IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2019

Project-Team CELESTE

mathematical statistics and learning

IN COLLABORATION WITH: Laboratoire de mathématiques d'Orsay de l'Université de Paris-Sud (LMO)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team CELESTE

Creation of the Project-Team: 2019 June 01

Keywords:

Computer Science and Digital Science:

A3.1.1. - Modeling, representation
A3.1.8. - Big data (production, storage, transfer)
A3.3. - Data and knowledge analysis
A3.3.3. - Big data analysis
A3.4. - Machine learning and statistics
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.3. - Reinforcement learning
A3.4.4. - Optimization and learning
A3.4.5. - Bayesian methods
A3.4.7. - Kernel methods
A3.5.1. - Analysis of large graphs
A5.9.2. - Estimation, modeling
A6. - Modeling, simulation and control
A6.1. - Methods in mathematical modeling
A6.2. - Scientific computing, Numerical Analysis & Optimization
A6.2.4. - Statistical methods
A6.3. - Computation-data interaction
A6.3.1. - Inverse problems
A6.3.3. - Data processing
A6.3.4. - Model reduction
A9.2. - Machine learning

Other Research Topics and Application Domains:

B1.1.4. - Genetics and genomics
B1.1.7. - Bioinformatics
B2.2.4. - Infectious diseases, Virology
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B3.4. - Risks
B4. - Energy
B4.4. - Energy delivery
B4.5. - Energy consumption
B5.2.1. - Road vehicles
B5.2.2. - Railway
B5.2.3. - Aviation
B5.5. - Materials
B5.9. - Industrial maintenance

B7.1. - Traffic management
B7.1.1. - Pedestrian traffic and crowds
B9.5.2. - Mathematics
B9.8. - Reproducibility

1. Team, Visitors, External Collaborators

Research Scientists

Kevin Bleakley [Inria, Researcher]
Gilles Celeux [Inria, Emeritus]
Matthieu Lerasle [CNRS, Researcher]
Gilles Stoltz [CNRS, Researcher, HDR]

Faculty Members

Sylvain Arlot [Team leader, Univ Paris-Sud, Professor]
Alexandre Janon [Univ Paris-Sud, Associate Professor]
Christophe Giraud [Univ Paris-Sud, Professor]
Christine Keribin [Univ Paris-Sud, Associate Professor, HDR]
Pascal Massart [Univ Paris-Sud, Professor]
Marie-Anne Poursat [Univ Paris-Sud, Associate Professor]
Patrick Pamphile [Univ Paris-Sud, Associate Professor]

Post-Doctoral Fellow

Evgenii Chzhen [Univ Paris-Sud, from Oct 2019]

PhD Students

Yvenn Amara Ouali [Univ Paris-Saclay, from Oct 2019]
Geoffrey Chinot [Univ Paris-Sud]
Olivier Coudray [PSA, from Oct 2019]
Remi Coulaud [SNCF, from Oct 2019]
Solenne Gaucher [Univ Paris-Sud]
Benjamin Goehry [Univ Paris-Sud]
Hedi Hadiji [Ministère de l'Enseignement Supérieur et de la Recherche, from Jun 2019]
Antoine Havet-Morel [Ecole polytechnique, until Aug 2019]
Malo Huard [Univ Paris-Sud, until May 2019]
Yann Issartel [Univ Paris-Saclay]
Perrine Lacroix [Univ Paris-Sud, from Oct 2019]
Guillaume Maillard [Univ Paris-Sud]
Timothee Mathieu [Ecole Normale Supérieure Cachan]
Minh-Lien Nguyen [Univ Paris-Sud]
El Mehdi Saad [Univ Paris-Sud, from Sep 2019]
Solene Thepaut [Univ Paris-Sud]

Interns and Apprentices

Ndeye Lo [Inria, from Jun 2019 until Aug 2019]
Mehdi Zadem [Inria, from Jul 2019 until Sep 2019]

2. Overall Objectives

2.1. Mathematical statistics and learning

Data science – a vast field that includes statistics, machine learning, signal processing, data visualization, and databases – has become front-page news due to its ever-increasing impact on society, over and above the important role it already played in science over the last few decades. Within data science, the statistical community has long-term experience in how to infer knowledge from data, based on solid mathematical foundations. The more recent field of machine learning has also made important progress by combining statistics and optimization, with a fresh point of view that originates in applications where prediction is more important than building models.

The CELESTE project-team is positioned at the interface between statistics and machine learning. We are statisticians in a mathematics department, with strong mathematical backgrounds behind us, interested in interactions between theory, algorithms and applications. Indeed, applications are the source of many of our interesting theoretical problems, while the theory we develop plays a key role in (i) understanding how and why successful statistical learning algorithms work – hence improving them – and (ii) building new algorithms upon mathematical statistics-based foundations

In the theoretical and methodological domains, CELESTE aims to analyze statistical learning algorithms – especially those which are most used in practice – with our mathematical statistics point of view, and develop new learning algorithms based upon our mathematical statistics skills.

A key ingredient in our research program is connecting our theoretical and methodological results with (a great number of) real-world applications. Indeed, CELESTE members work in many domains, including – but not limited to – neglected tropical diseases, pharmacovigilance, high-dimensional transcriptomic analysis, and energy and the environment.

3. Research Program

3.1. General presentation

Our objectives correspond to four major challenges of machine learning where mathematical statistics have a key role. First, any machine learning procedure depends on hyperparameters that must be chosen, and many procedures are available for any given learning problem: both are an estimator selection problem. Second, with high-dimensional and/or large data, the computational complexity of algorithms must be taken into account differently, leading to possible trade-offs between statistical accuracy and complexity, for machine learning procedures themselves as well as for estimator selection procedures. Third, real data are almost always corrupted partially, making it necessary to provide learning (and estimator selection) procedures that are robust to outliers and heavy tails, while being able to handle large datasets. Fourth, science currently faces a reproducibility crisis, making it necessary to provide statistical inference tools (p-values, confidence regions) for assessing the significance of the output of any learning algorithm (including the tuning of its hyperparameters), in a computationally efficient way.

3.2. Estimator selection

An important goal of CELESTE is to build and study procedures that can deal with general estimators (especially those actually used in practice, which often rely on some optimization algorithm), such as cross-validation and Lepski's method. In order to be practical, estimator selection procedures must be fully data-driven (that is, not relying on any unknown quantity), computationally tractable (especially in the high-dimensional setting, for which specific procedures must be developed) and robust to outliers (since most real data sets include a few outliers). CELESTE aims at providing a precise theoretical analysis (for new and existing popular estimator selection procedures), that explains as well as possible their observed behaviour in practice.

3.3. Relating statistical accuracy to computational complexity

When several learning algorithms are available, with increasing computational complexity and statistical performance, which one should be used, given the amount of data and the computational power available? This problem has emerged as a key question induced by the challenge of analyzing large amounts of data – the "big data" challenge. CELESTE wants to tackle the major challenge of understanding the time-accuracy trade-off, which requires providing new statistical analyses of machine learning procedures – as they are done in practice, including optimization algorithms – that are *precise enough* in order to account for differences of performance observed in practice, leading to general conclusions that can be trusted more generally. For instance, we study the performance of ensemble methods combined with subsampling, which is a common strategy for handling big data; examples include random forests and median-of-means algorithms.

3.4. Robustness to outliers and heavy tails (with tractable algorithms)

The classical theory of robustness in statistics has recently received a lot of attention in the machine learning community. The reason is simple: large datasets are easily corrupted, due to – for instance – storage and transmission issues, and most learning algorithms are highly sensitive to dataset corruption. For example, the lasso can be completely misled by the presence of even a single outlier in a dataset. A major challenge in robust learning is to provide computationally tractable estimators with optimal subgaussian guarantees. A second important challenge in robust learning is to deal with datasets where every (x_i, y_i) is slightly corrupted. In large-dimensional data, every single data point x_i is likely to have several corrupted coordinates, and no estimator currently has strong theoretical guarantees for such data. A third important challenge is that of robust estimator selection or aggregation. Even if several robust estimators can be built, the final aggregation or selection step in a user's routine is usually based on empirical means. This is not robust, and may damage the global performance of the procedure. Instead, we can consider more sophisticated types of aggregation of the base robust estimators built so far. A convenient framework to do so is called adversarial learning (also known as: prediction of individual sequences). Here, data is not assumed to be stochastic, and it could even be chosen by an adversary.

3.5. Statistical inference: (multiple) tests and confidence regions (including post-selection)

CELESTE considers the problems of quantifying the uncertainty of predictions or estimations (thanks to confidence intervals) and of providing significance levels (p-values, corrected for multiplicity if needed) for each "discovery" made by a learning algorithm. This is an important practical issue when performing feature selection – one then speaks of post-selection inference – change-point detection or outlier detection, to name but a few. We tackle it in particular through a collaboration with the Parietal team (Inria Saclay) and LBBE (CNRS), with applications in neuroimaging and genomics.

4. Application Domains

4.1. Neglected tropical diseases

CELESTE collaborates with Anavaj Sakuntabhai and Philippe Dussart (Pasteur Institute) on predicting dengue severity using only low-dimensional clinical data obtained at hospital arrival. Further collaborations are underway in dengue fever and encephalitis with researchers at the Pasteur Institute, including with Jean-David Pommier.

4.2. Pharmacovigilance

In pharmacovigilance, the goal is to detect, as soon as possible, potential associations between certain drugs and adverse effects, which appeared after the authorized marketing of these drugs. Preceding works showed the importance of defining an adapted methodology to deal with the size of the individual data (around 250000 reports, 2000 drugs, 4000 adverse effects) and their sparsity. We will explore several aspects from software point of view to statistical strategies as sub-sampling.

4.3. Electricity load consumption: forecasting and control

CELESTE has a long-term collaboration with EDF R&D about electricity consumption. An important problem is to forecast consumption. We currently work on an approach involving back and forth disaggregation (of the total consumption into the consumptions of well-chosen groups/regions) and aggregation of local estimates. We also work on consumption control by price incentives sent to specific users (volunteers), seeing it as a bandit problem.

4.4. Reliability

Collected product lifetime data is often non-homogeneous, affected by production variability and differing real-world usage. Usually, this variability is not controlled or observed in any way, but needs to be taken into account for reliability analysis. Latent structure models are flexible models commonly used to model unobservable causes of variability.

CELESTE currently collaborates with PSA Group. To dimension its vehicles, the PSA Group uses a reliability design method called Strength-Stress, which takes into consideration both the statistical distribution of part strength and the statistical distribution of customer load (called Stress). In order to minimize the risk of in-service failure, the probability that a severe customer will encounter a weak part must be quantified. Severity quantification is not simple since vehicle use and driver behaviour can be severe for some types of materials and not for others. The aim of the study is then to define a new and richer notion of the severity from the PSA databases resulting either from tests or client usages. This will lead to a more robust and accurate parts dimensioning method.

4.5. Spectroscopic imaging analysis of ancient materials

Ancient materials, encountered in archaeology and paleontology are often complex, heterogeneous and poorly characterized before physico-chemical analysis. A popular technique is to gather as much physico-chemical information as possible, is spectro-microscopy or spectral imaging, where a full spectra, made of more than a thousand samples, is measured for each pixel. The produced data is tensorial with two or three spatial dimensions and one or more spectral dimensions, and requires the combination of an "image" approach with a "curve analysis" approach. Since 2010 CELESTE (previously SELECT) collaborates with Serge Cohen (IPANEMA) on clustering problems, taking spatial constraints into account.

4.6. Forecast of dwell time during train parking at stations

This is a Cifre PhD in collaboration with SNCF.

One of the factors in the punctuality of trains in dense areas (and management crisis in the event of an incident on a line) is the respect of both the travel time between two stations and the parking time in a station. These depend, among other things, on the train, its mission, the schedule, the instantaneous charge, and the configuration of the platform or station. Preliminary internal studies at SNCF have shown that the problem is complex. From a dataset concerning line E of the Transilien in Paris, we will address prediction (machine learning) and modeling (statistics): (1) construct a model of station-hours, station-hours-type of train, by example using co-clustering techniques; (2) study the correlations between the number of passengers (load), up and down flows, and parking times, and possibly other variables to be defined; (3) model the flows or loads (within the same station, or the same train) as a stochastic process; (4) develop a realistic digital simulator of passenger flows and test different scenarios of incidents and resolution, in order to propose effective solutions.

4.7. Fatigue aided-design

This is a Cifre PhD in collaboration with PSA.

The digitalization of design is at the heart of the processes of automotive manufacturers departments, to enable them to reduce costs and development time. This also applies to reliability studies of certain components of the chassis of a vehicle, and the will is to drastically reduce the number of physical tests to tend towards an almost entirely digital design having only one phase of validation. Deterministic models, although developed from detailed design drawings, can predict behaviors different from those observed on the structure during testing. These deviations can be due to the more or less faithful discretization of the geometry, the uncertainties on some parameters of the model (such as the properties of the materials, the boundary conditions), or the random loadings undergone by the structure (Beck and Katafygiotis, 1998). It is important to make available new methods in addition to the classical finite element (FE) deterministic modeling, to enable the exploitation of the accumulated data over the years for all the projects: computation results, measurements and test data.

One of the objectives of this project is to propose a probabilistic modeling of the behavior of a structure starting from a FE model, taking into account the non assignable fluctuations of the model, in order to define a probabilistic criterion of rupture and its margins of confidence. The following three steps are envisaged: (1) Define relevant prior information using business experience feedback (REX) and use a Bayesian estimation to calibrate the parameters. This REX is consequent and will require advanced statistical processing of machine learning, and in particular in clustering to identify similarities or similar patterns among several models. The estimation will use Bayesian non-iterative methods (Celeux and Pamphile, 2019), which are less expensive and less unstable than conventional methods. This will test their effectiveness in this context. (2) Select important parameters (physical or modeling). (3) Define a probabilistic criterion of coaxial fatigue taking into account both the random behavior of the structure and the material (Fouchereau et al., 2014) extending the existing deterministic criteria (Dang-Van, 1993).

5. New Software and Platforms

5.1. BlockCluster

Block Clustering

KEYWORDS: Statistic analysis - Clustering package

SCIENTIFIC DESCRIPTION: Simultaneous clustering of rows and columns, usually designated by biclustering, co-clustering or block clustering, is an important technique in two way data analysis. It consists of estimating a mixture model which takes into account the block clustering problem on both the individual and variables sets. The blockcluster package provides a bridge between the C++ core library and the R statistical computing environment. This package allows to co-cluster binary, contingency, continuous and categorical data-sets. It also provides utility functions to visualize the results. This package may be useful for various applications in fields of Data mining, Information retrieval, Biology, computer vision and many more.

FUNCTIONAL DESCRIPTION: BlockCluster is an R package for co-clustering of binary, contingency and continuous data based on mixture models.

RELEASE FUNCTIONAL DESCRIPTION: Initialization strategy enhanced

- Participants: Christophe Biernacki, Gilles Celeux, Parmeet Bhatia, Serge Iovleff, Vincent Brault and Vincent Kubicki
- Partner: Université de Technologie de Compiègne
- Contact: Serge Iovleff
- URL: http://cran.r-project.org/web/packages/blockcluster/index.html

5.2. MASSICCC

Massive Clustering with Cloud Computing

KEYWORDS: Statistic analysis - Big data - Machine learning - Web Application

SCIENTIFIC DESCRIPTION: The web application let users use several software packages developed by Inria directly in a web browser. Mixmod is a classification library for continuous and categorical data. MixtComp allows for missing data and a larger choice of data types. BlockCluster is a library for co-clustering of data. When using the web application, the user can first upload a data set, then configure a job using one of the libraries mentioned and start the execution of the job on a cluster. The results are then displayed directly in the browser allowing for rapid understanding and interactive visualisation.

FUNCTIONAL DESCRIPTION: The MASSICCC web application offers a simple and dynamic interface for analysing heterogeneous data with a web browser. Various software packages for statistical analysis are available (Mixmod, MixtComp, BlockCluster) which allow for supervised and supervised classification of large data sets.

- Contact: Christophe Biernacki
- URL: https://massiccc.lille.inria.fr

5.3. Mixmod

Many-purpose software for data mining and statistical learning

KEYWORDS: Data mining - Classification - Mixed data - Data modeling - Big data

FUNCTIONAL DESCRIPTION: Mixmod is a free toolbox for data mining and statistical learning designed for large and highdimensional data sets. Mixmod provides reliable estimation algorithms and relevant model selection criteria.

It has been successfully applied to marketing, credit scoring, epidemiology, genomics and reliability among other domains. Its particularity is to propose a model-based approach leading to a lot of methods for classification and clustering.

Mixmod allows to assess the stability of the results with simple and thorough scores. It provides an easy-to-use graphical user interface (mixmodGUI) and functions for the R (Rmixmod) and Matlab (mixmodForMatlab) environments.

- Participants: Benjamin Auder, Christophe Biernacki, Florent Langrognet, Gérard Govaert, Gilles Celeux, Remi Lebret and Serge Iovleff
- Partners: CNRS - Université Lille 1 - LIFL - Laboratoire Paul Painlevé - HEUDIASYC - LMB
- Contact: Gilles Celeux
- URL: http://www.mixmod.org

6. New Results

6.1. Minimal penalty algorithms for model selection

Birgé and Massart proposed in 2001 the slope heuristics as a way to choose optimally from data an unknown multiplicative constant in front of a penalty. It is built upon the notion of minimal penalty, and it has been generalized since to some “minimal-penalty algorithms”. The survey [3] by S. Arlot reviews the theoretical results obtained for such algorithms, with a self-contained proof in the simplest framework, precise proof ideas for further generalizations, and a few new results. Explicit connections are made with residual-variance estimators —with an original contribution on this topic, showing that for this task the slope heuristics performs almost as well as a residual-based estimator with the best model choice— and some classical algorithms such as L-curve or elbow heuristics, Mallows' C_p, and Akaike's FPE. Practical issues are also addressed, including two new practical definitions of minimal-penalty algorithms that are compared on synthetic data to previously-proposed definitions. Finally, several conjectures and open problems are suggested as future research directions. This extensive survey is followed by a discussion by 13 authors, and a rejoinder in which another original result is proved (theoretical validation of the slope heuristics when all models in the collection are biased).

6.2. Kernel change-point detection

In collaboration with A. Celisse and Z. Harchaoui, S. Arlot worked on the change-point problem with data belonging to a general set. They built a penalty for choosing the number of change-points in the kernel-based method of Harchaoui and Cappé (2007). This penalty generalizes the one proposed by Lebarbier (2005) for one-dimensional signals. They prove in [4] a non-asymptotic oracle inequality for the proposed method, thanks to a new concentration result for some function of Hilbert-space valued random variables. Experiments on synthetic and real data illustrate the accuracy of our method, showing that it can detect changes in the whole distribution of data, even when the mean and variance are constant. This method has since been used successfully by several authors in various applied contexts.

6.3. A probabilistic method to characterize genomic alterations of tumors

Characterizing the genomic copy number alterations (CNA) in cancer is of major importance in order to develop personalized medicine. Single nucleotide polymorphism (SNP) arrays are still in use to measure CNA profiles. Among the methods for SNP-array analysis, the Genome Alteration Print (GAP) by Popova et al, based on a preliminary segmentation of SNP-array profiles, uses a deterministic approach to infer the absolute copy numbers profile. C. Keribin with Y. Liu, Y. Rozenholch and T. Popova developed a probabilistic model in [9] for GAP and define a Gaussian mixture model where centers are constrained to belong to a frame depending on unknown parameters such as the proportion of normal tissue. The estimation is performed using an expectation-maximization (EM) algorithm to recover the parameters characterizing the genomic alterations as well as the most probable copy number change of each segment and the unknown proportion of normal tissue. The tumor ploidy can be deduced from penalized model selection criterion. The model is tested on simulated and real data.

Surprisingly, the BIC selection criterion cannot recover the actual ploidy in the real data sets as slope heuristics do, even though all models are wrong. C. Keribin, in a discussion of S. Arlot's survey, gave some arguments to explain these behaviors [8].

6.4. New results for stochastic bandits

M. Brégère and G. Stoltz, in collaboration with P. Gaillard (Sierra team) and Y. Goude (EDF), provided a methodology in [5] based on a modeling by linear bandits, for managing (influencing) electricity consumption by sending tariff incentives. The main result is the very modeling of the problem: consumption is modeled as a generalized additive model based on the probabilistic allocation of tariffs picked and on the context (given by the type of day, hour of the day, weather conditions, etc.). Mathematical results are, on the other hand, direct extensions of earlier results for the LinUCB algorithm (see Li et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). Simulations on realistic data are provided: for bandit algorithms, one needs a data simulator, which we created based on an open data set consisting of households in London.

A second important result was obtained by H. Hadiji: he characterized the cost of adaptation to the unknown (Höderian) smoothness payoff functions in continuum-armed bandits [14]. He first rewrote and slightly extended the regret lower bounds exhibited by Locatelli and Carpentier (2018), and then exhibited an algorithm with matching regret upper bounds. This algorithm, unlike virtually all previous algorithms in X-armed bandits, which zoomed in as time passes, zooms out as time passes. This solves a problem that had been open for several years.

Also, H. Hadiji and G. Stoltz, in collaboration with P. Ménard (SequeL team) and A. Garivier, submitted a revised version of their results of simultaneous optimality (from both a distribution-dependent and a distribution-free viewpoints) for a variant of the KL-UCB algorithm in the case of vanilla K-armed stochastic bandits [22].

6.5. Robust risk minimization for machine learning

In collaboration with S. Minsker (USC), T. Mathieu worked on obtaining new excess risk bounds in robust empirical risk minimization. The method proposed in [29] is inspired from the robust risk minimization procedure using median-of-means estimators in Lecué, Lerasle and Mathieu (2018). The obtained excess risk are faster than the so-called "slow rate of convergence" obtained for the minimization procedure in Lecué, Lerasle and Mathieu (2018) and a slightly modified procedure achieves a minimax rate of convergence under low moment assumptions. Experiments on synthetic corrupted data and real dataset illustrate the accuracy of the method showing high performance in classification and regression tasks in a corrupted setting.

6.6. Optimal pair-matching

The sequential pair-matching problem appears in many applications (in particular for the internet) where one wants to discover, sequentially, good matches between pairs of individuals, for a given budget. C. Giraud, Y. Issartel, L. Lehéricy and M. Lerasle propose a formulation of this problem as a special bandit problem on graphs [23]. Formally, the set of individuals is represented by the nodes of a graph where the edges, unobserved at first, represent the potential good matches. The algorithm queries pairs of nodes and observes the presence/absence of edges. Its goal is to discover as many edges as possible with a fixed budget of queries. Pair-matching is a particular instance of multi-armed bandit problem in which the arms are pairs of individuals and the rewards are edges linking these pairs. This bandit problem is non-standard though, as each arm can only be played once.

Given this last constraint, sublinear regret can be expected only if the graph has some underlying structure. C. Giraud, Y. Issartel, L. Lehéricy and M. Lerasle show in [23] that sublinear regret is achievable in the case where the graph is generated according to a Stochastic Block Model (SBM) with two communities. Optimal regret bounds are computed for this pair-matching problem. They exhibit a phase transition related to the Kesten-Stigund threshold for community detection in SBMs. In practice, it is meaningful to constrain each node to be sampled less than a given amount of times, in order to avoid concentration of queries on a set of individuals. This setting is more challenging both on the statistical side and the algorithmic side. Optimal rates are also derived in this context, exhibiting how the regret deteriorates with this constraint.

6.7. Minimax estimation of network complexity in graphon model

In network analysis, the graphon model has attracted a lot of attention as a non-parametric model with some universal properties. However, this approach suffers from interpretability and identifiability issues in practice. A first solution to this problem was obtained by Y. Issartel: he introduces an identifiable and interpretable functional of the graphon, which measures the complexity of network [25]. It has simple interpretations on popular examples of random graphs: it matches the number of communities in stochastic block models; the dimension of the Euclidean space in random geometric graphs; the regularity of the link function in Hölder graphons. He also provides an estimation procedure of this complexity that is minimax optimal.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. ANR

Sylvain Arlot and Matthieu Lerasle are part of the ANR grant FAST-BIG (Efficient Statistical Testing for high-dimensional Models: application to Brain Imaging and Genetics), which is lead by Bertrand Thirion (Inria Saclay, Parietal).

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events: Organisation

8.1.1.1. Member of the Organizing Committees

Semaine SEME - Orsay 14/01-18/01-2019, https://www.math.u-psud.fr/seme2019/index.php. G. Stoltz was part of the organizing and scientific committees. C. Keribin was part of the scientific committee.

8.1.2. Scientific Events: Selection

8.1.2.1. Member of the Conference Program Committees

- S. Arlot was member of the steering committee of the 4th Junior Conference on Data Science and Engineering at Paris-Saclay (Sept. 2019), CentraleSupélec, Paris-Saclay campus, Gif-sur-Yvette.
- C. Keribin is member of steering committee of the the bi-montly Seminar of probability and statistics (Laboratoire de mathématiques d'Orsay)
- C. Giraud is co-organiser (with E. Kuhn) of the conference StatMathAppli at Fréjus (1-6 september 2019)
- C. Giraud is local member of scientific committee of the Institut Pascal (year around program)

8.1.2.2. Reviewer

We performed many reviews for various international conferences.

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

S. Arlot is associate editor of Annales de l'Institut Henri Poincaré B – Probability and Statistics.

8.1.3.2. Reviewer - Reviewing Activities

We performed many reviews for various international journals.

8.1.4. Invited Talks

- S. Arlot, Changepoint Workshop, Nov. 2019, Institut des Systèmes Complexes, Paris.
- S. Arlot, Thematic week "Data and Analytics for Short-Term Operations", Feb. 2019, Isaac Newton Institute for Mathematical Science, Cambridge, UK.
- C. Keribin, Some asymptotic properties of model selection criteria in the latent block model. 12th Scientific meeting CLADAG 2019, Cassino (Italie), September 11 – 13, 2019
- C. Giraud, Community detection and sequential learning, IOPS 2019, Bordeaux, June 19-21, 2019
- C. Giraud, Sequential learning in random graph, Genova, May 2019.
- P. Pamphile, Maintenance cost forecasting for a feet of vehicles, IMDR, Paris, April 2019

8.1.5. Research Administration

S. Arlot coordinates the math-AI (mathematics for artificial intelligence) program of the Labex Mathématique Hadamard and is member of the executive comittee of Fondation Mathématique Jacques Hadamard (FMJH).
S. Arlot is member of the steering committee of the Paris-Saclay Center for Data Science.
S. Arlot is member of the prefiguration group of the Computer Science Graduate School of University Paris-Saclay.
P. Massart is Director of the Fondation Mathématique Jacques Hadamard (FMJH).
C. Giraud has coordinated the math-SV (mathematics for life science) program of the Labex Mathématique Hadamard and is member of the executive comittee of Fondation Mathématique Jacques Hadamard (FMJH).
C. Giraud is member of the scientific committee of the Labex IRMIA (Strasbourg)

C. Giraud is local member of the scientific committee of the Pascal Institute (Saclay)
C. Giraud is member of the prefiguration group of the Mathematics Graduate School of University Paris-Saclay.
C. Giraud is member of the steering committee of the Mathematics Modelisation and Biodiversity chair.
C. Giraud is in charge of the whole master program in Mathematics of Paris Saclay.
C. Giraud is in charge of the Statistics and Machine Learning track in the master program in Mathematics of Paris Saclay.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Licence: S. Arlot, Probability and Statistics, 68h, L2, Université Paris-Sud

Master: S. Arlot, Statistical learning and resampling, 30h, M2, Université Paris-Sud

Master: S. Arlot, Probability and Statistics M2 seminar, 30h, M2, Université Paris-Sud

Master: S. Arlot, Preparation to French mathematics agrégation (statistics), 50h, M2, Université Paris-Sud

Master: C. Giraud, High-Dimensional Statistics, 45h, M2, Université Paris-Sud

Master: C. Giraud, Theoretical Guidelines in Data Analysis, 45h, M2, Université Paris-Sud

Master: C. Giraud, Lecture Group, 25h, M2, Université Paris-Sud

Master: C. Giraud, Mathematics for AI, 75h, M1, Université Paris-Sud

8.2.2. Supervision

PhD in progress: Guillaume Maillard, Aggregated cross-validation, started Sept. 2016, co-advised by S. Arlot and M. Lerasle

PhD in progress: El Mehdi Saad, Interactions between statistical and computational aspects in machine learning, started Sept. 2019, co-advised by S. Arlot and G. Blanchard

PhD in progress: Tuan-Binh Nguyen, Efficient Statistical Testing for High-Dimensional Models, co-advised by S. Arlot and B. Thirion

PhD in progress: Rémi Coulaud, Forecast of dwell time during train parking at station, started Oct. 2019, co-advised by G. Stoltz and C. Keribin, Cifre with SNCF

PhD in progress: Olivier Coudray, Fatigue aided-design, started Nov. 2019, co-advised by C. Keribin and P. Pamphile, Cifre with PSA

PhD: Solène Thépaut, Problèmes de clustering liés à la synchronie en écologie, Université Paris Saclay, Dec. 2019, C. Giraud.

PhD: Théophile Olivier, Le role de la diversité et des perturbations environnementales sur la stabilité temporelle des communautés animales en milieu naturel, Museum National Histoire Naturelle, Sep. 2019, co-advised by E. Porcher and C. Giraud.

PhD in progress: Yann Issartel, Non-parametric estimation in random networks, started Sep. 2017, C. Giraud.

PhD in progress: Solenne Gaucher, Sequential learning in random networks, started Sep. 2018, C. Giraud.

8.2.3. Juries

S. Arlot: referee for the HdR of Servane Gey, Université Paris Descartes, 07/02/2019.
S. Arlot: member of the PhD committee of Solène Thepaut, Université Paris-Sud, 06/12/2019.
C. Giraud: many HDR and PhD juries as referee or member of the committee

8.3. Popularization

8.3.1. Interventions

- Public exhibitions: S. Arlot is member of the steering committee of a general-audience exhibition about artificial intelligence, that is co-organized by Fermat Science (Toulouse), Institut Henri Poincaré (IHP, Paris) and Maison des Mathématiques et de l'Informatique (MMI, Lyon).

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] C. Keribin. *From clustering to co-clustering : a model based approach*, Université Paris Sud XI, November 2019, Habilitation à diriger des recherches, https://tel.archives-ouvertes.fr/tel-02397429

Articles in International Peer-Reviewed Journals

[2] S. Arlot. *Minimal penalties and the slope heuristics: a survey*, in "Journal de la Socie´te´ Française de Statistique", October 2019, vol. 160, n^{O} 3, pp. 1-106, https://arxiv.org/abs/1901.07277 , https://hal.archives-ouvertes.fr/hal-01989167

[3] S. Arlot. *Rejoinder on: Minimal penalties and the slope heuristics: a survey*, in "Journal de la Socie´te´ Française de Statistique", October 2019, vol. 160, n^{O} 3, pp. 158-168, https://arxiv.org/abs/1909.13499 , https://hal.archives-ouvertes.fr/hal-02300688

[4] S. Arlot, A. Celisse, Z. Harchaoui. *A Kernel Multiple Change-point Algorithm via Model Selection*, in "Journal of Machine Learning Research", December 2019, vol. 20, n^{O} 162, pp. 1–56, https://arxiv.org/abs/1202.3878 , https://hal.archives-ouvertes.fr/hal-00671174

[5] M. Brégère, P. Gaillard, Y. Goude, G. Stoltz. *Target Tracking for Contextual Bandits: Application to Demand Side Management*, in "Proceedings of Machine Learning Research", June 2019, vol. 97, pp. 754-763, https://arxiv.org/abs/1901.09532 , https://hal.archives-ouvertes.fr/hal-01994144

[6] G. Chinot, G. Lecué, M. Lerasle. *Robust Statistical learning with Lipschitz and convex loss functions*, in "Probability Theory and Related Fields", July 2019, https://arxiv.org/abs/1810.01090 [*DOI :* 10.1007_s00440-019-00931-3], https://hal.archives-ouvertes.fr/hal-01923033

[7] A. Havet, M. Lerasle, É. Moulines. *Density estimation for RWRE*, in "Mathematical Methods of Statistics", March 2019, https://arxiv.org/abs/1806.05839 [*DOI :* 10.3103/S1066530719010022], https://hal.archives-ouvertes.fr/hal-01815990

[8] C. Keribin. *A note on BIC and the slope heuristic*, in "Journal de la Socie´te´ Française de Statistique", 2019, https://hal.archives-ouvertes.fr/hal-02391310

[9] C. Keribin, Y. Liu, T. Popova, Y. Rozenholc. *A mixture model to characterize genomic alterations of tumors*, in "Journal de la Socie´te´ Française de Statistique", 2019, https://hal.archives-ouvertes.fr/hal-02391289

[10] G. LECUÉ, M. LERASLE. *Robust machine learning by median-of-means : theory and practice*, in "Annals of Statistics", February 2019, https://arxiv.org/abs/1711.10306 - 48 pages, 6 figures, https://hal.archives-ouvertes.fr/hal-01923036

Invited Conferences

[11] C. BIERNACKI, G. CELEUX, J. JOSSE, F. LAPORTE. *Dealing with missing data in model-based clustering through a MNAR model*, in "CRoNos & MDA 2019 - Meeting and Workshop on Multivariate Data Analysis and Software", Limassol, Cyprus, April 2019, https://hal.inria.fr/hal-02103347

International Conferences with Proceedings

[12] M. LERASLE, Z. SZABÓ, T. MATHIEU, G. LECUÉ. *MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, Proceedings of Machine Learning Research, June 2019, https://hal.archives-ouvertes.fr/hal-01705881

Conferences without Proceedings

[13] E. CHZHEN, C. DENIS, M. HEBIRI, L. ONETO, M. PONTIL. *Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification*, in "NeurIPS", Vancouver, Canada, December 2019, https://arxiv.org/abs/1906.05082 , https://hal-upec-upem.archives-ouvertes.fr/hal-02150662

[14] H. HADIJI. *Polynomial Cost of Adaptation for X -Armed Bandits*, in "Thirty-third Conference on Neural Information Processing Systems", Vancouver, France, December 2019, https://arxiv.org/abs/1905.10221 , https://hal.archives-ouvertes.fr/hal-02138492

[15] C. KERIBIN, C. BIERNACKI. *Co-clustering: model based or model free approaches*, in "62nd ISI World Statistics Congress 2019", Kuala Lumpur, Malaysia, August 2019, https://hal.archives-ouvertes.fr/hal-02399031

[16] C. KERIBIN, C. BIERNACKI. *Le modèle des blocs latents, une méthode régularisée pour la classification en grande dimension*, in "JdS 2019 - 51èmes Journées de Statistique de la SFdS", Nancy, France, June 2019, https://hal.archives-ouvertes.fr/hal-02391379

[17] C. KERIBIN. *Some Asymptotic Properties of Model Selection Criteria in the Matent Block Model*, in "CLADAG 2019 - 12th Scientific Meeting Classification and Data Analysis Group", Cassino, Italy, September 2019, https://hal.archives-ouvertes.fr/hal-02391398

[18] F. LAPORTE, C. BIERNACKI, G. CELEUX, J. JOSSE. *Modèles de classification non supervisée avec données manquantes non au hasard*, in "51e journées de statistique", Nancy, France, June 2019, https://hal.archives-ouvertes.fr/hal-02398984

Other Publications

[19] G. CELEUX, P. PAMPHILE. *Estimating parameters of the Weibull Competing Risk model with Masked Causes and Heavily Censored Data*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02410489

[20] G. CHINOT, G. LECUÉ, M. LERASLE. *Robust high dimensional learning for Lipschitz and convex losses*, June 2019, https://arxiv.org/abs/1905.04281 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02159943

[21] F. Ducros, P. Pamphile. *Maintenance cost forecasting for a fleet of vehicles*, February 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02055175

[22] A. Garivier, H. Hadiji, P. Menard, G. Stoltz. *KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01785705

[23] C. Giraud, Y. Issartel, L. Lehéricy, M. Lerasle. *Pair Matching: When bandits meet stochastic block model*, June 2019, https://arxiv.org/abs/1905.07342 - 57 pages, https://hal.archives-ouvertes.fr/hal-02159938

[24] A. Havet, M. Lerasle, É. Moulines, E. Vernet. *A quantitative Mc Diarmid's inequality for geometrically ergodic Markov chains*, July 2019, https://arxiv.org/abs/1907.02809 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02177452

[25] Y. Issartel. *On the Estimation of Network Complexity: Dimension of Graphons*, December 2019, https://arxiv.org/abs/1909.02900 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02413537

[26] A. Janon. *Global optimization using Sobol indices*, June 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02154121

[27] F. Laporte, C. Biernacki, G. Celeux, J. Josse. *Model-based clustering with missing not at random data. Missing mechanism*, July 2019, Working Group on Model-Based Clustering Summer Session, Poster, https://hal.archives-ouvertes.fr/hal-02398987

[28] G. Maillard, S. Arlot, M. Lerasle. *Aggregated Hold-Out*, September 2019, https://arxiv.org/abs/1909.04890 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02273193

[29] S. Minsker, T. Mathieu. *Excess risk bounds in robust empirical risk minimization*, December 2019, https://arxiv.org/abs/1910.07485 - working paper or preprint [*DOI :* 10.07485], https://hal.archives-ouvertes.fr/hal-02390397

[30] V. Robert, Y. Vasseur, V. Brault. *Comparing high dimensional partitions with the Coclustering Adjusted Rand Index*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01524832