390

ASTÉRISQUE

2017

SÉMINAIRE BOURBAKI
VOLUME 2015/2016
EXPOSÉ N° 1117

Francis BACH

Parcimonie et systèmes linéaires sous-determinés

SOCIÉTÉ MATHÉMATIQUE DE FRANCE
Publié avec le concours du CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Astérisque est un périodique de la Société Mathématique de France.

Numéro 390

Diffusion

Maison de la SMF	AMS
Case 916 - Luminy	P.O. Box 6248
13288 Marseille Cedex 9	Providence RI 02940
France	USA
christian.smf@cirm-math.fr	www.ams.org

Tarifs

Vente au numéro: 65 € ($ 97)

Abonnement électronique : 500 € ($ 750)

Abonnement avec supplément papier : 657 €, hors Europe : 699 € ($ 1049)

Des conditions spéciales sont accordées aux membres de la SMF.

Secrétariat : Nathalie Christiaën

Astérisque

Société Mathématique de France

Institut Henri Poincaré, 11, rue Pierre et Marie Curie

75231 Paris Cedex 05, France

Tél: (33) 01 44 27 67 99 • Fax: (33) 01 40 46 90 96

astsmf@ihp.fr • http://smf.emath.fr/

ISSN : print 0303-1179, electronic 2492-5926

ISBN 978-2-85629-855-8

Directeur de la publication: Stéphane Seuret

Séminaire BOURBAKI Juin 2016
68e année, 2015-2016, n° 1117, p. 409 à 426

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