10e Journées
STATISTIQUES DU SUD

TOULOUSE
DU 19 AU 21 JUIN 2024

Organisation

F. Bachoc, J. Chevallier, E. Claeys,
N. Enjalbert-Courrech, X. Mai, A. Mazoyer,
N. Peyrard, T. Rohmer, S. Yazzourh

AUDITORIUM J. HERBRAND
IRIT, CAMPUS UNIVERSITÉ PAUL SABATIER

LIEN
https://indico.math.cnrs.fr/event/jss2024

Mini-cours

Claire Boyer - Introduction aux modèles de diffusion

Emmanuel Rachelson - Introduction à l'apprentissage par renforcement

Exposés longs

Nicolas Chopin - Maud Delattre - Édith Gabriel - Sébastien Gerchinovitz

Jean Peyhardi - Anaïs Rouanet - Frédéric Richard - Mathieu Serrurier

Exposés courts

Marie Chion - Constantin Philippenko

Session Poster

https://indico.math.cnrs.fr/event/jss2024

Contents

Journées Statistiques du Sud

Ces journées sont un ensemble de groupes de travail ayant lieu dans le sud de la France. Leur but est de donner une vue d'ensemble des développements scientifiques récents en statistique et de promouvoir les échanges entre étudiant.es et chercheur.euses.

Au programme :

- Des mini-cours par Claire Boyer et Emmanuel Rachelson ;
- Des exposés par François Bachoc, Nicolas Chopin, Maud Delattre, Sébastien Gerchinovitz, Constantin Philippenko, Frédéric Richard, Anaïs Rouanet et Mathieu Serrurier ;
- Une session poster.

Depuis leur création en 2007, les JSS se sont tenues principalement dans les environs ensoleillés d'Avignon (2022), Barcelone (2014), Marseille (2009), Montpellier (2016), Nice (2007 et 2011) et Toulouse (2008, 2012 et 2024).

Comité d'organisation

François Bachoc	Juliette Chevallier	Emannuelle Claeys
Nicolas Enjalbert-Courrech	Xiaoyi Mai	Adrien Mazoyer
Nathalie Peyrard	Tom Rohmer	Sophia Yazzourh

Correspondants locaux des sites partenaires

Avignon	Florent Bonneu
Barcelone	Gábor Lugosi
Marseille	Christophe Pouet et Pierre Pudlo
Montpellier	Paul Bastide
Nice	Thomas Laloë
Toulouse	Juliette Chevallier et Adrien Mazoyer

Planning prévisionnel

Mercredi 19 juin

8:45–9:00	**Accueil**		
9:00–9:30	**Ouverture des journées**		
9:30–10:30	Exposé long	**Frédéric Richard** I2M, Université Aix-Marseille **Modération**: François Bachoc	Inference techniques for the analysis of Brownian image textures
10:30–11:00	**Pause café**		
11:00–12:30	Mini-Cours	**Emmanuel Rachelson** ISAE Supaéro **Modération**: Emmanuelle Claeys	Introduction à l'apprentissage par renforcement
12:30–14:00	**Repas**		
14:00–15:00	Exposé long	**Nicolas Chopin** ENSAE, IPP **Modération**: Juliette Chevallier	Unbiased estimation of smooth functions. Applications in statistic and machine learning
15:00–16:00	Exposé long	**Maud Delattre** INRAE, Unité MaIAGE **Modération**: Juliette Chevallier	A new preconditioned stochastic gradient algorithm for estimation in latent variable models
16:00–16:30	**Pause café**		
16:30-17:30	Exposé long	**François Bachoc** IMT, Université Paul Sabatier **Modération**: Juliette Chevallier	Gaussian processes with inequality constraints: Theory and computation
17:30	**Session poster – Cocktail**		

Jeudi 20 juin

8:45–9:00	**Accueil**		
9:00-9:30	Exposé court	**Constantin Philippenko** DI ENS, Inria Paris **Modération**: François Bachoc	Compressed and distributed least-squares regression: Convergence rates with applications to Federated Learning
9:30-10:30	Exposé long	**Mathieu Serrurier** IRIT, Université Paul-Sabatier **Modération**: François Bachoc	Building explainable and robust neural networks by using Lipschitz constraints and Optimal Transport
10:30–11:00	**Pause café**		
11:00–12:30	Mini-Cours	**Emmanuel Rachelson** ISAE Supaéro **Modération**: Emmanuelle Claeys	Introduction à l'apprentissage par renforcement
12:30–14:00	**Repas**		
14:00–15:00	Exposé long	**Anaïs Rouanet** ISPED, Université de Bordeaux **Modération**: Tom Rohmer	Nonparametric Bayesian mixture models for identifying clusters from longitudinal and cross-sectional data
15:00–16:00	Exposé long	**Jean Peyhardi** IMAG, Université de Montpellier **Modération**: Tom Rohmer	Polya urn models for multivariate species abundance data: Properties and application
16:00–16:30	**Pause café**		
16:30–18:00	Mini-Cours	**Claire Boyer** LPSM, Sorbonne Université **Modération**: Tom Rohmer	A primer on diffusion-based generative models
20:00	**Dîner de la conférence**		

Vendredi 21 juin

9:30–10:30	Exposé long	**Sébastien Gerchinovitz** IRT Saint Exupéry **Modération**: Xiaoyi Mai	Conformal prediction for object detection
10:30–11:00	**Pause café**		
11:00–12:30	Mini-Cours	**Claire Boyer** LPSM, Sorbonne Université **Modération**: Juliette Chevallier	A primer on diffusion-based generative models
12:30–13:00	**Mot de la fin**		

Mini-Cours

Claire Boyer
LPSM, Sorbonne Université

A primer on diffusion-based generative models

Emmanuel Rachelson
ISAE Supaéro

Introduction à l'apprentissage par renforcement

Exposés longs

Gaussian processes with inequality constraints: Theory & computation

Mer

François Bachoc, IMT, Université Paul Sabatier

In Gaussian process modeling, inequality constraints enable to take expert knowledge into account and thus to improve prediction and uncertainty quantification. Typical examples are when a black-box function is bounded or monotonic with respect to some of its input variables. We will show how inequality constraints impact the Gaussian process model, the computation of its posterior distribution and the estimation of its covariance parameters. An example will be presented, where a numerical flooding model is monotonic with respect to two input variables called tide and surge.
The talk will follow 3 parts. (1) An introduction to (constrained) Gaussian processes and their motivations in the field of computer experiments will be provided. (2) Theoretical results on the impact of the constraints on maximum likelihood estimation will be provided. (3) Focusing on numerical computations, an algorithm called MaxMod will be presented.

Unbiased estimation of smooth functions, Applications in statistic and machine learning

Mer

Nicolas Chopin, ENSAE, Institut Polytechnique de Paris

Given a smooth function f, we develop a general approach to turn Monte Carlo samples with expectation m into an unbiased estimate of $f(m)$. Specifically, we develop estimators that are based on randomly truncating the Taylor series expansion of f and estimating the coefficients of the truncated series. We derive their properties and propose a strategy to set their tuning parameters – which depend on m – automatically, with a view to make the whole approach simple to use. We develop our methods for the specific functions $f(x) = log(x)$ and $f(x) = \frac{1}{x}$, as they arise in several statistical applications such as maximum likelihood estimation of latent variable models and Bayesian inference for un-normalised models. Detailed numerical studies are performed for a range of applications to determine how competitive and reliable the proposed approach is.

A new preconditioned stochastic gradient algorithm for estimation in latent variable models

Mer

Maud Delattre, INRAE, Unité MaIAGE

Latent variable models are powerful tools for modeling complex phenomena involving in particular partially observed data, unobserved variables or underlying complex unknown structures. Inference is often difficult due to the latent structure of the model. To deal with parameter estimation in the presence of latent variables, well-known efficient methods exist, such as gradient-based and EM-type algorithms, but with practical and theoretical limitations. We propose as an alternative for parameter estimation an efficient preconditioned stochastic gradient algorithm. Our method includes a preconditioning step based on a positive definite Fisher information matrix estimate. We prove convergence results for the proposed algorithm under mild assumptions for very general latent variables models. We illustrate through relevant simulations the performance of the proposed methodology in a nonlinear mixed effects model and in a stochastic block model.

Conformal prediction for object detection

Ven

Sébastien Gerchinovitz, IRT Saint Exupéry

We address the problem of constructing reliable uncertainty estimates for object detection. We build upon classical tools from Conformal Prediction, which offer (marginal) risk guarantees when the predictive uncertainty can be reduced to a one-dimensional parameter. In this talk, we will first recall standard algorithms and theoretical guarantees in conformal prediction and beyond. We will then address the problem of tuning a two-dimensional uncertainty parameter, and will illustrate our method on an objection detection task. This is a joint work with Léo Andéol, Luca Mossina, and Adrien Mazoyer.

Polya urn models for multivariate species abundance data: properties and application

Jeu

Jean Peyhardi, IMAG, Université de Montpellier

This talk focuses on models for multivariate count data, with emphasis on species abundance data. Two approches emerge in this framework: the Poisson log-normal (PLN) and the Tree Dirichlet multinomial (TDM) models. The first uses a latent gaussian vector to model dependencies between species whereas the second models dependencies directly on observed abundances. The TDM model makes the assumption that the total abundance is fixed, and is then often used for microbiome datasets since the sequencing depth (in RNA seq) varies from one observation to another, leading to a total abundance that is not really interpretable. We propose to generalize TDM models in two ways: by relaxing the fixed total abundance assumption and by using Polya distribution instead of Dirichlet multinomial. This family of models corresponds to Polya urn models with a random number of draws and will be named Polya splitting distributions. In a first part I will present the probabilistic properties of such models, with focus on marginals and probabilistic graphical model. Then it will be shown that these models emerge as stationary distributions of multivariate birth death process under simple parametric assumption on birth-detah rates. These assumptions are related to the neutral theory of biodiversity that assumes no biological interaction between species. Finally the statistical aspects of Polya splitting models will be presented: the regression framework, the inference, the consideration of a partition tree structure and two applications on real data.

Nonparametric Bayesian mixture models for identifying clusters from longitudinal and cross-sectional data

Jeu

Anaïs Rouanet, ISPED, Université de Bordeaux

The identification of sets of co-regulated genes that share a common function is a key question of modern genomics. Bayesian profile regression is a semi-supervised mixture modelling approach that makes use of a response to guide inference toward relevant clusterings. Previous applications of profile regression have considered univariate continuous, categorical, and count outcomes. In this work, we extend Bayesian profile regression to cases where the outcome is longitudinal (or multivariate continuous), using multivariate normal and Gaussian process regression response models. The model is applied on budding-yeast data to identify groups of genes co-regulated during the Saccharomyces cerevisiae cell cycle. We identify four distinct groups of genes associated with specific patterns of gene expression trajectories, along with the bound transcriptional factors, likely involved in their co-regulation process.

Inference techniques for the analysis of Brownian image textures

Mer

Frédéric Richard, I2M, Université Aix-Marseille

In this talk, I will present some techniques for estimating the functional parameters of anisotropic fractional Brownian fields, and their application to the analysis of image textures. I will focus on a first approach based on the resolution of inverse problems which leads to a complete estimation of parameters. The formulation of these inverse problems comes from the fitting of the empirical semi-variogram of an image to the semi-variogram of a turning band field that approximates the anisotropic fractional Brownian field. It takes the form of a separable non-linear least square criterion which can be solved by a variable projection method, and extended to take into account additional penalties. Besides, I will also describe an alternate approach which uses neural networks to obtain accurate estimation of field features such as the field degree of regularity.

Building explainable and robust neural networks by using Lipschitz constraints and Optimal Transport

Jeu

Mathieu Serrurier, IRIT, Université Paul-Sabatier

The lack of robustness and explainability in neural networks is directly linked to the arbitrarily high Lipschitz constant of deep models. Although constraining the Lipschitz constant has been shown to improve these properties, it can make it challenging to learn with classical loss functions. In this presentation, we explain how to control this constant, and demonstrate that training such networks requires defining specific loss functions and optimization processes. To this end, we propose a loss function based on optimal transport that not only certifies robustness but also converts adversarial examples into provable counterfactual examples.

Exposé court

Compressed and distributed least-squares regression: Convergence rates with applications to Federated Learning

Jeu

Constantin Philippenko, DI ENS, Inria Paris

We investigate the impact of compression on stochastic gradient algorithms for machine learning, a technique widely used in distributed and federated learning.
We underline differences in terms of convergence rates between several unbiased compression operators, that all satisfy the same condition on their variance, thus going beyond the classical worst-case analysis. To do so, we focus on the case of least-squares regression (LSR) and analyze a general stochastic approximation algorithm for minimizing quadratic functions relying on a random field. More particularly, we highlight the impact on the convergence of the covariance of the additive noise induced by the algorithm. We consider weak assumptions on the random field, tailored to the analysis (specifically, expected Hölder regularity), and on the noise covariance, enabling the analysis of various randomizing mechanisms, including compression. We then extend our results to the case of federated learning.

Posters

Copula Integration for Genetic Selection Parameter Estimation in Bivariate Linear Mixed Models

Victoria Bruning, Tom Rohmer
INRAE

A general approximation lower bound in Lp norm, with applications to feed-forward neural networks

Armand Foucault, El Mehdi Achour, Sébastien Gerchinovitz, François Malgouyres
IMT, Université Paul Sabatier

Statistiques et machine learning pour la prédiction de sorties complexes avec application à la sûreté nucléaire

Florian Gossard, Jean Baccou, Francçois Bachoc
IMT, Université Paul Sabatier

Global sensitivity analysis with weighted Poincaré inequalities

David Heredia, Aldéric Joulin, Olivier Roustant
IMT, INSA Toulouse

Integration of medical knowledge into Reinforcement Learning for dynamic treatment regimes

Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre, Michael Kosorok
IMT, Université Paul Sabatier

Pré-Journée

Cette pré-journée est co-organisée avec la Fédération OcciMath et réservée aux étudiant.es en master et doctorat. Elle se compose d'exposés scientifiques courts donnés par des étudiant.es, et d'une séance informative autour de la thèse et l'après-thèse.

Planning – Salle Johnson (IMT, Université Paul Sabatier)

12:30–13:00	**Accueil & Introduction**		
13:00–13:30	Exposé court	**Hanna Bacave** INRAE, Unité MIAT **Modération**: Nicolas Enjalbert-Courrech	HSMM piloté par les observations pour l'estimation de la dynamique des adventices
13:30–14:00	Exposé court	**Lilit Hovsepyan** Le Mans Université, INRAE Toulouse **Modération**: Nicolas Enjalbert-Courrech	Fast inference in copula regression models with categorical explanatory variables using one-step procedures
14:00–14:30	Exposé court	**Sophia Yazzourh** IMT, Université Paul Sabatier **Modération**: Nicolas Enjalbert-Courrech	Bayesian outcome weighted learning
14:30–15:00	**Pause café**		
15:00–15:30	Exposé court	**Armand Foucault** IMT, Université Paul Sabatier **Modération**: Sophia Yazzourh	A general approximation lower bound in Lp norm, with applications to feed-forward neural networks
15:30–16:00	Exposé court	**Julien Demange-Chryst** IMT, ONERA **Modération**: Sophia Yazzourh	Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling
16:00–16:30	**Pause café**		
16:30-18:00	**Exposé & Table-ronde autour de la thèse et l'après-thèse**		

Programme scientifique

HSMM piloté par les observations pour l'estimation de la dynamique des adventices

Mar

Hanna Bacave, INRAE, Unité MIAT

Les adventices sont des plantes qui poussent spontanément dans les parcelles agricoles et qui entrent en compétition avec les cultures. Leur dynamique repose sur la colonisation et la dormance. La banque de graines n'étant jamais observée de manière naturelle, une modélisation de cette dynamique a été proposée dans le cadre des Hidden Markov Models (HMM). Ce modèle, appelé Observation Driven-HMM (OD-HMM) étend les HMM au cas où les probabilités de transition dépendent de l'observation courante pour tenir compte des nouvelles graines produites qui entrent dans la banque de graines. Cependant, pour plus de réalisme sur la distribution de la survie de la banque de graines, le cadre naturel serait celui des Hidden Semi-Markov Models (HSMM). Néanmoins la notion de durée de séjour dans l'état caché n'est plus adaptée dès lors que l'observation influence la chaîne cachée à chaque instant. En nous appuyant sur les deux cadres OD-HMM et HSMM, nous proposons un nouveau modèle général : l'OD-HSMM, permettant à la fois de tenir compte d'une influence des données sur la chaîne cachée et de s'affranchir de la loi du temps de séjour géométrique. Nous en présentons une version paramétrique à partir des paramètres clés de la dynamique d'une espèce adventice et nous discutons différentes approches pour leur estimation.

Variational autoencoder with weighted samples for high-dimensional non-parametric adaptive importance sampling

Mar

Julien Demange-Chryst, ONERA, IMT

Adaptive importance sampling is a well-known family of algorithms for density approximation, generation and Monte Carlo integration including rare event estimation. The main common denominator of this family of algorithms is to perform density estimation with weighted samples at each iteration. However, the classical existing methods to do so, such as kernel smoothing or approximation by a Gaussian distribution, suffer from the curse of dimensionality and/or a lack of flexibility. Both are limitations in high dimension and when we do not have any prior knowledge on the form of the target distribution, such as its number of modes. Variational autoencoders are probabilistic tools able to represent with fidelity high-dimensional data in a lower dimensional space. They constitute a parametric family of distributions robust faced to the dimension and since they are based on deep neural networks, they are flexible enough to be considered as non-parametric models. In this communication, we propose to use a variational autoencoder as the auxiliary importance sampling distribution by extending the existing framework to weighted samples. We integrate the proposed procedure in existing adaptive importance sampling algorithms and we illustrate its practical interest on diverse examples.

A general approximation lower bound in Lp norm, with applications to feed-forward neural networks

Mar

Armand Foucault, IMT, Université Paul Sabatier

We study the fundamental limits to the expressive power of neural networks. Given two sets F, G of real-valued functions, we first prove a general lower bound on how well functions in F can be approximated in $Lp(\mu)$ norm by functions in G, for any $p \geq 1$ and any probability measure μ. The lower bound depends on the packing number of F, the range of F, and the fat-shattering dimension of G. We then instantiate this bound to the case where G corresponds to a piecewise-polynomial feed-forward neural network, and describe in details the application to two sets F: Hölder balls and multivariate monotonic functions. Beside matching (known or new) upper bounds up to log factors, our lower bounds shed some light on the similarities or differences between approximation in Lp norm or in sup norm, solving an open question by DeVore et al. (2021). Our proof strategy differs from the sup norm case and uses a key probability result of Mendelson (2002).

One-Step estimation procedure in univariate and multivariate GLMs with categorical explanatory variables

Mar

Lilit Hovsepyan, Le Mans Université, Inrae Toulouse

Generalized linear models are commonly used for modeling relationships in both univariate and multivariate contexts, with parameters traditionally estimated via the maximum likelihood estimator (MLE). MLE, while efficient, often requires a Newton-Raphson type algorithm for computation, making it time-intensive particularly with large datasets or numerous variables. Although faster, alternative closed form estimators lack the efficiency. In this topic, we propose a fast and asymptotically efficient estimation of the parameters of generalized linear models with categorical explanatory variables. It is based on a one-step procedure where a single step of the gradient descent is performed on the log-likelihood function initialized from the explicit estimators. This work presents the theoretical results obtained, the simulations carried out and an application to car insurance pricing.

Multivariate GLMs are studied in many scientific contexts. In insurance sector actuaries and risk managers precisely, they allow to assess the joint probabilities of various events occurring simultaneously, such as multiple claims or correlated risks across different insurance policy types (e.g., life, property, and auto). Copula models provide flexible tools to model multivariate variables by distinguish marginal effects from the dependence structure. In this setting, the copula parameter which quantify the (non-linear) dependency of the coordinates and the parameters of the marginal distributions are unknown and have to be estimated jointly.

In order to infer the parameters, maximum likelihood estimators (MLE) can be used due to the asymptotic properties. However, MLE is generally not in closed-form expression and is consequently time consuming. An alternative procedure, called inference for margins estimators (IFM), has been proposed in (Xu 1996, Joe 1997, 2005). In the IFM procedure, parameters of the marginals are estimated separately and simultaneously and plug-in to obtain finally the copula parameter. Although, IFM-MLE can still be time-consuming for this reason in order to estimate the copula parameter, fast and asymptotically efficient OS-CFE are used to estimate the parameters of the marginals and plug-in to estimate the copula parameter with the IFM method.

Bayesian outcome weighted learning

Mar

Sophia Yazzourh, IMT, Université Paul Sabatier

L'un des objectifs principaux de la médecine de précision statistique est d'apprendre des règles de traitement individualisées optimales ou "Individualized Treatment Rules" (ITRs). La méthode "Outcome Weighted Learning" (OWL) propose pour la première fois, une approche basée sur la classification, ou l'apprentissage automatique, pour estimer les ITRs. Elle reformule le problème d'apprentissage des ITR optimales en un problème de classification pondérée, qui peut être résolu en utilisant des méthodes d'apprentissage automatique, telles que les machines à vecteurs de support. Dans cet article, nous introduisons une formulation bayésienne de l'OWL. En partant de la fonction objective de l'OWL, nous générons une pseudo-vraisemblance qui peut être exprimée comme un mélange d'échelles de distributions normales. Un algorithme de Gibbs sampling est développé pour échantillonner la distribution postérieure des paramètres. En plus de fournir une stratégie pour apprendre une ITR optimale, l'OWL bayésien offre (1) une approche méthodique pour la génération de règles de décision apprises sur données dispersées et (2) une approche probabiliste naturelle pour estimer l'incertitude des recommandations de traitement ITR elles-mêmes. Nous démontrons la performance de notre méthode à travers plusieurs études de simulation.

Exposé & Table-ronde autour de la thèse et l'après-thèse

Animation : Pierre Neuvial[1]
Vincent Baron[2], ***Juliette Chevallier***[3], ***Chifaa Dahik***[4], ***Sebastien Déjean***[1], ***Tom Rohmer***[5],
[1] IMT, CNRS - [2] Université Paul Sabatier - [3] IMT, INSA Toulouse - [4] Capgémini - [5] INRAE.

Partenaires institutionnels et Sponsors

Avec le soutien de :

- Institut de Mathématiques de Toulouse (IMT)
- Institut de Recherche en Informatique de Toulouse (IRIT)
- INSA Toulouse
- INRAE Occitanie-Toulouse (MathNum + GA)
- Centre International de Mathématiques et Informatique de Toulouse (CIMI)
- Société de Mathématiques Appliquées et Industrielles (SMAI)
- Société Française de Statistique (SFdS)
- Équipe projet "Apprentissage, Optimisation, Complexité" (AOC)
- ANR MASDOL
- ANR GAP