainte $Ax = y$. Nous allons remplacer la pénalité $x \mapsto \|x\|_0$ par une fonction convexe bien choisie.

Il est naturel de considérer l'*enveloppe convexe* de cette pénalité, i.e., la plus grande fonction convexe qui lui est inférieure. Si on se restreint aux vecteurs de norme $\ell_\infty$ bornée par un, alors cette enveloppe convexe est exactement la norme $\ell_1$ de $x$. Ceci donne le problème d'optimisation suivant [Tib, CDS] :

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \text{ tel que } Ax = y. \tag{2}$$

Dans le cas des matrices de rang faible, la pénalité non convexe naturelle est le rang de la matrice, et son enveloppe convexe sur l'ensemble des matrices dont les valeurs singulières sont bornées par un, est la *norme nucléaire*, égale à la somme des valeurs singulières [FHB]. Cette construction d'enveloppe convexe pour les méthodes parcimonieuses est plus générale et s'étend naturellement à des problèmes de parcimonie structurée, avec des structures de blocs ou d'arbre [Bac].

En termes computationnels, le programme convexe en Eq. (2) peut se résoudre en temps polynomial en $m$ et $n$. Une première solution est de le formuler comme un problème de *programmation linéaire* (objectif et contraintes linéaires), ce qui permet d'utiliser des algorithmes de simplexe [ST] ou de points intérieurs [NN]. Si ces reformulations permettent d'obtenir une solution avec haute précision, elles nécessitent la résolution de systèmes linéaires potentiellement grands. Pour les problèmes de grande taille, la formulation pénalisée $\min_{x \in \mathbb{R}^n} \|x\|_1 + \frac{1}{2\varepsilon}\|y - Ax\|^2$ est souvent considérée avec des méthodes du premier ordre utilisant la structure séparable du problème [BT, BJMO], la résolution se faisant alors à travers une succession de multiplications matrice-vecteur avec la matrice $A$.

### 1.2. Systèmes aléatoires et garanties d'unicité

Dans cette section seront présentés les deux résultats principaux pour l'hypothèse de parcimonie (vecteur $x$ creux), dans un ordre chronologique. Le premier a été obtenu par Emmanuel Candès et Terence Tao en 2005, et indique que si la matrice $A$ a des composantes gaussiennes indépendantes, alors la pénalisation par la norme $\ell_1$ permet

de retrouver $x^*$ à partir de $Ax^*$ pour tout $x^* \in \mathbb{R}^n$ $k$-creux, pour $k$ suffisamment petit (i.e., légèrement sous-linéaire en $m$).

THÉORÈME 1.1 ([CT1]). — *Soit $A$ une matrice dont les éléments sont des variables aléatoires gaussiennes indépendantes, de moyenne nulle et de variance $1/m$. Alors, si $k \leqslant c_1 \frac{m}{1+\log \frac{n}{m}}$, avec probabilité supérieure à $1-e^{-c_2 m}$, pour tout vecteur $x^* \in \mathbb{R}^n$ $k$-creux, l'Eq. (2) pour $y = Ax^*$ a comme solution unique $x^*$. Les constantes $c_1, c_2 \in \mathbb{R}_+^*$ sont universelles.*

Il est à noter que ce résultat n'est pas améliorable [CT2] en terme de nombre $k$ d'éléments non nuls que l'on peut retrouver : à un terme logarithmique près, il faut et il suffit d'un nombre de mesures $m$ proportionel au nombre $k$ de composantes à déterminer. Par ailleurs, la dimension ambiante des données $n$ n'intervient que logarithmiquement. La preuve de ce résultat est présentée en section 2 et met en jeu la notion d'*isométrie restreinte*, qui est un critère déterministe suffisant pour l'unicité de la solution pour tout $x^*$ $k$-creux. La matrice gaussienne est alors une matrice parmi d'autres satisfaisant ce critère avec forte probabilité.

Le résultat suivant, d'Emmanuel Candès et de Yaniv Plan, s'affranchit de ce passage par cette condition d'isométrie restreinte, et considère que chacune des $m$ lignes $a_i \in \mathbb{R}^n$ de $A$, $i = 1, \ldots, m$, est échantillonnée aléatoirement d'une loi isotrope (moyenne nulle et covariance égale à l'identité). Ce résultat met en valeur la notion de cohérence $\mu(A)$ définie telle que

$$\max_{i \in \{1,\ldots,n\}} |a_{ij}| \leqslant \mu(A),$$

avec forte probabilité ou presque sûrement. Cette cohérence sera large égale à $n$ si les lignes de $A$ sont prises uniformément à partir de la base canonique (multipliée par $n$ pour assurer l'isotropie), et par exemple égale à 1 pour des lignes correspondant à la base de Fourier discrète. Le théorème suivant montre que cette quantité contrôle le nombre de mesures nécessaires pour retrouver un signal donné.

THÉORÈME 1.2 ([CP]). — *Soit $x^*$ un vecteur fixe $k$-creux dans $\mathbb{R}^n$. Si les $m$ vecteurs $a_1, \ldots, a_m \in \mathbb{R}^n$ sont isotropes et indépendants, et si $y = Ax^*$, alors si $m \geqslant c_3$ $(1+\beta)\mu(A)k\log(n)$, avec probabilité au moins $1 - 5/n - e^{-\beta}$, $x^*$ est l'unique minimum de l'Eq. (2), où $c_3$ est une constante universelle.*

Comme pour le résultat précédent, le nombre de mesures n'est pas améliorable, car on peut trouver des exemples où l'agorithme échoue avec moins de mesures [CP]. Par rapport au résultat précédent, la garantie n'est que sur un signal fixe $k$-creux (i.e., pas sur tous les signaux $k$-creux), mais il permet un choix de matrices $A$ beaucoup plus flexible, qui inclut les bases de Fourier par exemple, mais aussi les matrices gaussiennes. Une esquisse de preuve sera présentée en section 3.

### 1.3. Résultats pour la complétion de matrices

Dans le cadre de la complétion de matrice, on cherche à estimer une matrice $X$ de taille $n_1 \times n_2$, à partir d'un certain nombre de ses éléments pris au hasard. Ceci correspond à une projection uni-dimensionnelle aléatoire (mais non gaussienne). Comme pour le cas des vecteurs creux dans la section précédente, une notion d'incohérence entre le signal (la matrice $X$ de rang faible) et les mesures (les formes linéaires accédant aux éléments de la matrice) est nécessaire et est introduite par [CR] ; $\mu(X)$ est maintenant le plus petit nombre tel que :

$$\max_{1\leqslant i\leqslant n_1} \frac{n_1}{r}\|\pi_{\text{colonnes}(X)}e_i\|_2^2 \leqslant \mu(X) \quad \text{et} \quad \max_{1\leqslant j\leqslant n_2} \frac{n_2}{r}\|\pi_{\text{lignes}(X)}e_j\|_2^2 \leqslant \mu(X),$$

où $r$ est le rang de $X$, $e_i$ le $i$-ème vecteur de la base canonique, et $\pi_{\text{colonnes}(X)}$ (resp. $\pi_{\text{lignes}(X)}$) la projection orthogonale sur l'espace engendré par les colonnes (resp. les lignes) de $X$. Ce paramètre mesure la corrélation entre les espaces engendrés par les lignes et les colonnes de $X$ avec les axes de coordonnées. Une matrice de forte cohérence sera plus difficile à estimer car elle aura un espace de colonnes (ou de lignes) trop aligné avec un des axes (dans le pire cas, tant qu'un élément particulier n'a pas été sélectionné, la matrice $X$ n'est pas estimable). Le théorème suivant permet une estimation précise de ce nombre de mesures.

Théorème 1.3 ([CR, CT3]). — *Soit $X^*$ une matrice fixe de taille $n_1 \times n_2$ et de rang $r$. Si on sélectionne aléatoirement et uniformément $m$ éléments de $X^*$ et que l'on minimise la norme nucléaire de $X$ avec contrainte d'avoir les mêmes valeurs pour les éléments observés, alors si $m \geqslant c_4\mu(X^*)r(n_1+n_2-r)\log^2(n_1+n_2)$, avec probabilité au moins $1-n^{-10}$, $X^*$ est l'unique minimum, où $c_4$ est une constante universelle.*

Le résultat est formellement proche de la situation de parcimonie (vecteurs creux), car la quantité $r(n_1+n_2-r)$ est le nombre de réels nécessaires pour représenter une matrice de rang $r$ (correspondant à $k$ pour un vecteur $k$-creux). Par ailleurs, au terme logarithmique près, ce résultat n'est pas améliorable [CT3].

### 1.4. Interprétation géométrique

Il existe une interprétation géométrique classique des résultats portant sur les normes induisant de la parcimonie (sur les éléments ou le spectre de la matrice). Les méthodes présentées minimisent une norme $\Omega$ sur un espace affine. Étant donné un élément $x^* \in \mathbb{R}^n$, on peut considérer le cône tangent à $x$, i.e.,

$$\mathcal{C} = \{h \in \mathbb{R}^n, \text{ tel qu'il existe } c > 0, \Omega(x+ch) \leqslant \Omega(x)\}. \tag{3}$$

Un tel cône est représenté en Figure 1 pour la norme $\ell_1$.

Le problème d'optimisation a alors une solution unique si et seulement si l'intersection du noyau de $A$ et du cône tangent est réduite à $\{0\}$. La raison intuitive pour laquelle minimiser la norme $\ell_1$ permet de retrouver des éléments creux vient du fait que le cône tangent est « étroit » pour ces vecteurs creux, et le noyau de $A$ ne va typiquement pas intersecter le cône tangent pour la plupart des matrices $A$ prises aléatoirement. Dans les

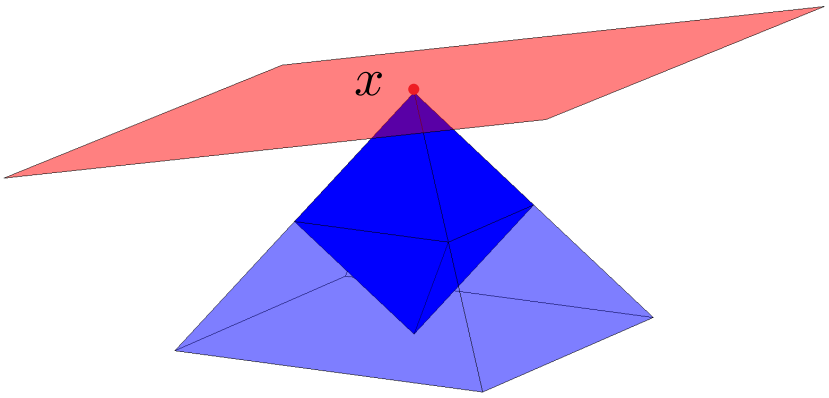


FIGURE 1. Boule $\ell_1$ (bleu foncé), cône tangent pour la norme $\ell_1$ (bleu clair) et sous-espace affine d'équation $Ax = b$ (rose). Quand l'espace affine est tangent à la boule, la solution est unique. De [Can1].

sections suivantes, nous allons montrer ceci précisément en utilisant des outils d'analyse convexe et de matrices aléatoires.

## 2. ANALYSE PAR ISOMÉTRIE RESTREINTE

On considère le problème d'optimisation en Eq. (2). On suppose donné un $x^* \in \mathbb{R}^n$ $k$-creux, pour $k \leqslant n$, on considère $y = Ax^*$ et on résout le problème en Eq. (2) ; on souhaite que $x^*$ soit la solution unique.

Si $A$ a des colonnes orthonormales, i.e., $A^\top A = I$, ce qui impose $m \geqslant n$, alors le problème a trivialement une solution unique $x^*$. Les premiers travaux dans ce domaine [DE, GN] ont considéré des situations s'écartant légèrement de cette situation idéale en faisant une hypothèse de *cohérence maximale*, i.e., la cohérence $\mu = \max_{i \neq j} \frac{|(A^\top A)_{ij}|}{(A^\top A)_{ii}^{1/2}(A^\top A)_{ii}^{1/2}}$, définie comme le cosinus maximal de l'angle entre deux des $n$ colonnes différentes de $A$, est supposée suffisamment faible. On peut alors montrer que si $k$ est plus petit que $\frac{1}{2}(1 + \frac{1}{\mu})$, alors $x^*$ est bien la solution unique (ce résultat sera un corollaire de résultats ci-dessous). Cependant, si l'on peut effectivement considérer des matrices $A$ pour lesquelles $m \leqslant n$, la cohérence maximale de la matrice $m \times n$ $A$ est toujours supérieure à $1/\sqrt{m}$ [SH], ce qui implique que la taille maximale des signaux que l'on peut retrouver est inférieure à une contante fois $m^{1/2}$ où $m$ est le nombre de mesures. Ceci reste très inférieur à la performance $m/\log(n/m)$ que nous atteindrons ci-dessous. On notera aussi la différence avec le théorème 1.2, qui utilise une notion de cohérence différente, et permettra d'obtenir des performances nettement supérieures.

### 2.1. Condition déterministe : isométrie restreinte

On définit $A_I$ la sous-matrice de $A$ de taille $m \times |I|$, obtenue en ne gardant que les colonnes indexées par $I$.

Définition 2.1. — *La matrice $A$ satisfait la propriété d'isométrie restreinte à l'ordre $k$ si il existe $\delta \in [0,1[$ tel que pour tout $I \subset \{1,\dots,n\}$ de cardinalité inférieure ou égale à $k$, les valeurs propres de $A_I^\top A_I$ sont comprises entre $(1-\delta)$ et $(1+\delta)$, i.e., pour tout $x \in \mathbb{R}^k$ :*

$$(1-\delta)\|x\|_2^2 \leqslant \|A_I x\|_2^2 \leqslant (1+\delta)\|x\|_2^2.$$

On notera $\delta_k$ la constante d'isométrie restreinte à l'ordre $k$ définie ci-dessus. On montre facilement qu'avec une cohérence de $\mu$, alors les valeurs propres de $A_I^\top A_I$ sont dans l'intervalle $[1-\mu(k-1), 1+\mu(k-1)]$, et donc la propriété d'isométrie restreinte à l'ordre $k$ est satisfaite pour $\mu < 1/(k-1)$, avec donc $\delta_k \leqslant \mu(k-1)$ (ceci permettra de retrouver les résultats existants sur la cohérence maximale, mais aussi de les généraliser). Une autre définition est qu'à une multiplication par une constante près, toutes les matrices $A_I^\top A_I$ sont inversibles.

Une première idée (computationellement non désirable) est de résoudre le problème suivant :

$$\min_{x \in \mathbb{R}^n} \|x\|_0 \text{ tel que } Ax = y. \tag{4}$$

Le lemme suivant donne une condition nécessaire et suffisante pour le bon fonctionnement de cette méthode.

Lemme 2.2. — *Les deux propriétés suivantes sont équivalentes, si $2k \leqslant n$ :*
*(i) $\alpha A$ satisfait la propriété restreinte à l'ordre $2k$, pour un $\alpha > 0$.*
*(ii) Pour tout $x^* \in \mathbb{R}^n$ de support de taille $k$, $x^*$ est l'unique minimiseur de $\|x\|_0$ tel que $Ax = Ax^*$.*

Preuve —

$(i) \Rightarrow (ii)$ : si $x$ est un minimiseur de support de taille inférieure à $k$, alors $x - x^*$ a un support de taille inférieure à $2k$ et $A(x-x^*) = 0$ ; comme la matrice $A$ restreinte au support de $x - x^*$ a un rang plein, ceci implique $x = x^*$.

$(ii) \Rightarrow (i)$ : tout vecteur $x \in \mathbb{R}^n$ $(2k)$-creux peut s'écrire $x = x_0 - x_1$ avec $x_0$ et $x_1$ $k$-creux. Si $Ax = 0$, alors $Ax_0 = Ax_1$ et donc $x_0 = x_1$ car ils sont tous les deux solutions uniques du même problème, d'où $x = 0$. Ceci implique que toutes les matrices $A_I^\top A_I$ sont inversibles pour tout $|I| \leqslant 2k$, et donc, après multiplication par un scalaire, $A$ satisfait la propriété d'isométrie restreinte.

Ainsi, la propriété d'isométrie restreinte est une condition nécessaire et suffisante pour que l'algorithme (cependant non implantable en pratique en temps polynomial [Nat]) donne lieu à une solution unique pour tout $x^* \in \mathbb{R}^n$. Il s'avère que pour que l'algorithme par optimisation convexe en Eq. (2) ait les mêmes garanties, il suffit d'une condition à peine plus stricte, i.e., on passe de $\delta_{2k} < 1$ à $\delta_{2k} < \sqrt{2} - 1$.

Théorème 2.3 ([Can2]). — *Si $2k \leqslant n$ et $A$ satisfait la condition d'isométrie restreinte avec ordre $2k$ et paramètre $\delta_{2k} = \delta < \sqrt{2} - 1$, alors, pour tout $x^* \in \mathbb{R}^n$ de support de taille $k$, $x^*$ est l'unique minimiseur de $\|x\|_1$ tel que $Ax = Ax^*$.*

Preuve —

(a) On montre d'abord que pour tous vecteurs $z, z'$ $k$-creux avec supports disjoints, $z^\top A^\top A z' \leqslant \delta \|z\|_2 \|z'\|_2$, ce qui se montre simplement par l'identité du parallélogramme en supposant les vecteurs normés : $z^\top A^\top A z' = \frac{1}{4}\|Az + Az'\|_2^2 - \frac{1}{4}\|Az - Az'\|_2^2 \leqslant \frac{1}{4}(1+\delta)\|z+z'\|_2^2 - \frac{1}{4}(1-\delta)\|z-z'\|_2^2 = \frac{1}{2}(1+\delta-1+\delta) + \delta z^\top z' = \delta$, car les supports de $z$ et $z'$ sont disjoints et donc $z^\top z' = 0$.

(b) On montre ensuite une propriété sur le noyau de $A$ : pour tout vecteur $z \in \mathbb{R}^n$ tel que $Az = 0$ et tout $I$ de cardinal inférieur à $k$, alors $\|z_I\|_1 \leqslant \rho \|z_{I^c}\|_1$, avec $\rho = \sqrt{2}\frac{\delta}{1-\delta}$. Il suffit de considérer $I = I_0$ composé des plus grands éléments de $z$ en valeur absolue, avec $I_1$ les $k$ plus grands suivants, $I_2$ les $k$ suivants, etc. On a alors, $\|z_{I_j}\|_2 \leqslant k^{1/2}\|z_{I_j}\|_\infty \leqslant k^{-1/2}\|z_{I_{j-1}}\|_1$ pour tout $j > 1$, car tous les éléments $z_{I_j}$ sont en valeur absolue inférieurs à tous les éléments $z_{I_{j-1}}$. Ceci implique $\sum_{j\geqslant 2}\|z_{I_j}\|_2 \leqslant k^{-1/2}\sum_{j\geqslant 1}\|z_{I_j}\|_1 = k^{-1/2}\|z_{I^c}\|_1$.
Par ailleurs, par isométrie restreinte, on a : $\|z_{I_0\cup I_1}\|_2^2 \leqslant (1-\delta)^{-1}\|A_{I_0\cup I_1} z_{I_0\cup I_1}\|_2^2$. De plus, en utilisant $Az = 0$ et $z = z_{I_0\cup I_1} + \sum_{j\geqslant 2} z_{I_j}$, on a $\|A_{I_0\cup I_1} z_{I_0\cup I_1}\|_2^2 = -(A_{I_0\cup I_1} z_{I_0\cup I_1})^\top \sum_{j\geqslant 2} A_{I_j} z_{I_j} \leqslant \sum_{j\geqslant 2}\left\{|(A_{I_0} z_{I_0})^\top A_{I_j} z_{I_j}| + |(A_{I_1} z_{I_1})^\top A_{I_j} z_{I_j}|\right\}$, qui est inférieur à $\delta(\|z_{I_0}\|_2 + \|z_{I_1}\|_2)\sum_{j\geqslant 2}\|z_{I_j}\|_2 \leqslant \delta\sqrt{2}\|z_{I_0\cup I_1}\|_2 k^{-1/2}\|z_{I^c}\|_1$ grâce à (a) et la décomposition de la norme $\|z_{I_0\cup I_1}\|_2^2$.
On obtient donc $\|z_{I_0\cup I_1}\|_2^2 \leqslant \frac{\sqrt{2}\delta}{1-\delta}\|z_{I_0\cup I_1}\|_2 k^{-1/2}\|z_{I^c}\|_1$, ce qui permet d'arriver au résultat attendu, i.e., $\|z_I\|_1 \leqslant k^{1/2}\|z_{I_0\cup I_1}\|_2 \leqslant \rho\|z_{I^c}\|_1$.

(c) Soit $x$ la solution de l'Eq. (2). Le vecteur $z = x - x^*$ est $(2k)$-creux et satisfait $Az = 0$. Donc, d'après (a), $\|z_I\|_1 \leqslant \rho\|z_{I^c}\|_1$ pour $I$ le support de $x^*$. On a donc, par optimalité de $x$ pour Eq. (2), $\|x_I\|_1 + \|x_{I^c}\|_1 = \|x\|_1 \leqslant \|x^*\|_1 = \|x_I^*\|_1$, car $I$ correspond aux $k$ éléments non nuls de $x^*$. Ceci implique par l'inégalité triangulaire que : $\|x_I^*\|_1 - \|z_I\|_1 + \|z_{I^c}\|_1 - \|x_{I^c}^*\|_1 \leqslant \|x_I\|_1 + \|x_{I^c}\|_1 \leqslant \|x_I^*\|_1$, et donc, $\|z_{I^c}\|_1 \leqslant \|z_I\|_1 \leqslant \rho\|z_{I^c}\|_1$. Pour $\delta < \sqrt{2} - 1$, on a $\rho < 1$, ce qui implique $z = 0$, et donc l'unicité de $x$.

On peut faire les observations suivantes :

- Le résultat précédent s'étend naturellement aux situations bruitées, c'est-à-dire, quand $x^*$ n'est pas $k$-creux ou quand $y$ est observé avec du bruit supplémentaire, avec un contrôle précis des erreurs commises [Can2].
- La condition déterministe d'isométrie restreinte permet d'assurer le succès de l'algorithme d'optimisation convexe. Cependant, étant donnée une matrice $A$, il n'existe pas d'algorithme en temps polynomial permettant de certifier qu'elle est satisfaite. En effet, le problème de « valeurs propres parcimonieuses », i.e., trouver la plus grande valeur propre de toutes les sous-matrices de taille $k$ est un problème computationellement difficile [BR], pour lequel des relaxations convexes existent [AGJL] mais ne permettent pas de retrouver les mêmes dépendances entre $k$, $m$ et $n$.

– On retrouve un résultat pour les garanties dépendant de la proprété de cohérence, car $\delta_{2k} \leqslant \mu(2k-1)$, ce qui montre que si $\mu(2k-1) < \sqrt{2}-1$, alors la solution de l'Eq. (2) est bien unique, ce qui correspond à la condition $k < \frac{\sqrt{2}-1}{2\mu} + \frac{1}{2}$, une condition à peine plus faible que les résultats existants [GN, DE].

### 2.2. Analyse probabiliste

Si il est difficile d'exhiber une matrice satisfaisant l'hypothèse d'isométrie restreinte, en échantillonnant les éléments de $A$ aléatoirement et indépendamment d'une distribution sous-gaussienne, on obtient de telles matrices avec forte probabilité si $k$ est suffisamment petit.

La seule propriété utilisée sera la concentration de $\|Ax\|_2^2$ autour de son espérance pour tout $x \in \mathbb{R}^n$. On fait l'hypothèse que $\mathbb{E}A = 0$ et $\mathbb{E}(A^\top A) = I$, de telle sorte que $\mathbb{E}\|Ax\|_2^2 = \|x\|_2^2$. On suppose donc que pour tout $x \in \mathbb{R}^n$,

$$\mathbb{P}\big(\big|\|Ax\|_2^2 - \|x\|_2^2\big| \geqslant \varepsilon\|x\|_2^2\big) \leqslant 2e^{-mc(\varepsilon)}, \tag{5}$$

pour une certaine constante $c(\varepsilon) > 0$ dépendant de $\varepsilon$. Dès que les éléments de $A$ sont pris aléatoirement et indépendants avec la même loi sous-gaussienne, la propriété en Eq. (5) est satisfaite. Par exemple, si cette distribution est gaussienne avec moyenne nulle et variance $1/m$, ceci est une conséquence de la concentration de carrés de gaussiennes [BLM]. Il existe d'autres matrices intéressantes ; en particulier quand les éléments de $A$ peuvent prendre deux valeurs distinctes $1/\sqrt{m}$ et $-1/\sqrt{m}$, ou alors trois valeurs, $+\sqrt{3/m}$, 0 et $-\sqrt{3/m}$, dont une nulle, ce qui peut avoir un intérêt algorithmique complémentaire. Pour toutes ces matrices, on peut choisir $c(\varepsilon) = \varepsilon^2/4 - \varepsilon^3/6$ [Ach].

On montre d'abord que pour tout $I$ ensemble à $k$ éléments, les valeurs propres de $A_I^\top A_I$ sont comprises entre $1-\delta$ et $1+\delta$ avec forte probabilité, en utilisant un argument classique de couverture de la sphère unité par des boules $\ell_2$. Nous reprenons ici la preuve de [BDDeVW], qui est elle-même très proche de la preuve du lemme de Johnson-Lindenstrauss [JL].

Proposition 2.4. — *Supposons que $A$ est aléatoire et satisfait la propriété de concentration en Eq. (5), alors pour tout $\delta \in ]0,1[$, les valeurs propres de $A_I^\top A_I$ sont comprises entre $\sqrt{1-\delta}$ et $\sqrt{1+\delta}$ avec probabilité supérieure à $1 - 2(12/\delta)^k e^{-mc(\delta/2)}$.*

Preuve — Sans perte de généralité, on se ramène à $I = \{1, \dots, k\}$. On choisit un ensemble de points $\mathcal{X}$ de la sphère unité en dimension $k$, de telle sorte que tout point de la sphère soit à distance inférieure à $\delta/4$ d'au moins un de ces points. Des arguments classiques de couverture montrent que l'on peut obtenir une telle distance avec au plus $(12/\delta)^k$ points. En considérant l'application de l'Eq. (5) aux éléments de $\mathcal{X}$ avec $\varepsilon = \delta/2$ et en utilisant la borne de l'union, avec la probabilité demandée, on peut approcher tout $x$ de norme 1 et de support inclus dans $I = \{1, \dots, k\}$ par un point $y$ bien choisi parmi l'ensemble de points précédents, pour obtenir $u = \max_{\|x\|_2=1, x_{I^c}=0} \|Ax\|_2 \leqslant \max_{\|x\|_2=1, x_{I^c}=0} \min_{y\in\mathcal{X}} \|A(x-y)\|_2 + \|Ay\|_2 \leqslant u\frac{\delta}{4} + 1 + \frac{\delta}{2}$, ce

qui implique que $u \leqslant \frac{1+\delta/2}{1-\delta/4} \leqslant 1+\delta$. L'autre inégalité se montre de manière équivalente et on obtient le résultat demandé.

Une fois obtenu que pour tout $I$ la propriété de valeurs propres restreintes est satisfaite, il ne reste plus qu'à dénombrer l'ensemble de ces ensembles à $k$ éléments parmi $n$.

THÉORÈME 2.5. — *Supposons que $A$ est aléatoire avec la propriété de concentration en Eq. (5) ; alors pour tout $\delta \in ]0,1[$, la matrice $A$ satisfait la propriété d'isométrie restreinte à l'ordre $k$ au niveau $\delta$ avec probabilité supérieure à $1-2e^{-c_2 m}$, dès que $k \leqslant c_1 \frac{m}{\log(n/k)}$, où les constantes $c_1, c_2$ ne dépendent que de $\delta$ (et de la distribution de $A$).*

PREUVE — Le résultat précédent montre qu'avec probabilité plus grande que $1-2(12/\delta)^k e^{-mc(\delta/2)}$, alors les valeurs propres de $A_I^\top A_I$ sont dans le bon intervalle. Comme il y a $\binom{n}{k} \leqslant (en/k)^k$ tels sous-ensembles, la probabilité que la propriété d'isométrie restreinte ne soit pas vérifiée est inférieure à $2(en/k)^k(12/\delta)^k e^{-mc(\delta/2)} = 2\exp\big[-mc(\delta/2)+k[\log(en/k)+\log(12/\delta)\big]$. En choisissant $k \leqslant c_1 \frac{m}{\log(n/k)}$, alors l'argument de l'exponentielle est inférieur à $-c_2 m$ dès que $c_2 \leqslant c(\delta/2) - c_1[1+(1+\log\frac{12}{\delta})/\log(n/k)]$, ce qui permet un choix de $c_1 > 0$. On note qu'on peut aussi considérer la condition $k \leqslant c_1 \frac{m}{\log(n/m)+1}$.

On peut faire les observations suivantes :

- Dans le cas gaussien, des résultats explicites plus précis peuvent être obtenus (voir par exemple [FR]).
- Il est possible d'obtenir des résultats similaires pour des sous-matrices aléatoires de la base de Fourier [RV, CT1]. On obtient alors un nombre de mesures $m$ devant dépasser une constante fois $k(\log n)^4$, et donc une dépendance linéaire dans le nombre de mesures, et toujours logarithmique dans la dimension ambiante (mais avec une puissance supérieure).

## 3. GARANTIES SANS ISOMÉTRIE RESTREINTE

Dans cette section, nous donnons les éléments de preuves principaux du théorème 1.2. La preuve est construite autour de la notion de certificat dual, commune à la plupart des problèmes d'optimisation convexe, que nous commençons par présenter.

Comme tout problème de programmation linéaire, on peut définir le problème dual comme suit :

$$\begin{aligned}
\min_{x\in\mathbb{R}^n} \|x\|_1 \text{ tel que } Ax = y &= \min_{x\in\mathbb{R}^n} \max_{z\in\mathbb{R}^m} \|x\|_1 - z^\top(Ax-y) \\
&= \max_{z\in\mathbb{R}^m} \min_{x\in\mathbb{R}^n} \|x\|_1 - z^\top(Ax-y) \\
&= \max_{z\in\mathbb{R}^m} z^\top y \text{ tel que } \|A^\top z\|_\infty \leqslant 1.
\end{aligned}$$

Le vecteur $x \in \mathbb{R}^n$ est optimal si et seulement si il existe $z \in \mathbb{R}^m$ tel que $\|A^\top z\|_\infty \leqslant 1$ et $\|x\|_1 = x^\top A^\top z$, ce qui implique que les composantes de $A^\top z$ correspondant aux éléments non nuls de $x$ doivent être égales aux signes correspondants (pris dans $\{-1, 1\}$). Ainsi, si $I$ est le support de $x$, alors on doit avoir $(A^\top z)_I = \mathrm{sign}(x_I)$ à l'optimalité. L'existence d'un tel $z$ montre que $x$ est bien solution. Pour être solution unique, on a la condition suffisante ci-dessous :

LEMME 3.1. — *Soit $x^* \in \mathbb{R}^n$ et $y = Ax^*$. On considère le support $I$ de $x^*$ défini par $I = \{i \in \{1, \dots, n\}, x_i^* \neq 0\}$. Si (a) le noyau de $A_I$ est égal à $\{0\}$, et (b) il existe $z \in \mathbb{R}^m$ tel que $(A^\top z)_I = \mathrm{sign}(x_I^*)$ et $\|(A^\top z)_{I^c}\|_\infty < 1$, alors $x^*$ est la solution unique du problème en Eq. (2).*

PREUVE — Si $x' = x^* + h$ est une autre solution, alors, $Ah = 0$ et $\|x^* + h\|_1 - \|x^*\|_1 = h_I^\top \mathrm{sign}(x_I^*) + \|h_{I^c}\|_1 > h_I^\top (A^\top z)_I + h_{I^c}^\top (A^\top z)_{I^c} = h^\top A^\top z = 0$, dès que $h_{I^c} \neq 0$. Ceci implique $h_{I^c} = 0$. Comme $A_I$ est supposée de rang égal à $|I|$, ceci implique $h = 0$, et donc le lemme.

L'idée maintenant est de montrer qu'il existe un tel certificat dual avec grande probabilité. La première approche consiste à construire à partir des données un bon candidat. La solution la plus simple est de considérer $z$ qui minimise $\|A^\top z\|_2$ avec la contrainte que $(A^\top z)_I = \mathrm{sign}(x_I^*)$. Si ce candidat $z \in \mathbb{R}^n$ satisfait $\|(A^\top z)_{I^c}\|_\infty < 1$, la preuve est terminée (des choix plus pertinents sont possibles [Gro]).

Ce choix de candidat est intéressant car on sait le calculer en formule analytique comme $z = A^\top A_I (A_I^\top A_I)^{-1} \mathrm{sign}(x_I^*)$, en faisant l'hypothèse que $A_I^\top A_I$ est inversible. On est donc amené à étudier (a) les conditions sous lesquelles $A_I^\top A_I$ est inversible pour le modèle de génération de la matrice $A$, et (b) s'assurer que l'on a bien $\|(A^\top z)_{I^c}\|_\infty < 1$. Ceci peut se faire à l'aide d'outils de matrices aléatoires, soit en utilisant des inégalités de Bernstein pour matrices [AW, Tro], soit en développant en série entière et bornant les moments d'ordre supérieur par des méthodes combinatoires [CRT].

## 4. MESURES GAUSSIENNES ET TRANSITION DE PHASE

Lorsque les mesures sont gaussiennes, i.e., quand chaque élément de $A$ est une variable gaussienne indépendante de moyenne nulle et de variance uniforme, il existe des connections intéressantes avec la géométrie convexe. En effet, dans le cas gaussien, le noyau de $A$ est uniformément distribué parmi tous les sous-espaces de dimension $n-m$.

Rappelons que la solution du problème de minimisation de norme $\Omega$ est unique si et seulement si le noyau $\mathrm{Ker}(A)$ de $A$ et le cône tangent $\mathcal{C}$, défini en Eq. (3), s'intersectent uniquement en $\{0\}$. Il suffit donc d'étudier l'intersection entre un cône donné et un sous-espace aléatoire, ce que les travaux de Gordon permettent d'obtenir très précisément et de manière très générale.

Nous aurons besoin de la notion de *largeur gaussienne* d'un ensemble $\mathcal{K}$, définie comme

$$w(\mathcal{K}) = \mathbb{E}_{g\sim\mathcal{N}(0,I)} \sup_{z\in\mathcal{K}\cap\mathbb{S}^{n-1}} g^\top z,$$

où $\mathcal{N}(0,I)$ est la loi normale de moyenne nulle et de matrice de covariance identité, et $\mathbb{S}^{n-1}$ est la sphère unité de $\mathbb{R}^n$. Nous avons alors le théorème suivant :

THÉORÈME 4.1 ([Gor]). — *Soit $\mathcal{K} \subset \mathbb{R}^n$ un cône et $A$ une matrice gaussienne. Si $m \geqslant (w(\mathcal{K})+t)^2+1$, alors* $\mathrm{Ker}(A)\cap\mathcal{K} = \{0\}$ *avec probabilité supérieure à* $1-\exp(-t^2/2)$.

Le théorème précédent implique immédiatement qu'un peu plus de $w(\mathcal{C})^2$ mesures gaussiennes sont nécessaires pour résoudre notre problème de parcimonie avec haute probabilité. Comme nous le verrons ci-dessous, un peu moins de mesures ne permettraient pas d'estimer notre signal.

Dans le cas de la norme $\ell_1$ et de son cône tangent, on peut montrer par un calcul explicite que $w(\mathcal{C})^2 \leqslant 2k\log(n/k)+2k$, et on retrouve le résultat de la section 2 dans le cas gaussien, mais le résultat est nettement plus général et s'applique à de nombreuses situations (i.e., de nombreuses normes) [CRPW] au-delà des vecteurs creux.

Un résultat encore plus marquant est que le carré de la largeur gaussienne (aussi appelé « dimension statistique' » [ALMCT]) permet de caractériser de manière très précise une transition de phase entre le succès et l'échec de la formulation convexe d'une minimisation de norme, dans un cadre très général.

THÉORÈME 4.2 ([ALMCT]). — *Soit $x^* \in \mathbb{R}^n$ un vecteur fixe, $\Omega$ une norme, et $\mathcal{C} \subset \mathbb{R}^n$ son cône tangent en $x^*$. Si $A$ est une matrice gaussienne $m \times n$ et $y = Ax^*$, alors pour tout $\varepsilon \in (0,1)$ :*

*(i) si $m \leqslant w(\mathcal{C})^2 - \sqrt{8\log(4/\varepsilon)}$, alors, $x^*$ est le minimum unique de $\Omega(x)$ avec la contrainte $Ax = y$ avec probabilité $\leqslant \varepsilon$,*

*(i) si $m \geqslant w(\mathcal{C})^2 + \sqrt{8\log(4/\varepsilon)}$, alors, $x^*$ est le minimum unique de $\Omega(x)$ avec la contrainte $Ax = y$ avec probabilité $\geqslant 1-\varepsilon$.*

Le théorème permet de caractériser de manière fine le phénomène de transition de phase pour les problèmes d'échantillonnage compressé résolu par optimisation convexe, complétant ainsi une série de travaux précédents montrant aussi une transition de phase dans ce cadre [DT, Wai]. Voir un exemple en Figure 2 pour la norme $\ell_1$ et des mesures gaussiennes.

## 5. CONCLUSIONS ET PERSPECTIVES

Dans ce texte, nous avons décrit certaines des contributions importantes de l'échantillonage compressé, en particulier à partir des travaux d'Emmanuel Candès. Ces travaux s'inscrivent dans une perspective plus large dont nous mentionnerons maintenant

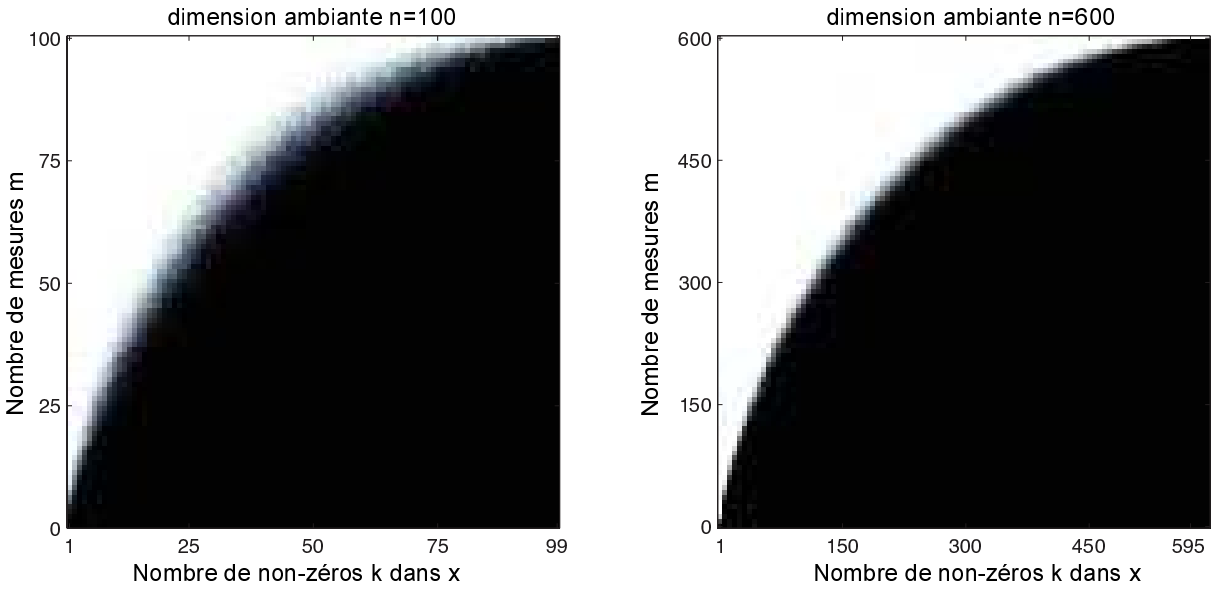


FIGURE 2. Exemple de transition de phase pour deux valeurs de $n$ ($n = 100$ à gauche, $n = 600$ à droite), mesures gaussiennes et norme $\ell_1$. La brillance de chacun des points correspond à la probabilité observée de succès, entre échec certain (noir) et succès certain (blanc). On pourra noter la transition de phase plus affirmée pour $n$ supérieur. De [ALMCT].

quelques éléments non exhaustifs. Pour plus de détails, on pourra consulter les ouvrages suivants [CGLP, FR, HTW, BVdG, Gir].

— **Estimation sans optimisation convexe** : Étant données les hypothèses faites sur la matrice $A$, il existe d'autres algorithmes permettant de retrouver une solution unique sans passer par l'optimisation convexe, avec des complexités algorithmiques et des résultats théoriques similaires [MZ, NV, TG].
— **Liens avec d'autres résultats en géométrie** : les différents résultats montrés en Sections 2, 3 et 4 ont utilisé des résultats en géométrie en haute dimension. Des connections supplémentaires existent comme le lien avec les largeurs de Gelfand ou de Kolmogorov [KT, CGLP].
— **Parcimonie en statistique** : Dans ce texte, nous avons porté l'accent sur les aspects de traitement du signal, où le vecteur $y$ est un vecteur de mesures et $A$ la matrice des projections donnant lieu à ces mesures. Les résultats théoriques principaux correspondent à des matrices $A$ aléatoires (e.g., gaussienne). Dans un cadre statistique comme l'application en génomique présentée en introduction, la matrice $A$ représente les données. Si un modèle stochastique accompagne souvent ces données il est nécessaire de prendre en compte des hypothèses plus fines, en particulier pour le bruit dans le vecteur de réponses $y$. Voir [BRT, HTW, BVdG, Gir] pour un traitement exhaustif.

## RÉFÉRENCES


[Ach] D. ACHLIOPTAS – *Database-friendly random projections : Johnson-Lindenstrauss with binary coins*, Journal of computer and System Sciences **66**, 4 (2003), 671–687.

[AW] R. AHLSWEDE, A. WINTER – *Strong converse for identification via quantum channels*, IEEE Transactions on Information Theory **48**, 3 (2002), 569–579.

[ALMCT] D. AMELUNXEN, M. LOTZ, M. McCOY, J. A. TROPP – *Living on the edge : phase transitions in convex programs with random data*, Information and Inference (2014).

[AGJL] A. D'ASPREMONT, L. EL GHAOUI, M. I. JORDAN, G. R. LANCKRIET – *A direct formulation for sparse PCA using semidefinite programming*, SIAM Review **49**, 3 (2007), 434–448.

[Bac] F. BACH – *Learning with submodular functions : A convex optimization Perspective*, Foundations and Trends in Machine Learning **6**, 2-3 (2013), 145–373.

[BJMO] F. BACH, R. JENATTON, J. MAIRAL, G. OBOZINSKI – *Optimization with sparsity-inducing penalties*, Foundations and Trends in Machine Learning **4**, 1 (2012), 1–106.

[BDDeVW] R. BARANIUK, M. DAVENPORT, R. DEVORE, M. WAKIN – *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation **28**, 3 (2008), 253–263.

[BT] A. BECK, M. TEBOULLE – *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences **2**, 1 (2009), 183–202.

[BR] Q. BERTHET, P. RIGOLLET – *Computational lower bounds for sparse PCA*, ArXiv preprint 1304.0828 (2013).

[BRT] P. J. BICKEL, Y. A. RITOV, A. B. TSYBAKOV – *Simultaneous analysis of Lasso and Dantzig selector*, The Annals of Statistics **37**, 4 (2009), 1705-1732.

[BLM] S. BOUCHERON, G. LUGOSI, P. MASSART – *Concentration inequalities : A nonasymptotic theory of independence*, Oxford University Press (2013).

[BVdG] P. BUHLMANN, S. VAN DE GEER – *Statistics for high-dimensional data : methods, theory and applications*, Springer (2011).

[Can1] E. J. CANDÈS – *Mathematics of sparsity (and a few other things)*, Proceedings of the International Congress of Mathematicians, Seoul, South Korea (2014).

[Can2] E. J. CANDÈS – *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus Mathématiques de l'Académie des Sciences **346**, 9 (2008), 589–592.

[CT1] E. J. CANDÈS, T. TAO – *Decoding by linear programming*, IEEE Transactions on Information Theory **51**, 12 (2005), 4203–4215. ISO 690

[CT2] E. J. CANDÈS, T. TAO – *Near-optimal signal recovery from random projections : Universal encoding strategies ?*, IEEE Transactions on Information Theory **52**, 12 (2006), 5406–5425.

[CT3] E. J. CANDÈS, T. TAO – *The power of convex relaxation : Near-optimal matrix completion*, IEEE Transactions on Information Theory **56**, 5 (2010), 2053–2080.

[CP] E. J. CANDÈS, Y. PLAN – *A probabilistic and RIPless theory of compressed sensing*, IEEE Transactions on Information Theory **57**, 11 (2011),7235–7254.

[CR] E. J. CANDÈS, B. RECHT – *Exact matrix completion via convex optimization*, Foundations of Computational mathematics, **9**, 6 (2009), 717–772.

[CRT] E. J. CANDÈS, J. ROMBERG, T. TAO – *Robust uncertainty principles : Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Information Theory **52**, 2 (2006), 489–509.

[CGLP] D. CHAFAI, O. GUEDON, G. LECUE, A. PAJOR – *Interactions between compressed sensing, random matrices, and high dimensional geometry*, Panoramas et synthèses **37** (2011).

[CRPW] V. CHANDRASEKARAN, B. RECHT, P. A. PARILLO, A. WILLSKY – *The convex geometry of linear inverse problems*, Foundations of Computational Mathematics **12**, 6 (2012), 805–849.

[CDS] S. S. CHEN, D. L. DONOHO, M. A. SAUNDERS – *Atomic decomposition by basis pursuit*. SIAM Review **43**, 1 (2001), 129–159.

[CDDeV] A. COHEN, W. DAHMEN, R. DEVORE – *Compressed sensing and best $k$-term approximation*, Journal of the American Mathematical Society **22**, 1 (2009), 211–231.

[DE] D. L. DONOHO, M. ELAD – *Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell_1$-minimization*, Proceedings of the National Academy of Sciences **100**, 5 (2003), 2197–2202.

[DT] D. DONOHO, J. TANNER – *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of the American Mathematical Society **22**, 1 (2009), 1–53.

[FHB] M. FAZEL, H. HINDI, S. P. BOYD – *A rank minimization heuristic with application to minimum order system approximation*, Proceedings of the American Control Conference **6** (2001), 4734–4739.

[FR] S. FOUCART, H. RAUHUT – *A mathematical introduction to compressive sensing*, Birkhäuser (2003).

[Gir] C. GIRAUD – *Introduction to high-dimensional statistics*, CRC Press (2014).

[Gor] Y. GORDON – *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$, In Geometric Aspects of Functional Analysis, Springer (2006), 84–106.

[GN] R. GRIBONVAL, M. NIELSEN – *Sparse representations in unions of bases*, IEEE Transactions on Information Theory **49**, 12 (2003), 3320–3325.

[Gro] D. GROSS – *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transactions on Information Theory **57**, 3 (2011), 1548–1566.

[HTW] T. HASTIE, R. TIBSHIRANI, M. WAINWRIGHT – *Statistical learning with sparsity : the Lasso and generalizations*, CRC Press (2015).

[JL] W. B. JOHNSON, J. LINDENSTRAUSS – *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary Mathematics **26**, 189–206.

[KT] B. S. KASHIN, V. N. TEMLYAKOV – *A remark on compressed sensing*, Mathematical Notes **82**, 5-6 (2007), 748–755.

[MZ] S. MALLAT, Z. ZHANG – *Matching pursuits with time-frequency dictionaries*. IEEE Transactions on Signal Processing **41**, 12 (1993), 3397–3415.

[Nat] B. K. NATARAJAN – *Sparse approximate solutions to linear systems*, SIAM Journal on Computing **24**, 2 (1995), 227–234.

[NV] D. NEEDELL, R. VERSHYNIN – *Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit*, Foundations of Computational Mathematics **9**, 3 (2009), 317–334.

[NN] Y. NESTEROV, A. NEMIROVSKII – *Interior-point Polynomial Algorithms in Convex Programming*, Society for industrial and Applied Mathematics (1994).

[RV] M. RUDELSON, R. VERSHYNIN – *On sparse reconstruction from Fourier and Gaussian measurements*, Communications on Pure and Applied Mathematics **61**, 8 (2008), 1025–1045.

[ST] D. A. SPIELMAN, S. H. TENG – *Smoothed analysis of algorithms : Why the simplex algorithm usually takes polynomial time*, Journal of the ACM (JACM) **51**, 3 (2004), 385–463.

[SH] T. STROHMER, R. W. HEATH – *Grassmannian frames with applications to coding and communication*, Applied and Computational Harmonic Analysis **14**, 3 (2003), 257–275.

[Tib] R. TIBSHIRANI – *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological) **58**, 1 (1996), 267-288.

[Tro] J. TROPP – *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics **12**, 4 (2012), 389–434.

[TG] J. TROPP, A. GILBERT – *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Transactions on Information Theory **53**, 12 (2007), 4655–4666.

[Wai] M. WAINWRIGHT – *Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso)*, IEEE Transactions on Information Theory **55**, 5 (2009), 2183–2202.

Francis BACH

INRIA
Département d'Informatique de
l'École normale supérieure
(UMR CNRS/ENS/INRIA)
2 rue Simone Iff
F–75012 Paris
*E-mail* : `francis.bach@ens.fr`