

**JOURNÉE GDR ISIS  
ESTIMATION ET TRAITEMENT STATISTIQUE EN GRANDE  
DIMENSION**

Télécom Paristech - 16 mai 2013 - Amphi Saphir

Organisation: R. Couillet, E. Duflos, J. Najim<sup>1</sup>

1. PRÉSENTATION

Les méthodes modernes d'acquisition de données engendrent souvent des jeux de données de grande dimension, qui nécessitent un traitement statistique approprié. L'objectif de la journée "estimation et traitement statistique en grande dimension" sera d'exposer des résultats récents sur cette thématique.

2. PROGRAMME PRÉVISIONNEL

9h30-10h30	Philippe Loubaton
10h30-11h10	Jérémie Mary
11h10-11h30	pause
11h30-12h30	Arnak Dalalyan
12h30-14h00	déjeuner
14h00-15h00	Alexandre d'Aspremont
15h00-15h40	Frédéric Pascal
15h40-16h00	pause
16h00-16h40	Arthur Tenenhaus
16h40-17h40	Michael Jordan

3. TITRES ET RÉSUMÉS DES EXPOSÉS

**Alexandre d'Aspremont (école polytechnique).**

*Titre.* Phase Recovery, MaxCut and Complex Semidefinite Programming.

---

<sup>1</sup>romain.couillet@supelec.fr, emmanuel.duflos@ec-lille.fr, najim@univ-mlv.fr

*Résumé.* Phase retrieval seeks to recover a signal  $x$  from the amplitude  $|Ax|$  of linear measurements. We cast the phase retrieval problem as a non-convex quadratic program over a complex phase vector and formulate a tractable relaxation (called PhaseCut) similar to the classical MaxCut semidefinite program. We solve this problem using a provably convergent block coordinate descent algorithm whose structure is similar to that of the original greedy algorithm in Gerchberg-Saxton, where each iteration is a matrix vector product. Numerical results show the performance of this approach on wavelet and molecular imaging examples.

[en collaboration avec Fajwel Fogel, Irène Waldspurger, Stéphane Mallat]

**Arnak Dalalyan (ENSAE).**

*Titre.* Learning Heteroscedastic Models by Convex Programming under Group Sparsity

*Résumé.* Popular sparse estimation methods based on  $l_1$ -relaxation, such as the Lasso and the Dantzig selector, require the knowledge of the variance of the noise in order to properly tune the regularization parameter. This constitutes a major obstacle in applying these methods in several frameworks-such as time series, random fields, inverse problems-for which the noise is rarely homoscedastic and its level is hard to know in advance.

In this talk, we propose a new approach to the joint estimation of the conditional mean and the conditional variance in a high-dimensional (auto-)regression setting. An attractive feature of the proposed estimator is that it is efficiently computable even for very large scale problems by solving a second-order cone program (SOCP). We present theoretical analysis and numerical results assessing the performance of the proposed procedure.

[en collaboration avec M. Hebiri, K. Meziani et J. Salmon]

**Michael Jordan (University of California, Berkeley).**

*Titre.* Inférence Statistique pour "Big Data"

*Résumé.* Dans cet exposé, nous présentons quelques résultats récents dans le domaine d'inférence pour "Big Data". Considérons, par exemple, le problème fondamental d'obtention d'intervalles de confiance pour les estimateurs. Le principe du "bootstrap" suggère d'échantillonner les données à plusieurs reprises pour obtenir des fluctuations et donc des intervalles de confiance, mais ceci est irréalisable à grande échelle. Si on fait appel à des sous-échantillons, on obtient des fluctuations qui ne sont pas à la bonne échelle. Nous présentons ici une approche nouvelle, le "bag of little bootstraps," qui circonviend ce problème et qui peut être appliquée à des données à grande échelle. Nous aborderons aussi le problème général de compromis (tradeoff) entre computation et inférence, utilisant une approche computationnelle basée sur les relaxations de problèmes convexes.

[En collaboration avec Ariel Kleiner, Purna Sarkar, Ameet Talwalkar, and Venkat Chandrasekaran.]

**Philippe Loubaton (université Paris Est).**

*Titre.* Grandes matrices aléatoires et traitement statistique du signal.

*Résumé.* Cet exposé a pour but de présenter un ensemble de résultats existants montrant les apports des grandes matrices aléatoires dans le cadre de problèmes de traitement statistique des signaux multivariés. Nous envisagerons un certain nombre de problèmes très classiques, mais étudiés dans le contexte inhabituel où le nombre d'échantillons prélevés est du même ordre de grandeur que la dimension des observations. Après une courte introduction ayant pour but de situer le domaine des grandes matrices aléatoires, nous commencerons par présenter la distribution de Marcenko-Pastur. Ensuite, nous aborderons les problèmes de détection d'une source dans un bruit, ainsi que les problèmes de localisation de sources par méthode sous-espace dans lesquels nous aurons à évoquer divers résultats et techniques. Nous conclurons l'exposé par un ensemble de questions plus ou moins ouvertes dans le contexte des grandes dimensions, tels que la séparation aveugle de sources ou divers traitements spatio-temporels.

**Jérémy Mary (Université de Lille).**

*Titre.* Recommendation systems from a bandit perspective.

*Résumé.* We are going to present recent and useful results for the exploration-exploitation dilemma. We will show how these algorithms rely and can complete matrix decomposition algorithms for recommendation. We will discuss how to evaluate an online policy using some batch data with a minimal bias. Some practical cases will be shown with a special emphasis on the Exploration & Exploitation challenge from IMLC'12

**Frédéric Pascal (supélec).**

*Titre.* Robust covariance matrices estimation and applications in signal processing.

*Résumé.* This talk deals with problems of covariance matrix estimation in signal processing. Under the widely used Gaussian assumption, the Sample Covariance Matrix (SCM) estimate provides optimal results in terms of estimation performance. However, when the data turn to be non-Gaussian, the resulting performance can be strongly degraded. To fill this gap, I will first introduce the general framework of the Robust Estimation Theory, and then, I will show some recent results, applied to radar detection as well as Direction-of-Arrival estimation. The second part of the presentation will be devoted to the Random Matrix Theory (RMT) applied to robust estimation problems. Particularly, I will show recent results in robust RMT that provide significant improvement in terms of estimation performance.

**Arthur Tenenhaus (supélec).**

*Titre.* Analyse canonique généralisée régularisée pour l'intégration de données de génétique multimodale.

*Résumé.* Par opposition, aux données standard décrites par un unique tableau individus  $\times$  variables, les données dites multi-blocs sont caractérisées par plusieurs tableaux interconnectés, potentiellement de grandes dimensions, hétérogènes. La nécessité de respecter l'intégrité de la structure multi-blocs des données dans l'analyse apparaît être de plus en plus clairement indispensable ; ceci requiert le développement de méthodes statistiques d'analyse de données adaptées. Dans ce contexte, l'Analyse Canonique Généralisée Régularisée (RGCCA) apparaît être une méthode appropriée et sera présentée au cours de cet exposé. Une application d'analyse de données de génétique multimodale (en collaboration avec l'Institut Gustave Roussy) sera présentée.