StatMathAppli - Fréjus

September 1-- 5, 2025

Scientific Program

Monday 01/09

09:00 - 10:30 : John Duchi – Lecture 1

11:00 - 12:20 : Contributed session – Privacy and Robustness

- Alexander Kent. Rate optimality and phase transition for user-level local differential privacy.
- Màté Kormos. Private double robust inference.
- Thomas Michel. DP-SPRT: differentially private sequential probability ratio tests.
- Richard Schwank. Robust score matching.

15:30 - 17:30 : Po-Ling Loh – Lecture 1

17:30 - 19:00 : Stéphane Robin – Some latent variable models in ecology

Tuesday 02/09

09:00 - 10:30 : Po-Ling Loh – Lecture 2

11:00 - 12:30 : John Duchi – Lecture 2

15:30 - 17:00 : Mathilde Mougeot – Interplay between data, physics and simulation models with a link to industrial use-cases

Poster Session.

17:30 - 18:00: Flash Talks. (2min)

18:15 - 19:15 Poster Discussion

Wednesday 03/09

09:00 - 10:30 : John Duchi – Lecture 3

11:00 - 13:00 : Po-Ling Loh – Lecture 3

Thursday 04/09

09:00 - 10:30 : Po-Ling Loh – Lecture 4

11:00 - 12:20 Contributed session – Robustness and optimisation

- Renaud Gaucher. A unified breakdown analysis for Byzantine robust Gossip.
- Laurentiu Marchis. On the benefits of accelerated optimization in robust and private estimation.
- El Mehdi Saad. New lower bounds for stochastic non-convex optimization through divergence decomposition.
- Laura Hucker. Comparing regularisation paths of (conjugate) gradient estimators in ridge regression.

15:30 - 17:00 : John Duchi – Lecture 4

17:30 - 18:50 : Contributed session – Applied statistics

- Matthieu Texier. Combining mixture models and Markov chains to explore spatio-temporal dynamics of child wasting in southern Madagascar.
- Tom Rohmer. Improving genetic parameter estimation for dependent traits under selection.
- Marina Gomtsyan. Variable selection in specific regression for count time series
- Christophe Ley. A versatile trivariate wrapped Cauchy copula with applications to toroidal and cylindrical data.

Friday 05/09

09:20 - 10:20 : Contributed session – Unsupervised learning

- Bertrand Even. Computational lower bounds for latent variables: clustering, sparse clustering and biclustering.
- Victor Thuot. Clustering items through bandit feedback: finding the right feature out of many
- Ibrahim Kaddouri. Clustering in slowly mixing Gaussian hidden Markov models.

10:40 - 11:40 : Contributed session – Statistical inference

- Paul Rognon-Vael. Improving variable selection properties by using external data.
- Sophia Loizidou. Optimal tests for symmetry on the torus.
- Philippe Berthet. Some recent applications of Gaussian couplings in empirical process theory.

Book of Abstracts

Monday 01/09. 11:00 - 12:20

Contributed session – Privacy and Robustness

- Alexander Kent. Rate optimality and phase transition for user-level local differential privacy.
- Màté Kormos. Private double robust inference.
- Thomas Michel. DP-SPRT: differentially private sequential probability ratio tests.
- Richard Schwank. Robust score matching.

Rate Optimality and Phase Transition for User-Level Local Differential Privacy

Alexander Kent[1] & Thomas B. Berrett[2] & Yi Yu[3]

Department of Statistics, University of Warwick, UK
[1] *alexander.kent@warwick.ac.uk*
[2] *tom.berrett@warwick.ac.uk*
[3] *yi.yu.2@warwick.ac.uk*

Given demands for rigorous data privacy guarantees from both a regulatory standpoint and from the concerns of the data subjects, definitions of privacy which can be theoretically validated are of great interest. One such method enjoying significant popularity in both academia and industry is that of *differential privacy* in which carefully calibrated noise is added to data to provide plausible deniability as to the true value. Differential privacy appears in both the *central model*, where a trusted aggregator has access to the data and releases a privatised output, and the *local model*, where each user adds noise before publishing their (now privatised) data to a potentially untrusted aggregator.

Referring to the traditional setting where each of the n data subjects hold a single data point as *item-level* privacy, a growing field of interest is that of *user-level* privacy where each of the n users holds T observations and wishes to maintain the privacy of their entire collection. We consider the model of *user-level local differential privacy*, which is relatively unexplored. Indeed, even for a problem as fundamental as univariate mean estimation, prior to this work the minimax rate of estimation was undetermined.

We aim to fill this gap, obtaining minimax optimal estimation rates for a range of canonical statistical estimation problems including univariate and multidimensional mean estimation, sparse mean estimation, and non-parametric density estimation. We first derive a general minimax lower bound, which shows that the risk cannot, in general, be made to vanish for a fixed number of users even when $T \to \infty$. We then derive matching, up to logarithmic factors, lower and upper bounds for the aforementioned canonical problems. In particular, with other model parameters held fixed, we observe phase transition phenomena in the minimax rates as T, the number of observations each user holds, varies.

In the case of (non-sparse) mean estimation and density estimation, we see that, for T below a phase transition boundary, the rate is the same as having nT users in the item-level setting. Different behaviour is however observed in the case of s-sparse d-dimensional mean estimation, wherein consistent estimation is impossible when d exceeds the number of observations in the item-level setting, but is possible in the user-level setting when $T \geq s \log(d)$, up to logarithmic factors.

Private Double Robust Inference

Máté Kormos [1] & Aad van der Vaart [2]

[1] *Department of Mathematics, Computer Science and Statistics, Universiteit Gent*
mate.kormos@ugent.be
[2] *Delft Institute of Applied Mathematics, TU Delft*
a.w.vandervaart@tudelft.nl

Privacy mechanisms preserve the privacy of individuals in a sample by injecting noise into their sensitive data in a controlled manner, revealing only the noisy, privatised data to the statistician for inference purposes. The inference of a parameter exhibits a rate double robustness property when the large-sample bias of an estimator of the parameter is characterised by the product of the estimation errors of two other, auxiliary (or nuisance), often infinite-dimensional, parameters. We propose a novel class of rate double robust parameters whose novelty lies in the potentially nonlinear but smooth dependence on a low-dimensional regression parameter. Among others, this includes average treatment effects. We show that the properties of the sensitive-data model carry over to the privatised-data model by a suitable choice of the privacy mechanism, which, in general, means a total-variationally private mechanism. Specifically, the double robustness property is retained. Hence, a correct parametric assumption about one nuisance parameter affords more flexible modelling and slower estimation of the other one, as is the well-known case in the nonprivate setting. To enable efficient estimation from the privatised sample, we leverage this by casting the estimation of the nuisance parameters as optimisation problems which translate directly to the privatised setting, and by developing a private method of moments estimator for parametric models.

DP-SPRT: Differentially Private Sequential Probability Ratio Tests

Thomas Michel & Debabrota Basu & Emilie Kaufmann

Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL
Lille, France
firstname.lastname@inria.fr

We revisit Wald's celebrated Sequential Probability Ratio Test for sequential tests of two simple hypotheses, under privacy constraints. We propose DP-SPRT, a wrapper that can be calibrated to achieve desired error probabilities and privacy constraints, addressing a significant gap in previous work. DP-SPRT relies on a private mechanism that processes a sequence of queries and stops after privately determining when the query results fall outside a predefined interval. This OutsideInterval mechanism improves upon naive composition of existing techniques like AboveThreshold, potentially benefiting other sequential algorithms. We prove generic upper bounds on the error and sample complexity of DP-SPRT that can accommodate various noise distributions based on the practitioner's privacy needs. We exemplify them in two settings: Laplace noise (pure Differential Privacy) and Gaussian noise (Rényi differential privacy). In the former setting, by providing a lower bound on the sample complexity of any ϵ-DP test with prescribed type I and type II errors, we show that DP-SPRT is near optimal when both errors are small and the two hypotheses are close. Moreover, we conduct an experimental study revealing its good practical performance.

Robust Score Matching

Richard Schwank [1] & Andrew McCormack [2] & Mathias Drton [1,3]

[1] *TU Munich, Germany*
richard.schwank@tum.de
[2] *University of Alberta, Canada*
mccorma2@ualberta.ca
[3] *Munich Center for Machine Learning, Germany*
mathias.drton@tum.de

Score matching is a parameter estimation procedure that does not require computation of distributional normalizing constants. In this work we utilize the geometric median of means to develop a robust score matching procedure that yields consistent parameter estimates in settings where the observed data has been contaminated. A special appeal of the proposed method is that it retains convexity in exponential family models. The new method is therefore particularly attractive for non-Gaussian, exponential family graphical models where evaluation of normalizing constants is intractable. Support recovery guarantees for such models when contamination is present are provided. Additionally, support recovery is studied in numerical experiments and on a precipitation dataset. We demonstrate that the proposed robust score matching estimator performs comparably to the standard score matching estimator when no contamination is present but greatly outperforms this estimator in a setting with contamination.

Tuesday 05/09. 17:30 - 19:15

Poster Session

17:30 - 18:00 : Flash Talks (2min)
18:15 - 19:15 : Poster Discussion

- Alberto Bordino: Density ratio permutation tests with connections to distributional shifts and conditional two-sample testing
- Maxim Fedotov: Projected and updated L^0 criteria for variable selection in regression models
- Jean-Baptiste Fermanian: Uncertainty reduction of class conditional conformal prediction via multi-inputs aggregation
- Maximilian Graf: Statistical inference for paired spatial Poisson processes with missing data
- Harold Guéneau: Money laundering detection: financial time series representation learning with a transformer by contrastive learning
- Anton Kutsenko: Complete tail asymptotics for branching processes
- Félix Laplante: Uniform nonparametric confidence bands for random cumulative distribution functions
- Romain Périer: Post hoc bounds for heterogeneous data
- Sylvain Procope-Mamert: Iterative forward scheme to construct proposals for sequential Monte Carlo Algorithms
- Paul Rosa: On L^2-posterior contraction rates in Bayesian nonparametric regression models
- Vincent Runge: DUST; a duality-based pruning method for exact multiple change-point detection
- Henning Stein: Gentle measurements of quantum states
- William Underwood: Model upgrading in survival analysis

Density Ratio Permutation Tests with connections to distributional shifts and conditional two-sample testing

Alberto Bordino [1,†] & Thomas B.Berrett [1]

[1] *Department of Statistics, University of Warwick, Coventry, UK*
[†] *alberto.bordino@warwick.ac.uk*

We introduce novel hypothesis tests to allow for statistical inference for density ratios. More precisely, we introduce the Density Ratio Permutation Test (DRPT) for testing $H_0 : g \propto rf$ based on independent data drawn from distributions with densities f and g, where the hypothesised density ratio r is a fixed function. The proposed test employs an efficient Markov Chain Monte Carlo algorithm to draw permutations of the combined dataset according to a distribution determined by r, producing exchangeable versions of the whole sample and thereby establishing finite-sample validity. Regarding the test's behaviour under the alternative hypothesis, we begin by demonstrating that if the test statistic is chosen as an Integral Probability Metric (IPM), the DRPT is consistent under mild assumptions on the function class that defines the IPM. We then narrow our focus to the setting where the function class is a Reproducing Kernel Hilbert Space, and introduce a generalisation of the classical Maximum Mean Discrepancy (MMD), which we term Shifted-MMD. For continuous data, assuming that a normalised version of $g - rf$ lies in a Sobolev ball, we establish the minimax optimality of the DRPT based on the Shifted-MMD. For discrete data with finite support, we characterise the complex permutation sampling distribution using a noncentral hypergeometric distribution, significantly reducing computational costs. We further extend our approach to scenarios with an unknown shift factor r, estimating it from part of the data using Density Ratio Estimation techniques, and derive Type-I error bounds based on estimation error. Additionally, we demonstrate how the DRPT can be adapted for conditional two-sample testing, establishing it as a versatile tool for assessing modelling assumptions on importance weights, covariate shifts and related scenarios, which frequently arise in contexts such as transfer learning and causal inference. Finally, we validate our theoretical findings through experiments on both simulated and real-world datasets.

Projected and Updated L0 Criteria for Variable Selection in Regression Models

Maxim Fedotov [1,*] & Gábor Lugosi [1,2] & David Rossell [1]

[1] *Department of Economics and Business, Universitat Pompeu Fabra*
[2] *ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain*
[*] *corresponding author: maxim.fedotov@upf.edu*

We propose Projected and Updated L0 Criteria for variable selection in regression models which attain good theoretical guarantees and are easier to compute than criteria based on M-estimators.

We show that the proposed approximate criteria are selection consistent under mild conditions whenever the corresponding exact criteria are selection consistent. Moreover, we argue that Projected and Updated L0 Criteria are not worse asymptotically than the standard L0 Criteria based on M-estimators, when the number of covariates is fixed and the sample size grows. These results hold for a fairly general family of regression models, including Generalized Linear Models as a special case.

In addition, we provide finite sample exponential upper bounds for the probability of mistake, i.e. selecting a wrong subset of covariates, in the classic framework of additive models with Gaussian noise. These bounds hold for a given dimensionality of the covariate space and any given number of samples exceeding a sufficient sample size. These results find a particularly appealing application in high-dimensional regimes where the number of covariates greatly exceeds the number of observations.

We argue that in the case of additive models Projected L0 criteria allow to rewrite the variable selection problem as a binary quadratic problem for which there exist practical solvers that brutally speed up computation.

These results provide a solid theoretical ground for establishing approximations for a wide range of L0 methods used for variable selection, including posterior model probabilities, information criteria like AIC, BIC, EBIC and so on.

Uncertainty reduction of class conditional conformal prediction via multi-inputs aggregation

Jean-Baptiste Fermanian [1] & Mohamed Hebiri [2] & Joseph Salmon [3]

[1] *Université de Montpellier - Inria*
jean-baptiste.fermanian@inria.fr
[2] *Université Gustave Eiffel*
mohamed.hebiri@univ-eiffel.fr
[3] *Université de Montpellier - Inria*
joseph.salmon@umontpellier.fr

Conformal prediction methods are statistical tools designed to quantify uncertainty and generate predictive sets with guaranteed coverage probabilities. This work introduces a refinement to these methods for classification tasks, specifically tailored for scenarios where multiple observations (multi-inputs) of a single instance are available at prediction time. Our approach is particularly motivated by applications in citizen science, where multiple images of the same plant or animal are captured by individuals. Our method integrates the information from each observation into conformal prediction, enabling a reduction in the size of the predicted label set while preserving the required class-conditional coverage guarantee. The approach is based on the aggregation of conformal p-values computed from each observation of a multi-input. By exploiting the exact distribution of these p-values, we propose a general aggregation framework using an abstract scoring function, encompassing many classical statistical tools. Knowledge of this distribution also enables refined versions of standard strategies, such as majority voting. We evaluate our method on simulated and real data, with a particular focus on Pl@ntNet, a prominent citizen science platform that facilitates the collection and identification of plant species through user-submitted images.

Statistical Inference for Paired Spatial Poisson Processes with Missing Data

Alexandra Carpentier [1] & **Maximilian Graf** [2] & Axel Munk [3] & Giacomo Nies [4]

[1] *Universität Potsdam*
[2] *Universität Potsdam*
graf9@uni-potsdam.de
[3] *Georg-August-Universität Göttingen*
[4] *Georg-August-Universität Göttingen*

We consider a model in which two types of points, such as different types of proteins, appear as spatially close pairs on a bounded domain. Each type forms an i.i.d. sample of size $N \sim \text{Poisson}(\lambda)$, yielding a paired Poisson point process. We study the problem of recovering the underlying pairs, including an extension of the model where some points are unobserved due to missing data.

We propose a procedure for reconstructing the true pairings from the observed data and analyze its statistical properties. In particular, we establish bounds on the expected number of mismatches under suitable assumptions on the distribution of the observations.

Money laundering detection: Financial time series representation learning with a transformer by contrastive learning

Harold Guéneau [1] & Alain Celisse [2] & Pascal Delange [3]

[1] *Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, Paris, France*
harold.gueneau@proton.me
[2] *Laboratoire SAMM, Université Paris 1 Panthéon-Sorbonne, Paris, France*
alain.celisse@univ-paris1.fr
[3] *Marble*
pascal@checkmarble.com

Money laundering is a major global issue, accounting for around 2 to 5% of the world's GDP annually, and regulators require financial institutions to combat it. However, legacy systems rely on rule-based approaches that are resource-intensive and inefficient, leading to a very high false-positive rate. The proportion of false positives in such systems can be between 95 to 98%. Therefore, there is opportunity for improvement in this area.

The aim of this work is to proposes a novel two-steps methodology for money laundering detection, leveraging temporal data with a transformer. The first phase is focused on developing robust representations, independent from the available fraud labels, through contrastive learning. The second phase uses these representations to generate a money-laundering scoring. This work then introduces a two-threshold approach, calibrated following the Benjamini-Hochberg procedure to ensure a controlled false-positive rate. Experiments show that the transformer is able to produce general representations that capture money-laundering patterns with minimal supervision from domain experts. Moreover, it shows how to leverage these representations to ensure a lower and controlled false-positive rate for further financial investigations.

Beyond the baseline setup, the present work also explores the improvement of positive and negative examples sampling in the contrastive learning phase described previously, with the goal of enhancing the representation quality. It introduces different noising procedures to improve the sampling diversity, including generative modeling techniques such as score-based diffusion models.

Complete tail asymptotics for branching processes

Anton A. Kutsenko [1]

[1] *University of Hamburg*
akucenko@gamail.com

For the density of the martingale limit of the Galton-Watson process, we give a complete tail asymptotic series. We discuss the conditions when the series converges everywhere, and the connections with holomorphic dynamics: The frequencies of oscillatory terms in the series form fractal structures in the complex plane. Some of the results are published in Journal of Statistical physics and in Journal of Fourier analysis and applications.

Uniform nonparametric confidence bands for random cumulative distribution functions

Application to dynamic prediction in joint modeling in survival analysis

Félix LAPLANTE [1] & Christophe AMBROISE [2] & Estelle KUHN [3]

[1] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.*
felixlaplante0@gmail.com
[2] *Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry-Courcouronnes, France.*
christophe.ambroise@univ-evry.fr
[3] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France.*
estelle.kuhn@inrae.fr

The ability to draw accurate dynamic outcome predictions from joint modeling of both longitudinal and time-to-event data constitutes a critical goal that has received considerable interest in recent years. Although the specific nature of events of interest may be diverse, the linear or nonlinear joint model typically targets patient-specific conditional survival probabilities of an individual remaining event-free beyond a prespecified horizon, conditioned on the survival up to the prediction time and on all biomarker and covariate measurements observed to the prediction date.

We present a comprehensive set of nonparametric methods that enable the transition from isolated pointwise confidence intervals to simultaneous confidence tubes that uniformly envelope the whole survival-probability curve. Such tubes permit rigorous control of family-wise error rates and yield valid error bounds for derived quantities such as survival quantiles or restricted mean survival time. Optimality of proposed constructions can also be shown in restricted cases. Applications include joint modeling, non time-homogeneous frailty models, ...

Post hoc bounds for heterogeneous data

Romain Périer [1,*] & Gilles Blanchard [1] & Sebastian Döhler [2] & Guillermo Durand [1] & Étienne Roquain [3]

[1] *Université Paris-Saclay, France*
[2] *Sorbonne Université, France*
[3] *Hochschule Darmstadt, Germany*
[*] *romain.perier@universite-paris-saclay.fr*

Ce travail revisite la construction de bornes post hoc pour les faux positifs dans un contexte de tests multiples, du point de vue des tests hétérogènes, qui apparaissent naturellement dans le cadre de tests discrets. Une grande littérature a récemment émergé autour de la construction de telles bornes, avec notamment la méta-méthode de l'interpolation sur des familles de référence. En tant que méta-méthode, elle est très flexible et permet de multiples constructions. Les constructions précédemment proposées sont cependant conçues pour le "pire cas possible", à savoir des p-valeurs homogènes de distribution uniforme sous l'hypothèse nulle, et lorsque que les distributions des p-valeurs sous la nulle sont hétérogènes super-uniformes, ces constructions peuvent induire une perte de capacité à détecter la présence de signal. Comme, dans ce contexte, les distributions sous la nulle sont connues, nous proposons ici d'exploiter cette information en modifiant certaines constructions existantes afin d'incorporer la connaissance de ces distributions et d'atteindre une meilleure puissance que les constructions agnostiques.

Iterated forward scheme to construct proposals for sequential Monte Carlo algorithms

Nicolas Chopin [1] & Maud Delattre [2] & Guillaume Kon Kam King [3] & Sylvain Procope-Mamert [4]

[1] *ENSAE-CREST, Institut Polytechnique de Paris, France, nicolas.chopin@ensae.fr*
[2] *Université Paris-Saclay, INRAE, MaIAGE, France, maud.delattre@inrae.fr*
[3] *Université Paris-Saclay, INRAE, MaIAGE, France, guillaume.konkamking@inrae.fr*
[4] *Université Paris-Saclay, INRAE, MaIAGE, France, sylvain.procope-mamert@inrae.fr*

Sequential Monte Carlo is a powerful set of algorithms used to sample from a sequence of distributions. It is useful notably for Bayesian inference with different types of models and real data applications. In particular, when we try to recover a hidden signal from sequentially produced data with state-space models, the canonically defined proposals known as the bootstrap particle filter are rarely well-behaved and need extra work to be turned into useful sampling algorithms. Previous works on iterated methods for the automated construction of sequential Monte Carlo proposals, which were based on a backward scheme, have shown how to gradually improve proposals to reach a global optimality criterion, but they require a good initial proposal and cannot be used online.

We introduce a forward scheme that bridges between local and global optimality. This scheme can be used online or as an initialization for a backward scheme. We implemented this scheme to perform inference of nonlinear models on different simulated and real data from the literature. On moderately challenging cases, we obtain results with performances comparable to state-of-the-art backward schemes. On more challenging cases, our forward scheme outperforms the backward schemes, showing results more robust to a lack of good initialization.

On L^2-posterior contraction rates in Bayesian nonparametric regression models

Paul Rosa

Statistical Laboratory, University of Cambridge
pfr25@cam.ac.uk

The nonparametric regression model with normal errors has been extensively studied, both from the frequentist and Bayesian viewpoint. A central result in Bayesian nonparametrics is that under assumptions on the prior, the data-generating distribution (assuming a true frequentist model) and a semi-metric $d(.,.)$ on the space of regression functions that satisfy the so called testing condition, the posterior contracts around the true distribution with respect to $d(.,.)$, and the rate of contraction can be estimated. In the regression setting, the semi-metric $d(.,.)$ is often taken to be the Hellinger distance or the empirical L^2 norm (i.e., the L^2 norm with respect to the empirical distribution of the design) in the present regression context. However, extending contraction rates to the "integrated" L^2 norm usually requires more work, and has previously been done for instance under sufficient smoothness or boundedness assumptions, which may not necessarily hold. In this work we show that, for priors based on truncated random basis expansions and in the random design setting, a high probability two sided inequality between the empirical L^2 norm and the integrated L^2 norm holds in appropriate spaces of functions of low frequencies, under mild assumptions on the underlying basis (which can be for instance a Fourier, wavelet or Laplace eigenfunction basis), allowing us to directly deduce an L^2 contraction rate from an empirical L^2 one without further assumption on the true regression function. We also discuss extensions to semi supervised learning on graphs, where the basis is estimated from the data itself.

DUST: A Duality-Based Pruning Method For Exact Multiple Change-Point Detection

Vincent Runge [1] & Charles Truong [2] & Simon Querné [3,4]

[1] *Université Paris-Saclay, CNRS, Univ Evry, Laboratoire de Mathématiques et Modélisation d'Evry, 91037, Evry-Courcouronnes, France.*
vincent.runge@univ-evry.fr
[2] *Centre Borelli, Université Paris-Saclay, CNRS, ENS Paris-Saclay, 4 avenue des Sciences, 91190, Gif-sur-Yvette, France*
[3] *Laboratoire de mathématiques de Versailles, Université Paris-Saclay, UVSQ, CNRS, 45 avenue des États-Unis,78000, Versailles, France*
[4] *IFPEN, 1-4 Av. du Bois Préau, 92852, Rueil-Malmaison, France*

We tackle the challenge of detecting multiple change points in large time series by optimising a penalised likelihood derived from exponential family models. Dynamic programming algorithms can solve this task exactly with at most quadratic time complexity. In recent years, the development of pruning strategies has drastically improved their computational efficiency. However, the two existing approaches have notable limitations: PELT struggles with pruning efficiency in sparse-change scenarios, while FPOP's structure is not adapted to multi-parametric settings. To address these issues, we introduce the DUal Simple Test (DUST) framework, which prunes candidate changes by evaluating a dual function against a threshold. This approach is highly flexible and broadly applicable to parametric models of any dimension. Under mild assumptions, we establish strong duality for the underlying non-convex pruning problem. We demonstrate DUST's effectiveness across various change-point regimes and models. In particular, for one-parametric models, DUST matches the simplicity of PELT with the efficiency of FPOP. Its use is especially advantageous for non-Gaussian models. Finally, we apply DUST to mouse monitoring time series under a change-in-variance model, illustrating its ability to recover the optimal change-point structure efficiently.

Gentle measurements of quantum states

Cristina Butucea [1] & Jan Johannes [2] & Henning Stein [3]

[1] *CREST, ENSAE, Institut Polytechnique de Paris, 91120 Palaiseau, France*
cristina.butucea@ensae.fr
[2] *Heidelberg University, 69120 Heidelberg, Germany*
johannes@math.uni-heidelberg.de
[3] *CREST, ENSAE, Institut Polytechnique de Paris, 91120 Palaiseau, France and Heidelberg University, 69120 Heidelberg, Germany*
henning.stein@math.uni-heidelberg.de

Gentle measurements of quantum states result in both a random variable and a non-collapsed post-measurement state which is at a prescribed trace-distance from the initial state. Unlike collapsed states, this can be further used in quantum computing. We introduce here locally gentle measurements of finite-dimensional quantum states and prove a quantum data processing inequality for such measurements. We introduce physically feasible gentle measurements and show optimal rates for learning and testing quantum states.

Model upgrading in survival analysis

William G. Underwood[1*], Oliver Y. Feng[2], Henry W. J. Reeve[3], Bhramar Mukherjee[4] and Richard J. Samworth[1]

[1]*Statistical Laboratory, University of Cambridge.*
[2]*Department of Mathematical Sciences, University of Bath.*
[3]*School of Artificial Intelligence, Nanjing University.*
[4]*Yale School of Public Health.*
[*]*wgu21@cam.ac.uk.*

Statistical models in biomedicine are regularly updated as new data, often with additional covariates, become available. We propose a general approach for combining existing 'external' estimators with a new data set in a time-to-event survival analysis setting. Our method consists of constructing convex combinations of the external relative risk estimators with a flexible family of models trained using the new data; we propose a framework based on reproducing kernels. The convex combination coefficients, along with regularisation parameters for the kernel estimators, are selected using cross-validation. We establish high-probability bounds for the L_2-error of our proposed aggregated estimator, showing that it achieves a rate of convergence that is at least as good as both the optimal kernel estimator and the best external model. Empirical results from simulation studies align with our theoretical results, and we also illustrate the improvements our method provides for cardiovascular disease risk modelling.

Thursday 04/09. 11:00 - 12:20

Contributed session – Robustness and optimization

- Renaud Gaucher. A unified breakdown analysis for Byzantine robust Gossip.
- Laurentiu Marchis. On the benefits of accelerated optimization in robust and private estimation.
- El Mehdi Saad. New lower bounds for stochastic non-convex optimization through divergence decomposition.
- Laura Hucker. Comparing regularisation paths of (conjugate) gradient estimators in ridge regression.

A Unified Breakdown Analysis for Byzantine Robust Gossip

Renaud Gaucher [1] & Aymeric Dieuleveut [2] & Hadrien Hendrikx [3]

[1] *CMAP, École Polytechnique & INRIA Grenoble*
renaud.gaucher@polytechnique.edu
[2] *CMAP, École Polytechnique*
aymeric.dieuleveut@polytechnique.edu
[3] *INRIA Grenoble*
hadrien.hendrikx@inria.fr

In decentralized machine learning, different devices communicate in a peer-to-peer manner to collaboratively learn from each other's data. Standard optimization approaches tackling this problem are known to be vulnerable to adversarial (or Byzantine) devices. Indeed: decentralized optimization algorithms typically rely on the so-called *gossip* communication routine, in which each device updates its parameter by averaging it with the ones of some of the other devices. But in the presence of Byzantines, this averaging step is non-robust, and the error introduced by the Byzantines on one honest device spread to the others along communication steps.

In our work, we investigate the notion of *breakdown point* of Byzantine-robust decentralized algorithms, ie the proportion of adversaries an algorithm can tolerate. Specifically, we propose a tight analysis of this breakdown point when devices communicate only with a small number of other devices.

We do so by showing an upper bound on the breakdown point of algorithms that rely on spectral quantities of the communication network. Then we propose a general framework for building robust gossip algorithms, coined F-RG, which combines gossip communications with any aggregation rule F that satisfies a robustness criterion that we introduce. The analysis of F-RG encompasses and refines existing SOTA algorithms, such as ClippedGossip and NNA. We additionally propose a practical robust aggregator relying on adaptive clipping, named CS_+, such that CS_+–RG has a breakdown point optimal up to a factor 2. Additionally, we show that robust communication is a sufficient building block for Byzantine-robust distributed SGD.

We give experimental evidence to validate the effectiveness of CS_+–RG and highlight the gap with existing algorithms, in particular against a novel attack tailored to disrupt decentralized communications.

On the Benefits of Accelerated Optimization in Robust and Private Estimation

Laurentiu Marchis [1] & Po-Ling Loh [2]

[1] *lam223@cam.ac.uk*
[2] *pll28@cam.ac.uk*

We study the advantages of accelerated gradient methods, specifically based on the Frank-Wolfe method and projected gradient descent, for privacy and heavy-tailed robustness. Our approaches are as follows: For the Frank-Wolfe method, our technique is based on a tailored learning rate and a uniform lower bound on the gradient of the ℓ_2-norm over the constraint set. For accelerating projected gradient descent, we use the popular variant based on Nesterov's momentum, and we optimize our objective over $\mathbb{R}^p$. These accelerations reduce iteration complexity, translating into stronger statistical guarantees for empirical and population risk minimization. Our analysis covers three settings: non-random data, random model-free data, and parametric models (linear regression and generalized linear models). Methodologically, we approach both privacy and robustness based on noisy gradients. We ensure differential privacy via the Gaussian mechanism and advanced composition, and we achieve heavy-tailed robustness using a geometric median-of-means estimator, which also sharpens the dependency on the dimension of the covariates. Finally, we compare our rates to existing bounds and identify scenarios where our methods attain optimal convergence.

New Lower Bounds for Stochastic Non-Convex Optimization through Divergence Decomposition

El Mehdi Saad [1] & Wei-Cheng Lee [2] & Francesco Orabona [3]

[1] *KAUST*
mehdi.saad@kaust.edu.sa
[2] *KAUST*
wei-cheng.lee@kaust.edu.sa
[3] *KAUST*
francesco.orabona@kaust.edu.sa

We study the minimax complexity of stochastic first-order methods for minimizing an L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ over a convex domain $\mathcal{X}$, when only unbiased gradient estimates with variance at most σ^2 are available. In particular, we focus on some structured non-convex classes defined by *Quasar-Convexity* (QC), *Quadratic Growth* (QG), and the *Restricted Secant Inequality* (RSI), characterized by parameters $\tau \in (0, 1]$ and $\mu > 0$ as follows

$$(\tau - \text{QC}) : \forall \boldsymbol{x} \in \mathbb{R}^d \quad f(\boldsymbol{x}) - f^* \leq \frac{1}{\tau}\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_p \rangle$$
$$(\mu - \text{QG}) : \forall \boldsymbol{x} \in \mathbb{R}^d \quad f(\boldsymbol{x}) - f^* \geq \frac{\mu}{2} \left\| \boldsymbol{x} - \boldsymbol{x}_p \right\|^2$$
$$(\mu - \text{RSI}) : \forall \boldsymbol{x} \in \mathbb{R}^d \quad \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_p \rangle \geq \mu \left\| \boldsymbol{x} - \boldsymbol{x}_p \right\|^2,$$

where $\boldsymbol{x}_p$ is the projection of $\boldsymbol{x}$ onto the set of global minimizers. Our goal is to characterize the smallest expected optimization error given by $\mathbb{E}\big[f(\hat{\boldsymbol{x}}_T) - f^*\big]$, for each of the classes above. where $f^* = \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ and $\hat{\boldsymbol{x}}_T$ is the algorithm's output.

By reformulating the optimization problem as a statistical function-identification problem, we construct a finite "hard" subclass and apply a *divergence decomposition* argument to bound the Kullback–Leibler divergence between any two feedback distributions. This yields sharp lower bounds (up to logarithmic factors) that depend on σ, L, τ, μ, and the ambient dimension d, which must exceed a logarithmic threshold in problem parameters. Concretely, for QC alone we show

$$\mathbb{E}[f(\hat{\boldsymbol{x}}_T)] - f^* = \Omega\Big(\frac{D\,\sigma}{\tau\sqrt{\ln(1/\tau)\,T}}\Big),$$

while adding QG, we prove the lower bound $\Omega\big(\sigma^2/(\mu\tau^2 \ln(1/\tau)\,T)\big)$, and for RSI we obtain $\Omega\big(L\sigma^2/(\mu^2 \ln(\kappa)\,T)\big)$ with $\kappa = L/\mu$. We further show that these lower bounds nearly match upper bounds achieved by SGD (up to log-factors). Finally, we present a specialized algorithm in the one-dimensional setting attaining $\tilde{O}(\sigma^2/(\mu\,T))$ (or $\tilde{O}(\sigma^2/(\mu\tau\,T))$ under QG+QC) with high probability. These results suggest that the logarithmic threshold for the ambient dimension for which the presented lower bounds hold is not a consequence of our analysis but rather that the dimension plays an intrinsic role in non-convex stochastic first-order optimization.

Comparing regularisation paths of (conjugate) gradient estimators in ridge regression

Laura Hucker [1] & Markus Reiß [2] & Thomas Stark [3]

[1] *Humboldt-Universität zu Berlin*
huckerla@math.hu-berlin.de
[2] *Humboldt-Universität zu Berlin*
mreiss@math.hu-berlin.de
[3] *Universität Wien*
thomas.stark@univie.ac.at

We consider standard gradient descent, gradient flow and conjugate gradients as iterative algorithms for minimizing a penalized ridge criterion in linear regression. While it is well known that conjugate gradients exhibit fast numerical convergence, the statistical properties of their iterates are more difficult to assess due to inherent nonlinearities and dependencies. On the other hand, standard gradient flow is a linear method with well known regularizing properties when stopped early. By an explicit non-standard error decomposition we are able to bound the prediction error for conjugate gradient iterates by a corresponding prediction error of gradient flow at transformed iteration indices. This way, the risk along the entire regularisation path of conjugate gradient iterations can be compared to that for regularisation paths of standard linear methods like gradient flow and ridge regression. In particular, the oracle conjugate gradient iterate shares the optimality properties of the gradient flow and ridge regression oracles up to a constant factor. Numerical examples show the similarity of the regularisation paths in practice.

Thursday 04/09. 17:30 - 18:50

Contributed session – Applied statistics

- Matthieu Texier. Combining mixture models and Markov chains to explore spatio-temporal dynamics of child wasting in southern Madagascar.
- Tom Rohmer. Improving genetic parameter estimation for dependent traits under selection.
- Marina Gomtsyan. Variable selection in specific regression for count time series
- Christophe Ley. A versatile trivariate wrapped Cauchy copula with applications to toroidal and cylindrical data.

Combining mixture models and Markov chains to explore spatio-temporal dynamics of child wasting in southern Madagascar

Matthieu Texier [1] & Pierre Masselot [2] & Nourddine Azzaoui [3] & Jacque Gardon [4] & Simon Carrière [5]

[1] *METIS, Sorbonne Université - matthieu.texier@sorbonne-universite.fr*
[2] *London School of Hygiene and Tropical Medicine - pierre.masselot@lshtm.ac.uk*
[3] *LMBP, Université Clermont Auvergne - nourddine.azzaoui@uca.fr*
[4] *HSM, Université de Montpellier, IRD - jacques.gardon@ird.fr*
[5] *HSM, Université de Montpellier, IRD - simon.carriere@ird.fr*

Malnutrition, particularly among children, remains a persistent global health concern. This study focuses on wasting, a critical form of malnutrition marked by significant variability across both space and time. In this contribution, we seek to insight wasting prevalence dynamics in data-poor areas of Southern Madagascar. Southern Madagascar faces severe malnutrition crisis, with recurrently high levels of Global Acute Malnutrition (GAM) among children. Some municipalities consistently report GAM rates (based on Middle Upper Arm Circumferences) exceeding 15% during lean seasons, corresponding to Phase 4 (Critical) of the IPC Acute Malnutrition framework.

We analyze data from municipal-level exhaustive surveys on child wasting prevalence, conducted by the Nutrition Cluster between 2018 and 2023 across 429 municipalities. This dataset is highly incomplete, with municipalities reporting data for only 1 to 16 out of 24 trimesters, and missing values are not missing at random (MNAR) across both space and time. This complexity rules out classical inference approaches and calls for advanced modeling that accounts for the spatial and temporal context in which the data are generated.

We propose a statistical space-time model that combines mixture models to characterize spatial heterogeneity and periodic, time-inhomogeneous Markov chains to model temporal evolution. Inferences were made using the Expectation-Maximization (EM) algorithm, combined with the Forward-Backward algorithm and Monte Carlo methods. We validate the inference procedure using data generated from theoretical sampling.

The model produces a spatial classification of child wasting severity, a standardized index by municipality, and a latent field of tercile probabilities across the study period. Our findings reveal distinct spatial patterns, capture expected seasonal trends (e.g. lean season peaks, post-harvest declines), and provide a preliminary estimate of the phenomenon's persistence.

This work offers valuable outputs for guiding targeted nutrition operations. It constitutes an initial step toward a more comprehensive framework for achieving an actionable understanding of the spatio-temporal variability of wasting.

Improving genetic parameter estimation for dependent traits under selection using a bivariate copula model and stochastic gradient descent approach

Tom Rohmer [1] & Victoria Brnüning[2] & Estelle Kuhn [3]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France*
tom.rohmer@inrae.fr
[2] *Institut Curie, PSL Research University, INSERM, U 1331, Mines Paris Tech, F-75005, Paris, France*
[3] *Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France*

In quantitative genetics and animal breeding, phenotypes are measured to improve traits of interest and are commonly modeled using mixed models that account for genetic and environmental effects. Because genetic effects are non-measurable, they are treated as latent variables with a covariance structure defined by pedigree relationships. Typically, multiple observed phenotypes are assumed to follow a joint Gaussian distribution, and estimation of variance components and genetic values is usually performed using restricted maximum likelihood (REML) under this assumption.

However, while individual phenotype components may appear Gaussian, their joint distribution can deviate from normality due to complex dependencies such as multivariate heavy-tailed distributions coupled with selection of reproducers. These deviations can lead to biased or inefficient estimates of genetic parameters.

To address this limitation, we propose an extension of the standard genetic model by incorporating copula functions, which offer a flexible way to model the joint distribution of phenotypes beyond Gaussian assumptions. We develop an estimation procedure combining stochastic gradient descent with a Monte Carlo Markov Chain step to accurately estimate variance components and predict genetic values under more general dependence structures.

We evaluate the performance of the proposed method through extensive simulations and apply it to a real dataset from pig breeding, where the assumption of joint normality for phenotypes is questionable. Our results demonstrate that the copula-based approach provides more robust and reliable estimates of genetic parameters compared to classical REML methods, particularly under conditions of non-Gaussian joint distributions.

Variable selection in a specific regression for count time series

Marina Gomtsyan [1]

[1] *LPSM, Sorbonne Université*
mgomtsian@lpsm.paris

Count time series occurring in various applications are often overdispersed, meaning their variance is much larger than the mean. In this work, we proposes a novel variable selection approach for processing such data. We are interested in sparse GLARMA models, see [2], for modeling discrete-valued time series. Our method consists in iteratively combining the estimation of the autoregressive moving average (ARMA) coefficients of GLARMA models and the overdispersion parameter with regularised methods designed to perform variable selection in regression coefficients of Generalised Linear Models (GLM).

We consider the negative binomial GLARMA model introduced in [1] with additional covariates. More precisely, given the past history $\mathcal{F}_{t-1} = \sigma(Y_s, s \leq t-1)$, we assume that

$$Y_t | \mathcal{F}_{t-1} \sim \mathrm{NB}\left(\mu_t^\star, \alpha^\star\right), \tag{1}$$

where $\mathrm{NB}(\mu, \alpha)$ denotes the negative binomial distribution with mean μ and overdispersion parameter α. In (1),

$$\mu_t^\star = \exp(W_t^\star) \text{ with } W_t^\star = \sum_{i=0}^{p} \beta_i^\star x_{t,i} + Z_t^\star, \tag{2}$$

with the $x_{t,i}$'s representing the p regressor variables ($p \geq 1$) and

$$Z_t^\star = \sum_{j=1}^{q} \gamma_j^\star E_{t-j}^\star \text{ with } E_t^\star = \frac{Y_t - \mu_t^\star}{\mu_t^\star + {\mu_t^\star}^2/\alpha^\star}. \tag{3}$$

Here $E_t^\star = 0$ for all $t \leq 0$ and $1 \leq q \leq \infty$. The vector $\boldsymbol{\beta}^\star$ is assumed to be sparse, *i.e.* a majority of its components is equal to zero. The goal of our method is to retrieve the indices of the nonzero components of $\boldsymbol{\beta}^\star$, also called active variables, from the observations $Y_1, \ldots, Y_n$.

We conducted numerous experiments to study the performance of the method on simulated data. Additionally, we applied the method to biological data, in particular to the study of the RNA-Seq time series. The aim was to see which non-coding genes have an influence on the expression of the coding genes at the different times.

The described method is covered in the paper *Variable selection in a specific regression time series of counts*. The implementation of the method is available in the NBtsVarSel package on CRAN (The Comprehensive R Archive Network).

References

[1] Richard A Davis, W. T. M. Dunsmuir, and Sarah B Streett. "Maximum likelihood estimation for an observation driven model for Poisson counts". In: *Methodology and Computing in Applied Probability* 7.2 (2005), pp. 149–159.

[2] Richard A Davis, W. T. M. Dunsmuir, and Ying Wang. "Modeling time series of count data". In: *Statistics Textbooks and Monographs* 158 (1999), pp. 63–114.

A versatile trivariate wrapped Cauchy copula with applications to toroidal and cylindrical data

Shogo Kato [1] & Christophe Ley [2] & Sophia Loizidou [3] & Kanti V. Mardia [4]

[1] *Institute of Statistical Mathematics*
skato@ism.ac.jp
[2] *Department of Mathematics, University of Luxembourg*
christophe.ley@uni.lu
[3] *Department of Mathematics, University of Luxembourg*
sophia.loizidou@uni.lu
[4] *Department of Statistics, University of Leeds*
K.V.Mardia@leeds.ac.uk

In this talk, we will present a new flexible distribution for data on the three-dimensional torus which we call a trivariate wrapped Cauchy copula. Our trivariate copula has several attractive properties. It has a simple form of density and desirable modality properties. Its parameters allow adjustable degree of dependence between every pair of variables and these can be easily estimated. The conditional distributions of the model are well studied bivariate wrapped Cauchy distributions. Furthermore, the distribution can be easily simulated. Parameter estimation via maximum likelihood for the distribution is given and we highlight the simple implementation procedure to obtain these estimates. We compare our model to its competitors for analyzing trivariate data and provide some evidence of its advantages. Another interesting feature of this model is that it can be extended to a cylindrical copula. We illustrate our trivariate wrapped Cauchy copula on data from protein bioinformatics of conformational angles, and our cylindrical copula on climate data related to buoy in the Adriatic Sea.

Friday 05/09. 09:20 - 10:20

Contributed session – Unsupervised learning

- Bertrand Even. Computational lower bounds for latent variables: clustering, sparse clustering and biclustering.
- Victor Thuot. Clustering items through bandit feedback: finding the right feature out of many
- Ibrahim Kaddouri. Clustering in slowly mixing Gaussian hidden Markov models.

Computational lower bounds for latent variables: clustering, sparse clustering and biclustering

Bertrand Even [1] & Christophe Giraud [2] & Nicolas Verzelen [3]

[1] *Université Paris-Saclay*
bertrand.even@universite-paris-saclay.fr
[2] *Université Paris-Saclay*
christophe.giraud@universite-paris-saclay.fr
[3] *INRAE Montpellier*
nicolas.verzelen@inrae.fr

In many high-dimensional problems, like sparse-PCA, planted clique, or clustering, the best known algorithms with polynomial time complexity fail to reach the statistical performance provably achievable by algorithms free of computational constraints. This observation has given rise to the conjecture of the existence, for some problems, of gaps – so called statistical-computational gaps – between the best possible statistical performance achievable without computational constraints, and the best performance achievable with poly-time algorithms. A powerful approach to assess the best performance achievable in poly-time is to investigate the best performance achievable by polynomials with low-degree. We build on the seminal paper of Schramm and Wein (2022) and propose a new scheme to derive lower bounds on the performance of low-degree polynomials in some latent space models. By better leveraging the latent structures, we obtain new and sharper results, with simplified proofs. We then instantiate our scheme to provide computational lower bounds for the problems of clustering, sparse clustering, and biclustering. We highlight a phenomenom which is that low-degree polynomials cannot leverage simultaneously the presence of a structure on the rows and on the columns of the data matrix.

Clustering Items through Bandit Feedback: Finding the Right Feature out of Many

Maximilian Graf[1] & Victor Thuot[2] & Nicolas Verzelen[2]

[1] *Institut für Mathematik, Universität Potsdam, Potsdam, Germany*
graf9@uni-potsdam.de
[2] *INRAE, Mistea, Institut Agro, Univ Montpellier, Montpellier, France*
victor.thuot@inrae.fr *nicolas.verzelen@inrae.fr*

We study the problem of clustering a set of items based on bandit feedback. Each of the n items is characterized by a feature vector, with a possibly large dimension d. The items are partitioned into two unknown groups, such that items within the same group share the same feature vector. We denote by $M = (M_{i,j})_{(i,j)\in[n]\times[d]}$ the matrix whose i-th row corresponds to the feature vector of item i. Our main assumption is that, there exist two distinct vectors $\mu_0, \mu_1 \in \mathbb{R}^d$, such that, for every item $i \in [n]$, the row $M_{i,\cdot}$ is either equal to μ_0 or μ_1.

We consider a sequential and adaptive setting in which, at each round, the learner selects an item I_t and a feature J_t, and observes a noisy evaluation X_t of the corresponding entry: $X_t = M_{I_t,J_t} + \epsilon_t$, where ϵ_t is a subGaussian noise. Given a prescribed probability of error δ, the learner's objective is to recover the correct partition of the items, while keeping the number of observations as small as possible.

We introduce the algorithm `BanditClustering` which relies on finding a relevant feature for the clustering task, leveraging the Sequential Halving algorithm. The procedure balances exploration of the matrix both (1) across features, to detect a discriminative feature, and (2) across items, to classify them accordingly.

With probability at least $1-\delta$, our algorithm successfully recovers the partition, and we derive an upper bound on the budget required. This bound depends on δ,n,d, the gap vector $\Delta = \mu_1 - \mu_0$, and the proportion of items in the smallest group. Furthermore, we obtain an instance-dependent lower bound, which is tight in some relevant cases. We conclude by validating our theoretical results through numerical experiments.

Clustering in Slowly Mixing Gaussian Hidden Markov Models

Ibrahim Kaddouri[1] & Mohamed Ndaoud[2]

[1] *Université Paris-Saclay*
ibrahim.kaddouri@universite-paris-saclay.fr
[2] *ESSEC Business School*
ndaoud@essec.edu

In this talk, I will discuss recent progress on understanding the problem of clustering under the hidden Markov model with Gaussian emissions, focusing on the regime where the hidden chain mixes slowly. We provide a precise characterization of how the Bayes risk depends on key model parameters and construct an adaptive Bayes-optimal and almost minimax optimal clustering procedure. Notably, our analysis reveals surprising and non-standard behavior of the Bayes risk in certain parameter regimes, offering new insights into the interplay between signal strength and temporal dependence.

Friday 05/09. 10:40 - 11:40

Contributed session – Statistical inference

- Paul Rognon-Vael. Improving variable selection properties by using external data.
- Sophia Loizidou. Optimal tests for symmetry on the torus.
- Philippe Berthet. Some recent applications of Gaussian couplings in empirical process theory.

Improving variable selection properties by using external data

Paul Rognon-Vael [1] & David Rossell [2] & Piotr Zwiernik [3]

[1] *Department of Economics and Business, Universitat Pompeu Fabra*
paul.rognon@gmail.com
[2] *Department of Economics and Business, Universitat Pompeu Fabra*
rosselldavid@gmail.com
[3] *Department of Statistical Sciences, University or Toronto*
piotr.zwiernik@upf.edu

Sparse high-dimensional signal recovery is only possible under certain conditions on the number of parameters, sample size, signal strength and underlying sparsity. We show that leveraging external information, as possible with data integration or transfer learning, allows to push these mathematical limits. Specifically, we consider external information that allows splitting parameters into blocks, first in a simplified case, the Gaussian sequence model, and then in the general linear regression setting. We show how external information dependent, block-based, ℓ_0 penalties attain model selection consistency under milder conditions than standard ℓ_0 penalties, and they also attain faster model recovery rates. We first provide results for oracle-based ℓ_0 penalties that have access to perfect sparsity and signal strength information. Subsequently, we propose an empirical Bayes data analysis method that does not require oracle information and for which efficient computation is possible via standard MCMC techniques.

Our results provide a mathematical basis to justify the use of data integration methods in high-dimensional structural learning.

Optimal tests for symmetry on the torus

Andreas Anastasiou [1] & Christophe Ley [2] & Sophia Loizidou [3]

[1] *University of Cyprus*
anastasiou.andreas@ucy.ac.cy
[2] *University of Luxembourg*
christophe.ley@uni.lu
[3] *University of Luxembourg*
sophia.loizidou@uni.lu

Several complex real-world data can be viewed as points on the hyper-torus, which is the cartesian product of circles. Over the past few years, this has motivated new proposals of distributions on the torus, both (pointwise) symmetric and sine-skewed asymmetric. In practice, it is relevant to know whether one should use the simpler symmetric models or the more convoluted yet more general asymmetric ones. So far, only parametric likelihood ratio tests have been defined to distinguish between a symmetric density and its sine-skewed counterpart. A new semi-parametric test is presented, a test which is valid not only under a given parametric hypothesis but also under a very broad class of symmetric distributions. A description of its construction and asymptotic properties under the null and alternative hypotheses will be presented. Using Stein's method, bounds for the rate of convergence of the test statistic are derived, and finite sample behavior (through Monte Carlo simulations) will be given, as well as an application of the test on protein data.

Some Recent Applications of Gaussian Couplings in Empirical Processes Theory

Philippe Berthet[1]

[1] *Institut de Mathématiques de Toulouse, UMR 5219*
Université de Toulouse, CNRS, F-31062 Toulouse, France
philippe.berthet@math.univ-toulouse.fr

We shall start by presenting general, non asymptotic rates of coupling of a sequence of empirical processes by a sequence of Gaussian processes, both indexed by functions. The mathematical benefit is to allow a stochastic process plug-in for any functional of the empirical measure. The emphasis will be on some – non asymptotical or asymptotical – consequences in mathematical statistics for new or underlooked notions that are indeed technically demanding. For each result discussed in the main part of the talk, applying a Donsker theorem is not possible and identifying the limit process is even not so obvious.

Firstly, we exhibit an upper bound of the accuracy of Monte-Carlo weighted bootstrap – using not necessarily independent weights. What about the initial bias versus additional precision compromise when bootstrapping a family of statistics $b_n \to +\infty$ times ?

Secondly, we control the empirical version of a bivariate rank-to-quantile transform that we define as a universal geometrical transport map between two planar distributions. How to describe the weak limits of the curvilinear quantile and rank processes ?

Thirdly, we derive limit theorems for statistical procedures improved by some auxiliary information. How to sharply evaluate the uniform decrease of variance and gaussian limit of the auxiliary information empirical process ?

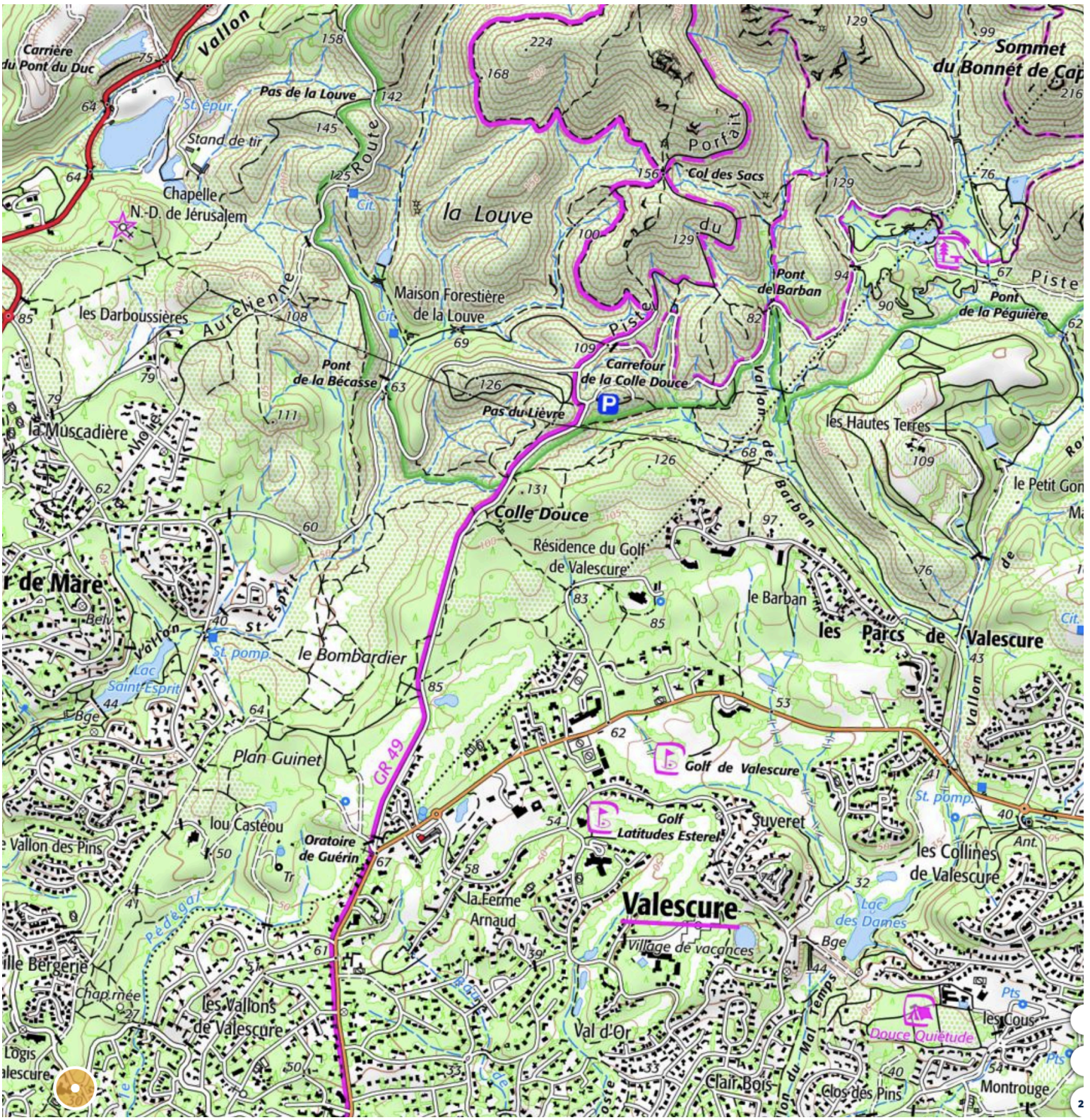
Lastly, we approximate the non-stationary empirical process, when the underlying measure is changing with each independent sample point. How to generalize the Lindeberg-Feller theorem to the infinite dimensional, empirical measure level ?

Maps of Esterel trails and St Raphael

Villa Clythia is the orange point.

permanent link to geoportail maps:

https://www.geoportail.gouv.fr/carte?c=6.768424000000001,43.444939000000005&z=17&l0=GEOGRAPHICALGRIDSYSTEMS.MAPS.SCAN25TOUR.CV::GEOPORTAIL:OGC:WMTS(1)&permalink=yes


Carrière du Pont du Duc
Vallon
Pas de la Louve
St. épur.
Stand de tir
Chapelle N.-D. de Jérusalem
Route
la Louve
Col des Sacs
Sommet du Bonnet de Cap
Piste du Portail
Pont de Barban
Pont de la Péguière
Piste
les Darboussières
Aurélienne
Maison Forestière de la Louve
Pont de la Bécasse
Carrefour de la Colle Douce
Pas du Lièvre
la Muscadière
Colle Douce
Résidence du Golf de Valescure
les Hautes Terres
le Petit Gon
le Barban
les Parcs de Valescure
le Bombardier
St. Esprit
Lac Saint-Esprit
St. pomp.
Plan Guinet
GR 49
Golf de Valescure
Golf Latitudes Esterel
Suveret
les Collines de Valescure
Oratoire de Guérin
Vallon des Pins
la Ferme Arnaud
Valescure
Village de vacances
Lac des Dames
les Vallons de Valescure
Val d'Or
Clair Bois
Clos des Pins
Douce Quiétude
les Clous
Montrouge


Nécropole Militaire d'Indochine
Pagode Hong Hien
Tple bouddh.
le Logis de Valescure
le Hameau de Valescure
Villa Magali
Village de vac.
les Clairettes
Palais des sports
Gymn.
St. pomp.
Chap.
Coll.
Vallon de Valescure
D 100
D 37
Pin Bernard
la Remonte
GR 49
les Arènes
Gare auto-train
Ant.
Lyc.
Pisc.
Écl.
Mus.
St-Sébastien
les Gazeaux
Plateau Notre-Dame
Gare routière
Gare
Capitainerie
Mon.
Feu
Fréjus
Base nautique
Plage de
Stèle
Casino
Plage du Veillat
35,6 C
ST-RAPHAËL