Statistical learning,
February 1 - 5, 2016

Françis Bach: Large-scale machine learning and convex optimization.

Many machine learning and signal processing problems are traditionally cast as convex optimization problems. A common difficulty in solving these problems is the size of the data, where there are many observations ("large n") and each of these is large ("large p"). In this setting, online algorithms such as stochastic gradient descent which pass over the data only once, are usually preferred over batch algorithms, which require multiple passes over the data. Given n observations/iterations, the optimal convergence rates of these algorithms are $O(1/\sqrt{n})$ for general convex functions and reaches $O(1/n)$ for strongly-convex functions.

In this tutorial, I will first present the classical results in stochastic approximation and relate them to classical optimization and statistics results. I will then show how the smoothness of loss functions may be used to design novel algorithms with improved behavior, both in theory and practice: in the ideal infinite-data setting, an efficient novel Newton-based stochastic approximation algorithm leads to a convergence rate of $O(1/n)$ without strong convexity assumptions, while in the practical finite-data setting, an appropriate combination of batch and online algorithms leads to unexpected behaviors, such as a linear convergence rate for strongly convex problems, with an iteration cost similar to stochastic gradient descent.

Philippe Besse: Apprentissage et données massives.

Le phénomène "big data" envahie les médias depuis plus de trois ans à tel point qu'une "nouvelle" science a vu le jour sous l'appellation de "science des données". Nous ferons rapidement le point sur les principaux enjeux émergeant dans la recherche en Mathématiques et Statistique avec les saisies et stockages de données massives en lien avec une forme de "datafication" de notre quotidien. Quels sont les nouveaux (ou pas) paradigmes de cette "nouvelle" science ? Quels en sont les principaux enjeux économiques ? Nous nous focaliserons donc sur la valorisation des données massives qui fait nécessairement appel à des algorithmes d'apprentissage machine / statistique. L'avalanche de données génère une avalanche de nouvelles technologies (Hadoop, Spark...) et aussi la large diffusion de certaines plus anciennes (R, Pyhton). Quelles sont les méthodes d'apprentissages compatibles avec ces technologies et qui (scalable) supportent le passage à l'échelle volume ? Quelles sont celles finalement mises en exploitation ? Ces questions seront illustrées en explicitant, à l'interface maths-stat-info, les implémentations et utilisation de quelques méthodes sur des usages parmi les plus courants : classification non supervisée, discrimination, fouille de texte, recommandation.

Gilles Blanchard: Is adaptive early stopping possible in statistical inverse problems?

We consider a standard setting of statistical inverse problem, taking the form of the Gaussian sequence model with D observed noisy coefficients. Consider the simple family of "keep or kill" estimators depending on a cutoff index k_0.

For the choice of the cutoff index, there exist a number of well-known methods achieving oracle adaptivity (i.e. data-dependent choice of k_0 whose performance is comparable to the unknown optimal one), such as penalization and Lepski's method. However, they have in common that the estimators for all values of k_0 have to be computed first and compared to each other in some way. Contrast this to an "early stopping" approach where we would like to compute iteratively the estimators for $k_0 = 1, 2, ...$ and have to decide to stop at some point without being allowed to compute the following estimators. Is oracle adaptivity possible then? This question is motivated by settings where computing estimators for larger k_0 requires more computational cost; furthermore some form of early stopping is most often used in practice. We propose a precise mathematical formulation of this question and provide upper and lower bounds on what is achievable.
(Joint work with M. Hoffmann and M. Reiss)

Sébastien Bubeck: Entropy, geometry, and a CLT for Wishart matrices.

Wishart matrices appear in many areas of (applied) mathematics, e.g. as covariance matrices in statistics, or as a model of a random mixed quantum state in physics. In this talk I will prove a new central limit theorem for high-dimensional Wishart matrices, using a now well-understood information theoretic machinery (which will be reviewed). I will discuss an application of this result to the problem of finding geometry in random networks. Several (new) conjectures will be mentioned too.
Joint work with Shirshendu Ganguly.

Peter Buhlmann: The power of heterogeneous large-scale data for high-dimensional causal inference.

We present a novel methodology for causal inference based on an invariance principle. It exploits the advantage of heterogeneity in larger datasets, arising from different experimental conditions (i.e. an aspect of "Big Data"). Despite fundamental identifiability issues, the method comes with statistical confidence statements leading to more reliable results than alternative procedures based on graphical modeling. We also discuss applications in biology, in particular for large-scale gene knock-down experiments in yeast where computational and statistical methods have an interesting potential for prediction and prioritization of new experimental interventions.

Stéphane Canu: Mixed integer programming for sparse and non convex machine learning.

Many problems in machine learning can be recast as mixed 0-1 integer programs (MIP). This the case of classification, clustering, variable selection, outlier detection, 3d image reconstruction, low rank matrix factorisation. Despite the computational complexity of integer optimization, hardware improvements combined with algorithmic advances and better formulation makes it possible to consider this kind of approach for solving moderate size machine learning problems. We will introduce MIP and discussed their theoretical and practical interests. In particular, we will see how to do regression and SVM with both variable selection and outlier detection in this framework. Practical implementations issues will be also discussed and illustrated on both synthetic and real datasets.

Stéphane Chrétien: A Lagrangian viewpoint on Robust PCA.

Robust PCA was introduced by Candes, Li, Ma et Wright for performing the Singular Value Decomposition of a data matrix corrupted by outliers at unknown positions. In this talk, we will present an elementary approach to the analysis of RPCA based on Lagrange duality and the recent work of Amelunxen, Lotz, McCoy, and Tropp. If time permits, we will also show how to extend the analysis to certain time series models.

Emilie Devijver: Block-diagonal covariance selection for high-dimensional Gaussian graphical models.

Gaussian graphical models are widely utilized to infer and visualize networks of dependencies between continuous variables. However, inferring the graph is difficult when the sample size is small compared to the number of variables. To reduce the number of parameters to estimate in the model, we propose a non-asymptotic model selection procedure supported by strong theoretical guarantees based on an oracle inequality and a minimax lower bound. The covariance matrix of the model is approximated by a block-diagonal matrix. The structure of this matrix is detected by thresholding the sample covariance matrix, where the threshold is selected using the slope heuristic. Based on the block-diagonal structure of the covariance matrix, the estimation problem is divided into several independent problems: subsequently, the network of dependencies between variables is inferred using the graphical lasso algorithm in each block. The performance of the procedure is illustrated on simulated and real data.

Stéphane Gaiffas: Statistical learning with Hawkes processes and new matrix concentration inequalities.

We consider the problem of unveiling the implicit network structure of user interactions in a social network, based only on high-frequency timestamps. Our inference is based on the minimization of the least-squares loss associated with a multivariate Hawkes model, penalized by ℓ_1 and trace-norms. We provide a first theoretical analysis of the generalization error for this problem, that includes sparsity and low-rank inducing priors. This result involves a new data-driven concentration inequality for matrix martingales in continuous time with observable variance, which is a result of independent interest. A consequence of our analysis is the construction of sharply tuned penalizations, that leads to a data-driven scaling of the variability of information available for each users. Numerical experiments illustrate the significant improvements achieved by the use of such data-driven penalizations.

Pierre Geurts: Random forests variable importances: towards a better understanding and large-scale feature selection.

Random forests are among the most popular supervised machine learning methods. One of their most practically useful features is the possibility to derive from the ensemble of trees an importance score for each input variable that assesses its relevance for predicting the

output. These importance scores have been successfully applied on many problems, notably in bioinformatics, but they are still not well understood from a theoretical point of view. In this talk, I will present our recent works towards a better understanding, and consequently a better exploitation, of these measures. In the first part of my talk, I will present a theoretical analysis of the mean decrease impurity importance in asymptotic ensemble and sample size conditions. Our main results include an explicit formulation of this measure in the case of ensemble of totally randomized trees and a discussion of the conditions under which this measure is consistent with respect to a common definition of variable relevance. The second part of the talk will be devoted to the analysis of finite tree ensembles in a constrained framework that assumes that each tree can be built only from a subset of variables of fixed size. This setting is motivated by very high dimensional problems, or embedded systems, where one can not assume that all variables can fit into memory. We first consider a simple method that grows each tree on a subset of variables randomly and uniformly selected among all variables. We analyse the consistency and convergence rate of this method for the identification of all relevant variables under various problem and algorithm settings. From this analysis, we then motivate and design a modified variable sampling mechanism that is shown to significantly improve convergence in several conditions.

Claire Lacour: About the Goldenshluger-Lepski methodology for bandwidth selection.

In this talk we consider the problem of estimating a density with kernel estimators. A classical issue is the choice of the bandwidth. Here we focus on the Goldenshluger-Lepski selection method, which is based on pairwise comparisons between estimators with respect to some loss function. The method also involves a penalty term than typically needs to be large enough in order that the method works (in the sense that one can prove some oracle type inequality for the selected estimator). In the case of the quadratic loss, we study the procedure for different values of the tuning parameters. In particular, we show that a degenerate case, where all the estimators are compared to the overfitted one, works surprinsingly well. We also give a minimal value of the penalty, beyond which the procedure fails, that brings to light a phase transition phenomenon for penalty calibration.

Matthieu Lerasle: Sub-Gaussian mean estimators.

In this presentation I will present a new estimation procedure for the mean of a real valued random variable from an i.i.d. sample. The resulting estimators have a sub-Gaussian behavior even when the underlying distribution is heavy-tailed. I will also show various impossibility results and more generally discuss possibilities and limitations of estimating the mean.
This is joint work with L. Devroye, G. Lugosi and R. I. Oliveira.

Clément Levrard: Reconstruction simpliciale de variétés via l'estimation des plans tangents.

Je présenterai une procédure de reconstruction de variétés développée par l'équipe Geometrica de INRIA (le complexe de Delaunay tangentiel), et montrerai, dans un cadre statistique, plusieurs de ses propriétés en tant qu'estimateur d'une variété source. Une attention particulière sera portée au lien entre vitesse d'estimation des plans tangents et vitesse de convergence du Delaunay tangentiel, ainsi que sur la robustesse à un certain type de bruit.

Sébastien Loustau: Quantization, Learning and Games with OPAC.

We present an unsupervised counterpart of the problem of prediction with expert advices. After some historical considerations, we present sparsity regret bounds for the problem of online clustering. A PAC-Bayesian algorithm called Online PAc Clustering (OPAC) illustrates these theoretical results over simulated and real time generated dataset from e-commerce. Similar results are presented in the batch setting thanks to the standard online to batch conversion.

Stéphane Mallat: Understanding (or not) Deep Neural Networks.

Deep convolutional networks provide state of the art classifications and regressions results over many high-dimensional problems, with spectacular results. We review their architecture, which scatters data with a cascade of linear filter weights and non-linearities. A mathematical framework is introduced to analyze their properties, with many open questions. Computations of invariants involve multiscale contractions with wavelets, the linearization of hierarchical symmetries, and sparse separations. Applications are shown for image and audio processing as well as quantum energy regressions.

André Mas: Eigenvalue-free risk bounds for PCA projectors.

We focus on the PCA of a n-sample of Hilbert-valued data and prove several non asymptotic results related to the difference between estimated and true eigenprojectors. We derive first a lower bound. Then we give upper bounds for the mean square risk for single projectors as well as for projectors associated to the k first eigenvalues. The main point is that these rates do not depend on the spacings between eigenvalues or on their rate of decrease. These results may be applied in several directions where dimension reduction through PCA is at work : functional data inference and nonparametric regression are mentioned.

Patricia Reynaud-Bouret: Estimation of local independence graphs via Hawkes processes and link with the functional neuronal connectivity.

Functional connectivity in Neuroscience is considered as one of the main features of the neural code. It is nowadays possible to obtain the spike activities of tens to hundreds of neurons simultaneously. The question is : can we infer the functional connectivity thanks

to those data ? We use a previous work done with N.R. Hansen (Copenhagen) and V. Rivoirard (Dauphine) to implement data-driven lasso estimates based on Hawkes processes to reconstruct a local independence graph, based on Bernstein type-inequalities for martingales. In a current collaboration with C. Tuleau-Malot (Nice), V. Rivoirard (Dauphine) and neurobiologists of Paris 6 (Thomas Bessaih, Regis Lambert and Nathalie Leresche), we are currently trying to understand to which extent this reconstructed graph is indeed the functional connectivity graph as thought by the biologists, in particular when the underlying process is not a Hawkes process, as it will be the case on real data..

Eric Sibony : A novel multi resolution framework for the statistical analysis of ranking data.

Though the statistical analysis of ranking data has been a subject of interest over the past centuries, especially in economics, psychology or social choice theory, it has been revitalized in the past 15 years by recent applications such as recommender or search engines and is receiving now increasing interest in the machine learning literature. Numerous modern systems indeed generate ranking data, representing for instance ordered results to a query or user preferences. Each such ranking usually involves a small but varying subset of the whole catalog of items only. The study of the variability of these data, i.e. the statistical analysis of incomplete rankings , is however a great statistical and computational challenge, because of their heterogeneity and the related combinatorial complexity of the problem. Whereas many statistical methods for analyzing full rankings (orderings of all the items in the catalog) are documented in the dedicated literature, partial rankings (full rankings with ties) or pairwise comparisons , only a few approaches are available today to deal with incomplete ranking, relying each on a very strong specic assumption. It is the purpose of this talk to introduce a novel general framework for the statistical analysis of incomplete rankings. It is based on a representation tailored to these specic data, whose construction shall also be explained, that fits with the natural multiscale structure of incomplete rankings and provides a new decomposition of rank information with a multiresolution analysis interpretation (MRA). We show that the MRA representation naturally allows to overcome both the statistical and computational challenges without any structural assumption on the data. It therefore provides a general and flexible framework to solve wide variety of statistical problems, where data are of the form of incomplete rankings.

Gilles Stoltz: Robust sequential learning with applications to the forecasting of electricity consumption and of exchange rates.

Sometimes, you feel you're spoilt for choice: there are so many good predictors that you could use! Why select and focus on just one? I will review the framework of robust online aggregation (also known as prediction of individual sequences or online aggregation of expert advice). This setting explains how to combine base forecasts provided by ensemble methods. No stochastic modeling is needed and the performance achieved is comparable to the one of the best (constant convex combination of) base forecast(s). I will illustrate the technology on various data sets, including electricity consumption and exchange rates. More importantly, I will point out open issues, both on the theoretical and on the practical sides.

Alexandre Tsybakov: Oracle inequalities for network models and sparse graphon estimation.

Inhomogeneous random graph models encompass many network models such as stochastic block models and latent position models. We consider the problem of statistical estimation of the matrix of connection probabilities based on the observations of the adjacency matrix of the network. Taking the stochastic block model as an approximation, we construct estimators of network connection probabilities for sparse networks and to find predictions that satisfy oracle inequalities with respect to the block constant oracle. As a consequence, we derive optimal rates of estimation of the connection probability matrix. Two ways of construction of the estimators are explored – modifications of the least squares method and exponentially weighted estimators. For exponentially weighted estimators, we obtain sharp oracle inequalities. As a consequence, we derive optimal rates of estimation of the probability matrix. These results cover the important setting of sparse networks. Another consequence consists in establishing upper bounds on the minimax risks for graphon estimation in the L_2 norm when the probability matrix is sampled according to a graphon model. These bounds include additional terms accounting for the "agnostic" error induced by the variability of the latent unobserved variables of the graphon model. In this setting, the optimal rates are influenced not only by the bias and variance components as in usual nonparametric problems but also include the third component, which is the agnostic error. The results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).
This is a joint work with Olga Klopp and Nicolas Verzelen.