RESEARCH CENTRE

Saclay - Île-de-France

IN PARTNERSHIP WITH:

CNRS, Université Paris-Saclay

2021 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, until Jul 2021]
- Evgenii Chzhen [CNRS, Researcher, from Oct 2021]
- 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 Fellows

- Pierre Andrieu [Univ Paris-Saclay, from Oct 2021]
- Evgenii Chzhen [Univ Paris-Saclay, until Sep 2021]
- Pierre Humbert [Inria, from Sep 2021]

PhD Students

- Yvenn Amara-Ouali [Univ Paris-Saclay]
- Filippo Antonazzo [Inria, co-advised with Modal team (Inria Lille)]
- Emilien Baroux [Groupe PSA]
- Antoine Barrier [ENS Lyon]
- Simon Briend [Univ Paris-Saclay, from Sep 2021, co-advised with Pompeu Fabra University]
- Samy Clementz [Univ Paris 1, from Sep 2021]
- Olivier Coudray [Groupe PSA]
- Remi Coulaud [SNCF, CIFRE]
- Solenne Gaucher [Univ Paris-Saclay]
- Karl Hajjar [Univ Paris-Saclay]
- Perrine Lacroix [Univ Paris-Saclay]
- Etienne Lasalle [Inria, co-advised with Datashape team]
- Timothee Mathieu [Ecole normale supérieure Paris-Saclay, until Aug 2021]

- Binh Nguyen [Inria, co-advised with Parietal team]
- Louis Pujol [Inria, from Nov 2021, co-advised with Datashape team]
- El Mehdi Saad [Univ Paris-Saclay, co-advised with Datashape team]
- Gayane Taturyan [IRT SystemX]

Interns and Apprentices

- Nawel Arab [Inria, from Apr 2021 until Sep 2021]
- Diane Iris Berenger [Inria, from Apr 2021 until Jul 2021]

Administrative Assistant

- Laurence Fontana [Inria]

External Collaborators

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

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, reliability, 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-scale 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, the imprudent use of machine learning algorithms may lead to unfair and discriminatory decisions on individuals, often inheriting or even amplifying data biases; such biases must be taken into account in order to build algorithms with fairness guarantees on their decisions. 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 to provide a precise theoretical analysis (for new and existing popular estimator selection procedures) that describes 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 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 fairness and accuracy of a learned model – more accurate models are less fair. One of our main research directions is to provide a theoretical study of these types of trade-offs. The goal is to provide user-friendly statistical quantification of such trade-offs and build statistically optimal algorithms in this context. A special attention will be paid to the online learning setting.

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, and outlier detection, to name but a few. We tackle this 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. Other 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 consumption 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 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 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, with 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

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. A CIFRE PhD is in progress on this topic, in collaboration with the SNCF (Rémi Coulaud).

5 Social and environmental responsibility

5.1 Footprint of research activities

As 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, 2021 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 2021.

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 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 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 from several direct collaborations with companies such as EDF (forecasting and control of electricity load consumption), SNCF (punctuality of trains in densely-populated regions, one Cifre contract ongoing) and Stellantis (automobile reliability, with two Cifre contracts ongoing).

6 Highlights of the year

6.1 Awards

- Margaux Brégère: “Coup de cœur du jury” in the Smart Grids 2021 Thesis Prizes
- Margaux Brégère: Finalist in the Amies Thesis Prize, 2021
- Evegnii Chzhen, with Nicolas Schreuder (Univ. of Genoa): runner-up best student paper award at UAI 2021 [15]

7 New results

7.1 Model selection criteria for the Latent Block Model

Participants: Christine Keribin, Diane-Iris Béranger.

The Latent Block Model (LBM) is a model-based approach for co-clustering: the simultaneous clustering of rows and columns of a data matrix. After results obtained last year on the consistency and asymptotic normality of the maximum likelihood estimator for LBM, we continued the theoretical study of this model: C. Keribin supervised M2 student Diane-Iris Béranger during a four month internship hosted by Inria-Celeste on model selection criteria for LBM. They extended to LBM the results of Wang et al (2017) on stochastic block models for clustering random graphs. They provided the asymptotic behavior of the likelihood ratio, and conducted numerical experiments to illustrate it in underestimation cases. Furthermore, they defined a penalty term which leads to a consistant criterion for model selection in LBM. However, the BIC penalty does not fulfill this condition, which is only sufficient. BIC consistency therefore remains a conjecture only.

7.2 Data analysis of bacterial interactions in complex ecosystems

Participants: Christine Keribin.

Datasets in biology are often faced with the problem of having many more variables than individuals or units. In the supervised context, well-known regularized methods involving lasso or ridge penalization bring answers. The ridge approach uniformly shrinks the estimates, while the lasso can be viewed as a variable selection method and is then often used for its ease of interpretation. However, variables in biology are often correlated and it could be better to select groups of variables rather than individual variables. The group-lasso is one method, but it requires knowing the group *a priori*. C. Keribin (in collaboration with Béatrice Laroche, INRAE- MaIAGE) has started to study the added value of using co-clustering as an unsupervised method to define a way of qualifying groups of variables and their interactions along with classification or regression tasks. Several use cases underlie this study, involving the determination of color of cheese with regards to bacteria and yeast composition of the flora, or the dynamics of bacteria flora in the presence or absence of a pathogen.

7.3 Fast rates for prediction with limited expert advice

Participants: E.M. Saad.

E.M. Saad, in collaboration with G. Blanchard, investigated in [9] the problem of minimizing the excess generalization error with respect to the best expert prediction in a finite family in the stochastic setting, under limited access to information. They assumed that the learner only has access to a limited number of expert advice per training round, as well as for prediction. Assuming that the loss function is Lipschitz and strongly convex, they showed that if we are allowed to see the advice of only one expert per round for T rounds in the training phase, or to use the advice of only one expert for prediction in the test phase, the worst-case excess risk is $\Omega(1/\sqrt{T})$ with probability lower bounded by a constant. However, if we are allowed to see at least two actively chosen experts' advice per training round and use at least two experts for prediction, a fast rate of $\mathcal{O}(1/T)$ can be achieved. We designed novel algorithms achieving this rate in this setting and that where the learner has a budget constraint on the total number of observed experts' advice, and gave precise instance-dependent bounds on the number of training rounds and queries needed to achieve a given generalization error precision.

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

Participants: Filippo Antonazzo, Christine Keribin.

Clustering is impacted by the regular increase in sample sizes, the latter providing opportunity to reveal previously-undiscoverable information. However, the sheer volume of data leads to issues related to computational resource requirements, not to mention high energy consumption. Resorting to binned data depending on an adaptive grid is expected to help with such green computing issues, while not overly 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 by F. Antonazzo and C. Keribin (in collaboration with C. Biernacki) in [16], with a numerical illustration of its advantages.

In a second step, F. Antonazzo and C. Keribin (in collaboration with C. Biernacki) focus on discovery of tiny but possible high-valued clusters which were not visible with more modest sample sizes. In this case, clustering is dependent on computational limits due to the high volume of data, possibly requiring extremely high memory and computation resources. In addition, the classical subsampling strategy, often adopted to overcome such limitations, is expected to fail to discover clusters in highly imbalanced cluster cases. Our proposal first consists in drastically compressing the data volume by only preserving its marginal-bin values, thus discarding the cross-bin ones. Despite this extreme information loss, we nevertheless prove an identifiability property for the diagonal mixture model, and also introduce a specific EM-like algorithm associated with a composite likelihood approach. The latter is much more frugal than a regular but unfeasible EM algorithm you would normally use on such marginal-bin data, while preserving all consistency properties. Finally, numerical experiments highlight that this proposed method outperforms subsampling both in controlled simulations and in various real applications where imbalanced clusters may typically appear, such as image segmentation, hazardous asteroid recognition, and fraud detection [19].

7.5 Optimality of variational inference for a stochastic block model with missing links

Participants: Solenne Gaucher.

Variational methods are extremely popular in the analysis of network data. Statistical guarantees obtained for these methods typically provide asymptotic normality for the problem of estimation of global

model parameters under the stochastic block model. In [24], S. Gaucher (in collaboration with O. Klopp) considers the case of networks with missing links, which is important in applications, and show that the variational approximation to the maximum likelihood estimator converges at the minimax rate. This provides the first minimax optimal and tractable estimator for the problem of parameter estimation for the stochastic block model with missing links. The theoretical results are complemented with numerical studies of simulated and real networks, which confirm the advantages of this estimator over other current methods.

7.6 Massive data in structural geology

Participants: Christine Keribin, Nawal Arab.

C. Keribin initiated a collaboration with Antonio Benedicto (GEOPS Paris-Saclay). Data in structural geology are essentially treated manually and machine learning could contribute to real advances in this field. To conduct an exploratory study, they co-supervised the five-month long M2-Datascience internship of the student Nawel Arab (supported by Inria-Celeste) and M2-Geology student Muchan Chai (supported by GEOPS). The goal was to integrate heterogeneous databases (mineralogy, structural data, geochemistry, lithology, radiometry) to determine types of structures and more generally features that could predict uranium fields. Heterogeneity (due to the types of variables as well as the location of the measures) and an extremely unbalanced situation (very few mineralized samples) made the learning tricky. An integrated database was built, and standard machine learning methods tested. While the preliminary results are not completely convincing for the moment, some interesting points have been identified to be studied further.

7.7 Localization in 1D non-parametric latent space models from pairwise affinities

Participants: Christophe Giraud.

In [25], C. Giraud (in collaboration with Y. Issartel and N. Verzelen) considers the problem of estimating latent positions in a one-dimensional torus from pairwise affinities. The observed affinity between a pair of items is modeled as a noisy observation of a function $f(x_i, x_j)$ of the latent positions x_i, x_j of the two items on the torus. The affinity function f is unknown, and it is only assumed to fulfill some shape constraints ensuring that $f(x, y)$ is large when the distance between x and y is small, and vice-versa. This non-parametric modeling offers good flexibility to fit data. An estimation procedure is introduced that provably locates all of the latent positions with a maximum error in the order of $\log(n)/n$, with high probability. This rate is proven to be minimax optimal. A computationally efficient variant of the procedure is also analyzed under some more restrictive assumptions. These general results can be used for the problem of statistical seriation, leading to new bounds for the maximum error in an ordering.

7.8 Algorithmic fairness

Participants: Evgenii Chzhen, Christophe Giraud, Gilles Stoltz.

Many decision problems are of a sequential nature, and efforts are needed to better handle fairness in such settings. In [8], E. Chzhen, C. Giraud, and G. Stoltz have introduced a unified approach to sequential fair learning in the presence of sensitive and non-sensitive contexts. The introduced approach translates the problem of fair learning as an approachability problem. It relies on Blackwell's approachability framework and, hence, it inherits the main appealing features of the Blackwell's result: a generic way to produce necessary and sufficient conditions when learning is possible, and a tractable algorithm in the

latter case. Using this framework, the authors provided several (im)possibility results unifying previous results, and obtaining brand new ones. In particular, the authors provide a complete description of the trade-off between the demographic parity constraint and the group-wise calibration performance criterion.

In [15], N. Schreuder (Univ. of Genoa) and E. Chzhen study the problem of binary classification under demographic parity constraint with controlled abstention rate. They provide an efficient post-processing algorithm and exhibit distribution-free fairness, abstention, and risk guarantees. An empirical study on real data demonstrates that a very low level of abstention suffices in order to bypass the trade-off between fairness and risk. Thus, this work proposes a third parameter, which, in the presence of humans in the loop, can lighten the burden of additional constraints. Furthermore, a `python` implementation is provided by the authors: here.

7.9 Classification with the F-score: minimax optimality

Participants: Evgenii Chzhen.

In [2], E. Chzhen studies the problem of binary classification with the F-score as the performance measure. We propose a post-processing algorithm for this problem which fits a threshold for any score base classifier to yield a high F-score. The post-processing step involves only unlabeled data and can be performed in logarithmic time. Finite sample post-processing bounds for the proposed procedure are derived. Furthermore, it is shown that the procedure is minimax rate optimal, when the underlying distribution satisfies classical nonparametric assumptions. This result improves upon previously known rates for the F-score classification and bridges the gap between standard classification risk and the F-score. Finally, the generalization of this approach to the set-valued classification is provided.

7.10 Set-valued classification: advantages of the semi-supervised approach

Participants: Evgenii Chzhen.

In collaboration with C. Denis and M. Hebiri [3], E. Chzhen studies study supervised and semi-supervised algorithms in the set-valued classification framework with controlled expected size. While the former can use only labeled samples, the latter are able to make use of additional unlabeled data. The theoretical analysis implies that if no further assumption is made, there is no supervised method that outperforms the semi-supervised estimator proposed in this work – the best achievable rate for any supervised method is parametric even if the margin assumption is extremely favorable; on the contrary, the developed semi-supervised estimator can achieve faster rate of convergence provided that sufficiently many unlabeled samples are available. We also show that under additional smoothness assumption, supervised methods are able to achieve faster rates and the unlabeled sample cannot improve the rate of convergence. Finally, a numerical study supports our theory and emphasizes the relevance of the theoretical assumptions from an empirical perspective.

7.11 Spatially relaxed inference on high-dimensional linear models

Participants: T.-B. Nguyen.

T.-B. Nguyen, in collaboration with J.-A. Chevalier, B. Thirion, and J. Salmon, considered the inference problem for high-dimensional linear models, when covariates have an underlying spatial organization reflected in their correlation. A typical example of such a setting is high-resolution imaging, in which neighboring pixels are usually very similar. Accurate point and confidence interval estimation is not

possible in this context with many more covariates than samples, not to mention high correlation between covariates. This calls for a reformulation of the statistical inference problem that takes into account the underlying spatial structure: if covariates are locally correlated, it is acceptable to detect them up to a given spatial uncertainty. They thus proposed to rely on the δ-FWER, that is, the probability of making a false discovery at a distance greater than δ from any true positive. With this target measure in mind, they studied the properties of ensemble clustered inference algorithms which combine three techniques: spatially constrained clustering, statistical inference, and ensembling to aggregate several clustered inference solutions. They showed that ensembled clustered inference algorithms control the δ-FWER under standard assumptions for δ equal to the largest cluster diameter. They complemented the theoretical analysis with empirical results, demonstrating accurate δ-FWER control and decent power achieved by such inference algorithms.

7.12 Training Integrable Parameterizations of Deep Neural Networks in the Infinite-Width Limit

Participants: Karl Hajjar, Christophe Giraud.

To theoretically understand the behavior of trained deep neural networks, it is necessary to study the dynamics induced by gradient methods from a random initialization. However, the nonlinear and compositional structure of these models make these dynamics difficult to analyze. To overcome these challenges, large-width asymptotics have recently emerged as a fruitful viewpoint and have led to practical insights on real-world deep networks. For two-layer neural networks, it has been understood via these asymptotics that the nature of the trained model radically changes depending on the scale of the initial random weights, ranging from a kernel regime (for large initial variance) to a feature learning regime (for small initial variance). For deeper networks, more regimes are possible. In [26], K. Hajjar, L. Chizat and C. Giraud study in detail a specific choice of a "small" initialization corresponding to "mean-field" limits of neural networks, called integrable parameterizations (IPs). First, they show that under a standard i.i.d. zero-mean initialization, integrable parameterizations of neural networks with more than four layers start at a stationary point in the infinite-width limit, and no learning occurs. Then, they propose various methods to avoid this kind of behavior, and analyze in detail the resulting dynamics. Theoretical results were confirmed by numerical experiments on image classification tasks.

7.13 Characteristics and Mortality Risk Factors for Covid ICU Patients in the French West Indies

Participants: Kevin Bleakley.

Guadeloupe, a French West Indies island, was heavily affected by the first two large Covid waves. The therapeutic approach taken was different for the two waves in the ICU. We aimed to compare the two different periods in terms of characteristics and outcomes, and to evaluate risk factors associated with 60-day mortality in our overall cohort.

Patients were treated during the first wave with a combination of Hydroxychloroquine and Azithromycin, and during the second wave with dexamethasone and reinforced anticoagulation. We found that overall mortality at day 60 was high (45%) and not different between the two waves. In patients under mechanical ventilation, risk factors associated with death in a multivariate analysis were a high number of comorbidities, a high SOFA (Sequential Organ Failure Assesment) score, and—surprisingly—a delay in starting invasive mechanical ventilation after admission to the ICU [7].

7.14 Hierarchical transfer learning with applications for electricity load forecasting

Participants: Solenne Gaucher.

The recent abundance of data on electricity consumption at different scales opens new challenges and highlights the need for new techniques to leverage information present at finer scales in order to improve forecasts at wider scales. In [34], S. Gaucher (in collaboration with A. Antoniadis and Y. Goude) takes advantage of the similarity between this hierarchical prediction problem and multi-scale transfer learning. They develop two methods for hierarchical transfer learning, based respectively on the stacking of generalized additive models and random forests, and on the use of aggregation of experts. They apply these methods to two problems of electricity load forecasting at a national scale, using smart meter data in the first case, and regional data in the second. For these two use cases, they compare the performances of their methods to that of benchmark algorithms, and investigate their behaviour using variable importance analysis. Their results demonstrate the interest of the two methods, both of which lead to a significant improvement in predictions.

7.15 Fatigue data-based design

Participants: Olivier Coudray, Christine Keribin, Patrick Pamphile, Pascal Massart.

Deterministic fatigue criteria are used to identify critical zones for fatigue failure of mechanical parts. While these criteria prove to be effective on experimental test data with standardized specimens, they are less effective for bench tests with prototypes: the variability inherent to tests on prototypes is poorly addressed by deterministic criteria, not to mention errors due to numerical simulations. We therefore propose to use statistical methods, (1) to improve the deterministic criteria; (2) to build new fatigue criteria [10]. While an observed failure during testing suggests the presence of a critical zone, the absence of failure does not necessarily imply a safe zone. We have related this setting to PU learning (learning from positive and unlabeled data); indeed, it can be seen as a semi-supervised situation, with a special case of label noise where only a fraction of the positive instances is labeled. Some results exist for when the labeling noise is constant (the selected completely at random assumption). We are interested in the case where the probability of being labeled may depend on the covariates (selection bias, selected at random assumption). In this context, we have provided upper and lower bounds on the minimax risk, proving that the upper bound on the excess risk is almost optimal. In addition, we have quantified the impact of label noise on PU learning compared to the standard classification setting. [23].

7.16 Analysis Of Real-Life Multi-Input Loading Histories For The Reliable Design Of Vehicle Chassis

Participants: Emilien Baroux, Patrick Pamphile.

In order to reliably design automotive structures, engineers must determine and justify validation conditions and levels. These must be derived from a thorough understanding of the structural damage induced by in-service loading conditions. Based on variable amplitudes and multiple input loading histories applied on car axles, E. Baroux and P. Pamphile (in collaboration with B. Delattre, A. Constantinescu, and I. Raoult) propose in [12] a multidimensional description of pseudo-damage for the design of weak points of a car chassis. A classification of drivers allowed us to identify driving profiles that are more or less damaging to the structure. This allows the design office to identify critical points of the structure specific to certain driving events (sharp turns, potholes, pavements, etc.) Moreover, we propose a damage reconstruction of a critical driving profile using track tests. The design office can then set up bench tests representative of custumer driving profiles.

8 Bilateral contracts and grants with industry

Participants: Christine Keribin, Patrick Pamphile, Gilles Stoltz.

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.
- G. STOLTZ: Ongoing contract with BNP Paribas (10 kE), on stochastic bandits under budget constraints, for an application to loan management.

9 Partnerships and cooperations

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

9.1 International initiatives

Christophe Giraud is part of the DFG/ANR PCRI project ASCAI ("Segmentation, clustering, et seriation actifs et passifs: vers des fondations unifiées en IA"), which is jointly lead by Alexandra Carpentier (Postdam University) and Nicolas Verzelen (INRA Montpellier)

9.2 National initiatives

9.2.1 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).

9.2.2 Other

Kevin Bleakley worked partially for IRT SystemX inside the Confiance.ai program.

10 Dissemination

Participants: Equipe Celeste.

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)

Member of the organizing committees

- Christophe Giraud is co-organizer with Estelle Kuhn of StatMathAppli Frejus, 2022
- Christophe Giraud is organizing a "statistical learning" session for the joint AMS/EMS/SMF conference, Grenoble 2022

10.1.2 Scientific events: selection

Chair of conference program committees

- Christine Keribin is VP of the scientific committee of Federated Learning Workshops, in collaboration with Owkin and Accenture Lab (two live-streamed events and one full day in person): workshop webpage.

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*
- 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, Colloquium, Department of Statistics and Actuarial Science, The University of Iowa, 18/03/2021
S. Arlot, International Conference on Statistics and Related Fields (ICON STARF), University of Luxembourg, 14/07/2021
S. Arlot, Séminaire MODAL'X, Université Paris Nanterre, 23/09/2021
S. Arlot, Séminaire de probabilités de Lyon, ENS Lyon, 25/11/2021
K. Bleakley, Parietal Inria Team seminar, 01/06/2021
E. Chzhen, DATAIA seminar, DATAIA, 22/06/2021
E. Chzhen, Séminaire Rennais de statistique, ENSAI Rennes, 05/10/2021
E. Chzhen, ML-MTP : Machine Learning in Montpellier, Theory & Practice, Université de Montpellier, 25/10/2021
C. Giraud, Weierstrass Insitute seminar, Berlin, 17/11/2021
C. Giraud, ETH Foundations of Data Science monthly seminar, ETH Zurich, 02/12/2021
C. Keribin, MHC2021 international conference, Université Paris-Saclay, 03/06/2021
C. Keribin, Working Group on Mode Based Clustering WGMBC2021, Athènes, 27/10/2021
C. Keribin, Séminaire Entropie-Mots-Stats, Caen, 19/11/2021
G. Stoltz, Séminaire de l'équipe MIA de l'AgroParisTech, 12/04/2021

10.1.5 Scientific expertise

- G. Stoltz: HCERES panel member for Laboratoire Raphaël Salem, Rouen

10.1.6 Research administration

- S. Arlot is a member of the board of the Computer Science Graduate School of University Paris-Saclay.
- S. Arlot is a member of the board of the Computer Science Doctoral School (ED STIC) of University Paris-Saclay.
- C. Giraud is a member of the Scientific Committe 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. Keribin is member of the board of the Computer Science Doctoral School (ED MSTIC) of Paris-Est Sup.
- M-A. Poursat is in charge of the bachelor program in computer science and mathematics of University Paris-Saclay

10.2 Teaching - Supervision - Juries

10.2.1 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.

- Masters: S. Arlot, Statistical learning and resampling, 30h, M2, Université Paris-Sud
- Masters: S. Arlot, Preparation for French mathematics agrégation (statistics), 50h, M2, Université Paris-Sud
- 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: M-A Poursat, applied statistics, 21h, in M1 Artificial Intelligence and statistical learning, 42h, in M2 Bioinformatics Université paris-Saclay
- Licence: M-A Poursat, statistical inference, 72h, in LDD3 Université Paris Saclay
- Masters: G. Stoltz, Sequential Learning and Optimization, 18h, M2 Université Paris-Saclay

10.2.2 Supervision

- PhD defended on Jan. 2021: Timothée Mathieu, M-estimation and Median of Means applied to statistical learning, started Sept. 2018, co-advised by G. Lecué (ENSAE) and M. Lerasle
- PhD defended on Aug. 2021: Vincent Divol, Contributions to geometric inference on manifolds and to the statistical study of persistence diagrams, started Oct. 2018, co-advised by F. Chazal (INRIA Datashape) and P. Massart
- PhD defended on Dec. 2021: Tuan-Binh Nguyen, Efficient Statistical Testing for high-dimensional Models, started Oct. 2018, co-advised by S. Arlot and B. Thirion (INRIA Parietal)

- PhD in progress: Solenne Gaucher, Sequential learning in random networks, started Sept. 2018, C. Giraud.
- PhD in progress: 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 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 (INRIA Datashape)
- PhD in progress: Perrine Lacroix, High-dimensional linear regression applied to gene interactions network inference, started Sept. 2019, co-advised by P. Massart & M.-L. Martin-Magniette (INRAE)
- PhD in progress: 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 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: Louis Pujol, CYTOPART - Flow cytometry data clustering, started Nov. 2019, co-advised by P. Massart and M. Glisse (INRIA Datashape)
- 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
- 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, C. Giraud and L. Chizat.
- 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: Simon Briend, Statistical theory for overparametrized deep neural networks, started Sept. 2021, co-advised by C. Giraud and G. Lugosi
- 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)

10.2.3 Juries

- S. Arlot: member of the HdR committee of Mohamed Hebiri, Université Gustave Eiffel, 07/01/2021.
- S. Arlot: referee for the PhD of TrungTin Nguyen, Université de Caen Normandie, 14/12/2021.
- C. Keribin: referee for the PhD of Gabriel Frisch, Université de Technologie de Compiègne, 18/10/2021
- C. Keribin: member of the PhD jury for Rem-Sophia Mouradi, Institut National Polytechnique de Toulouse, 16/03/2021
- G. Stoltz: referee for the PhD of David Obst, Aix-Marseille Université, 10/12/2021.
- G. Stoltz: president of the PhD jury for Rémy Garnier, Université de Cergy, 08/12/2021.

10.3 Popularization

10.3.1 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). The exhibition was inaugurated at MMI Lyon in September 2021.

11 Scientific production

11.1 Publications of the year

International journals

[1] I. Bournaud and P. Pamphile. 'Un dispositif d'accompagnement dans la transition lycée-université (IUT) : enjeux et effets'. In: *Revue internationale de pédagogie de l'enseignement supérieur* 37.2 (Mar. 2021). URL: https://halshs.archives-ouvertes.fr/halshs-03174559.

[2] E. Chzhen. 'Optimal Rates for Nonparametric F-Score Binary Classification via Post-Processing'. In: *Mathematical Methods of Statistics* (7th Sept. 2021). URL: https://hal-upec-upem.archives-ouvertes.fr/hal-02123314.

[3] E. Chzhen, C. Denis and M. Hebiri. 'Minimax semi-supervised set-valued approach to multi-class classification'. In: *Bernoulli* (21st Nov. 2021). URL: https://hal-upec-upem.archives-ouvertes.fr/hal-02112918.

[4] C. Keribin. 'Cluster or co-cluster the nodes of oriented graphs?' In: *Journal de la Société Française de Statistique* 162.1 (2021), p. 24. URL: https://hal.inria.fr/hal-03139333.

[5] 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.

[6] S. Minsker and T. Mathieu. 'Excess risk bounds in robust empirical risk minimization'. In: *Information and Inference* (23rd Apr. 2021). URL: https://hal.archives-ouvertes.fr/hal-02390397.

[7] J.-D. Pommier, F. Martino, K. Bleakley, L. Flurin, F. V. Roy, M. Carles, M. Valette and L. Camous. 'A Tale of a Two Waves Epidemic: Characteristics and Mortality Risk Factors for COVID-19 ICU Patients in the French West Indies'. In: *Journal of Biology and Today's World* (5th Apr. 2021). DOI: 10.21203/rs.3.rs-200243/v1. URL: https://hal.inria.fr/hal-03536693.

International peer-reviewed conferences

[8] E. Chzhen, C. Giraud and G. Stoltz. 'A Unified Approach to Fair Online Learning via Blackwell Approachability'. In: 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Vol. 34. Advances in Neural Information Processing Systems. Virtual conference, Australia, 2021. URL: https://hal.archives-ouvertes.fr/hal-03268279.

[9] E. M. Saad and G. Blanchard. 'Fast rates for prediction with limited expert advice'. In: Neural Information Processing Systems (NeurIPS) 2021. Vol. 34. NeurIPS 2021 proceedings. [Online], United States, 6th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03405899.

National peer-reviewed Conferences

[10] O. Coudray, P. Bristiel, M. Dinis, C. Keribin and P. Pamphile. 'Fatigue Data-Based Design: statistical methods for the identification of critical zones'. In: SIA Simulation Numérique. Online, France, 7th Apr. 2021. URL: https://hal.inria.fr/hal-03483277.

Conferences without proceedings

[11] F. Antonazzo, C. Biernacki and C. Keribin. 'Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach'. In: Working Group - Model-based Clustering. Athens, Greece, 25th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505673.

[12] 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: Fatigue Design 2021. SENLIS, France, 17th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03507704.

[13] C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. M. Lourdelle and A. Sportisse. 'Dealing with missing data in model-based clustering through a MNAR model'. In: The 14th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Zakopane, Poland, 11th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03505659.

[14] C. Biernacki, C. Boyer, G. Celeux, J. Josse, F. Laporte, M. Marbac Lourdelle, A. Sportisse and V. Vandewalle. 'Impact of Missing Data on Mixtures and Clustering'. In: MHC2021 - Mixtures, Hidden Markov Models, Clustering. Orsay, France, 2nd June 2021. URL: https://hal.archives-ouvertes.fr/hal-03505664.

[15] N. Schreuder and E. Chzhen. 'Classification with abstention but without disparities'. In: Conference on Uncertainty in Artificial Intelligence (UAI 2021). Virtual conference, Australia, 27th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03152091.

Scientific book chapters

[16] F. Antonazzo, C. Biernacki and C. Keribin. 'A binned technique for scalable model-based clustering on huge datasets'. In: *Book of Short Papers of the 5th international workshop on Models and Learning for Clustering and Classification MBC2 2020, Catania, Italy*. 26th Oct. 2021, pp. 11–16. URL: https://hal.archives-ouvertes.fr/hal-03097284.

Reports & preprints

[17] Y. Amara-Ouali, M. Fasiolo, Y. Goude and H. Yan. *Daily peak electrical load forecasting with a multi-resolution approach*. 7th Dec. 2021. URL: https://hal.inria.fr/hal-03469721.

[18] Y. Amara-Ouali, Y. Goude, P. Massart, J.-M. Poggi and H. Yan. *A review of electric vehicle load open data and models*. 2nd Apr. 2021. URL: https://hal.inria.fr/hal-03028375.

[19] F. Antonazzo, C. Biernacki and C. Keribin. *Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach*. 17th Dec. 2021. URL: https://hal.archives-ouvertes.fr/hal-03485364.

[20] A. Antoniadis, S. Gaucher and Y. Goude. *Hierarchical transfer learning with applications for electricity load forecasting*. 19th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03429702.

[21] 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 datasets*. 6th Dec. 2021. URL: https://hal.uvsq.fr/hal-03104488.

[22] E. Chzhen, C. Denis, M. Hebiri and T. Lorieul. *Set-valued classification – overview via a unified framework*. 1st Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03154625.

[23] 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.

[24] S. Gaucher and O. Klopp. *Optimality of variational inference for stochastic block model with missing links*. 4th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03393160.

[25] C. Giraud, Y. Issartel and N. Verzelen. *Localization in 1D non-parametric latent space models from pairwise affinities*. 6th Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03317631.

[26] K. Hajjar, L. Chizat and C. Giraud. *Training Integrable Parameterizations of Deep Neural Networks in the Infinite-Width Limit*. 20th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03388815.

[27] P. Lacroix, M. Gallopin and M.-L. Martin. *A comprehensive review of variable selection in high-dimensional regression for molecular biology*. 5th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03366851.

[28] E. Lasalle. *Heat diffusion distance processes: a statistically founded method to analyze graph data sets*. 5th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03366848.

[29] G. Maillard. *Aggregated hold out for sparse linear regression with a robust loss function*. 4th May 2021. URL: https://hal.archives-ouvertes.fr/hal-02485694.

[30] P. Pamphile and G. Celeux. *Bayesian Estimations for Weibull Competing Risk Model with Masked Causes and Heavily Censored Data*. 27th Aug. 2021. URL: https://hal.inria.fr/hal-02410489.

[31] L. Pujol. *ISDE : Independence Structure Density Estimation*. 12th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03401530.

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

11.2 Other

Educational activities

[33] F. Antonazzo, C. Biernacki and C. Keribin. 'Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach'. Doctoral. France, 29th Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03505670.

11.3 Cited publications

[34] S. Gaucher, Y. Goude and A. Antoniadis. 'Hierarchical transfer learning with applications for electricity load forecasting'. In: *arXiv preprint arXiv:2111.08512* (2021).