



High Dimensional Statistics, theory and practice ECAS – SFdS course

October 1-6, 2017 - La Villa Clythia - CAES du CNRS, Fréjus, France

<http://ecas2017.sfds.asso.fr/>

Main topic

This course has presented an introduction and recent advances in high dimensional statistics, with a special emphasis on main concepts of variable selection, nonparametric estimation, supervised and non-supervised classification and multiple testing, addressing theoretical, methodological and practical aspects.

Organizers

This meeting is part of the ECAS (ecas.fenstats.eu) courses (15 courses since 1987) which are intended to achieve postgraduate training in special areas of statistics for both researchers and teachers at universities. Young researchers especially from all European countries are encouraged to participate and to contribute to an international audience.

The French statistical society (SFdS, sfds.asso.fr) was also involved in the ECAS initiative and in particular is associated with the organization and the administration of the course.

So, the course benefits from the Board of ECAS and the Council of SFdS. The coordination and organization of the school was ensured by Jean-Michel Poggi (<http://www.math.u-psud.fr/~poggi/>), President of ECAS and one of the former SFdS President.

Supports

The course has benefited from the financial support of the *Fondation mathématique Jacques Hadamard* (FMJH, <http://www.fondation-hadamard.fr/>) and the *Agence pour les mathématiques en interaction avec l'entreprise et la société* (AMIES, <http://www.agence-maths-entreprises.fr/a/>).



The school has also benefited from the indirect SFdS funding through two grants (fees + accommodation + meals) for two PhD students.

In addition, the course has benefited from the direct support of the SFdS Permanent Office (secretary), webmaster and treasurer.

Audience The ideal number of attendees (including 3 speakers and 1 organizer) was reached: 42, mainly doctoral students, postdocs and junior academics, and marginally senior academics and practitioners. A doctoral training session is intended for PhD students and it has consisted of relatively longer and more basic lectures as compared with specialized schools or conferences mainly designed for the training of advanced colleagues.

Detailed program

Classical statistical methods developed during the last century were suitable when the number of observations is much larger than the number of parameters to infer. Unfortunately, many fields such as astronomy, genetics, medicine or neuroscience produce large and complex data sets whose dimension is much larger than the number of experimental units. Such data are said to be *high-dimensional*. To face with this challenging curse of dimensionality, new methodologies have been developed based on sparsity assumptions. The goal of these courses is to present the main concepts and ideas on some selected topics of high-dimensional statistics such as variable selection, nonparametric estimation, supervised and non-supervised classification, or multiple testing. Theoretical aspects are motivated by applicable developments of presented methods.

Course 1 (Vincent Rivoirard): Estimation in the high-dimensional setting

The goal of this course is to present modern statistical tools and some of their theoretical properties for estimation in the high-dimensional setting, including:

- Wavelets and thresholding rules
- Penalized estimators: model selection procedures, Ridge and Lasso estimates
- Generalizations and variations of the Lasso estimate: Group-Lasso, Fused-Lasso, elastic-net and Dantzig selectors. Links with Bayesian rules.
- Statistical properties of Lasso estimators: study in the classical regression model. Extensions for the generalized linear model.

Course 2 (Franck Picard): Empirical properties of penalized estimators in high-dimensional linear models

In this course, we illustrate and explore the empirical properties of penalized estimators in linear models (Gaussian regression and generalized linear models). Using R we learn how to build relevant simulation designs to assess the performance of the LASSO and its derivatives. We also focus on the importance of penalty calibration in practice, using different methods like cross validation or the BIC for instance. Most importantly we show how penalization can be used in the low dimensional as well as in the high dimensional settings. These simulation studies constitute a good framework to explore the inherent difficulties of high dimensional statistics in practice, what should be expected, and what should not be! Another aspect of the course is to learn penalized unsupervised methods like sparse PCA and sparse clustering.

Course 3 (Tristan Mary-Huard): Supervised classification in the high-dimensional setting

In this course, we first introduce the basics about supervised classification along with some classical supervised methods (logistic regression, linear discriminant analysis). We then focus on the high dimensional setting and introduce regularized methods (ridge logistic regression, SVM...). We will investigate how these algorithms can be cast in a general framework of penalized convex risk minimization and will provide some theoretical guarantees about their performances. Practical questions such as model selection via cross-validation are also introduced.

Lecturers

Vincent Rivoirard is Professor, University Paris Dauphine, France, since 2010, and he was Associate Professor at University of Orsay (2003-2010). Graduated at ENS Cachan, he obtained his Ph.D. in Statistics at University Paris Diderot (supervisor: Dominique Picard).

Its research areas include High-dimensional statistics, Non-parametric estimation, Inverse problems, Frequentist and Bayesian approaches, Wavelets and approximation theory, Multivariate point processes.

Personal webpage: <https://www.ceremade.dauphine.fr/~rivoirar/>

Franck Picard is CNRS Research Director since 2016. Graduated at AgroParisTech, he did his PhD in Statistics applied to genomic data, University Orsay (supervisors: J.-J. Daudin, S. Robin). He developed segmentation models for univariate and multivariate series applied to the mapping of genomic copy number variations. He also developed probabilistic models for random graphs based on mixture models, variational inference and model selection. Recent projects involve high dimensional statistics with regularization and selection based on generalized lasso penalties, and wavelet-based non-parametric regression. He is currently heading the *High dimensional statistics for Genomics* group in the Biometry and Evolutionary Biology Lab, Lyon.

Personal webpage: <http://pbil.univ-lyon1.fr/members/fpicard/>

Tristan Mary-Huard spent 6 years as an assistant professor at AgroParisTech, after graduating with PhD in statistics applied to genomics. He is now full researcher at INRA, both in the Mathematics and Plant Genetics departments. His research areas include model selection and variable selection in classification and regression with applications to human and plant genomics, and the development of statistical approaches for data analysis in plant genetics.

Personal webpage:

<https://www6.inra.fr/mia-paris/Equipes/Membres/Tristan-Mary-Huard>

Program Overview

	9:00-10:15	10:45-12:00	12:05-12:30	16:00 -16:25	16:30-18:30	18:45-19:10
Monday 2	Rivoirard	Rivoirard	<i>Flash presentations</i>	<i>Flash presentations</i>	Picard	<i>Welcome reception</i>
Tuesday 3	Rivoirard	Rivoirard	<i>Flash presentations</i>	<i>Flash presentations</i>	Picard	<i>Flash presentations</i>
Wednesday 4	Rivoirard	Mary-Huard		<i>Free time</i>		
Thursday 5	Mary-Huard	Mary-Huard	<i>Flash presentations</i>	<i>Flash presentations</i>	Picard	<i>Flash presentations</i>
Friday 6	Mary-Huard	Mary-Huard				

All the participants have been invited to present in a short speech of 5 minutes, their thesis or their research motivating presence in this course (Flash presentations sessions). More than 30 such presentations have been given.

Practical information

Doctoral students *Fee: 220 € before May 31, 2017 (320 € after)*

University (or similar) *Fee: 420 € before May 31, 2017 (550 € after)*

Other *Fee: 650 € before May 31, 2017 (850 € after)*

Accommodation and meals: 475 €

On-site accommodation and meals is offered for 475 € for 5 nights including breakfast, lunch and dinner (from Sunday evening, until Friday after lunch).

Scientific impact

The course was extremely successful, merging very different kind of attendees (in addition to the three lecturers and the organizer) in terms of level of qualification (16 PhD students, 10 Post-doc and young academics but also 8 more experienced and senior researchers as well as 4 practitioners from business and industry), various nationalities (France, Belgium, Spain, Portugal, Deutschland, Ukrainian, Palestinian, Italian, Chinese) and addressing very different complementary needs and skills.

The course has addressed theoretical, methodological and practical aspects in excellent conditions with a very active group during the course but also outside of the lecture room during the week with a lot of informal and fruitful discussions.

Organization

The organization has benefited from the support of the permanent office of the SFdS as well from the team of CAES-CNRS center.

The technical conditions as well as accommodation and meals were excellent with the highest level of professionalism.

Paris, October 27, 2017
Jean-Michel Poggi