RESEARCH CENTRE

Saclay - Île-de-France

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2020 ACTIVITY REPORT

Project-Team CELESTE

mathematical statistics and learning

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

DOMAIN

Applied Mathematics, Computation and Simulation

THEME

Optimization, machine learning and statistical methods

Contents

Project-Team CELESTE

Creation of the Project-Team: 2019 June 01

Keywords

Computer sciences and digital sciences

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

B9.9. – Ethics

1 Team members, visitors, external collaborators

Research Scientists

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

Faculty Members

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

Post-Doctoral Fellow

- Evgenii Chzhen [Univ Paris-Saclay]

PhD Students

- Yvenn Amara-Ouali [Univ Paris-Saclay]
- Emilien Baroux [Groupe PSA, from Jul 2020]
- Margaux Bregere [EDF, until Oct 2020]
- Geoffrey Chinot [ENSAE, until Aug 2020]
- Olivier Coudray [Groupe PSA]
- Remi Coulaud [SNCF, CIFRE]
- Solenne Gaucher [ENSAE]
- Hedi Hadiji [Ministère de l'Enseignement Supérieur et de la Recherche]
- Karl Hajjar [Univ Paris-Saclay, from Oct 2020]
- Malo Huard [Univ Paris-Saclay]
- Yann Issartel [Univ Paris-Saclay]
- Perrine Lacroix [Univ Paris-Saclay]
- Guillaume Maillard [Univ Paris-Saclay]
- Timothee Mathieu [École Normale Supérieure de Cachan]
- El Mehdi Saad [Univ Paris-Saclay]

Technical Staff

- Benjamin Auder [CNRS, Engineer]

Interns and Apprentices

- Cecile Poulain [Inria, from Mar 2020 until Aug 2020]

Administrative Assistant

- Laurence Fontana [Inria, from Oct 2020]

External Collaborators

- Claire Lacour [Univ Paris-Est Marne La Vallée]
- Matthieu Lerasle [CNRS, HDR]

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—Covid-19, 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 Covid-19

We collaborate with researchers at the Pasteur Institute and the University Hospital of Guadeloupe on the development of a rapid test for Covid-19 severity prediction as well as risk modeling and outcome prediction for patients admitted to ICU units.

4.3 Electricity load consumption: forecasting and control

CELESTE has a long-term collaboration with EDF R&D on 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 thus to define a new and richer notion of “severity” from PSA databases, resulting either from tests or client usages. This will lead to more robust and accurate parts dimensioning methods. Two CIFRE theses are in progress on such subjects:

Olivier COUDRAY, “Fatigue Data-based Design: Probabilistic Modeling of Fatigue Behavior and Analysis of Fatigue Data to Assist in the Numerical Design of a Mechanical Part”. Here, we are seeking to build probabilistic fatigue criteria to identify the critical zones of a mechanical part.

Emilien BAROUX, “Reliability dimensioning under complex loads: from specification to validation”. Here, we seek to identify and model the critical loads that a vehicle can undergo according to its usage profile (driver, roads, climate, etc.).

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 the SNCF.

One of the factors in the punctuality of trains in dense areas (and management crises 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 the SNCF have shown that the problem is complex. From a dataset concerning the line E of the Transilien in Paris, we aim to 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 Algorithmic fairness

Machine learning algorithms make pivotal decisions, which influence our lives on a daily basis, using data about individuals. Recent studies show that imprudent use of these algorithms leads to unfair and discriminating decisions, often inheriting or even amplifying disparities present in data. The goal of this research program is to design and analyze novel tractable algorithms that, while still optimizing prediction performance, mitigate or remove unfair decisions of the learned predictor. A major challenge in the machine learning fairness literature is to obtain algorithms which satisfy fairness and risk guarantees simultaneously. Several empirical studies suggest that there is a trade-off between fairness and accuracy of a learned model – more accurate models are less fair. A theoretical study of these types of trade-offs is among the main directions of this research project. The goal is to provide user-friendly statistical quantification of these trade-offs and build statistically optimal algorithms in this context.

5 Social and environmental responsibility

5.1 Footprint of research activities

Influenced in particular by the Covid-19 pandemic in 2020, the carbon emissions of Celeste team members related to their jobs were very low and came essentially from:

- limited levels of transport to and from work, and not from travel to conferences.
- electronic communication (email, Google searches, Zoom meetings, online seminars, etc.).
- the carbon emissions embedded in their personal computing devices (construction), either laptops or desktops.
- electricity for personal computing devices and for the workplace, plus also water, heating, and maintenance for the latter. Note that only 7.1% (2018) of France's electricity is not sourced from nuclear energy or renewables so team member carbon emissions related to electricity are minimal.

In terms of magnitude, the largest per capita ongoing emissions (excluding flying) are likely simply to be those from buying computers that have a carbon footprint from their construction, in the range of 100 kg Co2-e each. In contrast, typical email use per year is around 10 kg Co2-e per person, and a Zoom call comes to around 10g Co2-e per hour per person, while web browsing uses around 100g Co2-e per hour. Consequently, 2020 was a very low carbon year for the Celeste team. To put this in the context of work travel by flying, one return Paris-Nice flight corresponds to 160 kg Co2-e emissions, which likely dwarfs the total emissions of any one Celeste team member's work-related emissions in 2020.

The approximate (rounded for simplicity) Co2-e values cited above come from the book, "How Bad are Bananas" by Mike Berners-Lee (2020) which estimates carbon emissions in everyday life.

5.2 Impact of research results

In addition to the long-term impact of our theoretical works—which is of course impossible to assess immediately—we are involved in several applied research projects which aim at having a short/mid-term positive impact on society.

First, we collaborate with the Pasteur Institute and the University Hospital of Guadeloupe on medical issues related to some neglected tropical diseases and to Covid-19.

Second, the broad use of artificial intelligence/machine learning/statistics nowadays comes with several major ethical issues, one being to avoid making unfair or discriminatory decisions. Our theoretical work on algorithmic fairness has already led to several "fair" algorithms that could be widely used in the short term (one of them is already used for enforcing fair decision-making in student admissions at the University of Genoa).

Third, we expect short-term positive impact on society thanks to several direct collaborations with companies such as EDF (forecasting and control of electricity load consumption), SNCF (punctuality of trains in densely-populated regions, 1 Cifre contract ongoing) and the PSA group (reliability, with 2 Cifre contracts ongoing).

6 Highlights of the year

6.1 Awards

S. Arlot is junior member of Institut Universitaire de France (IUF) since September 2020.

The paper [2], first-authored by E. Chzhen, has been selected for an oral presentation at NeurIPS 2020 (1.1% of submitted works accepted).

7 New software and platforms

7.1 New software

7.1.1 BlockCluster

Name: 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 Contributions: Initialization strategy enhanced

URL: http://cran.r-project.org/web/packages/blockcluster/index.html

Authors: Parmeet Bhatia, Serge Iovleff, Vincent Brault

Contacts: Christophe Biernacki, Gilles Celeux, Serge Iovleff

Participants: Christophe Biernacki, Gilles Celeux, Parmeet Bhatia, Serge Iovleff, Vincent Brault, Vincent Kubicki

Partner: Université de Technologie de Compiègne

7.1.2 MASSICCC

Name: 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.

URL: https://massiccc.lille.inria.fr

Contact: Christophe Biernacki

7.1.3 Mixmod

Name: 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.

URL: http://www.mixmod.org

Authors: Christophe Biernacki, Florent Langrognet, Gérard Govaert, Gilles Celeux

Contacts: Christophe Biernacki, Gilles Celeux

Participants: Benjamin Auder, Christophe Biernacki, Florent Langrognet, Gérard Govaert, Gilles Celeux, Remi Lebret, Serge Iovleff

Partners: CNRS, Université Lille 1, LIFL, Laboratoire Paul Painlevé, HEUDIASYC, LMB

8 New results

8.1 Aggregated Hold-Out

Aggregated hold-out (Agghoo) is a method which averages learning rules selected by hold-out (that is, cross-validation with a single split).

G. Maillard, S. Arlot and M. Lerasle provided in [11] the first theoretical guarantees on Agghoo, ensuring that it can be used safely: Agghoo performs at worst like hold-out when the risk is convex. The same holds true in classification with the 0-1 risk, with an additional constant factor. For hold-out, oracle

inequalities are known for bounded losses, as in binary classification. They show that similar results can be proved, under appropriate assumptions, for other risk-minimization problems. In particular, an oracle inequality holds true for regularized kernel regression with a Lipschitz loss, without requiring that the Y variable or the regressors be bounded. Numerical experiments show that aggregation brings a significant improvement over hold-out and that Agghoo is competitive with cross-validation.

In another paper [33], G. Maillard studied aggregated hold out for sparse linear regression with a robust loss function. Sparse linear regression methods generally have a free hyperparameter which controls the amount of sparsity, and is subject to a bias-variance tradeoff. This article considers the use of aggregated hold-out to aggregate over values of this hyperparameter, in the context of linear regression with the Huber loss function. Aggregated hold-out (Agghoo) is a procedure which averages estimators selected by hold-out (cross-validation with a single split). In the theoretical part of the article, it is proved that Agghoo satisfies a non-asymptotic oracle inequality when it is applied to sparse estimators which are parametrized by their zero-norm. In particular, this includes a variant of the Lasso introduced by Zou, Hastie and Tibshirani. Simulations are used to compare Agghoo with cross-validation. They show that Agghoo performs better than CV when the intrinsic dimension is high and when there are confounders correlated with the predictive covariates.

In his Ph.D. thesis [4], G. Maillard obtained more precise results in a specific setting, showing that Agghoo then strictly improves the performance of *any* model selection procedure. This is a remarkable result, which is to the best of our knowledge the first result of that kind. It required the use of several advanced mathematical results to be proved.

8.2 Online Orthogonal Matching Pursuit

Greedy algorithms for feature selection are widely used for recovering sparse high-dimensional vectors in linear models. In classical procedures, the main emphasis is put on the sample complexity, with little or no consideration of the computation resources required. E.M. Saad and S. Arlot, in collaboration with G. Blanchard proposed in [34] a novel online algorithm, called Online Orthogonal Matching Pursuit (OOMP), for online support recovery in the random design setting of sparse linear regression. Our procedure selects features sequentially, alternating between allocation of samples only as needed to candidate features, and optimization over the selected set of variables to estimate the regression coefficients. Theoretical guarantees about the output of this algorithm are proven and its computational complexity is analysed.

8.3 Aggregation of Multiple Knockoffs

T.-B. Nguyen and S. Arlot, in collaboration with J.-A. Chevalier and B. Thirion, developped an extension of the knockoff inference procedure, introduced by Barber and Candès [2015]. This new method, called aggregation of multiple knockoffs (AKO), addresses the instability inherent to the random nature of knockoff-based inference. Specifically, AKO improves both the stability and power compared with the original knockoff algorithm while still maintaining guarantees for false discovery rate control. They provided in [13] a new inference procedure, prove its core properties, and demonstrate its benefits in a set of experiments on synthetic and real datasets.

8.4 New results for stochastic bandits

G. Stoltz and H. Hadiji (see [30]) studied adaptation to the range for stochastic bandit problems with finitely many arms, each associated with a distribution supported on a given finite range $[m, M]$. They do not assume that the range $[m, M]$ is known, and show that there is a cost for learning this range. Indeed, a new trade-off between distribution-dependent and distribution-free regret bounds arises, which prevents one from simultaneously achieving the typical $\ln T$ and $\sqrt{T}$ bounds. For instance, a $\sqrt{T}$ distribution-free regret bound may only be achieved if the distribution-dependent regret bounds are at least of order $\sqrt{T}$. We exhibit a strategy for achieving the rates for regret indicated by the new trade-off.

8.5 Finite continuum-armed bandits

The finite continuum-armed bandit problem arises in many applications where an agent must allocate a finite budget T between a larger number of N actions described by covariates, and each action can only be taken once. Focusing on a nonparametric setting, where the mean reward is an unknown function of a one-dimensional covariate, [28] propose an optimal strategy for this problem. Under natural assumptions on the reward function, the optimal regret scales as $O(T^{1/3})$ up to poly-logarithmic factors when the budget T is proportional to the number of actions N. When T becomes small compared to N, a smooth transition occurs. When the ratio T/N decreases from a constant to $N^{-1/3}$, the regret increases progressively up to the $O(T^{1/2})$ rate encountered in classical continuum-armed bandits.

8.6 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 their paper [36] 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 a real dataset illustrate the accuracy of the method, showing high performance in classification and regression tasks in a corrupted setting.

8.7 Fairness: statistical guarantees and efficient methods

Until very recently results on algorithmic fairness were almost exclusively focused on classification problems. Yet, in a lot of application domains, continuous outputs are more valuable even if the underlying problem is that of classification (*e.g.,* credit scoring). In collaboration with C. Denis, M. Hebiri (Univ. Gustave Eiffel), L. Oneto (Univ. Genoa), M. Pontil (Istituto Italiano di Tecnologia, Univ. College London), E. Chzhen proposed a post-processing regression method which enjoys risk and fairness finite sample guarantees in [18]. Their approach is based on a carefully chosen discretization of the signal space, essentially reducing the problem of regression to a problem of multi-class classification. Later, in [19] a connection between the problem of finding the optimal fair regression (in the sense of Demographic Parity) and the Wasserstein barycenter problem is derived. This connection allows us to build a data-driven post-processing method, which avoids the discretization step using the theory of optimal transport. This algorithm enjoys distribution-free fairness guarantees. Under additional assumptions, risk guarantees are also derived. A statistical minimax framework is proposed by E. Chzhen and N. Schreuder (CREST, ENSAE) in [27]. This framework is built upon the earlier established connection of fair regression and the optimal transport theory, and allows us to study partially fair predictions. Within the proposed setup, Chzhen and Schreduer quantify the trade-off between Demographic Parity fairness and squared risk by obtaining a characterization of the Pareto frontier. Finally, they derive a general-problem dependent lower bound on the risk of any partially fair prediction and confirm its tightness on a Gaussian regression model with systematic group-dependent bias.

8.8 Should the clustering of graphs be bipartite?

When clustering the nodes of a graph, a unique partition of the nodes is usually built, whether the graph is undirected or directed. While this choice is pertinent for undirected graphs, it is debatable for directed graphs because it implies that no difference is made between the clusters of source and target nodes. Defining two different clusterings for source and target nodes leads to considering a kind of bipartite clustering. We examine this question in the context of probabilistic models with latent variables, and compare the use of the stochastic block model (SBM) and the latent block model (LBM). We analyze and discuss this comparison through simulated and real data sets and provide recommendation [32].

8.9 Stastical analyses of standardized micropatterned cells

Live imaging of lysosomal secretion monitored by total internal reflection fluorescence imaging of VAMP7-pHluorin is a straightforward way to explore secretion from this compartment. Taking advantage of cell culture on micropatterned surfaces to normalize cell shape, we employed a variety of statistical tools to perform a spatial analysis of secretory patterns. Using Ripley's K function and a statistical test based on nearest neighbor distance (NND) we confirmed that secretion from lysosomes is not a random process but shows significant clustering [9].

8.10 Comparison of dengue case classification schemes and evaluation of biological changes in different dengue clinical patterns

The World Health Organization (WHO) proposed guidelines on dengue clinical classification in 1997 and more recently in 2009 for the clinical management of patients. The WHO 1997 classification defines three categories of dengue infection according to severity: dengue fever (DF), dengue hemorrhagic fever (DHF), and dengue shock syndrome (DSS). Alternative WHO 2009 guidelines provide a cross-sectional classification aiming to discriminate dengue fever from dengue with warning signs (DWWS) and severe dengue (SD). In this study we performed a comparison of the two dengue classifications both from a biological and statistical point of view [7].

8.11 Consistency and asymptotic normality of Latent Block Model estimators

The Latent Block Model (LBM) is a model-based method to cluster simultaneously the d columns and n rows of a data matrix. Parameter estimation in LBM is a difficult and multifaceted problem. Although various estimation strategies have been proposed and are now well understood empirically, theoretical guarantees about their asymptotic behavior is rather sparse and most results are limited to the binary setting. We have proved in [6] theoretical guarantees in the valued settings. We show that under some mild conditions on the parameter space, and in an asymptotic regime where $log(d)/n$ and $log(n)/d$ tend to 0 when n and d tend to infinity, (1) the maximum-likelihood estimate of the complete model (with known labels) is consistent and (2) the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models. This equivalence allows us to transfer the asymptotic consistency, and under mild conditions, asymptotic normality, to the maximum likelihood estimate under the observed model. Moreover, the variational estimator is also consistent and, under the same conditions, asymptotically normal.

8.12 A quantitative McDiarmid's inequality for geometrically ergodic Markov chains

We state and prove in [8] a quantitative version of the bounded difference inequality for geometrically ergodic Markov chains. Our proof uses the same martingale decomposition as in an earlier result but compared to this paper the exact coupling argument is modified to fill a gap between the strongly aperiodic case and the general aperiodic case.

8.13 Robust machine learning by median-of-means: theory and practice

We introduce in [10] new estimators for robust machine learning based on median-of-means (MOM) estimators of the mean of real valued random variables. These estimators achieve optimal rates of convergence under minimal assumptions on the dataset. The dataset may also have been corrupted by outliers on which no assumption is granted. We also analyze these new estimators with standard tools from robust statistics. In particular, we revisit the concept of breakdown point. We modify the original definition by studying the number of outliers that a dataset can contain without deteriorating the estimation properties of a given estimator. This new notion of breakdown number, that takes into account the statistical performances of the estimators, is non-asymptotic in nature and adapted for machine learning purposes. We proved that the breakdown number of our estimator is of the order of (number of observations)*(rate of convergence). For instance, the breakdown number of our estimators for the problem of estimation of a d-dimensional vector with a noise variance σ^2 is $\sigma^2 d$ and it becomes

$\sigma^2 s \log(d/s)$ when this vector has only s non-zero component. Beyond this breakdown point, we proved that the rate of convergence achieved by our estimator is (number of outliers) divided by (number of observation). Besides these theoretical guarantees, the major improvement brought by these new estimators is that they are easily computable in practice. In fact, basically any algorithm used to approximate the standard Empirical Risk Minimizer (or its regularized versions) has a robust version approximating our estimators. As a proof of concept, we study many algorithms for the classical LASSO estimator. A byproduct of the MOM algorithms is a measure of depth of data that can be used to detect outliers.

8.14 A binned technique for scalable model-based clustering on huge datasets

Clustering is impacted by the regular increase of sample sizes which provides opportunity to reveal information previously out of scope. However, the volume of data leads to some issues related to the need of many computational resources and also to high energy consumption. Resorting to binned data depending on an adaptive grid is expected to give proper answer to such green computing issues while not harming the quality of the related estimation. After a brief review of existing methods, a first application in the context of univariate model-based clustering is provided in [12], with a numerical illustration of its advantages. Finally, an initial formalization of the multivariate extension is done, highlighting both issues and possible strategies.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

- G. Stoltz: New contract with BNP Paribas (10 kE), on stochastic bandits under budget constraints, for an application to loan management. New contract with EDF R&D on studying the Covid-19 impact on electricity demand (with Solenne Gaucher as research engineer).

- C. KERIBIN and P. PAMPHILE. OpenLabIA Inria-Groupe PSA collaboration contract. 85 KE.

- A. CONSTANTINESCU and P. PAMPHILE. Collaboration contract with Groupe PSA. 95 KE.

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria associate team not involved in an IIL

C. Keribin collaborates with Christophe Biernacki (INRIA-Modal) on unsupervised learning of huge datasets with limited computer resources. A co-advised thesis (DGA grant) is ongoing.

10.2 National initiatives

10.2.1 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).

Sylvain Arlot and Christophe Giraud are part of the ANR Chair-IA grant Biscotte, which is led by Gilles Blanchard (Université Paris Saclay).

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees C. Giraud: Co-organizer with Estelle Kuhn of the conference "StatMathAppli", to occur in August 2021.

11.1.2 Scientific events: selection

Member of the conference program committees

- S. Arlot: aera chair for AISTATS 2021
- C. Keribin: Scientific VP for Federated learning workshops (with SFdS and Owkin)
- C. Giraud: in charge of the "High-dimensional statistics" session of the Bernoulli-IMS symposium, August 2020
- C. Giraud: in charge of a special session "Statistical learning" in the AMS-SMF-EMS joint international meeting, July 2021.
- C. Giraud: program committee COLT, July 2021

Reviewer We performed many reviews for various international conferences.

11.1.3 Journal

Member of the editorial boards

- S. Arlot: associate editor for *Annales de l'Institut Henri Poincaré B – Probability and Statistics*
- G. Stoltz: associate editor for *Mathematics of Operations Research*

Reviewer - reviewing activities We performed many reviews for various international journals.

11.1.4 Invited talks

S. Arlot, Statistics seminar of LPSM, Paris, 01/12/2020.
C. Keribin, Statistics seminar AgroParisTech, Paris, 18/05/2020
C. Keribin, Statistics seminar INRAE-MaIAGE, Jouy en Josas, 14/12/2020
C. Keribin, ERCIM-CMStatistics, online, 20/12/2020
E. Chzhen, Le Seminaire Palaisien, online, 06/10/2020
E. Chzhen, Statistics seminar, AgroParisTech, Paris, 02/03/2020
E. Chzhen, Stat·Eco·ML Seminar at ENSAE Paris, Palaiseau, 05/02/2020

11.1.5 Leadership within the scientific community

C. Keribin is President of the MALIA (Machine Learning and IA) group of the French Statistical Society (SFdS).

11.1.6 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 (temporary) board of the Computer Science Graduate School of University Paris-Saclay.
- S. Arlot is member of the (temporary) board of the Mathematics Graduate School of University Paris-Saclay.
- S. Arlot is member of the board of the Computer Science Doctoral School (ED STIC) of University Paris-Saclay.
- 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 steering committee of the Mathematics Graduate School of University Paris-Saclay.
- C. Giraud is in charge of the whole master program in Mathematics of Paris Saclay.
- C. Keribin is elected member of the steering committee of Labex LMH and FMJH foundation
- C. Keribin is elected member of CAC Paris-Saclay
- C. Keribin is member of the jury for awarding the Paris-Saclay Idex and the FMJH Sophie Germain scholarships
- C. Keribin is in charge of Master 1 Applied Mathematics and Master 2 Datascience of Paris-Saclay
- P. Massart is Director of the Fondation Mathématique Jacques Hadamard (FMJH).

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

Most of the team members (especially Professors, Associate Professors and Ph.D. students) teach several courses at University Paris-Saclay, as part of their teaching duty. We mention below some of the classes in which we teach.

- 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 Probability and Statistics, 45h, M2, Université Paris-Saclay
- Master: C. Giraud, Mathematics for AI, 75h, M1, Université Paris-Saclay
- Master: C. Keribin, unsupervised and supervised learning, M1, 42h, Université Paris-Saclay

11.2.2 Supervision

- PhD defended on December 4, 2020: Hédi Hadiji, Sur quelques questions d'adaptation dans des problèmes de bandits stochastiques, started September 2017, co-advised by G. Stoltz and P. Massart
- PhD defended on December 10, 2020: Margaux Brégère, Algorithmes de bandits stochastiques pour la gestion de la demande électrique, started October 2017, manuscript: [21], co-advised by G. Stoltz, P. Gaillard and Y. Goude
- PhD defended on November 24, 2020: Yann Issartel, Inférence sur des graphes aléatoires, started Sep. 2017, advised by C. Giraud.
- PhD defended on September 23, 2020: Malo Huard, Apprentissage et prévision séquentiels : bornes uniformes pour le regret linéaire et séries temporelles hiérarchiques, started October 2016, advised by G. Stoltz.
- PhD defended on September 20, 2020: 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, started Oct. 2018, 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 data-based design, started Nov. 2019, co-advised by C. Keribin and P. Pamphile, Cifre with Groupe PSA
- PhD in progress: Filippo Antonnazo, Unsupervised learning of huge datasets with limited computer resources, started Nov. 2019, co-advised by C. Biernacki (INRIA-Modal) and C. Keribin, DGA grant
- PhD in progress: Solenne Gaucher, Sequential learning in random networks, started Sep. 2018, C. Giraud.
- PhD in progress: Karl Hajjar, analyse dynamique de réseaux de neurones, started Oct. 2020, C. Giraud and L. Chizat.
- PhD in progress: Emilien Baroux, Reliability dimensioning under complex loads: from specification to validation, started July. 2020, co-advised by A. Constantinescu and P. Pamphile , CIFRE with Groupe PSA

11.2.3 Juries

- S. Arlot: referee for the HdR of Erwan Scornet, Université Paris-Saclay, 17/12/2020.
- S. Arlot: member of the HdR committee of Victor-Emmanuel Brunel, Institut Polytechnique de Paris, 15/09/2020.
- S. Arlot: member of the HdR committee of Guillem Rigaill, Université Paris-Saclay, 18/09/2020.
- S. Arlot: member of the PhD committee of Baptiste Barreau, Université Paris-Saclay, 15/09/2020.
- S. Arlot: member of the PhD committee of Gautier Appert, Institut Polytechnique de Paris, 29/10/2020.
- C. Giraud: many HDR and PhD juries as referee or member of the committee
- G. Stoltz: reviewer of the PhD manuscripts by Vincent Margot (Sorbonne Université, October 2020) and Julien Seznec (Inria Lille, December 2020), and of the HDR manuscript by Emilie Kaufmann (Inria Lille, November 2020)

- C. Keribin: reviewer of the PhD manuscript by Léa Longepierre (Sorbonne University, July 2020)
- C. Keribin: member of the PhD committee of Eva Lawrence (Université Paul Sabatier, December 2020)

11.3 Popularization

11.3.1 Interventions

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

12 Scientific production

12.1 Major publications

[1] G. Celeux and P. Pamphile. 'Estimating parameters of the Weibull Competing Risk model with Masked Causes and Heavily Censored Data'. working paper or preprint. Oct. 2020. URL: https://hal.inria.fr/hal-02410489.

[2] E. Chzhen, C. Denis, M. Hebiri, L. Oneto and M. Pontil. 'Fair Regression with Wasserstein Barycenters'. In: *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*. Vancouver / Virtuel, Canada, Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02866811.

[3] H. Hadiji and G. Stoltz. 'Adaptation to the Range in K-Armed Bandits'. working paper or preprint. Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02794382.

[4] G. Maillard. 'Hold-out and Aggregated hold-out'. Theses. Université Paris-Saclay, Sept. 2020. URL: https://tel.archives-ouvertes.fr/tel-02971403.

[5] P. Pamphile and F. Ducros. 'Maintien en conditions opérationnelles d'une flotte de véhicules : estimation du besoin en pièce de rechange'. In: *E-congrès 2020 Lambda λμ22 - 22e Congrès de Maîtrise des Risques et Sûreté de Fonctionnement λμ22*. Le Havre / Virtual, France, Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-03080355.

12.2 Publications of the year

International journals

[6] V. Brault, C. Keribin and M. Mariadassou. 'Consistency and Asymptotic Normality of Latent Blocks Model Estimators'. In: *Electronic journal of statistics* 14.1 (24th Mar. 2020), pp. 1234–1268. DOI: 10.1214/20-EJS1695. URL: https://hal.archives-ouvertes.fr/hal-01511960.

[7] P. Dussart, V. Duong, K. Bleakley, C. Fortas, P. L. Try, K. S. Kim, R. Choeung, S. In, A.-C. Andries, T. Cantaert, M. Flamand, P. Buchy and A. Sakuntabhai. 'Comparison of dengue case classification schemes and evaluation of biological changes in different dengue clinical patterns in a longitudinal follow-up of hospitalized children in Cambodia'. In: *PLoS Neglected Tropical Diseases* 14.9 (2020), e0008603. DOI: 10.1371/journal.pntd.0008603. URL: https://hal.archives-ouvertes.fr/hal-03033376.

[8] A. Havet, M. Lerasle, É. Moulines and E. Vernet. 'A quantitative Mc Diarmid's inequality for geometrically ergodic Markov chains'. In: *Electronic Communications in Probability* (10th Feb. 2020). URL: https://hal.archives-ouvertes.fr/hal-02177452.

[9] H. Lachuer, P. Mathur, K. Bleakley and K. Schauer. 'Quantifying Spatiotemporal Parameters of Cellular Exocytosis in Micropatterned Cells'. In: *Journal of visualized experiments : JoVE* 163 (2020). DOI: 10.3791/60801. URL: https://hal.archives-ouvertes.fr/hal-03033396.

[10] G. Lecué and M. Lerasle. 'Robust machine learning by median-of-means : theory and practice'. In: *Annals of Statistics* (26th May 2020). DOI: 10.1214/19-AOS1828. URL: https://hal.archives-ouvertes.fr/hal-01923036.

[11] G. Maillard, S. Arlot and M. Lerasle. 'Cross-validation improved by aggregation: Agghoo'. In: *Journal of Machine Learning Research* 22.20 (Feb. 2021), pp. 1–55. URL: https://hal.archives-ouvertes.fr/hal-03094497.

International peer-reviewed conferences

[12] F. Antonazzo, C. Biernacki and C. Keribin. 'A binned technique for scalable model-based clustering on huge datasets'. In: MBC2 - Models and Learning for Clustering and Classification. Journal ADAC - Advances in Data Analysis and Classification, Catania, Italy, 2nd Sept. 2020. URL: https://hal.archives-ouvertes.fr/hal-03097284.

[13] T.-B. Nguyen, J.-A. Chevalier, B. Thirion and S. Arlot. 'Aggregation of Multiple Knockoffs'. In: *Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2020*. ICML 2020 - 37th International Conference on Machine Learning. Proceedings of the ICML 37th International Conference on Machine Learning, 119. Vienne / Virtual, Austria, 12th July 2020. URL: https://hal.archives-ouvertes.fr/hal-02888693.

National peer-reviewed Conferences

[14] F. Antonazzo, C. Biernacki and C. Keribin. 'Estimation of univariate Gaussian mixtures for huge raw datasets by using binned datasets'. In: JDS2020. Nice, France, 25th May 2020. URL: https://hal.archives-ouvertes.fr/hal-03082437.

[15] O. Coudray, P. Bristiel, M. Dinis, C. Keribin and P. Pamphile. 'Characterization of critical areas for mechanical part fatigue design'. In: E-congrès 2020 Lambda $\lambda\mu$22 - 22e Congrès de Maîtrise des Risques et Sûreté de Fonctionnement $\lambda\mu$22. Le Havre / Virtual, France, 12th Aug. 2020. URL: https://hal.inria.fr/hal-03121282.

[16] O. Coudray, C. Keribin, P. Pamphile, M. Dinis and P. Bristiel. 'Characterization of critical areas for mechanical part fatigue design'. In: SFdS2020 - 52èmes Journées de Statistiques de la Société Française de Statistique. Nice, France, 25th May 2020. URL: https://hal.inria.fr/hal-03079350.

[17] R. Coulaud, C. Keribin and G. Stoltz. 'Quels modèles pour le temps de stationnement des trains en Île de France ?' In: SFdS 2020 - 52èmes Journées de Statistiques de la Société Française de Statistiques. Nice, France, 25th May 2020. URL: https://hal.inria.fr/hal-03065339.

Conferences without proceedings

[18] E. Chzhen, C. Denis, M. Hebiri, L. Oneto and M. Pontil. 'Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Virtuel, Canada, 6th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02501190.

[19] E. Chzhen, C. Denis, M. Hebiri, L. Oneto and M. Pontil. 'Fair Regression with Wasserstein Barycenters'. In: NeurIPS 2020 - 34th Conference on Neural Information Processing Systems. Vancouver / Virtuel, Canada, 6th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-02866811.

[20] P. Pamphile and F. Ducros. 'Maintien en conditions opérationnelles d'une flotte de véhicules : estimation du besoin en pièce de rechange'. In: E-congrès 2020 Lambda $\lambda\mu$22 - 22e Congrès de Maîtrise des Risques et Sûreté de Fonctionnement $\lambda\mu$22. Le Havre / Virtual, France, 12th Aug. 2020. URL: https://hal.archives-ouvertes.fr/hal-03080355.

Doctoral dissertations and habilitation theses

[21] M. Brégère. 'Stochastic bandit algorithms for demand side management'. Université Paris-Saclay, 10th Dec. 2020. URL: https://hal.archives-ouvertes.fr/tel-03059605.

[22] G. Chinot. 'Localization methods with applications to robust learning and interpolation'. Institut Polytechnique de Paris, 22nd June 2020. URL: https://tel.archives-ouvertes.fr/tel-02886789.

[23] M. Huard. 'Sequential learning and prediction : uniform regret bounds and hierarchical time series'. Université Paris-Saclay, 23rd Sept. 2020. URL: https://tel.archives-ouvertes.fr/tel-02957602.

[24] Y. Issartel. 'Inference on random networks'. Faculté des sciences d'Orsay, Université Paris-Saclay, 24th Nov. 2020. URL: https://hal.inria.fr/tel-03041741.

Reports & preprints

[25] Y. Amara-Ouali, Y. Goude, P. Massart, J.-M. Poggi and H. Yan. *A review of electric vehicle load open data and models*. 27th Nov. 2020. URL: https://hal.inria.fr/hal-03028375.

[26] G. Celeux and P. Pamphile. *Estimating parameters of the Weibull Competing Risk model with Masked Causes and Heavily Censored Data*. 15th Oct. 2020. URL: https://hal.inria.fr/hal-02410489.

[27] E. Chzhen and N. Schreuder. *A minimax framework for quantifying risk-fairness trade-off in regression*. 16th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03073960.

[28] S. Gaucher. *Finite Continuum-Armed Bandits*. 2nd Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02975304.

[29] H. Hadiji, S. Gerchinovitz, J.-M. Loubes and G. Stoltz. *Diversity-Preserving K-Armed Bandits, Revisited*. 5th Oct. 2020. URL: https://hal.archives-ouvertes.fr/hal-02957485.

[30] H. Hadiji and G. Stoltz. *Adaptation to the Range in K-Armed Bandits*. 10th Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02794382.

[31] M. Huard, R. Garnier and G. Stoltz. *Hierarchical robust aggregation of sales forecasts at aggregated levels in e-commerce, based on exponential smoothing and Holt's linear trend method*. 5th June 2020. URL: https://hal.archives-ouvertes.fr/hal-02794320.

[32] C. Keribin. *Cluster or co-cluster the nodes of oriented graphs?* 11th Feb. 2021. URL: https://hal.inria.fr/hal-03139333.

[33] G. Maillard. *Aggregated hold out for sparse linear regression with a robust loss function*. 26th Feb. 2020. URL: https://hal.archives-ouvertes.fr/hal-02485694.

[34] E. M. Saad, G. Blanchard and S. Arlot. *Online Orthogonal Matching Pursuit*. 14th Feb. 2021. URL: https://hal.archives-ouvertes.fr/hal-03141061.

Other scientific publications

[35] E. Chzhen and N. Schreuder. *An example of prediction which complies with Demographic Parity and equalizes group-wise risks in the context of regression*. Vancouver, Canada, 12th Dec. 2020. URL: https://hal.archives-ouvertes.fr/hal-03073964.

12.3 Cited publications

[36] S. Minsker and T. Mathieu. 'Excess risk bounds in robust empirical risk minimization'. working paper or preprint. Dec. 2019. URL: https://hal.archives-ouvertes.fr/hal-02390397.