RESEARCH CENTRE

Inria Saclay Center
at Université Paris-Saclay

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2022 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

Inria

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
A3.4.6. – Neural networks
A3.4.7. – Kernel methods
A3.4.8. – Deep learning
A3.5.1. – Analysis of large graphs
A6.1. – Methods in mathematical modeling
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
B4. – Energy
B4.4. – Energy delivery
B4.5. – Energy consumption
B5.2.1. – Road vehicles
B5.2.2. – Railway
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]
- Evgenii Chzhen [CNRS, Researcher]
- Gilles Stoltz [CNRS, Senior Researcher, HDR]

Faculty Members

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

Post-Doctoral Fellows

- Pierre Andrieu [UNIV PARIS-SACLAY, until Aug 2022, Co-advised with Sarah Cohen-Boulakia (LISN)]
- Pierre Humbert [INRIA]

PhD Students

- Yvenn Amara Ouali [UNIV PARIS-SACLAY, until Sep 2022, Co-advised with Yannig Goude (EDF)]
- Filippo Antonazzo [Inria, until Sep 2022, Co-advised with Christophe Biernacki (Modal)]
- Emilien Baroux [GROUPE PSA, CIFRE, Co-advised with Andrei Constantinescu (LMS)]
- Antoine Barrier [ENS Lyon, Co-advised with Aurélien Garivier]
- Samy Clementz [UNIV PARIS 1 PANTHEON SORBONNE, Co-advised with Alain Celisse]
- Olivier Coudray [GROUPE PSA, CIFRE]
- Rémi Coulaud [SNCF, CIFRE]
- Solenne Gaucher [UNIV PARIS-SACLAY, until Jun 2022]
- Karl Hajjar [UNIV PARIS SACLAY, Co-advised with Lénaïc Chizat (EPFL)]
- Perrine Lacroix [UNIV PARIS SACLAY, Co-advised with Marie-Laure Martin-Magniette (Inrae)]
- Etienne Lasalle [UNIV PARIS-SACLAY, Co-advised with Frédéric Chazal (Datashape)]
- Leonardo Martins-Bianco [Inria, from Apr 2022, Co-advised with Zacharie Naulet (LMO); M2 intern from April to August, then PhD student]
- Chiara Mignacco [UNIV PARIS-SACLAY, from Sep 2022]

- Louis Pujol [Inria, Co-advised with Marc Glisse (Datashape)]
- El Mehdi Saad [UNIV PARIS SACLAY, Co-advised with Gilles Blanchard (Datashape)]
- Gayane Taturyan [IRT SystemX, Co-advised with Jean-Michel Loubes (IMT) and Mohamed Hebiri (Univ Paris-Est)]

Administrative Assistants

- Aissatou-Sadio Diallo [Inria, from May 2022]
- Laurence Fontana [Inria, until May 2022]

External Collaborators

- Claire Lacour [UNIV PARIS EST]
- Matthieu Lerasle [ENSAE]
- Jean-Michel Poggi [UNIV PARIS CITE, from Mar 2022]

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 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, 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. Therefore, we tackle several major challenges of machine learning with our mathematical statistics point of view (in particular the algorithmic fairness issue), always having in mind that modern datasets are often high-dimensional and/or large-scale, which must be taken into account at the building stage of statistical learning algorithms. For instance, there often are trade-offs between statistical accuracy and complexity which we want to clarify as much as possible.

In addition, most theoretical guarantees that we prove are non-asymptotic, which is important because the number of features p is often larger than the sample size n in modern datasets, hence asymptotic results with p fixed and $n \to +\infty$ are not relevant. The non-asymptotic approach is also closer to the real-world than specific asymptotic settings, since it is difficult to say whether $p = 1000$ and $n = 100$ corresponds to the setting $p = 10n$ or $p = n^{3/2}$.

Finally, a key ingredient in our research program is connecting our theoretical and methodological results with (a great number of) real-world applications. This is the reason why a large part of our work is devoted to industrial and medical data modeling on a set of real-world problems coming from our long-term collaborations with several partners, as well as various opportunistic one-shot collaborations.

3 Research program

3.1 General presentation

We split our research program into four research axes, distinguishing problems and methods that are traditionally considered part of mathematical statistics (e.g., model selection and hypothesis testing, see section 3.2) from those usually tackled by the machine learning community (e.g., multi-armed bandits, deep learning, clustering and pairwise-data inference, see section 3.3). Section 3.4 is devoted to industrial and medical data modeling questions which arise from several long-term collaborations and more recent research contracts. Finally, section 3.5 is devoted to algorithmic fairness, a theme of Celeste which we want to specifically emphasize. Despite presenting mathematical statistics, machine learning, and data modeling as separate axes, we would like to make clear that these axes are strongly interdependent in our research and that this dependence is a key factor in our success.

3.2 Mathematical statistics

One of our main goals is to address major challenges in machine learning in which mathematical statistics naturally play a key role, in particular in the following two areas of research.

Estimator selection. Any machine learning procedure requires a choice for the values of hyper-parameters, and one must also choose among the numerous procedures available for any given learning problem; both situations correspond to an estimator selection problem. High-dimensional variable (feature) selection is another key estimator selection problem. Celeste addresses all such estimator selection problems, where the goal is to select an estimator (or a set of features) minimizing the prediction/estimation risk, and the corresponding non-asymptotic theoretical guarantee—which we want to prove in various settings—is an oracle inequality.

Statistical reproducibility. Science currently faces a reproducibility crisis, making it necessary to provide statistical inference tools (hypotheses tests, confidence regions) for assessing the significance of the output of any learning algorithm in a computationally efficient way. Our goal here is to develop methods for which we can prove upper bounds on the type I error rate, while maximizing the detection power under this constraint. We are particularly interested in the variable selection case, which here leads to a multiple testing problem for which key metrics are the family-wise error rate (FWER) and the false discovery rate (FDR).

3.3 Theoretical foundations of machine learning

Our distinguishing approach (compared to peer groups around the world) is to offer a statistical and mathematical point of view on machine-learning (ML) problems and algorithms. Our main focus is to provide theoretical guarantees for certain ML problems, with special attention paid to the statistical point of view, in particular minimax optimality and statistical adaptivity. In the areas of deep learning and big data, computationally-efficient optimization algorithms are essential. The choice of the optimization algorithm has been shown to have a dramatic impact on generalization properties of predictors. Such empirical observations have led us to investigate the interplay between computational efficiency and statistical properties. The set of problems we tackle includes online learning (stochastic bandits and expert aggregation), clustering and co-clustering, pairwise-data inference, semi-supervised learning, and the interplay between optimization and statistical properties.

3.4 Industrial and medical data modeling

Celeste collaborates with industry and with medicine/public health institutes to develop methods and apply results of a broadly statistical nature—whether they be prediction, aggregation, anomaly detection, forecasting, and so on—in relationship with pressing industrial and/or societal needs (see sections 4 and 5.2). Most of these methods and applied results are directly related to the more theoretical subjects examined in the first two research axes, including for instance estimator selection, aggregation, and

supervised and unsupervised classification. Furthermore, Celeste is positioned well for problems with data requiring unconventional methods—for instance, non asymptotic analysis and data with selection bias—, and in particular problems that can give rise to technology transfers in the context of Cifre Ph.D.s.

3.5 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 may lead to unfair and discriminatory decisions, often inheriting or even amplifying disparities present in data. The goal of Celeste on this topic 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 the fairness and accuracy of a learned model: more accurate models are less fair. We are focused on providing user-friendly statistical quantification of such trade-offs and building statistically-optimal algorithms in this context, with special attention paid to the online learning setting. Relying on the strong mathematical and statistical competency of the team, we approach the problem from an angle that differs from the mainstream computer science literature.

4 Application domains

4.1 Neglected tropical diseases

Celeste collaborates with researchers at Institut Pasteur on encephalitis in South-East Asia, especially with Jean-David Pommier.

4.2 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 consumption of well-chosen groups/regions) and aggregation of local estimates.

4.3 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 Stellantis. To dimension its vehicles, Stellantis 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 Stellantis’s databases, resulting either from tests or client usages. This will lead to more robust and accurate parts dimensioning methods. Two CIFRE theses (one recently defended, the other in progress) tackle such subjects:

Olivier Coudray, “A statistical point of view on fatigue criteria : from supervised classification to positive-unlabeled learning” [24]. 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.).

4.4 Exploiting a data-rich environment for railway operations, with SNCF–Transilien

We have an on-going collaboration with SNCF–Transilien to exploit rich datasets of railway operations and passenger flows, obtained by automatic recording devices (for passenger flows, these correspond to infra-sensors at the door level). We tackle two problems. First, we model and estimate the dwell time of trains, as well as their delay to scheduled arrival; both are important factors to control to guarantee punctuality. Second, we model and forecast passenger movements inside train coaches, so as to be able to provide incoming passengers with information on crowding of coaches. Both series of problems come with new results described in section 7.11. They correspond to the final year of the CIFRE PhD of Rémi Coulaud between our Celeste team and SNCF–Transilien [25].

5 Social and environmental responsibility

5.1 Footprint of research activities

Still influenced by the aftermath of the Covid-19 pandemic, 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 a small amount for essentially land travel to conferences in France and Europe.
- 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 10 g Co2-e per hour per person, while web browsing uses around 100g Co2-e per hour. Consequently, 2022 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) kg 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 work—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 on neglected tropical diseases; encephalitis in particular, with implications in global health strategies.

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 from several direct collaborations with companies such as EDF (forecasting and control of electricity load consumption), SNCF (punctuality of trains and better passenger information on crowding inside train coaches) and Stellantis (automobile reliability, with two Cifre contracts).

6 New software and platforms

6.1 New software

6.1.1 CRT-Logit

Keywords: Hypothesis testing, Variable selection, High Dimensional Data, Statistical learning, Classification

Functional Description: This packages runs the algorithm CRT-Logit to perform the conditional randomization test to identify relevant variables for a classification model.

URL: https://github.com/tbng/crt_logit

Publication: hal-03680792

Contact: Tuan Binh Nguyen

6.1.2 pysarpu

Keywords: Statistical learning, Semi-supervised classification

Functional Description: Implementation of the estimation procedure of a PU learning model (joint estimation of the classifier and the propensity function) under parametric assumptions using the EM (Expectation Maximization) algorithm. Several classification models (logistic regression, linear discriminant analysis) and propensity models (logistic, probit, Gumbel) are implemented.

URL: https://pysarpu.readthedocs.io/en/latest/index.html

Publication: hal-03526889

Contact: Christine Keribin

Partner: STELLANTIS

6.1.3 binMixtC

Keywords: Clustering, EM algorithm

Functional Description: Software implementing a data compression scheme preserving bin-marginal values with a specific EM-like algorithm.

URL: https://github.com/Fili75/binMixtC

Contact: Christine Keribin

6.1.4 gsbm

Keywords: Missing data, Outlier detection, Stochastic block model

Functional Description: Given an adjacency matrix drawn from a Generalized Stochastic Block Model with missing observations, this package robustly estimates the probabilities of connection between nodes and detects outliers node

URL: https://cran.r-project.org/web/packages/gsbm/index.html

Publication: hal-02386940

Contact: Solenne Gaucher

Partner: Ecole des Ponts ParisTech

6.1.5 DBABST

Keywords: Classification, Algorithmic fairness, Demographic parity

Functional Description: Post-processing algorithm for binary classification with abstention and DP constraints

URL: https://github.com/evgchz/dpabst

Publication: hal-03152091

Contact: Evgenii Chzhen

6.1.6 HiDimStat

Keywords: Statistical inference, High Dimensional Data, Variable selection

Functional Description: The HiDimStat package provides statistical inference methods to solve the problem of support recovery in the context of high-dimensional and spatially-structured data.

URL: https://github.com/ja-che/hidimstat/blob/main/README.md

Publications: hal-02888693, hal-01815255, hal-03887430, hal-03179728, hal-03023228

Contact: Tuan Binh Nguyen

6.2 New platforms

Participants: Alexandre Janon.

- A. Janon : Deployment of a R (Shiny) server for the LAGUN platform in uncertainty quantification. Link.

7 New results

7.1 Conditional Randomization Test for Sparse Logistic Regression in High-Dimension

Participants: Sylvain Arlot.

Identifying the relevant variables for a classification model with correct confidence levels is a central but difficult task in high-dimension. Despite the core role of sparse logistic regression in statistics and machine learning, it still lacks a good solution for accurate inference in the regime where the number of features p is as large as or larger than the number of samples n. In [19], in collaboration with B. Nguyen and B. Thirion, we tackle this problem by improving the Conditional Randomization Test (CRT). The original CRT algorithm shows promise as a way to output p-values while making few assumptions on the distribution of the test statistics. As it comes with a prohibitive computational cost even in mildly high-dimensional problems, faster solutions based on distillation have been proposed. Yet, they rely on unrealistic hypotheses and result in low-power solutions. To improve this, we propose CRT-logit, an algorithm that combines a variable-distillation step and a decorrelation step that takes into account the geometry of ℓ^1-penalized logistic regression problem. We provide a theoretical analysis of this procedure, and demonstrate its effectiveness on simulations, along with experiments on large-scale brain-imaging and genomics datasets.

7.2 Robust Kernel Density Estimation with Median-of-Means principle

Participants: Pierre Humbert.

In [17], in collaboration with Batiste Le Bars (Inria Magnet) and Ludovic Minvielle (ENS Paris-Saclay), we introduce a robust nonparametric density estimator combining the popular Kernel Density Estimation method and the Median-of-Means principle (MoM-KDE). This estimator is shown to achieve robustness to any kind of anomalous data, even in the case of adversarial contamination. In particular, while previous works only prove consistency results under known contamination model, this work provides finite-sample high-probability error-bounds without a priori knowledge on the outliers. Finally, when compared with other robust kernel estimators, we show that MoM-KDE achieves competitive results while having significant lower computational complexity.

7.3 Stochastic bandits

Participants: Antoine Barrier, Zhen Li, Gilles Stoltz, Solenne Gaucher, Christophe Giraud .

In [18], we revisit the problem of contextual bandits with budget constraints (called contextual bandits with knapsacks), and apply it to conversion models. This is a joint work with BNP Paribas with a research contract, and our running example is the clever use of a discount budget to grant loans; conversions correspond to whether or not a customer takes the loan offer. We provide a strategy in a direct primal formulation, where previous contributions in the literature rather suggested strategies based on dual formulations of the problem, with tuning issues (on how to set the dual variables).

In [31], we tackle the problem of best-arm identification [BAI] with a fixed budget T, a problem that remains vastly unexplored: the literature is rich as far as BAI with fixed confidence is concerned. We survey existing results, both upper bounds, typically based on successive-rejects-type algorithms, and lower bounds, and extend them to the case of non-parametric models consisting, e.g., of the set of all distributions over a compact interval. We introduce new information-theoretic quantities measuring the hardness of these problems. We however still could not close the $\ln K$ gap between upper and lower bounds, though we reduced it.

In [16] we address discriminative bias in linear bandits and quantified the price of unfair evaluations in the worst case and the gap-minimax setting. The results revealed a transition between a regime where the problem is as difficult as its unbiased counterpart, and a regime where it can be much harder. Unlike previously-mentioned contributions, which were model-free, this work postulated an explicit source of discriminating bias.

7.4 Constant regret for sequence prediction with limited advice

Participants: El Mehdi Saad.

In [42], in collaboration with Gilles Blanchard (Datashape), we consider the problem of cumulative regret minimization for individual sequence prediction with respect to the best expert in a finite family under limited access to information. We provide a strategy that combines two experts and observes at least two experts' predictions in each round. We prove that this strategy allows having a constant bound on the regret (an upper bound independent of the time horizon T), both in expectation and with high probability with respect to the player's randomization. Finally, we show that if the player is restricted to playing one expert or observing one expert's prediction per round, her regret is lower bounded by $\sqrt{T}$ for some sequences.

7.5 High Dimension and Estimation Challenges in Model-Based Co-Clustering

Participants: Christine Keribin.

In collaboration with Christophe Biernacki (Inria Modal) and Julien Jacques (Université de Lyon 2), we have written a survey [32] on model-based co-clustering. This problem can be seen as a particularly valuable extension of model-based clustering for three main reasons: (1) while allowing parsimoniously a drastic reduction of both the number of lines/individuals and columns/variables of a data set, (2) it also allows interpretability of such a resulting reduced data set since initial individuals and features meaning is preserved in this latter; (3) moreover it benefits from the powerful mathematical statistics theory for both estimation and model selection. Hence, many authors produced new advances on this topic in the recent years, and we offer a general update of the related literature. In addition, it is the opportunity to pass two messages, supported by specific research materials: (1) co-clustering still requires some new and motivating researches for fixing some well-identified estimation issues, (2) co-clustering is probably one of the most promising clustering approaches to be addressed in the (very) high dimensional setting, which corresponds to the global trend on modern data sets.

7.6 Optimization

Participants: Karl Hajjar, Evgenii Chzhen.

Symmetries are expected to play an important role in the effectiveness of Neural Networks. We have described in [37] a class of symmetries that are preserved during the learning process.

Leveraging statistical ideas, we have developed zero-order optimization algorithms using ℓ_1-randomized noisy function evaluations [14]. We have shown that the dual-averaging algorithm with ℓ_1-randomized noisy function evaluations improves the convergence rates of the previously best-performing constructions.

7.7 A statistical point of view on fatigue criteria: from supervised classification to positive-unlabeled learning

Participants: Olivier Coudray, Christine Keribin, Patrick Pamphile.

Celeste currently collaborates with the automobile manufacturer Stellantis. In Olivier Coudray's Ph.D. [20], the challenge for Stellantis was to identify critical zones of a mechanical part. This is an unsupervised classification problem with a selection bias in the data (during a test, the absence of the start of a crack does not necessarily mean that the zone is safe). We proposed to use a semi-supervised classification method called positive-unlabelled (PU) learning [34]. The optimality of the speed of convergence of the classifier (in the minimax sense) was obtained in this work. The PU learning classifier was then implemented on simulated datasets to compare its performance to other classification methods and on Stellantis datasets [software pysarpu, section 6.1.2].

7.8 Reliable design under complex loads: from specification to validation

Participants: Emilien Baroux, Patrick Pamphile.

Celeste currently collaborates with the automobile manufacturer Stellantis. In Emilien Baroux's Ph.D., co-advised with Andréï Constantinescu, the challenge for Stellantis is to identify critical areas for vehicule chassis fatigue, and in particular to identify severe fatigue profiles (driver and road type) for vehicule chassis. We started by creating a multi-axial mechanical damage measurement system that allowed us to comprehensively understand cases of locally critical loads (torsion, bending, etc.), independent of the vehicle model. During a driving campaign, damage measurements were taken simultaneously in the longitudinal, vertical, and horizontal axes, as well as on the four wheels, for 50 vehicles. We further proposed factorial analyses and unsupervised classification methods to construct a robust severity distribution for fatigue design tasks [46].

7.9 Daily peak electrical load forecasting with a multi-resolution approach

Participants: Yvenn Amara-Ouali.

In the context of smart grids and load balancing, daily peak load forecasting has become a critical activity for stakeholders in the energy industry. An understanding of peak magnitude and timing is paramount for the implementation of smart grid strategies such as peak shaving. In [3], in collaboration with Matteo Fasiolo, Yannig Goude and Hui Yan, we proposed a modelling approach which leverages high-resolution and low-resolution information to forecast daily peak demand size and timing. The resulting multi-resolution modelling framework can be adapted to different model classes and are implemented via generalised additive models and neural networks. Experimental results on real data from the UK electricity market confirm that the predictive performance of the proposed modelling approach is competitive with that of low- and high-resolution alternatives.

7.10 A benchmark of electric vehicle load and occupancy models for day-ahead forecasting on open charging session data

Participants: Yvenn Amara-Ouali.

The development of electric vehicles (EV) is a major lever towards low carbon transportation. It comes with increasing numbers of charging infrastructures which can be smartly managed to control the CO2 cost of EV electricity consumption or used as flexible assets for grid management. To achieve that, an efficient day-ahead forecast of charging behaviours is required at different spatial resolutions (e.g., household and public stations). In [15], in collaboration with Yannig Goude, Bachir Hamrouche and Matthew Bishara, we proposed an extensive benchmark of 14 models for both load and occupancy day-ahead forecasts, covering 8 open charging session datasets of different types (residential, workplace and public stations) is proposed. Two modelling approaches are compared: direct and bottom-up. The direct approach forecasts the aggregated load (resp. occupation) directly of an area/station whereas the bottom-up approach models each individual EV charging session before aggregating them. Both machine learning models and statistical models are considered. Results show that direct approaches reach better performances than bottom-up approaches. The different approaches used can lead to an improved performance of direct approaches when using an adaptive aggregation strategy.

7.11 Exploiting a data-rich environment for railway operations, with SNCF–Transilien

Participants: Rémi Coulaud, Christine Keribin, Gilles Stoltz.

As mentioned in Section 4.4, two sources of data may and should be mixed: railway operations (e.g., scheduled and observed arrival and departure times of trains in stations) and passenger flows (e.g.,

numbers of alighting and boarding passengers in stations). We model dwell times, the difference between departure and arrival times, in [8], based on various machine-learning models (linear regression, random forests and XGBoost, neural networks). Typically, the literature was only using one of the two sources of data at a time (just because only one such source was available for each given problem). By combining the two sources, we are able to identify the most critical source (railway operations) and quantify the added value of the other source (passenger flows: to help modeling critical situations, like delayed trains). To be able to exploit this initial complex modeling of dwell time, we need to forecast variables like numbers of passengers alighting and boarding. We do so in [21] by introducing simple bi-autoregressive models, which we call L-shaped as they exploit the past both in terms of previous trains at a given station and of previous stations of a given train ride.

A second series of work is for now only described in the PhD manuscript by Rémi Coulaud [25] and is currently being finalized (and thus, will be described in greater details in next year's report). It deals with the modeling and forecasting of passengers' movements inside communicating train coaches. The model exhibited relies on an inhomogeneous Markov chain modeling, using coach loads as latent spaces.

7.12 SEAE: The South-East Asia Encephalitis Project

Participants: Kevin Bleakley.

In [13], K. Bleakley was the statistician/modeler/machine learning lead for the large-scale 3-year SEAE encephalitis project in South-East Asia, looking for patterns in the (relatively "big") dataset relating environmental variables to encephalitis causes and outcomes. His work ranged across (multiple) statistical testing, machine learning (trees, random forests, and logistic regression), PCA, data visualisation, survival analysis, and missing data. Several interesting risk factors for severe encephalitis were uncovered, and steps were made to start looking for new, unknown causes (bacterial, viral, etc.) of encephalitis in South-East Asia. This work was published in the renowned journal, Lancet Infection Diseases (IF: 71); related work is ongoing.

7.13 Study of demographic parity constraint

Participants: Evgenii Chzhen, Solenne Gaucher.

In [35], we have shown several fundamental characterizations of the optimal classification function under the demographic parity constraint. In the awareness framework, akin to the classical unconstrained classification case, we have shown that maximizing accuracy under this fairness constraint is equivalent to solving a corresponding regression problem followed by thresholding at level $1/2$. We have extended this result to linear-fractional classification measures (e.g., F-score, AM measure, balanced accuracy, etc.). These results further deepen our understanding of fairness constraints and their impact on decision making.

8 Bilateral contracts and grants with industry

Participants: Alexandre Janon, Christine Keribin, Patrick Pamphile, Jean-Michel Poggi, Gilles Stoltz, Yvenn Amara-Ouali.

8.1 Bilateral contracts with industry

- C. Keribin and P. Pamphile. OpenLabIA Inria-Groupe Stellantis collaboration contract. 85 KE.

- A. Constantinescu and P. Pamphile. Collaboration contract with Stellantis. 95 KE.
- C. Keribin and G. Stoltz. Ongoing contrat with SNCF (45 kE), on the modeling and forecasting of dwell time.
- J.M. Poggi, Analyse et modélisation des biais des modèles numériques du NO2 en vue de la fusion de données de réseaux de mesure hétérogènes, ATMO NORMANDIE, 20KE.
- Y. Amara-Ouali, P. Massart, J-M. Poggi, Modélisation spatio-temporelle de la charge des véhicules électriques, EDF, 20KE.
- G. Stoltz: Ongoing contract with BNP Paribas (2 x 10 kE), on stochastic bandits under budget constraints, with applications to loan management.
- A. Janon: Contract with INSERM Toulouse (3,3 kE), on variable selection for identification of link between microbiotal dysbiosis and type-2 diabetes.

9 Partnerships and cooperations

9.1 National initiatives

Participants: Sylvain Arlot, Kevin Bleakley, Christophe Giraud, Matthieu Lerasle.

9.1.1 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).

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

Christophe Giraud is part of the ANR ASCAI: Active and batch segmentation, clustering, and seriation: toward unified foundations in AI, with Potsdam University, Munich University, Montpellier INRAE.

9.1.2 Other

K. Bleakley works at 1/3-time (*disponibilité*) with IRT SystemX under the umbrella of Confiance.AI on the subject of anomaly detection in high-dimensional time series data for French industry.

10 Dissemination

Participants: The full Celeste team.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- C. Keribin is president of the Specialized Group *MAchine Learning et Intelligence Artificielle* (MALIA) of the French Statistical Society (SFdS)
- J.-M. Poggi is President of ENBIS (European Network for Business and Industrial Statistics)

Member of the organizing committees

- S. Arlot is member of the scientific committee of the Séminaire Palaisien
- E. Chzhen is co-organizer of DATAIA institutional seminar
- C. Giraud is co-organizer with Estelle Kuhn of StatMathAppli Frejus, 2022
- C. Giraud is organizing a “statistical learning” session for the joint AMS/EMS/SMF conference, Grenoble 2022
- A. Janon is co-organizer of UQSay seminar
- C. Keribin is co-organizer FrENBIS satellite event “Statistical learning for temporal data, new horizons and industrial applications” during JDS2022 (Lyon, 17/06/2022).
- J.-M. Poggi is organizer of the ECAS-ENBIS course, Text Mining: from basics to deep learning tools, Trondheim, June 26, 2022

10.1.2 Scientific events: selection

Member of the conference program committees

- A. Janon is member of the program commitee of the Mascot-Num 2023 conference in Le Croisic.
- J.-M. Poggi is member program commitee of ENBIS-22 Conference, Trondheim, June 26-30, 2022

Reviewer

- We performed many reviews for various international conferences.

10.1.3 Journal

Member of the editorial boards

- S. Arlot: Associate Editor for *Annales de l’Institut Henri Poincaré B – Probability and Statistics*
- C. Giraud: Action Editor for JMLR
- C. Giraud: Associate Editor for ESAIM-proc
- P. Massart: Associate editor for Panoramas et Synthèses (SMF), Foundations and Trends in Machine Learning, and Confluentes Mathematici
- J.-M. Poggi: Associate Editor Advances in Data Analysis and Classification
- J.-M. Poggi: Associate Editor JDSSV J. Data Science, Statistics and Visualization
- J.-M. Poggi: Associate Editor for Journal of Statistical Software
- J.-M. Poggi: co-Editor of the book "Interpretability for Industry 4.0: Statistical and Machine Learning Approaches", Springer, nov. 2022
- G. Stoltz: Associate Editor for *Mathematics of Operations Research*

Reviewer - reviewing activities

- We performed many reviews for various international journals.

10.1.4 Invited talks

S. Arlot, Journées Statistiques du Sud, Avignon, 02/06/2022
E. Chzhen, Re-thinking High-dimensional Mathematical Statistics, Oberwolfach, 20/05/2022
E. Chzhen, New trends in statistical learning II, Porquerolles, 14/06/2022
E. Chzhen, Computational Statistics and Machine Learning, Genoa, 12/07/2022
E. Chzhen, Workshop on Ethical AI, Paris, 29/09/2022
E. Chzhen, MADSTAT seminar, Toulouse, 13/10/2022
C. Giraud, ASCAI, Montpellier, 29/02/2022
C. Giraud, Séminaire INRIA Paris Centre, 14/04/2022
C. Giraud, ETH-FDS seminar series, Zurich, 08/06/2022
C. Giraud, IMS, London, 28/06/2022
C. Giraud, course in the summer school "Geometry and Statistics", Cargèse, 05-09/09/2022
C. Giraud, Van Dantzig seminar, Amsterdam, 16/12/2022
C. Keribin, JSTAR 2022 (Rennes)
C. Keribin, CMStatistics Londres 2022
J.-M. Poggi, Mathmet 2022, Paris, France, November 2022
J.-M. Poggi, Compstat 2022, Bologna, August 2022
J.-M. Poggi, ISBIS 2022, June 2022, Naples, Italy
J.-M. Poggi, CISEM 2022, Mahdia (Tunisie) mai 2022
J.-M. Poggi, Seminar Department of Mathematics, Université du Luxembourg, October 2022

10.1.5 Leadership within the scientific community

G. Stoltz: main contact point [2018–2022] for the author, G. Favre, of a sociologic report on collaborations of mathematicians with companies, commissioned by AMIES (agence maths-entreprises, which is an Inria–CNRS–Université Grenoble Alpes entity).

10.1.6 Research administration

- S. Arlot is member of the council of the Computer Science Graduate School (GS ISN) of University Paris-Saclay.
- S. Arlot is member of the council of the Computer Science Doctoral School (ED STIC) of University Paris-Saclay.
- E. Chzhen is member of Bibliothèque Jacques Hadamard scientific committee
- C. Giraud is member of the Scientific Committee of labex IRMIA+, Strasbourg
- C. Giraud is in charge of the whole Masters program in mathematics for University Paris-Saclay
- C. Giraud is member of the board of the Mathematics Graduate School of University Paris-Saclay
- C. Giraud is senior member of CCUPS (Commission Consultative Université Paris Saclay)
- C. Keribin is in charge of the M1-Applied Mathematics and M2-Data-Science program in the master of the mathematical school
- C. Keribin is member of the board of the Computer Science Doctoral School (ED MSTIC) of Paris-Est Sup.
- C. Keribin is vice-president of the Math - CCUPS (Commission consultative de l'Université Paris–Saclay).
- C. Keribin is member of the council of the mathematics department.
- C. Keribin is co-president of the scholarship allocation committee MixtAI of the SaclAI school.
- P. Massart is director of the Fondation Mathématique Jacques Hadamard.

- M-A. Poursat is in charge of the M1-Mathematics and artificial intelligence program in the master of the mathematical school

10.1.7 Service to the academic community

- K. Bleakley: translation into English of the webpages of the LMO's website dedicated to research activities
- C. Giraud: coordinator of computing ressources at the Institut Mathématiques d'Orsay (10 engineers)
- C. Keribin: selection committee for an assistant professor, UTC, May 2022
- C. Keribin: selection committee for an assistant professor, ENSMM Besançon, May 2022
- C. Keribin: member of the follow-up committee for PhD student Tom Guedon (Inrae)
- C. Keribin: member of the follow-up committee for for PhD student Anderson Augusma (Laboratoire d'informatique de Grenoble)
- G. Stoltz: co-head of the committee working out a new version of the LMO's website, 2019–2022, with the year 2022 dedicated to the Intranet
- G. Stoltz: selection committee for a professor position, Université Rennes 2, May 2022
- G. Stoltz: selection committee for a professor position, Université Paris 1, May 2022

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

Most of the team members (especially Professors, Associate Professors and PhD 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.

- Masters: S. Arlot, Statistical learning and resampling, 30h, M2, Université Paris-Saclay
- Masters: S. Arlot, Preparation for French mathematics agrégation (statistics), 25h, M2, Université Paris-Saclay
- Masters: E. Chzhen, Fairness and Privacy in Machine Learning, 18h, M2 ENSAE
- Masters: E. Chzhen, Statistical Theory of Algorithmic Fairness, 20h, M2 Université Paris-Saclay
- Masters: C. Giraud, High-Dimensional Probability and Statistics, 45h, M2, Université Paris-Saclay
- Masters: C. Giraud, Mathematics for AI, 75h, M1, Université Paris-Saclay
- Masters: C. Keribin, unsupervised and supervised learning, M1, 42h, Université Paris-Saclay
- Masters: C. Keribin, Cours accéléré en statistiques, M2, 21h, Université Paris-Saclay
- Masters: C. Keribin, Modélisation statistique, M1, 2 x 20h, Université Paris-Saclay
- Masters: C. Keribin, Internship supervision for M1-Applied Mathematics and M2-DataScience, Université Paris-Saclay
- Masters: M-A Poursat, applied statistics, 21h, M1 Artificial Intelligence, Université Paris-Saclay
- Masters: M-A Poursat, statistical learning, 42h, M2 Bioinformatics, Université Paris-Saclay
- Masters: M-A Poursat, méthodes de classification, 24h, M1, Université Paris-Saclay
- Licence: M-A Poursat, inférence statistique, 72h, L3, Université Paris-Saclay
- Masters: G. Stoltz, Sequential Learning and Optimization, 18h, M2 Université Paris-Saclay

10.2.2 Supervision

- PhD defended on June 2022: Solenne Gaucher, Sequential learning in random networks, started Sept. 2018, C. Giraud.
- PhD defended on Sep. 2022: Yvenn Amara-Ouali, Spatio-temporal modelization of electrical vehicles load, started Oct. 2019, co-advised by P. Massart, J.M. Poggi (Université de Paris) and Y. Goude (EDF R& D)
- PhD defended on Sep. 2022: 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 defended on Nov. 2022: 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 defended on Dec. 2022: Etienne Lasalle, Statistical foundations of topological data analysis for graph structured data, started Sept. 2018, co-advised by F. Chazal (INRIA Datashape) and P. Massart
- PhD defended on Dec. 2022: Olivier Coudray, Fatigue data-based design, started Nov. 2019, co-advised by C. Keribin and P. Pamphile, Cifre with Groupe PSA
- PhD defended on Dec. 2022: Louis Pujol, CYTOPART - Flow cytometry data clustering, started Nov. 2019, co-advised by P. Massart and M. Glisse (INRIA Datashape)
- PhD defended on Dec. 2022: El Mehdi Saad, Interactions between statistical and computational aspects in machine learning, started Sept. 2019, co-advised by S. Arlot and G. Blanchard (INRIA Datashape)
- PhD defended on Dec. 2022: Perrine Lacroix, High-dimensional linear regression applied to gene interactions network inference, started Sept. 2019, co-advised by P. Massart and M.-L. Martin-Magniette (INRAE)
- PhD in progress: Emilien Baroux, Reliability dimensioning under complex loads: from specification to validation, started July. 2020, co-advised by A. Constantinescu (LMS) and P. Pamphile, CIFRE with Groupe PSA
- PhD in progress: Antoine Barrier, started Sept. 2020, Best Arm Identification, co-advised by G. Stoltz and A. Garivier (ENS Lyon)
- PhD in progress: Karl Hajjar, analyse dynamique de réseaux de neurones, started Oct. 2020, co-advised by C. Giraud and L. Chizat (EPFL).
- PhD in progress: Samy Clementz, Data-driven Early Stopping Rules for saving computation resources in AI, started Sept. 2021, co-advised by S. Arlot and A. Celisse
- PhD in progress: Gayane Taturyan, Fairness and Robustness in Machine Learning, started Nov. 2021, co-advised by E. Chzhen, J.-M. Loubes (Univ. Toulouse Paul Sabatier) and M. Hebiri (Univ. Gustave Eiffel)
- PhD in progress: Leonardo Martins-Bianco, Disentangling the relationships between different community detection algorithms, started October 2022, co-advised by C. Keribin and Z. Naulet (Univ. Paris-Saclay)
- PhD in progress: Chiara Mignacco, Aggregation (orchestration) of reinforcement learning policies, started October 2022, co-advised by G. Stoltz and Matthieu Jonckheere (LAAS Toulouse)

10.2.3 Juries

We participated to many PhD committees (too many to keep an exact record), at University Paris-Saclay as well as at other universities, and we refereed several of these PhDs.

10.3 Popularization

10.3.1 Articles and contents

K. Bleakley gave an interview for Inria's "News and Events" outreach on his work on Encephalitis in South-East Asia in collaboration with the Pasteur Institute. The resulting news article was published here.

10.3.2 Education

Christophe Giraud produces educational videos on his YouTube channel "High-dimensional probability and statistics".

10.3.3 Interventions

Christine Keribin was invited speaker at the Ateliers de la Statistique de la SFdS for an introductory lecture to machine learning (2022).

11 Scientific production

11.1 Major publications

[1] E. Chzhen and N. Schreuder. 'A minimax framework for quantifying risk-fairness trade-off in regression'. In: *Annals of Statistics* 50.4 (25th Aug. 2022), pp. 2416–2442. URL: https://hal.archives-ouvertes.fr/hal-03073960.

[2] J. D. Pommier, C. Gorman, Y. Crabol, K. Bleakley, H. Sothy, K. Santy, H. T. T. Tran, L. V. Nguyen, E. Bunnakea, C. S. Hlaing et al. 'Childhood encephalitis in the Greater Mekong region (the SouthEast Asia Encephalitis Project): a multicentre prospective study'. In: *The Lancet global health* 10.7 (July 2022), e989–e1002. DOI: 10.1016/S2214-109X(22)00174-7. URL: https://hal.archives-ouvertes.fr/hal-03823946.

11.2 Publications of the year

International journals

[3] Y. Amara-Ouali, M. Fasiolo, Y. Goude and H. Yan. 'Daily peak electrical load forecasting with a multi-resolution approach'. In: *International Journal of Forecasting* (11th July 2022). URL: https://hal.inria.fr/hal-03469721.

[4] M. Brégère and M. Huard. 'Online hierarchical forecasting for power consumption data'. In: *International Journal of Forecasting* 38.1 (Jan. 2022), pp. 339–351. DOI: 10.1016/j.ijforecast.2021.05.011. URL: https://hal.archives-ouvertes.fr/hal-03884826.

[5] G. Celeux, S. X. Cohen, A. Grimaud and P. Gueriau. 'Hierarchical clustering of spectral images with spatial constraints for the rapid processing of large and heterogeneous data sets'. In: *SN Computer Science* 3.194 (22nd Mar. 2022). DOI: 10.1007/s42979-022-01074-4. URL: https://hal.uvsq.fr/hal-03104488.

[6] G. Chinot and M. Lerasle. 'On the robustness of the minimim l2 interpolator'. In: *Bernoulli* (2022). DOI: 10.48550/arXiv.2003.05838. URL: https://hal.archives-ouvertes.fr/hal-03874519.

[7] E. Chzhen and N. Schreuder. 'A minimax framework for quantifying risk-fairness trade-off in regression'. In: *Annals of Statistics* 50.4 (25th Aug. 2022), pp. 2416–2442. URL: https://hal.archives-ouvertes.fr/hal-03073960.

[8] R. Coulaud, C. Keribin and G. Stoltz. 'Modeling dwell time in a data-rich railway environment: with operations and passenger flows data'. In: *Transportation research. Part C, Emerging technologies* 146 (2023), p. 103980. URL: https://hal.archives-ouvertes.fr/hal-03651835.

[9] A. Garivier, H. Hadiji, P. Ménard and G. Stoltz. 'KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints'. In: *Journal of Machine Learning Research* (2022). URL: https://hal.archives-ouvertes.fr/hal-01785705.

[10] G. Maillard. 'Aggregated hold out for sparse linear regression with a robust loss function'. In: *Electronic Journal of Statistics* 16.1 (2022), pp. 935–997. URL: https://hal.archives-ouvertes.fr/hal-02485694.

[11] T. Mathieu. 'Concentration study of M-estimators using the influence function'. In: *Electronic Journal of Statistics* 16.1 (1st Jan. 2022), pp. 3695–3750. DOI: 10.1214/22-ejs2030. URL: https://hal.archives-ouvertes.fr/hal-03757720.

[12] M.-L. J. Nguyen, C. Lacour and V. Rivoirard. 'Adaptive Greedy Algorithm for Moderately Large Dimensions in Kernel Conditional Density Estimation'. In: *Journal of Machine Learning Research* 23.254 (2022). URL: https://hal.archives-ouvertes.fr/hal-02085677.

[13] J. D. Pommier, C. Gorman, Y. Crabol, K. Bleakley, H. Sothy, K. Santy, H. T. T. Tran, L. V. Nguyen, E. Bunnakea, C. S. Hlaing et al. 'Childhood encephalitis in the Greater Mekong region (the SouthEast Asia Encephalitis Project): a multicentre prospective study'. In: *The Lancet global health* 10.7 (July 2022), e989–e1002. DOI: 10.1016/S2214-109X(22)00174-7. URL: https://hal.science/hal-03823946.

International peer-reviewed conferences

[14] A. Akhavan, E. Chzhen, M. Pontil and A. B. Tsybakov. 'A gradient estimator via L1-randomization for online zero-order optimization with two point feedback'. In: NeurIPS 2022 - Thirty-sixth Conference on Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03782218.

[15] Y. Amara-Ouali, Y. Goude, B. Hamrouche and M. Bishara. 'A benchmark of electric vehicle load and occupancy models for day-ahead forecasting on open charging session data'. In: e-Energy '22: The Thirteenth ACM International Conference on Future Energy Systems. Virtual Event, France: ACM, 29th June 2022, pp. 193–207. DOI: 10.1145/3538637.3538850. URL: https://hal.inria.fr/hal-03880426.

[16] S. Gaucher, A. Carpentier and C. Giraud. 'The price of unfairness in linear bandits with biased feedback'. In: NeurIPS 2022. New Orleans, United States, 29th Nov. 2022. URL: https://hal.science/hal-03611628.

[17] P. Humbert, B. Le Bars and L. Minvielle. 'Robust Kernel Density Estimation with Median-of-Means principle'. In: Proceedings of the 39th International Conference on Machine Learning (ICML). Vol. 162. Proceedings of Machine Learning Research. Baltimore, United States, 17th July 2022. URL: https://hal.archives-ouvertes.fr/hal-02882092.

[18] Z. Li and G. Stoltz. 'Contextual Bandits with Knapsacks for a Conversion Model'. In: Thirty-sixth Conference on Neural Information Processing Systems. Vol. 35 (NeurIPS 2022). Advances in Neural Information Processing Systems. New Orleans, United States, 2022. URL: https://hal.archives-ouvertes.fr/hal-03683289.

[19] B. T. Nguyen, B. Thirion and S. Arlot. 'A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension'. In: NeurIPS 2022. Vol. 35. Advances in Neural Information Processing Systems. New Orleans, United States, 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03680792.

National peer-reviewed Conferences

[20] O. Coudray, C. Keribin and P. Pamphile. 'Convergence rates for Positive-Unlabeled learning under Selected At Random assumption: sensitivity analysis with respect to propensity'. In: CAp&RFIAP 2022 - Conférence sur l'Apprentissage automatique. Vannes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03738277.

Conferences without proceedings

[21] R. Coulaud, C. Keribin and G. Stoltz. 'One-Station-Ahead Forecasting of Dwell Time, Arrival Delay and Passenger Flows on Trains Equipped with Automatic Passenger Counting (APC) Device'. In: WCRR 2022 - World Congress on Railway Research. Birmingham, United Kingdom, 6th June 2022. URL: https://hal.archives-ouvertes.fr/hal-03835496.

Doctoral dissertations and habilitation theses

[22] Y. Amara-Ouali. 'Statistical modelling of electric vehicle charging behaviours'. Université Paris-Saclay, 22nd Sept. 2022. URL: https://theses.hal.science/tel-03850949.

[23] F. Antonazzo. 'Unsupervised learning of huge data sets with limited computed resources'. Université de Lille, 30th Sept. 2022. URL: https://theses.hal.science/tel-03846222.

[24] O. Coudray. 'A statistical point of view on fatigue criteria : from supervised classification to positive-unlabeled learning'. Université Paris-Saclay, 8th Dec. 2022. URL: https://theses.hal.science/tel-03934858.

[25] R. Coulaud. 'Modeling and forecasting of railway operations variables and passenger flows for dense traffic areas'. Université Paris-Saclay, 30th Nov. 2022. URL: https://theses.hal.science/tel-03934383.

[26] S. Gaucher. 'Contributions to stochastic bandits and link prediction problems'. Université Paris-Saclay, 27th June 2022. URL: https://theses.hal.science/tel-03727474.

[27] P. Lacroix. 'Contributions to variable selection in high-dimension and its uses in biology'. Université Paris-Saclay, 16th Dec. 2022. URL: https://theses.hal.science/tel-03940928.

[28] E. Lasalle. 'Contributions to statistical analysis of graph-structured data'. Université Paris-Saclay, 5th Dec. 2022. URL: https://theses.hal.science/tel-03941869.

[29] E. M. Saad. 'Contributions to Frugal Learning'. Université Paris-Saclay, 9th Dec. 2022. URL: https://theses.hal.science/tel-03940730.

Reports & preprints

[30] A. Antoniadis, S. Gaucher and Y. Goude. *Hierarchical transfer learning with applications for electricity load forecasting*. 10th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03429702.

[31] A. Barrier, A. Garivier and G. Stoltz. *On Best-Arm Identification with a Fixed Budget in Non-Parametric Multi-Armed Bandits*. 30th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03792668.

[32] C. Biernacki, J. Jacques and C. Keribin. *A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges*. 5th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03769727.

[33] K. Bleakley, C. Fortas, V. Duong, S. Ly, T. Cantaert, H. Sothy, D. Laurent, P. Buchy, P. Dussart and A. Sakuntabhai. *Multivariate evidence-based pediatric dengue severity prediction at hospital arrival*. 1st Dec. 2022. URL: https://hal.inria.fr/hal-03881145.

[34] O. Coudray, C. Keribin, P. Massart and P. Pamphile. *Risk bounds for PU learning under Selected At Random assumption*. 14th Jan. 2022. URL: https://hal.inria.fr/hal-03526889.

[35] S. Gaucher, N. Schreuder and E. Chzhen. *Fair learning with Wasserstein barycenters for non-decomposable performance measures*. 2nd Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03767640.

[36] H. Hadiji and G. Stoltz. *Adaptation to the Range in K-Armed Bandits*. 9th June 2022. URL: https://hal.archives-ouvertes.fr/hal-02794382.

[37] K. Hajjar and L. Chizat. *On the symmetries in the dynamics of wide two-layer neural networks*. 30th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03829400.

[38] G. Maillard. *Local asymptotics of cross-validation around the optimal model*. 30th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03263396.

[39] A. Nitanda, G. Chinot and T. Suzuki. *Gradient Descent can Learn Less Over-parameterized Two-layer Neural Networks on Classification Problems*. 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03874553.

[40] L. Pujol. *ISDE : Independence Structure Density Estimation*. 5th May 2022. URL: https://hal.archives-ouvertes.fr/hal-03401530.

[41] L. Pujol. *Nonparametric estimation of a multivariate density under Kullback-Leibler loss with ISDE*. 5th May 2022. URL: https://hal.archives-ouvertes.fr/hal-03660157.

[42] E. M. Saad and G. Blanchard. *Constant regret for sequence prediction with limited advice*. 4th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03797597.

[43] A. Sportisse, M. Marbac, C. Biernacki, C. Boyer, G. Celeux, F. Laporte and J. Josse. *Model-based Clustering with Missing Not At Random Data*. 2022. URL: https://hal.science/hal-03494674.

[44] S. Thépaut and N. Verzelen. *Optimal Estimation of Schatten Norms of a rectangular Matrix*. 2nd Dec. 2022. URL: https://hal.inria.fr/hal-03882348.

Other scientific publications

[45] O. Coudray, C. Keribin and P. Pamphile. 'Convergence rates for PU learning under the SAR assumption: influence of propensity'. In: CAp&RFIAP 2022 - Conférence sur l'Apprentissage automatique. Vannes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03738282.

11.3 Cited publications

[46] E. Baroux, B. Delattre, A. Constantinescu, P. Pamphile and I. Raoult. 'Analysis Of Real-Life Multi-Input Loading Histories For The Reliable Design Of Vehicle Chassis'. In: *Procedia Structural Integrity* 38 (2022), pp. 497–506.