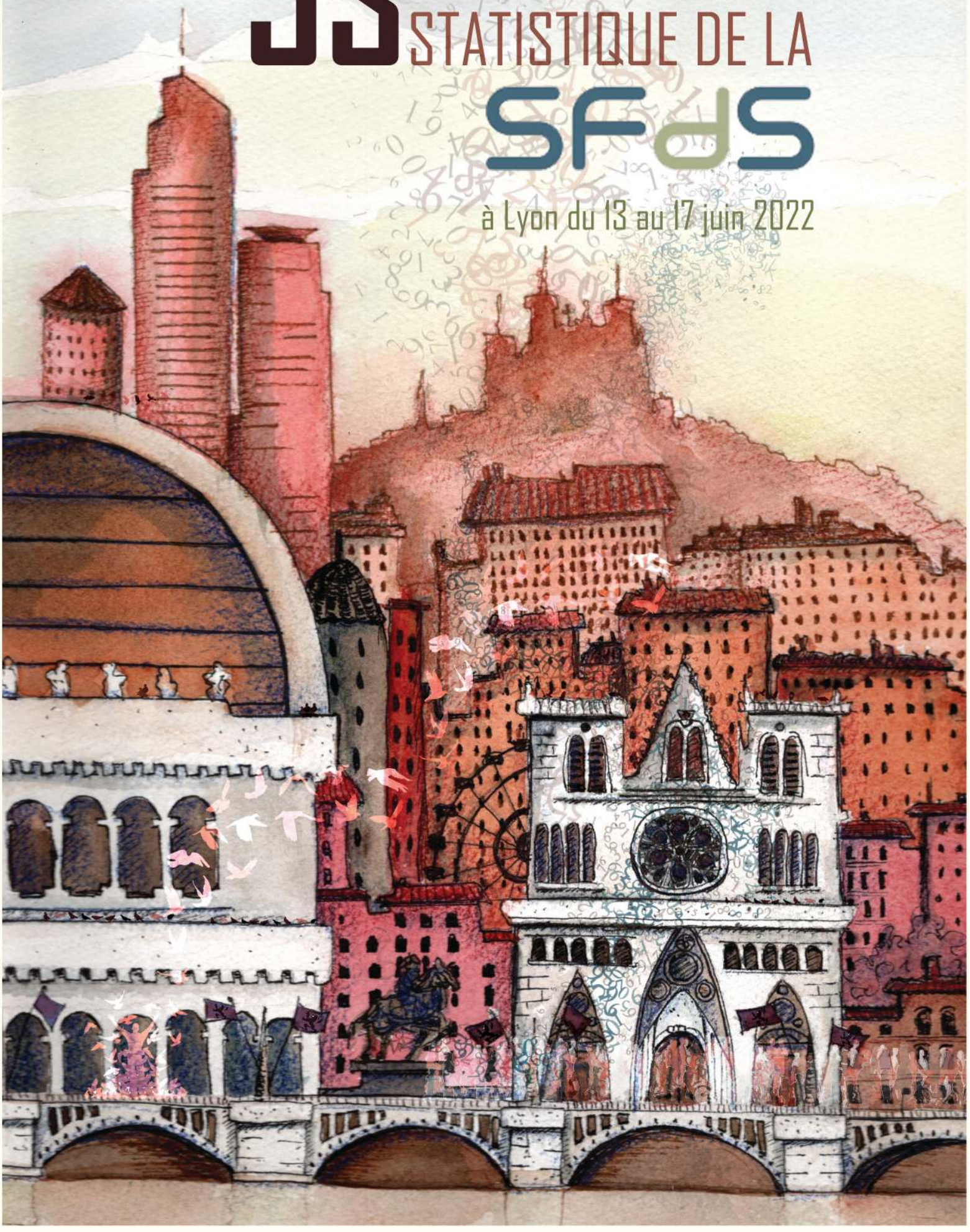
53es JOURNÉES DE STATISTIQUE DE LA SFdS

à Lyon du 13 au 17 juin 2022



53es Journées de statistique de la SFdS

Lyon, 13 au 17 juin 2022

C'est avec un immense plaisir que nous vous accueillons pour les 53èmes journées de statistique, congrès annuel de la société française de statistique (SFdS), qui se dérouleront du 13 au 17 juin 2022 à l'université Claude Bernard Lyon 1. Après plusieurs années de chamboulements majeurs des vies sociales et scientifiques, cet événement revêt une importance particulière, celle du retour à un rassemblement d'une communauté humaine impatiente de pouvoir interagir directement avec des collègues partageant un intérêt commun pour la "chose" statistique. Cette édition 2022 sera rythmée par les exposés scientifiques patiemment sélectionnés par le comité scientifique, mais aussi par des événements sociaux rassemblant notre communauté scientifique (apéritif, assemblée générale, activités sociales et soirée du mercredi). Nous souhaitons donc remercier chaleureusement l'ensemble des comités scientifique et d'organisation, dont le travail a permis l'élaboration de cette semaine de conférences et débats qui nous permettront de "faire société", à nouveau.

Anne-Laure Fougères, Mathilde Mougeot, Franck Picard

Image de couverture réalisée par Laure Savetier

Comité scientifique

Pascal Ardilly	Insee, Lyon
Nicolas Bousquet	EDF
Julie Delon	CNRS Université de Paris
Alain Célisse	Université Paris 1 Panthéon-Sorbonne
Alice Cleynen	CNRS, Montpellier
Nicolas Champagnat	INRIA, U Lorraine
Arthur Charpentier	UQAM
Amandine Marrel	CEA, cadarache
Mathilde Mougeot	ENSIIE, ENS Paris-Saclay, **présidente**
Madalina Olteanu	Université Paris Dauphine
Nelly Pustelnik	CNRS, Lyon
Nicolas Verzelen	INRA, Montpellier

Comité d'organisation

Jean-Baptiste Aubin
Caroline Bayart
Jérémie Becker
Christophette Blanchet
Yohann De Castro
Stéphane Chrétien
Gabriela Ciuperca
Jairo Cugliari
Anne-Béatrice Dufour
Thibault Espinasse
Anne-Laure Fougères (co-présidente)
Irène Gannaz
Aurélien Garivier
Rémi Gribonval
Sonia Guérin-Hamdi
Céline Helbert (co-trésorière)
Laurent Jacob
Julien Jacques, Clélia Lopez
Clément Marteau
Esterina Masiello (co-trésorière)
Cécile Mercadier
Franck Picard (co-président)
Antoine Rolland
Pierre Ribereau
Didier Rullière
Mathieu Sart
Vivian Viallon

Table des matières

Informations pratiques **4**

13 juin 2022 **6**

10h00 Ouverture des Journées . . . 9

10h10 – 11h10 : PLENIERE : Prix Laplace – Marc Hallin . . . 9

11h10 – 11h30 Pause . . . 10

11h30 – 13h10 . . . 10

SFdS - Royal Statistical Society . . . 10
Statistique spatiale . . . 13
Statistique bayésienne . . . 15
Applications : société & industrie . . . 18
Apprentissage non supervisé . . . 21

12h50 – 14h20 Déjeuner . . . 24

14h20 – 15h40 . . . 24

Extrêmes . . . 24
Statistique mathématique . . . 27
Statistique mathématique 2 . . . 30
SFdS - Biométrie . . . 32
SFdS - MALIA - SSFAM . . . 34

15h40 – 16h00 Pause . . . 36

16h00 – 17h00 . . . 36

PLENIERE : Clémentine Prieur . . . 36
PLENIERE : Aurélien Bellet . . . 37

17h00 – 17h20 Pause . . . 37

17h20 – 18h40 . . . 38

Graphes – Réseaux . . . 38
Séries temporelles . . . 41
Statistique mathématique – grande dimension . . . 43
Applications : biologie et santé . . . 46
Planification – Plans d'expérience . . . 49

18h35-19h00 : Prix ENSAI - Enora Alaoui, Pierre Barbe, Félix Lucas et Victoria Mas . . . 51

19h00 – 20h30 Coktail de Bienvenue . . . 52

14 juin 2022 **52**
9h00 – 10h00 PLENIERE : Francoise Berthoud 55
10h00 – 10h10 Pause 55
10h10 – 12h10 55
SFdS - Environnement 55
SFdS - Enseignement 57
Application : santé & société 60
Données Fonctionnelles 63
Statistique mathématique 66
12h10 – 13h20 Déjeuner 68
Déjeuner scientifique 68
13h20 – 14h20 69
PLENIERE ANNULÉE : Chloé-Agathe Azencott 69
PLENIERE : Nicolas Papadakis 70
14h20 – 14h30 Pause 70
14h30 – 16h00 70
SFdS - Fiabilité 71
SFdS - Jeunes Statisticiens 74
Agrégation experts 75
Apprentissage non supervisé 78
Transport optimal 81
16h00 – 16h10 Pause 83
16h10 – 16h30 COMPUTO 83
16h30 – 16h40 Pause 84
16h40 – 18h00 AG SFdS 84

15 juin 2022 **84**
9h00 – 10h00 PLENIERE : Conférence Le Cam – Markus Reiss 87
10h00 – 10h10 Pause 87
10h10 – 11h50 87
SFdS - Società Italiana di Statistica 88
SFdS - ENBIS 89
SFdS - Sport 92
Statistique mathématique 95
11h50 – 12h00 Pause 98
12h00 – 13h00 98
PLENIERE : Li-Chun Zhang 98
PLENIERE : Emilie Kaufmann 99
13h00 – 14h00 Déjeuner 100
14h00 – 19h00 Programme Social 100
19h00 – 21h00 Soirée festive 100

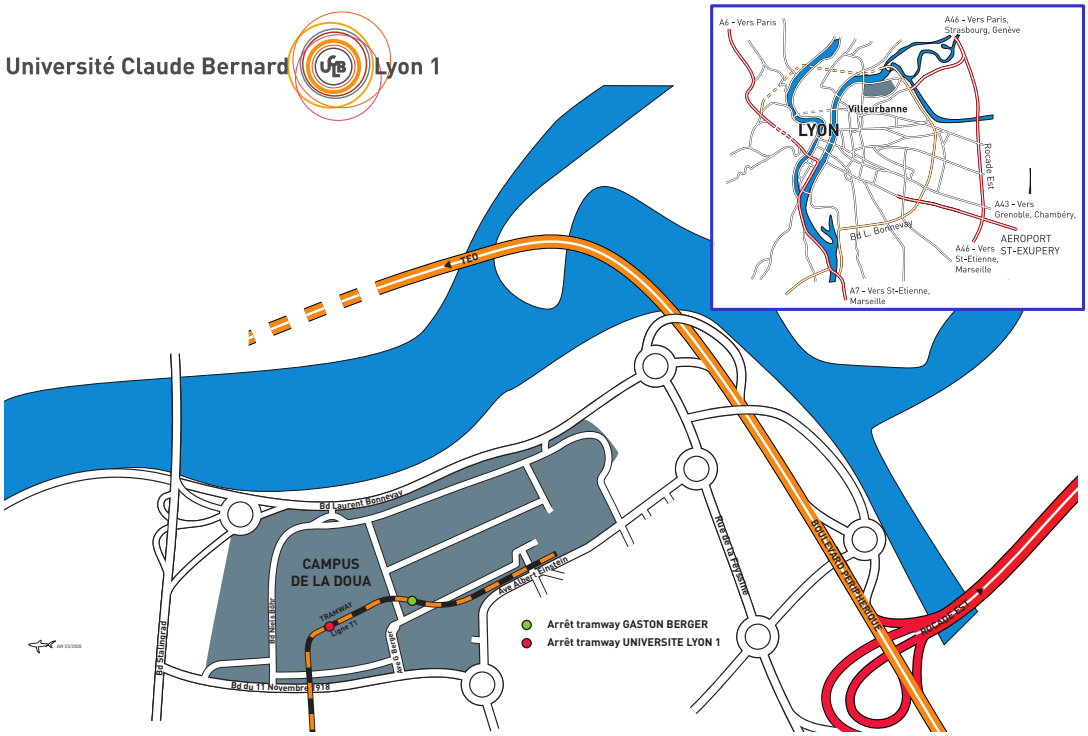
16 juin 2022 **101**
9h00 – 10h00 . . . 105
PLENIERE : Samory Kpotufe . . . 105
PLENIERE : Pierre Chainais . . . 106
10h00 – 10h20 Pause . . . 106
10h20 – 11h40 . . . 106
SFdS - Statistique mathématique . . . 106
SFdS - BFA . . . 107
Applications . . . 110
Extrêmes & Prediction . . . 113
Survie . . . 116
11h40 – 12h00 Pause . . . 119
12h00 – 13h10 . . . 119
SFdS - AMIES . . . 119
ML - Graphes - Réseaux . . . 121
Méta-modèles – Analyse de sensibilité . . . 124
ML & Regret . . . 127
Statistique mathématique . . . 129
13h30 – 14h30 Déjeuner . . . 133
14h30 – 15h30 . . . 133
PLENIERE : Frédéric Chazal . . . 133
PLENIERE : Gersende Fort . . . 133
15h30 – 15h50 Pause . . . 134
15h50 – 17h20 . . . 134
Séries Temporelles . . . 134
ML et Extrêmes . . . 136
Méthodes Génératives - ML - Deep . . . 139
Processus - Statistique mathématique . . . 142
Statistique mathématique – Valeurs Manquantes . . . 145
17h20 – 17h30 Pause . . . 148
17h30 – 18h00 Clôture . . . 148
18h00 – 18h30 Pause . . . 148
18h30 – 19h30 . . . 148
Rencontre Jeunes statisticiens . . . 148
Café de la Statistique . . . 149

Liste des auteurs **151**

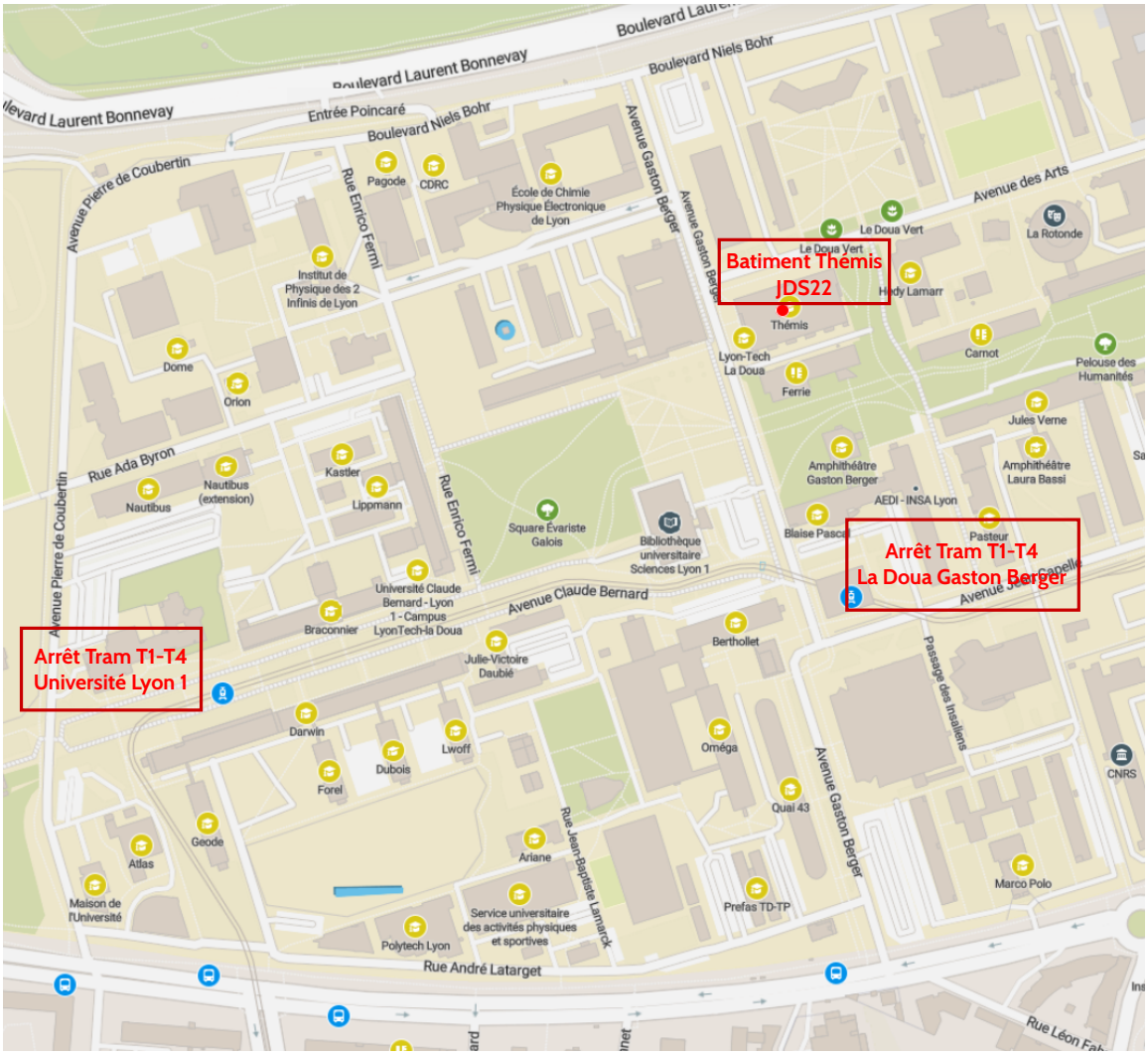
Sponsors **157**

Informations pratiques

Les JDS 2022 se dérouleront à Villeurbanne, sur le campus de l'université Claude Bernard, Lyon 1, au sein du bâtiment Thémis. Le site universitaire est accessible en transports en commun (trams T1 et T4). Les repas seront servis sur l'esplanade en face du bâtiment Thémis (entre la bibliothèque universitaire et le square Evariste Galois). Les tickets repas seront remis à l'accueil.



Bâtiment Thémis, 11 Avenue Gaston Berger, 69100 Villeurbanne



13 juin 2022

Programme

10h00 Ouverture des Journées **9**

10h10 – 11h10 : PLENIERE : Prix Laplace – Marc Hallin **9**

11h10 – 11h30 Pause **10**

11h30 – 13h10 . . . . . **10**

SFdS - Royal Statistical Society 10

Catey Bunce 10

Roger Beecham 12

Victoria Cornelius 13

Statistique spatiale 13

Denis Allard 13

Camille Frevent [et al.] 14

Solange Pruilh [et al.] 14

Cécile Spychala [et al.] 15

Statistique bayésienne 15

Marion Naveau [et al.] 15

Florence Forbes [et al.] 16

Louise Alamichel [et al.] 17

Mounia Zaouche [et al.] 17

Feriel Bouhadjera [et al.] 18

Applications : société & industrie 18

Axel Potier [et al.] 19

Corentin Duprey [et al.] 19

Yunjiao Lu [et al.] 20

Eric Parent [et al.] 20

Apprentissage non supervisé 21

Messan Martial Amovin-Assagba [et al.] 21

Aurélie Fischer [et al.] 22

Alex Mourer [et al.] 22

Sara Rejeb [et al.] 23

Christine Thomas-Agnan 24

12h50 – 14h20 Déjeuner **24**

14h20 – 15h40 . . . . . **24**
Extrêmes 24
Michael Allouche [et al.] 25
Benjamin Bobbia [et al.] 25
Alexis Boulin 26
Matthieu Garcin 26
Cambyse Pakzad [et al.] 26
Statistique mathématique 27
Marie Du Roy De Chaumaray [et al.] 27
Julien Gibaud [et al.] 28
Perrine Lacroix 28
Mathurin Massias [et al.] 29
Louis Pujol 29
Statistique mathématique 2 30
Felix Cheysson 30
Iqraa Meah [et al.] 30
Zaher Mohdeb [et al.] 31
Yves Ismaël Ngounou Bakam [et al.] 31
Ahmed Zaoui [et al.] 32
SFdS - Biométrie 32
Anaïs Rouanet 32
Clement Massonaud [et al.] 33
Marie Verbanck 34
SFdS - MALIA - SSFAM 34
Mathieu Carrière [et al.] 35
Samuel Vaiter 35
Julie Digne 35
15h40 – 16h00 Pause **36**
16h00 – 17h00 . . . . . **36**
PLENIERE : Clémentine Prieur 36
PLENIERE : Aurélien Bellet 37
17h00 – 17h20 Pause **37**
17h20 – 18h40 . . . . . **38**
Graphes – Réseaux 38
Pierre Barbillon 38
Rémi Boutin [et al.] 39
Clément Mantoux [et al.] 39
Etienne Lasalle 40
Tâm Le Minh 40
Séries temporelles 41
Joseph De Vilmarest [et al.] 41
Echarif El Harfaoui [et al.] 41
Mohamed Djemaa Sadoun [et al.] 42
Amélie Rosier [et al.] 43

Statistique mathématique – grande dimension 43
Louna Alsouki [et al.] 43
Samy Clementz [et al.] 44
Florian Dussap 44
Jean-Baptiste Fermanian 45
Oskar Laverny 45
Applications : biologie et santé 46
Benjamin Hivert [et al.] 46
Charles-Elie Rabier [et al.] 47
Wencan Zhu [et al.] 47
Tom Rohmer [et al.] 48
Johann Kuhn [et al.] 48
Planification – Plans d'expérience 49
Guillaume Chennetier [et al.] 49
Clément Gauchy [et al.] 50
Loïc Iapteff [et al.] 50
Frederique Leblanc 51
Astrid Jourdan 51
18h35-19h00 : Prix ENSAI - Enora Alaoui, Pierre Barbe, Félix Lucas et Victoria Mas . . 51
19h00 – 20h30 Coktail de Bienvenue 52

10h00 Ouverture des Journées

10h10 – 11h10

PLENIERE : Prix Laplace – Marc Hallin

(Amphi 9 - 10h10-11h10)

Quantiles, Profondeur et Transports de Mesures

Marc Hallin

Univ. libre de Bruxelles

Le concept univarié de fonction quantile—l'inverse d'une fonction de répartition—joue un rôle fondamental en Statistique et en Analyse des Données aussi bien qu'en Calcul des Probabilités. En dimension supérieure à un, malheureusement, l'inversion de la fonction de répartition traditionnelle ne mène à aucun des résultats qui font des versions empiriques de ces concepts (quantiles empiriques, rangs) des outils statistiques de première importance. La raison fondamentale en est que, contrairement à la droite, l'espace réel en dimension d>1 n'est pas ordonné de façon canonique. Les concepts de profondeur (deont l'exemple le plus connu est la profondeur au sens de Tukey) ont été introduits pour pallier ce problème. Les concepts de profondeur, malheureusement, ne mènent pas à des notions jouissant des propriétés attendues d'une notion de quantile. En particulier, la probabilité d'une région délimitée par un contour de profondeur donnée n'est pas indépendante de la loi sous-jacente, ce qui contredit à l'essence même d'une région quantile. Fondée sur des idées de transports de mesures, un concept de fonction de répartition multivariée nouveau, dit "center-outward," donne lieu à une notion de contours quantiles multivariés présentant toutes les propriétés de la notion univariée. Sa version empirique, de même, conduit à des rangs et des signes multivariés et des tests qui étendent au contexte multivarié la théorie des tests de rangs univariés et des R-estimateurs associés au nom de Hajek. Les contours quantiles correspondants peuvent être interprétés comme une version "transformation-retransformation" des contours de profondeur traditionnels.

11h10 – 11h30 Pause

11h30 – 13h10

SFdS - Royal Statistical Society

(Amphi 7 - 11h30-12h50)

Tips for Applied Medical Statisticians from Florence Nightingale - Data Visualisation

Catey Bunce* [1]

[1] Royal Marsden NHS Foundation Trust – Royaume-Uni

Resume
Catey Bunce is an applied medical statistician working at the Royal Marsden NHS Foundation Trust. She holds an honorary associate professorship at the London School of Hygiene Tropical Medicine and is an honorary consultant in applied medical statistics at Moorfields Eye Hospital. She is co lead of the National Institute for Health Research Medical Statistics group and an ambassador for the Royal Statistical Society championing the message that better data = better research = better healthcare.

Key words: Statistical Communication, medical statistics

Abstract: 12th May 2020 was the bicentenary of the birth of Florence Nightingale. Whilst many are aware of Florence Nightingale as a pioneer of modern nursing, less may be aware that she is an inspirational female leader of applied medical statistics (Cohen 1984). Florence Nightingale recommended by Farr was the first female member of the Royal Statistical Society (Spiegelhalter, 1999). Whilst there have been criticisms of Nightingale's work (Lezzoni 1996 a,

b) this has been forcefully rebutted by Vandenbroucke and Vandenbroucke-Grauls (1996) and Vandenbroucke (2003). McDonald (2014) identified that much of the criticism levied against Nightingale was entirely lacking in hard evidence and Beyersmann and Schrade (2015) argue that Nightingale and Farr suggested forms for reporting hospital mortality that were conceptually more complete than many competing risk analyses today. I believe that Florence Nightingale was inspirational in her ability to communicate statistical concepts to those without statistical training and she was perhaps ahead of her peers in using data visualisation in this. Whilst statisticians may see patterns when presented with tables of data, people from other professions or from other academic disciplines may need different images for them to comprehend the same message. Florence Nightingale used her iconic polar chart or Rose diagram to make the importance of improved ventilation and sanitation in military hospitals leap off the page (Figure 1). It was a powerful before and after visualisation which cleverly directed readers towards one interpretation of the data and not another. (Harford 2020). It persuaded senior doctors and politicians of the need for better sanitation and as result death rates fell.

There is today evidence of poor statistics within medical research (Smith 2014). There is evidence also that people can suffer harm because of statistical misunderstandings (Bunce 2019). It is important to note that even if the medical research has been conducted and reported correctly, its findings may be miscommunicated. Better communication between statistician and non-statistician has been encouraged for many years (Greenfield 1993). Data visualisation is a vital tool for communication and is one that statisticians need to engage with in order that the underlying statistical messages are accurately conveyed.

Ironically, 2020, the year celebrating the bicentenary of Florence Nightingale's birth, was the year that the COVID-19 pandemic struck the UK and other countries. Tens of millions of lives were potentially at risk and so were billions of people's livelihoods. (Harford 2020). Politicians needed to make vital decisions and to make these decisions quickly using data-informed forecasts. 2020 was arguably the year that data visualisation came into its element and a year that very much highlighted the value of statistical literacy and proper risk communication. (Palmeiro-Silva 2021). Whilst some researchers used conventional methods – for example, polar charts by countries, others realised that the rate of change in information meant that static images did not adequately convey key messages. (Figure 2).

Animations and interactive tools were developed which provided moving stories illustrating to all how the pandemic was evolving and responding to changes in policy. Whilst Nightingale and Farr had access to a single static dataset which they understood well in terms of data definitions and data quality – statisticians and data scientists in 2020 had multiple data sources. Data definitions were at times unclear and missing data was not always highlighted. Transparency in relation to the robustness of the data was at times lacking and this lack of openness was not conducive to maintaining public confidence. (Royal Statistical Society 2020)

My talk will simply reflect upon the use of data visualisation over time and the challenges facing applied medical statisticians in providing messages that are clear to non-statisticians but do not compromise on statistical integrity to communicate simple and powerful messages. (Blastland 2020)

Bibliography

Beyersmann J and Schrade C (2017) Florence Nightingale, William Farr and competing risks J R Statist Soc A 180 Part 1 pp 285-293

Blastland M, Freeman ALJ, van der Linden S, Marteau TM, Spiegelhalter D. (2020) Five rules for evidence communication. Nature. 2020 Nov;587(7834):362-364.

Bunce C, Stratton IM, Elders A, Czanner G, Dore C, Freemantle N; Ophthalmic Statistics Group (2019) Ophthalmic Statistics Note 13: Method Agreement Studies in Ophthalmology – please don't carry on correlating ... Br J Ophthalmol 2019 Sep 103(9) 1201-1293

Cohen IB (1984) Florence Nightingale Scient Am 250 128-137

Greenfield A (1993) Communicating Statistics J R Stat Soc (A) 156 Part 2 pp 287-297

Harford T (2020) How to Make the World Add Up The Bridge Street Press

McDonald L (2014) Florence Nightingale, Statistics and the Crimean War J R Statistic Soc A 177 569-586

Palmeiro-Silva YK, Weinstein-Oppenheimer C, Henríquez-Roldán CF, Bangdiwala S (2021) Statistical literacy and risk communication for COVID-19 vaccination: a scoping review Rev Panam Salud Publica 45 https://doi.org/10.26633/RPSP.2021.108

Royal Statistical Society (2020) Royal Statistical Society Submission to the Joint Health and Social Care Committee and Science Technology Committee Inquiry: 'Coronovirus: lessons learnt' 4.12.2020 https://rss.org.uk/policy-campaigns/search-policy-documents/covid-19/ date accessed 25/2/2022

Smith R (2014) Medical Research – still a scandal The BMJ 31 Jan 2014 https://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/dateaccessed25/2/2022

Spiegelhalter D J (1999) Surgical audit: statistical lessons from Nightingale and Codman JR Statist Soc A 162 45-48

Vandenbroucke J (2003) Continuing controversies over 'risk and rates' more than a century after William Farr's 'On Prognosis' Socl Prev Med 48 216-218

Vandenbroucke JP and Vandenbroucke-Grauls C (1996) In defence of Farr and Nightingale Ann Intern Med 125 1014

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy

Roger Beecham * [1]

[1] School of Geography, University of Leeds – Royaume-Uni

Concerns around the replicability of published scientific findings has prompted much introspection into the way in which scientific knowledge is produced. To address issues of data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that research findings should only be claimed through pre-registered confirmatory data analyses. Pre-registration studies are, though,

somewhat inimical to the more informal research environments typical of modern applied data analysis ('Data Science'). In this talk I enumerate some of these challenges and demonstrate, through an analysis of road crash data in the UK, how nascent visualization techniques can be used to navigate and inject statistical rigour into contemporary data analysis environments.

Data visualisation of adverse events in randomised clinical trials

Victoria Cornelius* [1]

[1] Imperial College London – Royaume-Uni

Methods to analyse efficacy outcomes in randomised controlled trials (RCTs) are well established. With substantial improvements in statistical software it's now trivial to undertake advanced statistical modelling and present this data well. Despite this progress, the analysis and presentation of adverse events (AEs) in trial publications has seen very little progress.
AE data is particularly difficult to analyse due to its multi-faceted nature. One of these features is the large number of AE outcomes that get recorded in a trial. There is a lack of guidance on what and how to visually display complex AE data.
We previously undertook a methodology review to identify statistical methods specifically developed to analyze AE data. (Phillips et al 2020) This current paper examines two visual analysis methods identified to be suitable for any adverse events collected during a trial, and one new approach proposed by the authors for pre-specified harm outcomes that is particularly valuable for multi-arm studies We explore their value using data from a COVID-19 treatment trial and COVID-19 vaccination trial.

Statistique spatiale

(Amphi 8 - 11h30-12h50)

Fully nonseparable Gneiting covariance functions for multivariate space-time data

Denis Allard*[1]

[1] Biostatistique et Processus Spatiaux – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UR0546 – Site Agroparc Domaine St Paul 84914 Avignon cedex 9, France

Nous élargissons la classe de Gneiting des fonctions de covariance spatio-temporelle en introduisant une classe très générale de fonctions de covariance multivariées entièrement non séparables. Chaque composante est modélisée par une fonction de covariance spatiale de la famille Matérn avec ses propres paramètres de lissage et d'échelle et, contrairement à tous les modèles actuellement disponibles, sa propre fonction de corrélation dans le temps. Ce modèle est illustré sur un jeu de données météorologiques trivariées. Notre nouveau modèle mène à un meilleur ajustement et de meilleurs scores prédictifs comparé à un modèle plus parcimonieux avec une fonction de corrélation temporelle commune.

Investigating spatial scan statistics for multivariate functional data

Camille Frevent [*1,2], Mohamed-Salem Ahmed [2,3], Sophie Dabo-Niang [1,4], Michaël Genin [2]

[1] MOdel for Data Analysis and Learning – Inria Lille - Nord Europe – France
[2] METRICS – Univ. Lille, CHU Lille, ULR 2694 - METRICS : Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille – France
[3] Société d'Alicante, Seclin – Société d'Alicante, Seclin – France
[4] Laboratoire Paul Painlevé - UMR 8524 – Université de Lille : UMR8524, Centre National de la Recherche Scientifique : UMR8524 – France

Ce travail introduit de nouvelles statistiques de scan spatiales pour des données fonctionnelles multivariées indexées dans l'espace. Les méthodes sont dérivées d'un test de la MANOVA pour données fonctionnelles, une adaptation du test d'Hotelling, et une extention multivariée du test de Wilcoxon. Dans une étude de simulation les performances des méthodes sont investiguées. Nous les avons ensuite appliquées sur des données fonctionnelles multivariées mesurées avec une très fine résolution spatiale pour détecter des clusters de pollution dans le nord de la France en octobre 2021.

Estimation de processus de mélanges spatio-temporels pour détecter des changements dynamiques de populations

Solange Pruilh [*2,1], Stéphanie Allassonnière [1], Anne-Sophie Jannot [1]

[1] Health data- and model- driven Knowledge Acquisition – Inria de Paris, Centre de Recherche des Cordeliers, Université de Paris – France
[2] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France

Le suivi des populations est un défi dans de nombreux domaines tels que la santé publique ou l'écologie. Nous proposons une méthode pour modéliser et surveiller les distributions de populations dans l'espace et le temps, afin de construire un système d'alerte de changements spatio-temporels dans la répartition des données. Ce pipeline, appelé *spatio-temporal mixture process*, combine un suivi temporel avec des estimations par algorithmes de type Expectation-Maximisation (EM). Nous améliorons les algorithmes existants pour éviter des fausses alertes dues à des maxima locaux et à des données à forte variabilité.

Modélisation spatiale des accidents routiers de Besançon (France) à l'aide de processus de Cox log-Gaussien

Cécile Spychala*[1], Clément Dombry[1], Camelia Goga[1]

[1] Université de Franche-Comté – Laboratoire de Mathématiques de Besançon – France

Dans un objectif de prévention et/ou d'anticipation des accidents routiers, la modélisation statistique de la dépendance spatiale et des facteurs de risque potentiels représente un atout majeur. L'intérêt de cette étude se porte sur la localisation géoréférencée des accidents. Nous avons croisé ces événements avec des covariables caractérisant la zone géographique d'étude (socio-démographiques et infrastructures par exemple). Après une sélection de variables (aggrégation de modèles de Poisson), la survenue des accidents a été modélisée par un processus de Cox log-Gaussien spatial. Les résultats de cette analyse permettent l'identification des principaux facteurs de risques d'accident et l'identification des zones critiques. Les données mises en application sont les accidents routiers s'étant produits entre 2017 et 2019 dans la CAGB (communauté urbaine de Besançon).

Statistique bayésienne

(Amphi 9 - 11h40-13h10)

Sélection de variables bayésienne en grande dimension dans les modèles non-linéaires à effets mixtes utilisant l'algorithme SAEM

Marion Naveau*[1], Guillaume Kon Kam King[2], Laure Sansonnet[3], Maud Delattre[4]

[1] Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement – Université Paris-Saclay, INRAE, MaIAGE, Université Paris-Saclay, INRAE, MIA-Paris – France
[2] Mathématiques et Informatique Appliquées du Génome à l'Environnement [Jouy-En-Josas] – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UPR1404 – France
[3] UMR MIA-Paris – AgroParisTech, INRA - Université Paris-Saclay – France
[4] INRAE – MaIAGE, INRAE, Université Paris-Saclay – France

Les données de grande dimension, avec beaucoup plus de covariables que d'observations, comme les données génomiques par exemple, sont maintenant couramment analysées. Dans ce contexte, il est souvent souhaitable de pouvoir se concentrer sur les quelques covariables les plus pertinentes grâce à une procédure de sélection de variables. La question de la sélection de variables en grande dimension est largement documentée dans les modèles de régression standard, mais il existe encore peu d'outils pour y répondre dans le cadre des modèles à effets mixtes non linéaires. Dans ce travail, nous abordons la sélection de variables sous un angle bayésien et proposons une procédure de sélection combinant l'utilisation de priors spike-and-slab et l'algorithme SAEM. Comme pour la régression Lasso, l'ensemble des covariables pertinentes est sélectionné en explorant une grille de valeurs pour le paramètre de pénalisation. L'approche proposée est plus rapide qu'un algorithme MCMC classique et montre de très bonnes performances de sélection sur des données simulées.

Mixture of experts posterior surrogates for approximate Bayesian computation

Florence Forbes*[1], Hien Duy Nguyen[2], Trungtin Nguyen[1], Julyan Arbel[1]

[1] STATIFY team – Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France, INRIA Greoble-Rhône Alpes, Univ. Grenoble-Alpes, Laboratoire Jean Kuntzman – France
[2] Senior Lecturer, School of Mathematics and Physics, University of Queensland, St. Lucia, Queensland – Australie

Les procédures de calcul bayésien approché (ABC) reposent sur l'évaluation de l'écart entre les données simulées et les données observées. Cet écart est souvent évalué en comparant des statistiques résumées plutôt que directement les données. Le choix d'une distance et de résumés appropriés est donc une étape cruciale qui peut affecter la qualité des approximations. Dans ce travail, nous introduisons une étape d'apprentissage préliminaire dans laquelle des lois de substitution, issues d'un modèle de mélange d'experts, sont construites pour approximer les lois a posteriori visées. Ces lois a posteriori de substitution sont ensuite utilisées à la place des statistiques résumées et comparées à l'aide de métriques entre distributions. On montre que la quasi-loi a posteriori résultante converge vers la vraie loi a posteriori, dans des conditions standard. Des expériences sur des données synthétiques et réelles montrent que notre approche est particulièrement performante lorsque la loi a posteriori est multimodale.

Bayesian nonparametric mixtures inconsistency for the number of clusters

Louise Alamichel[*1,2], Daria Bystrova[1,2], Julyan Arbel[1,2], Guillaume Kon Kam King[3]

[1] Statify, Inria Grenoble - Rhône-Alpes – Institut National de Recherche en Informatique et en Automatique – France
[2] Laboratoire Jean Kuntzmann – Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique, Université Grenoble Alpes, Institut polytechnique de Grenoble - Grenoble Institute of Technology – France
[3] Mathématiques et Informatique Appliquées du Génome à l'Environnement [Jouy-En-Josas] – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UPR1404 – France

Bayesian nonparametric mixture models are often employed for modeling complex data. While these models are well-suited for density estimation, their application for clustering has some limitations. Miller and Harrison, JMLR (2014), proved posterior inconsistency in the number of clusters when the true number of clusters is finite for Dirichlet process and Pitman–Yor process mixture models. In this work, we extend this result to additional Bayesian nonparametric priors such as Gibbs-type processes and finite-dimensional representations of them. The latter include the Dirichlet multinomial process and the recently emerged Pitman–Yor and normalized generalized gamma multinomial processes. We show that mixture models based on these processes are also inconsistent in the number of clusters.

Pedestrian traffic mapping over time at the scale of a city

Mounia Zaouche[*1], Nikolai Bode[1]

[1] Université de Bristol – Royaume-Uni

La marche est une bonne alternative aux déplacements motorisés car elle permet de réduire les embouteillages et la pollution atmosphérique, mais aussi de renforcer la cohésion sociale. L'activité des piétons est par ailleurs un indicateur du dynamisme et de la prospérité financière d'une ville. La compréhension de la distribution du trafic dans le temps et l'espace est une question fondamentale en ingénierie des transports. Cependant la cartographie du trafic piéton à de grandes échelles spatio-temporelles bien que nécessaire reste rare. Cette étude vise à utiliser des données de comptage de piétons localisées pour cartographier le trafic piéton dans le temps à l'échelle d'une ville en utilisant la modélisation bayésienne. Nous avons utilisé la méthode d'inférence puissante Integrated Nested Laplace Approximations (INLA) combinée à l'approche Stochastic Partial Differential Equation (INLA-SPDE) pour exploiter un grand nombre de données de comptage horaire enregistrées par plusieurs capteurs installés dans le centre-ville de Melbourne, mais aussi pour obtenir des résultats a posteriori rapides et précis. Nous avons développé et testé deux modèles spatio-temporels sophistiqués décrivant la variabilité spatiale au sein des données mais intégrant également différents termes d'interaction, y compris des inter-

actions spatio-temporelles. Les critères de l'erreur quadratique moyenne ou Root Mean Square Error (RMSE) et du Deviance Information Criterion (DIC) utilisés pour analyser et comparer leurs performances ont permis de mettre en lumière l'un de ces modèles. Ce dernier a permis d'observer une amélioration de plus de 39% en termes de RMSE en comparaison avec l'écart-type des données utilisées. Des cartes de prédiction et d'incertitude dans le temps ont finalement été produites.

Bayesian functional linear regression estimation. Extension to scalar and categorical covariates

Feriel Bouhadjera [*1], Meili Baragatti [1], Nadine Hilgert [1], Nathalie Smits [2], Paul-Marie Grollemund [3]

[1] MISTEA – Université Montpellier, Institut Agro, INRAE – France
[2] ABSys – Université Montpellier, Institut Agro, INRAE – France
[3] LMBP – Université Clermont Auvergne, CNRS – France

Nous considérons un modèle de régression linéaire où la variable à expliquer est réelle et les co-variables sont catégorielles, scalaires et fonctionnelles et agissent de manière additive dans le modèle. Notre objectif est d'estimer les paramètres de ce modèle, tout en conservant un caractère interprétable pour la partie du modèle incluant les variables fonctionnelles. Ce travail est une extension du modèle Bliss (Bayesian functional Linear regression with Sparse Step function), voir Grollemund et al. (2019), développé dans un cadre Bayésien et qui ne contient que des variables fonctionnelles. Nous proposons d'étendre la méthode Bliss à un modèle plus général contenant également des co-variables catégorielles et scalaires. Dans la suite, nous explicitons le modèle étendu et expliquons comment estimer les paramètres d'une manière interprétable. Une illustration est faite sur des données simulées et un jeu de données réelles sur le dépérissement de la vigne.

Applications : société & industrie

(Amphi 10 - 11h50-13h10)

Modélisation de la Substitution entre Articles pour Optimiser leur Réapprovisionnement

Axel Potier [*1], Christophe Biernacki [2,3], Julien Favre [1], Matthieu Marbac-Lourdelle [4], Vincent Vandewalle [5]

[1] ADEO – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[2] UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille – Univ. Lille, CNRS – France
[3] Inria Lille-Nord Europe – Inria Lille - Nord Europe, CNRS – France
[4] ENSAI – Ecole Nationale de la Statistique et de l'Analyse de l'Information – France
[5] Université de LILLE – Université de Lille, Droit et Santé – France

Dans le commerce de détail, l'optimisation des stocks est un problème courant nécessitant un compromis entre le risque de rupture et les coûts générés par le sur-stock. Nous proposons une extension du modèle du vendeur de journaux (Newsvendor problem) prenant en compte le phénomène de substitution entre articles. Dans cette extension nous définissons un modèle de substitution probabiliste réaliste et générique, et nous montrons comment celui-ci intervient dans le calcul de la fonction de profit afinde déterminer les stocks optimaux. L'approche est illustrée sur des données simulées similaires aux données de vente rencontrées dans la société Adeo.

Méthode de détection automatique des pannes sur avion : outil d'aide à la décision

Corentin Duprey [*1], Anne-Françoise Yao [1], Nourddine Azzaoui [1], Arnaud Guillin [2], Julie Galant

[1] Laboratoire de Mathématiques Blaise Pascal - Clermont Auvergne – Université Clermont Auvergne : UMR6620, Centre National de la Recherche Scientifique : UMR6620 – France
[2] LMBP – Université Clermont Auvergne, CNRS : UMR1234 – France

Le Flight Control No Dispatch (FCND) est une panne complexe récurrente sur les avions E190 et E170 de la marque Embraer. Une panne complexe est un aléa dont les causes peuvent être multiples. Une mauvaise identification du système défaillant peut mener à des conséquences sur l'exploitation d'une compagnie aérienne : retards ou annulations. La direction de la maintenance de HOP Air France agit pour améliorer ces méthodes de résolution. Nos travaux de recherche visent à proposer aux équipes de maintenance un outil d'aide à la décision. Il permet de cibler le système mis en cause lors d'une panne complexe. Les prises de décision sont accélérées et nous améliorons notre anticipation dans la gestion des pannes. Notre approche utilise des méthodes de classification supervisées et non-supervisée. Dans cet exposé, nous présenterons la procédure mise en place à cet effet.
Mots-clés. Détection de pannes, clustering, classification supervisée, Text mining.

Abstract. Flight Control No Dispatch (FCND) is a recurrent and complex failure on E190 and E170 aircraft of the Embraer brand. A complex failure is a fault with multiple causes. A wrong identification of the faulty system can lead to consequences on the operation of an airline company: delays or cancellations. The maintenance department of HOP Air France is working to improve these resolutions methods. Our research work aims to provide maintenance teams with a decision support tool that targets the system involved in a complex failure, thus accelerating decision-making and improving anticipation in the management of failures. Our approach uses both supervised and unsupervised learning. In this presentation, we will explain the procedure implemented for this purpose.
Keywords. Fault detection, clustering, supervised learning, Text mining.

Number of zero velocity points: a critical parameter for handwriting model estimation towards dysgraphia diagnosis assistance

Yunjiao Lu [*1], Jerome Boutet [2], Vincent Brault [1], Caroline Jolly [3], Etienne Labyt [2,4], Raphaël Lambert [2], Jean-Charles Quinton [1]

[1] Laboratoire Jean Kuntzmann – Université Grenoble Alpes, Centre National de la Recherche Scientifique - CNRS, Institut polytechnique de Grenoble (Grenoble INP) – France
[2] CEA-Grenoble – Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble – France
[3] Laboratoire de Psychologie et NeuroCognition – Université Pierre Mendès France - Grenoble 2, Université Joseph Fourier - Grenoble 1, Université Savoie Mont Blanc, Centre National de la Recherche Scientifique : UMR5105, Université Grenoble Alpes – France
[4] Mag4Health – Mag4Health – France

Une nouvelle voie pour l'analyse de l'écriture manuscrite afin d'aider à diagnostiquer la dysgraphie chez les enfants, en utilisant des modèles oscillatoires dynamiques, a été étudiée. Le *Parsimonious Oscillatory Model of Handwriting* (POMH) s'est avéré capable de reconstruire la trace de l'écriture manuscrite à condition qu'elle soit suffisamment lisse. Pour les enfants souffrant de dysgraphie, POMH conduit à une plus grande déviation lors de la tentative de reconstruction du signal. Cette différence de performance de POMH est censée servir de discriminant pour la détection automatique des enfants dysgraphiques. Un paramètre critique pour POMH est le nombre de points de vélocité nulle dans l'écriture d'un symbole. Les relations entre l'âge, le nombre de points de vélocité nulle et la performance de POMH sont étudiées comme travail préliminaire pour établir un algorithme de classification pour l'aide au diagnostic de la dysgraphie.

Prévision probabiliste d'ensemble en environnement

Eric Parent [*1], Jacques Bernier

[1] AgroParisTech, INRA (MIA) – AgroParisTech – INRA-UMR 518 MIA F 75005, Paris, France

La prévision d'ensemble se fonde sur des scénarios obtenus en modifiant les conditions et les paramètres initiaux d'un code déterministe décrivant un mécanisme physique (modèles météorologiques régionaux, modèles climatiques, modèles de précipitation et de transport solide...). Ces scénarios variés traduisent l'espoir du physiciens de représenter l'incertitude due à l'état initial du système (par exemple la température de l'atmosphère, l'humidité du sol, etc.) ou découlant de la connaissance incomplète des paramètres. En propageant les perturbations des conditions limites à travers le code numérique intégro-différentiel complexe simulant le comportement du système, on obtient un ensemble de trajectoires hypothétiques de la variable à prédire (par exemple, la température de l'aéroport, la pluie, le débit des cours d'eau...), quantités également connues sous le nom de *membres*. Dans quelle mesure les informations véhiculées par les différents membres de l'ensemble peuvent-elles aider à construire une prévision probabiliste de la quantité à prévoir ? La littérature statistique fourmille de méthodes prêtes à l'emploi, connues sous le nom de techniques de "post-traitement" des ensembles, mais peu d'entre elles traitent formellement des spécificités des ensembles : (1) la trajectoire de la Nature peut ne pas être exactement celle générée par le code déterministe, (2) les membres de chaque ensemble visent à se comporter, à cause du principe même de leur construction, comme l' échantillonnage d'un modèle de type échangeable. En statistique, DeFinetti et ses successeurs ont démontré que la propriété d'échangeabilité impose un modèle à effet aléatoire. L'effet aléatoire résume parcimonieusement l'information véhiculée par les membres de l'ensemble vis à vis la variable à prédire. Dans cet article, nous construisons un modèle normal échangeable pour les températures. Les performances des modèles proposés sont vérifiées et comparées à celles d'autres techniques de post-traitement sur une longue série d'enregistrements de températures. Les résultats montrent la robustesse des structures prédictives parcimonieuses construites, ce qui plaide en faveur de l'échangeabilité et de la cohérence probabiliste lors de la modélisation des membres d'un ensemble environnemental. Finalement, nous suggérons d'autres développements bayésiens hiérarchiques *ad hoc* : données environnementales avec valeurs nulles (précipitations) ou structures conjointes de séries hydrométéorologiques (pluies-débits).

Apprentissage non supervisé

(Amphi 11 - 11h50-13h10)

Clustering et détection d'anomalies dans les données fonctionnelles

Messan Martial Amovin-Assagba*[1,2], Irène Gannaz[3], Julien Jacques[2]

[1] Arpege Master K – Université Lumière - Lyon II – France

[2] Entrepôts, Représentation et Ingénierie des Connaissances – Université Lumière - Lyon 2 : EA3083 –

France
[3] Institut Camille Jordan [Villeurbanne] – Ecole Centrale de Lyon, Université de Lyon, Université Claude Bernard Lyon 1, Institut National des Sciences Appliquées de Lyon, Institut National des Sciences Appliquées : UMR5208, Université Jean Monnet [Saint-Etienne], Centre National de la Recherche Scientifique – France

Ce travail propose une méthode permettant simultanément de grouper en clusters et de détecter des anomalies dans des données fonctionnelles multivariées. Les données fonctionnelles sont décomposées dans une base de fonctions de dimension finie. La méthode repose ensuite sur des modèles de mélanges gaussiens contaminés parcimonieux sur cette décomposition. Un algorithme ECM est utilisé pour l'inférence du modèle. La performance du modèle est illustrée sur des données simulées.

Clustering spectral adaptatif

Aurélie Fischer*[1], Mathilde Mougeot[2,3], Ilaria Giulini[1]

[1] Laboratoire de Probabilités, Statistiques et Modélisations – Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France
[2] Ecole Nationale Supérieure dÍnformatique pour l'Industrie et l'Entreprise – France
[3] CB - Centre Borelli - UMR 9010 – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9010, Ecole Normale Supérieure Paris-Saclay – France

Nous nous intéressons au clustering spectral pour des réalisations i.i.d. d'une loi de probabilité inconnue dont le support est une union de composantes compactes connexes, dans un espace de Hilbert. Nous décrirons un algorithme, basé sur la modification de la procédure par Ng, Jordan et Weiss (2001) proposée dans Giulini (2016). Un grand avantage de la méthode est qu'elle permet d'estimer automatiquement le nombre de classes. Concernant les garanties théoriques, comme démontré dans Giulini (2016), la convergence de la procédure peut être caractérisée en termes de convergence d'opérateurs de Gram. Les performances de cette méthode adaptative seront illustrées par des expériences numériques approfondies sur des données simulées et des données réelles se présentant sous forme d'images.

Sparse and group-sparse clustering for mixed data. An illustration of the vimpclust package

Marie Chavent[1], Alex Mourer*[2], Madalina Olteanu[3]

[1] CQFD (INRIA Bordeaux - Sud-Ouest) – CNRS : UMR5251, INRIA – France
[2] Statistique, Analyse et Modélisation Multidisciplinaire (SAmos-Marin Mersenne) – Université Paris 1 Panthéon-Sorbonne : EA4543 – France

[3] CEntre de REcherches en MAthématiques de la DEcision – Centre National de la Recherche Scientifique : UMR7534 / URA749, Université Paris Dauphine-PSL – France

High-dimensional data may often contain both numerical and categorical features, and in some cases features may be available as natural groups (repeated measurements, categories of features, ...). Clustering this kind of data raises several issues: how to simultaneously deal with numerical and categorical features? how to build meaningful clusters of the input entities? how to select the most informative features or groups of features for the clustering? In the k-means framework, one may rely on a penalised version of the between-cluster variance, and find both the best partitioning of the data, and the most informative features or groups of features. The present manuscript illustrates sparse k-means and group-sparse k-means for mixed data, using the `vimpclust` package. The example provided on a small real-life dataset shows how feature selection may be directly combined with clustering, and provide a meaningful selection while preserving the quality of the clustering.

Cartes auto-organisatrices pour l'exploration des données partiellement observées et l'imputation des données manquantes

Sara Rejeb*[1], Catherine Duveau[2], Tabea Rebafka[1]

[1] LPSM – Sorbonne Université UPMC Paris VI, Université de Paris, CNRS : UMR8001 – France
[2] Safran Aircraft Engines – Villaroche – France

Les cartes auto-organisatrices de Kohonen sont des réseaux de neurones qui accomplissent deux tâches simultanément : visualiser des données multidimensionnelles par projection sur une carte bidimensionnelle et discrétiser l'espace de façon ordonnée. L'algorithme classique de Kohonen développé pour optimiser ces cartes est conçu uniquement pour des données complètes. Cependant, dans de nombreuses applications, les données contiennent naturellement de nombreuses données manquantes.
Dans ce travail, nous proposons alors une extension des cartes auto-organisatrices pour des données partiellement observées. Notre approche consiste à traiter simultanément les deux problèmes de l'apprentissage d'une carte auto-organisatrice et de l'imputation des données manquantes. Nous définissons une nouvelle fonction de perte, qui vise à optimiser simultanément la carte et également l'imputation des valeurs manquantes. Les simulations numériques illustrent de bonne performance en termes de qualité de la carte et d'imputation des données manquantes comparé à l'état de l'art. Une implémentation de la méthode en R est disponible sur le CRAN dans le package *missSOM*.

Détection d'atypiques pour données de composition avec la méthode ICS

Christine Thomas-Agnan*[1]

[1] Toulouse School of Economics – Toulouse School of Economics – 1 Esplanade de l'université, 31080 Toulouse Cedex 06, France

La méthode ICS (Invariant Coordinate Selection) est une méthode statistique introduite par Tyler et al. (2009) basée sur la diagonalisation jointe de deux matrices de dispersion. Une approche de cette méthode basée sur un modèle et appelée ICA (Invariant Coordinate Analysis) a déja été adaptée aux données de composition par Muehlmann et al. (2021). Dans une approche sans modèle, ICS est aussi efficace pour détecter des valeurs atypiques, voir Nordhausen et Ruiz-Gazen (2022). Nous proposons de développer une version d'ICS pour détecter des atypiques dans les données de composition. Cette méthode est d'abord introduite dans un espace de coordonnées ilr associé à une matrice de contraste et suit la procédure de Archimbaud et al. (2018b). Nous montrons ensuite que la technique est indépendante du choix de matrice de contraste et peut être entièrement définie dans le simplexe. Pour cela nous établissons d'abord un ensemble de propriétés des matrices satisfaisant la propriété "somme-zero" et introduisons une définition de la distance de Mahalanobis adaptée au simplexe ainsi que la classe des M-estimateurs de dispersion en une étape. Nous définissons également la famille de distributions elliptiques dans le simplexe. Nous montrons ensuite comment interpréter les résultats directement dans le simplexe pour deux jeux de données simulées ainsi qu'un jeu de données de parts de marché pour le marché automobile.

12h50 – 14h20 Déjeuner

14h20 – 15h40

Extrêmes

(Amphi 7 - 14h20-15h40)

On the estimation of extreme quantiles with neural networks

Michael Allouche*[1], Stéphane Girard[2], Emmanuel Gobet[3]

[1] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[2] Statify – Laboratoire Jean Kuntzmann, INRIA, Université Grenoble Alpes, Institut polytechnique de Grenoble (Grenoble INP) – France
[3] Ecole Polytechnique [Palaiseau] – Ecole Polytechnique – École Polytechnique, 91128 Palaiseau Cedex, France

In this communication, we propose a new parametrization for one-hidden layer neural networks able to estimate extreme quantiles, starting from data from heavy-tailed distributions. We provide an analysis of the uniform error between the extreme log-quantile and its neural network approximation. Numerical experiments are conducted on simulated data to compare the performance of our method with other estimators from the literature.

Théorème de Donsker pour des mesures empiriques locales sur des régions aléatoires

Benjamin Bobbia*[1], Clément Dombry[2], Davit Varron[2]

[1] Ecole supérieur d'ingénieur Léonard de Vinci – Association Léonad de Vinci – France
[2] Laboratoire de mathématiques de Besançon – Université de Besançon – France

Les processus empiriques sont aujourd'hui des objets bien connus. Une des raisons qui a poussées au d'eveloppement de l'étude des processus empiriques est qu'il est possible, dans de nombreuses modélisations, d'écrire les estimateurs comme images de mesures empiriques. Dans ce travail nous regardons le cas particulier des mesures empiriques locales, c'est-à-dire, la mesure empirique construite sur un sous-échantillon d'observations conditionnées à un être dans une certaine partie de l'espace. De nombreux résultats existent pour ce type de mesure, mais que peut-on dire si la partie de l'espace en question dépend des données ? Il est possible de s'en sortir au prix d'une grande technicité et de conditions de régularité, le propos de ce travail et de présenter un cadre général pour l'étude de ces mesures empiriques particulières permettant d'obtenir des résultats asymptotiques à moindre cout (technique et conditions). Les résultats sont présentés ici au travers du prisme de la théorie de valeurs extrêmes et nous proposons un exemple qui met en lumière la pertinence de cette approche notamment lorsqu'elle est couplée avec des méthodes issues du transport optimal.

Non-parametric estimator of a multivariatemadogram for missing-data and extreme valueframework

Alexis Boulin*[1]

[1] Laboratoire J.A. Dieudonné – UMR CNRS 7351, Université Côte d'Azur – France

Modéliser la dépendance entre maxima est un sujet d'intérêt dans les domaines d'application d'analyse du risque. Dans cet objectif, la copule de valeurs extrêmes, caractérisée par le madogramme, peut être utilisée comme une description de la structure de dépendance. Concrètement, la famille des distributions à valeurs extrêmes est très riche et survient naturellement comme la limite composante par composante des maxima préalablement normalisés. Dans cette présentation, nous étudions l'estimation non paramétrique du madogramme lorsque les données sont absentes complètement au hasard. Nous fournissons un théorème de la limite centrale fonctionnelle pour l'estimateur considéré du madogramme, correctement normalisé, vers un processus Gaussien tendu pour lequel la fonction de covariance dépend des probabilités de perte de la donnée. L'expression explicite de la variance asymptotique est aussi donnée. Nos résultats sont illustrés dans une étude numérique lorsque la taille d'échantillon est finie.

Méthode de plug-in pour l'estimateur non-paramétrique du coefficient de dépendance de queues

Matthieu Garcin*[1]

[1] Pôle Universitaire Léonard de Vinci – Research Center Pôle Léonard de Vinci – France

Une expression théorique est donnée pour l'erreur quadratique moyenne asymptotique d'un estimateur non-paramétrique du coefficient de dépendance de queues, en fonction d'un seuil qui définit quel rang délimite les queues d'une distribution. Nous proposons une nouvelle méthode pour sélectionner de manière optimale ce seuil. Il combine l'erreur quadratique moyenne théorique de l'estimateur avec une estimation paramétrique de la copule reliant les observations dans les queues. À l'aide de simulations, nous comparons cette méthode semi-paramétrique avec d'autres approches proposées dans la littérature, y compris l'algorithme de recherche de plateau.

Estimation fonctionnelle de L^1-expectiles multivariés extrêmes à queue lourde

Cambyse Pakzad*[1], Elena Di Bernardino[1], Thomas Laloe[1]

[1] Laboratoire J.A. Dieudonné – CNRS UMR 7351 – France

Les expectiles et les quantiles peuvent être interprétés comme des solutions de problèmes de minimisation convexes mais seuls les expectiles constituent une mesure de risque, invariante par loi, cohérente au sens d'Artzner et al (1998) et élicitable au sens de Gneiting (2011). Nous étudions ici une généralisation au cas multivarié, à savoir le L^1-expectile. Nous adressons la question de son estimation dans le cadre extrême, avec des distributions sous-jacentes à queue lourde et lorsqu'une covariable fonctionnelle est disponible. Pour cela, nous avons recours à une approche de minimisation qui nécessite l'estimation de différents objets tels que l'indice de queue conditionnel ou les fonctions de dépendance de queue bivariées. Nous abordons la question de la vitesse de convergence de notre méthode d'estimation. Pour ce faire, nous explorons en particulier la théorie de la variation régulière.

Statistique mathématique

(Amphi 8 - 14h20-15h40)

Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models

Marie Du Roy De Chaumaray[*1,2], Matthieu Marbac[2,3]

[1] Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz] – Centre de Recherche en Économie et STatistique (CREST) – France
[2] Centre de Recherche en Économie et Statistique (CREST) – France
[3] ENSAI – Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Cet article s'intéresse à l'estimation du modèle complet dans des modèles de mélange non-paramétriques finis. Il présente une approche pour sélectionner le nombre de composantes ainsi que le sous-ensemble de variables discriminantes, c'est-à-dire le sous-ensemble de variables ayant des distributions différentes pour les composantes du mélange. L'approche proposée repose sur la discrétisation de chaque variable en B intervalles et sur une pénalisation de la log-vraisemblance qui en résulte. Pour un choix judicieux du terme de pénalité, on montre la consistance de l'estimateur du modèle (nombre de composantes et sous-ensemble de variables discriminantes) lorsque le nombre d'intervalles tend vers l'infini avec la taille de l'échantillon.

Modèles linéaires généralisés multivariés à composantes supervisées et facteurs latents, avec partitionnement thématique des variables explicatives

Julien Gibaud*[1], Xavier Bry[1], Catherine Trottier[1,2]

[1] IMAG – CNRS, Université de Montpellier – France
[2] Université Paul-Valéry - Montpellier 3 – Université Paul-Valéry - Montpellier 3 – France

A l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) a été conçue pour trouver, au sein de très nombreuses covariables redondantes, des composantes explicatives conjointement supervisées par plusieurs réponses, ce qui est nécessaire dans un contexte de grande dimension. Plus tard, SCGLR fut améliorée pour chercher des composantes au sein de variables explicatives partitionnées en thèmes. Dans ce travail, nous proposons d'étendre cette méthode en modélisant la matrice de variance-covariance conditionnelle des réponses de telle sorte que la covariance conditionnelle des réponses soit principalement expliquée par quelques facteurs. Nous chercherons donc non seulement à extraire des thèmes des composantes explicatives, mais aussi à identifier des blocs dans la matrice de variance-covariance des réponses conditionnellement à ces composantes. Après la linéarisation du modèle, un algorithme combinant EM et celui de SCGLR thématique est proposé afin d'estimer les paramètres du modèle.

Calibrations de pénalités pour la régression linéaire gaussienne en grande dimension

Perrine Lacroix*[1,2]

[1] Institute of Plant Sciences - Paris Saclay (IPS2), INRA-CNRS-Université Paris Sud, Labex Saclay Plant Science, Orsay, UMR 518 INRA AgroParisTech Mathématiques et Informatique appliquées (MIA), Paris. France. – Institut national de la recherche agronomique (INRA) : UMR518 – France
[2] Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8628 – France

Dans un contexte de grande dimension, une approche classique pour estimer le paramètre inconnu en régression linéaire gaussienne est de minimiser les moindres carrés pénalisés. Pour obtenir une inégalité oracle sur le risque prédictif, la théorie développée par (Birgé et Massart, 2001) fournit des formes de pénalité connues à constantes multiplicatives près. Cependant, le groupe de variables sélectionnées ne correspond pas aux variables dites actives, c'est-à-dire celles intervenant réellement dans la régression linéaire pour expliquer Y. Ainsi, contrôler la qualité de prédiction n'est pas suffisant pour limiter la sélection des variables dites inactives (par opposition aux variables actives). C'est pourquoi, dans un premier temps et sous un modèle simplifié, nous proposons une étude théorique et une heuristique dépendante des données permettant de calibrer l'une des constantes de la pénalité pour éviter la sélection de variables inactives tout en conservant une bonne qualité de prédiction. Dans un second temps, nous proposons un algorithme qui étend

le principe de l'heuristique de pente pour calibrer les deux constantes restantes tout en assurant un contrôle du risque prédictif.

Beyond L1: Faster and Better Sparse Models with skglm

Mathurin Massias*[1], Quentin Bertrand[2], Quentin Klopfenstein[3]

[1] Inria – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[2] MILA – Canada
[3] Université du Luxembourg – Luxembourg

Nous proposons un nouvel algorithme rapide pour les modèles linéaires généralisés parcimonieux avec pénalités séparables non convexes. Notre algorithme est capable de résoudre des problèmes avec des millions de variables et d'observations en quelques secondes, en s'appuyant sur la descente de coordonnées, les working sets et l'accélération d'Anderson. Une version longue est disponible à : https://arxiv.org/abs/2204.07826.

ISDE : Estimation de Densité Multidimensionnelle sous Hypothèse de Structure D'indépendance

Louis Pujol*[1]

[1] Laboratoire de Mathématiques d'Orsay – CNRS, Université Paris Sud, Université Paris Saclay, INRIA Saclay Ile-de-France – France

Nous nous intéressons au problème de l'estimation de densité d'un point de vue pratique et dans un contexte multidimensionnel. Pour une classe de régularité de type Hölder sur des densités d-dimensionnelles, la convergence du risque minimax en perte L_2 au carré est d'autant plus lente que la dimension est élevée. Pour y remédier, on peut ajouter l'hypothèse de structure d'indépendance , consistant à supposer que la densité inconnue peut se décomposer comme un produit de certaines de ses marginales de dimension au plus k. Nous présentons ISDE (Independence Structure Density Estimation), un algorithme permettant d'estimer simultanément une partition des variables et un estimateur de densité comme produit de marginales s'appuyant sur cette partition en maximisant un critère de log-vraisemblance. Nous prouvons l'efficacité de la méthode sur données simulées et son intérêt sur des données réelles issues d'expériences de cytométrie de masse.

Statistique mathématique 2

(Amphi 9 - 14h30-15h50)

Comparaison d'arbres CART par sous-échantillonnage sans remplacement

Felix Cheysson*[1]

[1] Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France

Les arbres CART sont des méthodes non paramétriques intéressantes pour les problèmes de classification et de régression, car ils peuvent souvent modéliser des relations complexes entre covariables et résultats sans connaissances préalables, et le processus de décision s'apparente à la façon dont les humains prennent des décisions. Cependant, ces méthodes sont très sensibles à l'ensemble d'apprentissage, ce qui peut poser des problèmes lors de la comparaison des arbres construits à partir de deux échantillons différents.
Nous proposons un test pour la comparaison d'arbres CART, reposant sur la théorie des U-statistiques, où le seuil critique du test d'hypothèse est estimé par sous-échantillonnage sans remplacement. Nous prouvons notamment un théorème central limite pour cet estimateur. Une courte étude de simulation et une application aux données Covid-19 illustrent les performances de la méthode proposée.

Online multiple-testing with super-uniformity reward

Iqraa Meah*[1], Etienne Roquain[2], Sebastian Döhler[3]

[1] Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France
[2] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France
[3] Hochschule Darmstadt, University of Applied Sciences – Allemagne

Le test multiple en ligne fait référence au contexte où un nombre potentiellement infini d'hypothèses sont testées, et les p-valeurs sont disponibles une par une, de manière séquentielle. Cela diffère du contexte classique où le nombre d'hypothèses à tester $m < \infty$ est connu à l'avance, et les p-valeurs sont toutes disponibles ensemble. Les méthodes en ligne existantes peuvent souffrir d'une perte de puissance significative lorsque les p-valeurs sont obtenues à partir de tests discrets. Pour résoudre ce problème, nous introduisons une méthode de "récompense" basée sur les fonctions de répartitions des p-valeurs sous l'hypothèse nulle. Nous prouvons que les procédures

récompensées gardent le contrôle de l'erreur de type I tout en permettant plus de découvertes, et illustrons leurs performances sur des simulations et sur une application en biologie.

Test de linéarité dans un modèle de régression non paramétrique

Zaher Mohdeb[*1,2], Abdelkader Mokkadem[3]

[1] Faculté de Génie des Procédés, Université Salah Boubnider de Constantine 3 – Algérie
[2] Laboratoire de Mathématiques et Sciences de la Décision, Université frères Mentouri, Constantine – Algérie
[3] Laboratoire de Mathématiques de Versailles – Université de Versailles-Saint-Quentin-en-Yvelines (UVSQ), Paris, France. – France

Une procédure de test d'hypothèse linéaire sur la fonction de régression f dans un modèle de régression non paramétrique est proposée. Plus précisément, on teste l'hypothèse que f est un élément de E, où E est un espace vectoriel de dimension finie. En supposant que les fonctions considérées sont höldériennes d'ordre plus grand que $1/2$ et on obtient le comportement asymptotique du test proposé, on a donc ainsi le niveau et la puissance asymptotique du test. Une étude par simulation a été menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposé.

Multi-sample comparison of multivariate copulas and clustering

Yves Ismaël Ngounou Bakam[*1], Denys Pommeret[2]

[1] Institut de Mathématiques de Marseille – Centre National de la Recherche Scientifique : UMR7373, Ecole Centrale de Marseille : UMR7373, Aix Marseille Université : UMR7373 – France
[2] ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, – ISFA – France

The comparison of copulas is a major challenge in copula modeling. Copulas are still an important topic with many applications in finance, actuarial science, environmental data, ecology, and so on. They can be used to model the dependence structures of multivariate observations.
In the two-sample case, Rémillard and Scaillet (2009) proposed a test to compare two nonparametric copulas, that is to test $H_0 : C_1 = C_2$, where C_1 and C_2 are two copulas observed on two iid samples, which may be paired.
To our knowledge, there is no extension to the K sample case. However, the increasing amount of data requires sometimes more comprehensive analyzes. It is in this sense that we propose an equality test of K copulas simultaneously when K populations are observed. We propose to test the following hypothesis:

$$H_0 : C_1 = ... = C_K,$$

From K iid samples, possibly paired. It is therefore a generalization of Rémillard and Scaillet

(2009). However, we obtain the exact asymptotic distribution of the test statistic and the convergence of the test. The idea of the test is to transform the observations to uniform laws, then to use the decomposition of the density of the copula in the Legendre polynomials orthogonal basis. Returning to the copula function we obtain what are called copula coefficients which characterize each copula. The test then allows to simultaneously comparing these coefficients. The number of involved copula coefficients is automatically selected by a data driven selection. In addition, we suggest a clustering algorithm to classify populations with similar forms of dependence structure. A simulation study, analyse the level and the power of the test to show the good behavior of our test procedure and its performances to Rémillard and Scaillet (2009) approach within two-sample case. We provide some illustrations of the test and clustering through a real datasets in insurance and finance to demonstrate the method.

Consistance de l'intervalle de confiance plug-in en régression

Ahmed Zaoui*[1], Christophe Denis[1], Mohamed Hebiri[1]

[1] LAMA-Laboratoire d'Analyse et de Mathématiques Appliquées – Université Gustave Eiffel : UMR8050 – France

Nous nous intéressons au problème de régression hétéroscédastique. L'objectif est de construire un ensemble de confiance auquel la variable réponse Y appartient avec une grande probabilité. L'ensemble optimal repose sur le seuillage de la densité conditionnelle. La construction de cet ensemble repose sur une procédure d'estimation semi-supervisée en deux étapes. Dans une première étape, à partir d'un premier échantillon de données étiquetées, nous estimons la densité conditionnelle. Dans une deuxième étape, nous calibrons le seuil à l'aide d'un deuxième échantillon de données non étiquetées. L'ensemble construit offre des garanties de consistance et de bonnes performances numérique.

SFdS - Biométrie

(Amphi 10 - 14h30-15h50)

Modélisation conjointe des relations temporelles entre marqueurs longitudinaux multivariés de progression de la Maladie d'Alzheimer et événements cliniques

Anaïs Rouanet*[1]

[1] Bordeaux population health – Université de Bordeaux, Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED), Institut National de la Santé et de la Recherche Médicale : U1219 – France

La richesse des données de biomarqueurs désormais disponibles dans les cohortes sur le vieillissement offre l'opportunité de mieux appréhender les mécanismes complexes de la maladie d'Alzheimer (MA), ce qui est essentiel à l'amélioration de stratégies de prévention et de prise en charge des patients. Cependant, les outils statistiques actuels ne permettent pas de modéliser conjointement les multiples biomarqueurs de la MA tout en capturant leurs relations temporelles. Nous proposons un nouvel outil de modélisation conjointe pour décrire la dynamique complexe des différentes dimensions impliquées dans la progression de la MA et appréhender leurs relations temporelles causales. Ce modèle causal dynamique combine un modèle mixte multivarié avec équations de différence pour expliquer l'évolution dans le temps de chaque dimension en fonction des caractéristiques des autres. Les associations possibles avec le diagnostic de la MA et le décès sont prises en compte via une approche de modélisation conjointe à effets aléatoires partagés. La méthodologie est appliquée à la cohorte prospective française PAQUID pour décrire les relations temporelles entre la cognition, la dépression et l'autonomie fonctionnelle, en lien avec les deux principaux événements cliniques de la progression de la MA : le diagnostic de démence et le décès. Ce travail a pour but d'aider à comprendre l'interaction complexe entre les biomarqueurs de la MA et à identifier les cibles pertinentes capables de ralentir la progression à chaque stade de la maladie.

Evaluation par la modélisation des stratégies de rappel vaccinal contre la Covid-19 dans une population partiellement vaccinée

Clement Massonaud[1], Jonathan Roux[1], Vittoria Colizza[2], Pascal Crepey*[1]

[1] EHESP – Univ Rennes, EHESP, REPERES (Pharmacology and health services research) - EA 7449, F-35000 Rennes, France – France
[2] INSERM – Institut National de la Santé et de la Recherche Médicale - INSERM U1219 – France

Alors que les preuves montrant que l'immunité vaccinale contre le COVID-19 diminue avec le temps et l'apparition de nouveaux variants, plusieurs pays mettent en œuvre des campagnes de rappel. L'objectif de cette étude est d'analyser la morbidité et la mortalité associée à différentes stratégies de primo-vaccination et de rappel contre la COVID-19, en utilisant la France comme cas d'étude. Nous utilisons un modèle compartimental déterministe, structuré par âge, calibré sur les données d'admission à l'hôpital et validé par rapport aux données de séroprévalence en France. Ce modèle permet d'analyser l'impact des stratégies de rappel vaccinal sur la morbidité et la mortalité en supposant une diminution de l'immunité et une transmissibilité accrue du virus pendant l'hiver. L'efficacité des stratégies de rappel ciblant différentes tranches d'âge varie selon les niveaux de transmissibilité du virus, et selon la perte d'immunité supposée pour chaque tranche d'âge. Si la baisse de l'immunité touche toutes les tranches d'âge, les personnes âgées de 30 à 49 ans devraient être boostées en priorité, même pour des niveaux de transmissibilité faibles. Si la réduction de l'immunité est limitée aux personnes âgées de plus de 65 ans, le rappel vaccinal des personnes plus jeunes ne devient efficace qu'au-delà de certains niveaux de transmissibilité. L'augmentation de la couverture vaccinale primaire doit rester une priorité pour

réduire la morbidité et la mortalité dues à la COVID-19. Si un plateau de primo-vaccination a été atteint, le renforcement de l'immunité dans les groupes d'âge les plus jeunes pourrait prévenir plus d'hospitalisations et de décès que le renforcement de l'immunité des personnes âgées, en particulier dans des conditions d'augmentation la transmissibilité du SARS-CoV-2, ou face à de nouveaux variants.

Quantifying horizontal pleiotropy in human genetic variation using GWAS summary statistics

Marie Verbanck*[1,2]

[1] EA 7537 - BioSTM (Biostatistique, Traitement et Modélisation des données biologiques) – Université de Paris – France
[2] Institute for Personalized Medicine, Icahn school of medicine at Mount Sinai, NYC – États-Unis

Invitation de la part de la Societe Française de Biometrie

Horizontal pleiotropy, where one variant has independent effects on multiple traits, is crucial for our understanding of the genetic architecture of human traits. Here we have developed a method to model and quantify horizontal pleiotropy using publicly-available summary statistics from published genome-wide association studies. We developed a method to quantify horizontal pleiotropy through a variant-level pleiotropy quantification and perform validation using simulations. When applied to 1,564 medical phenotypes measured in 337,119 humans from the UK Biobank, our pleiotropy score detected a significant excess of horizontal pleiotropy. This signal of horizontal pleiotropy was pervasive throughout the human genome and across a wide range of phenotypes, but was especially prominent in regions of high linkage disequilibrium and among highly polygenic phenotypes. We identified thousands of loci with extreme horizontal pleiotropy, a majority of which had never been reported in any published GWAS. These findings suggest that horizontal pleiotropy is pervasive in human genetic variation, and has significant implications for our understanding of the genetic architecture of complex traits and disease.

SFdS - MALIA - SSFAM

(Amphi 11 - 14h30-15h50)

Analyse statistique de la topologie de Mapper pour des filtres stochastiques

Mathieu Carrière*[1], Bertrand Michel[2]

[1] DataShape – CRISAM - Inria Sophia Antipolis - Méditerranée – France
[2] Laboratoire de Mathématiques Jean Leray – École Centrale de Nantes, Ecole Centrale de Nantes – France

Mapper est un descripteur commun issu de l'analyse topologique de données, et utilisé dans une grande variété d'applications de la science des données, allant de la biologie algorithmique à la visualisation. Ceci dit, sa statistique, et plus précisément sa stabilité et vitesse de convergence vers sa version limite, appelée espace de Reeb, est encore mal comprise et constitue une question ouverte de la communauté. Dans cette présentation, nous proposons des bornes supérieures sur l'espérance de la distance entre Mapper et espace de Reeb. Notre approche inclut en particulier le cas où le filtre utilisé dans le calcul de Mapper est lui-même stochastique, comme c'est le cas pour les fonctions propres d'une ACP.

Graph Neural Networks on Large Random Graphs

Samuel Vaiter*[1]

[1] Laboratoire J.A. Dieudonné – Centre National de la Recherche Scientifique, Université Côte d'Azur (UCA) – France

Graph Neural Networks (GNNs) are deep architectures defined over graph data that have garnered a lot of attention in recent years. In this talk, I will give some insight on how such architectures behave on random graphs. I will first give non-asymptotic convergence bounds of GNNs toward "continuous" equivalents as the number of nodes grows. Then, I will show their stability to small deformations of the underlying random graph model, a crucial property in traditional CNNs. Finally, I will discuss universality and approximation power with respect to traditional graph tools. This is joint work with Nicolas Keriven and Alberto Bietti.

Apprentissage pour l'amélioration de surfaces définies par des nuages de points

Julie Digne*[1]

[1] Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Université Claude Bernard

Lyon 1, Centre National de la Recherche Scientifique : UMR5205 – France

Les formes 3D acquises par un scanner laser sont souvent données sous la forme d'un ensemble de points qui peuvent être bruités ou avoir une résolution limitée et des données manquantes. Dans cet exposé, j'expliquerai comment la prise en compte de l'auto-similarité des surfaces peut aider à surmonter ces problèmes. En effet, de nombreuses surfaces présentent des textures et des structures géométriques répétées qui peuvent être exploitées pour améliorer l'acquisition de surface. En concevant des descripteurs de forme locaux et en agrégeant leurs informations, il est possible de débruiter une surface ou d'en calculer la super-résolution basée sur des données. Les nuages de points peuvent aussi comporter des parties non 2-variété : par exemple, les câbles de rue peuvent être considérés comme des courbes, en fonction de la précision d'acquisition. Nous définissons des Local Probing Fields adaptés à ce cas particulier et les analysons à l'aide d'algorithmes d'apprentissage de dictionnaires. Cela nous permet de revoir diverses tâches de traitement de forme telles que le débruitage et le rééchantillonnage de forme.

15h40 – 16h00 Pause

16h00 – 17h00

PLENIERE : Clémentine Prieur

(Amphi 7 - 16h00-17h00)

(Non)linear dimension reduction of input parameter space using gradient information

Clémentine Prieur

Univ. Grenoble

Many problems that arise in uncertainty quantification, e.g., integrating or approximating multivariate functions, suffer from the curse of dimensionality. The cost of computing a sufficiently accurate approximation grows indeed dramatically with the dimension of input parameter space. It thus seems important to identify and exploit some notion of low-dimensional structure as, e.g., the intrinsic dimension of the model. A function varying primarily along a a low dimensional manifold embedded in the high-dimensional input parameter space is said of low intrinsic dimension. In that setting, algorithms for quantifying uncertainty focusing on the most relevant features of input parameter space are expected to reduce the overall cost. Our presentation goes from global sensitivity analysis to (non)linear gradient-based dimension reduction, generalizing the active subspace methodology.

PLENIERE : Aurélien Bellet

(Amphi 10 - 16h00-17h00)

Differentially Private Machine Learning

Aurélien Bellet

INRIA, Lille

Personal data is being collected at an unprecedented scale by businesses and public organizations, driven by the progress of data science and machine learning. While such data can be turned into useful knowledge about the global population by computing aggregate statistics or training machine learning models, this can also lead to undesirable disclosure of personal information. We must therefore deal with two conflicting objectives: maximizing the utility of data while protecting the privacy of individuals whose data is used in the analysis.

In this talk, I will present differential privacy (DP), a statistical definition of privacy which comes with rigorous guarantees as well as an algorithmic framework that allows the design of practical privacy-preserving algorithms. I will then discuss the application of DP to machine learning, and some related open questions.

17h00 – 17h20 Pause

17h20 – 18h40

Graphes – Réseaux

(Amphi 7 - 17h20-18h35)

Détection des structures communes dans une collection de réseaux

Pierre Barbillon*[1,2]

[1] Mathématiques, Informatique Appliquées (MIA) – AgroParisTech, Institut national de la recherche agronomique (INRA) : UMR0518 – 16 rue Claude Bernard, 75005 Paris, France
[2] AgroParisTech/INRA – Institut national de la recherche agronomique (INRA) – Paris, France

Soit une collection de réseaux constituée d'un ensemble de réseaux qui ne partagent pas de nœuds mais qui décrivent le même type d'interactions observées dans des situations ou des contextes différents. Une hypothèse est que les réseaux de la collection partagent une structure commune puisque la nature des interactions est la même. Nous proposons de nous appuyer sur le modèle à blocs stochastiques (SBM) pour identifier la structure commune de la collection. Le SBM est un modèle probabiliste qui suppose l'existence de variables latentes représentant les groupes de nœuds (blocs) du réseau et dont les paramètres fournissent une description succincte de la structure du réseau à l'échelle mésoscopique. Nous appelons colSBM notre extension du SBM à une modélisation conjointe d'une collection de réseaux. Les réseaux de la collection sont supposés être des réalisations indépendantes de différents SBM, qui partagent - à travers des paramètres communs - la même structure de connectivité, éventuellement aux proportions de blocs et/ou un facteur de densité près.
Les paramètres du modèle sont estimés et les blocs latents sont retrouvés à l'aide d'un algorithme EM variationnel. Nous utilisons un critère ad-hoc, basé sur la vraisemblance de classification intégrée pour sélectionner le nombre de blocs et évaluer l'adéquation du consensus trouvé entre les structures des différents réseaux. Des applications sur des réseaux trophiques sont présentées afin d'illustrer l'intérêt de ces modèles.

Clustering networks with textual edges by combining the Embedded Topic Model and the Stochastic Block model. Variational inference and derivation of a model selection criterion.

Rémi Boutin*[1], Pierre Latouche[1], Charles Bouveyron[2,3]

[1] Mathématiques Appliquées Paris 5 – Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Centre National de la Recherche Scientifique : UMR8145, Université de Paris : UMR8145 – France
[2] Laboratoire J.-A. Dieudonné, UMR CNRS 7531, Université Côte d'Azur – CNRS : UMR7531 – France
[3] Equipe Maasai, Inria Sophia Antipolis – Institut National de Recherche en Informatique et en Automatique – France

Communication networks (emails, social networks, IOT) are now ubiquitous and their analysis has become a strategic field to secure our numerical lives. Unfortunately, most of communication networks come with textual data on the edges which are not incorporated in classical model. In this paper, we introduce the embedded topics for the stochastic block model (ETSBM) in order to simultaneously perform clustering on the nodes while modeling the topics used between the different clusters. ETSBM extends both the stochastic bloc model and the embedded topic model. The main motivation of this work is to allow a simultaneous analysis of texts and node interactions. A variational-Bayes expectation-maximisation algorithm (VBEM) combined with a stochastic gradient descent (SGD) is used to perform inference. A model selection criterion is also derived. The methodology is evaluated on synthetic data and on a real world dataset.

Modèles génératifs de graphes : application à la connectivité cérébrale

Clément Mantoux*[1,2,3], Stanley Durrleman[1,2], Stéphanie Allassonnière[4,5]

[1] Algorithms, models and methods for images and signals of the human brain – Sorbonne Université : UM75, Inria de Paris, Institut du Cerveau et de la Moëlle Epinière = Brain and Spine Institute – France
[2] Institut du Cerveau et de la Moëlle Epinière = Brain and Spine Institute – Institut National de la Santé et de la Recherche Médicale : U1127, CHU Pitié-Salpêtrière [APHP], Sorbonne Université : UM75, Centre National de la Recherche Scientifique : UMR7225, CHU Pitié-Salpêtrière [AP-HP] – France
[3] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[4] Université Paris-Descartes – Centre de Recherche des Cordeliers – France
[5] Health data- and model- driven Knowledge Acquisition – Inria de Paris, Centre de Recherche des Cordeliers – France

La connectivité fonctionnelle est l'étude de réseaux d'interactions entre les régions du cerveau humain. Pour chaque individu, on peut ainsi mesurer une matrice de connectivité, qui varie beaucoup d'un sujet à l'autre. En raison du faible nombre d'observations et de leur grande

dimension, de nombreux modèles statistiques classiques ne peuvent être employés en pratique pour décrire cette variabilité. Nous présentons un modèle hiérarchique simple pour contourner ces difficultés en s'appuyant sur la structure de rang faible des matrices de connectivité. Nous montrons que notre modèle est identifiable, et que son estimateur est consistant et asymptotiquement normal. Enfin, nous appliquons cet algorithme à une base de données de matrices de connectivité cérébrale. Nous montrons que notre modèle donne une description précise, compacte et interprétable de la distribution des données.

Analyse statistique de graphes, via des processus de diffusion de la chaleur.

Etienne Lasalle*[1,2]

[1] Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8628 – France
[2] DataShape – Inria Saclay - Ile de France – France

Ce travail porte sur la comparaison de données de graphes, potentiellement pondérés et de tailles différentes. Lorsqu'on travaille avec des graphes pondérés, on peut interpréter les poids des arêtes comme des conductivités thermiques. Dès lors, on peut comparer les graphes en comparant leur répartition de chaleur après un temps de diffusion t. Ce paramètre d'échelle t doit être minutieusement choisi pour s'assurer des comparaisons pertinentes. A l'opposé de précédents travaux considérant un temps de diffusion fixé arbitrairement ou choisi à partir des données, on propose de prendre en compte tout le processus de diffusion. Pour cela, on définit des processus à valeurs réelles indexés par tous les temps de diffusion dans (0,T), en concaténant les comparaisons faites aux différentes échelles. Dans cet exposé, nous commencerons par présenter ces processus de comparaison de graphes et leurs propriétés statistiques. Puis, nous montrerons comment en dériver des tests à deux échantillons consistants. Nous présenterons quelques applications sur des jeux de données synthétiques et réels. On s'intéressera notamment aux données provenant de graphes d'activations de réseaux de neurones.

Convergence en loi des U-statistiques sur une matrice exchangeable ligne-colonne

Tâm Le Minh*[1]

[1] MIA-Paris – Université Paris Saclay, INRAe, AgroParisTech, UMR MIA-Paris, 75005, Paris, France – France

Les U-statistiques sont utilisées pour estimer les paramètres d'une population en moyennant une fonction d'un sous-ensemble sur tous les sous-ensembles de cette population. Notre travail porte

sur une population formée par les valeurs d'une matrice échangeable ligne-colonne. On considère les U-statistiques issues de fonctions sur des quadruplets, c'est-à-dire des sous-matrices de taille 2 x 2. On démontre un résultat de convergence faible pour ces U-statistiques et on établit un Théorème Central Limite dans le cas où la matrice est dissociée. Les matrices échangeables ligne-colonne sont une représentation naturelle des réseaux bipartites échangeables, nous appliquons donc les résultats à l'inférence statistique pour les réseaux.

Séries Temporelles

(Amphi 8 - 17h20-18h35)

Estimation Adaptative des Variances dans un Modèle Espace-État

Joseph De Vilmarest[*1,2], Olivier Wintenberger[3]

[1] EDF Labs – EDF Recherche et Développement – France
[2] Laboratoire de Probabilités, Statistique et Modélisation – Université Paris Diderot - Paris 7 : UMR8001, Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001 – France
[3] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France

Nous considérons la prévision adaptative de séries temporelles. Nous étudions un modèle espace-état linéaire gaussien dans lequel les variances sont inconnues, et évoluent potentiellement au cours du temps. Nous représentons les variances comme des variables latentes auxiliaires dans un modèle espace-état augmenté en mode tracking. L'inférence repose sur le cadre online Variational Bayes, qui consiste à approcher la loi jointe a posteriori par la meilleure distribution produit, la meilleure au sens de la divergence de Kullback-Leibler (KL). Nous obtenons l'algorithme Viking, qui optimise la divergence KL à chaque étape de manière récursive. Cet algorithme est robuste car il fonctionne sur des données mal spécifiées, au sens où elles ne sont pas générées par le modèle espace-état considéré.

On multi-change-points tests based on multi-samples U-Statistics for weakly dependent observations

Echarif El Harfaoui[*1], Michel Harel[2], Joseph Ngatchou-Wandji[3]

[1] Université Chouaib Doukkali – Maroc
[2] Université de Limoges – INPE – France
[3] Université de Lorraine – IECL – France

In this paper, we study multi-change-points detection test using multi-sample tests based on U-statistics for absolutely regular observations. Our results extend the results of Ngatchou-Wandji et al. (2022) based on U-statistics for absolutely regular observations only in the case of one change-point detection. The asymptotic distributions of the test statistics under the null hypothesis and under the local alternatives are given explicitly and the tests are shown to be consistent.

On estimating in generalized integer-valued GARCHX models with structural breaks

Mohamed Djemaa Sadoun *[1,2], Abderaouf Khalfi [1]

[1] RECITS Laboratory, Research Center in Applied Economics for Development (CREAD) – Algérie
[2] Operational Research Department, University of Sciences and Technology Houari Boumediene (USTHB) – Algérie

Nous abordons le problème de l'estimation des paramètres dans un modèle INGARCH généralisé à changements structurels incluant des co-variables exogènes (ci-après noté GCP-INGARCHX). Cette classe de modèles appartient aux modèles de type "observation-driven" avec changement de régime où le changement de régime est entraîné par certains points de rupture survenant dans le temps. Les estimateurs des moindres carrés conditionnels (MCC) et du maximum de vraisemblance conditionnelle (MVC) des paramètres du modèle sous-jacent sont obtenus dans les deux cas: lorsque les points de rupture sont connus ou non. L'estimation de ces points de rupture est réalisée par une méthode d'estimation hors ligne. Une étude de simulation et une application sur données réelles sont fournies pour évaluer la qualité du modèle.
Mots-clés. Processus de comptage à valeurs entières, modèle INGARCH non stationnaire, estimateurs du (MCC) et (MVC), méthode d'estimation hors ligne.

Abstract. We deal with the estimation problem in a generalized INGARCH model with structural changes including exogenous covariates (hereafter referred to as GCP-INGARCHX). This class of models belongs to the observation-driven type models with regimes change where the regime-switching is driven by certain failure points occuring in the time. The conditional least squares (CLS) and the conditional maximum likelihood (CML) estimators of the underlying parameters are obtained for both the cases that the break points are known or not. An off-line estimation method is used to estimate the breaks points. A simulation study and an application on real data set are provided to assess the performance of the model.
Keywords. Integer-valued process of counts, no-stationary INGARCH model, (CLS) estimators, (CML) estimators, off-line estimation method.

Tight Risk Bound For High Dimensional Time Series Completion

Amélie Rosier*[1], Nicolas Marie[2], Pierre Alquier[3]

[1] Modélisation aléatoire de Paris X – Université Paris Nanterre : EA3454 – France
[2] Université Paris Nanterre. Laboratoire Modal'X. – Université Paris Nanterre. Laboratoire Modal'X (EA 3454). – France
[3] RIKEN Center for Advanced Intelligence Project [Tokyo] – Japon

Initially designed for independent datas, low-rank matrix completion was successfully applied in many domains to the reconstruction of partially observed high-dimensional time series. However, there is a lack of theory to support the application of these methods to dependent datas. In this presentation, we propose a general model for multivariate, partially observed time series. We show that the least-square method with a rank penalty leads to reconstruction error of the same order as for independent datas. Moreover, when the time series has some additional properties such as periodicity, the rate can actually be faster than in the independent case. This is a collaborative work with Pierre Alquier and Nicolas Marie.

Statistique mathématique – grande dimension

(Amphi 9 - 17h20-18h35)

Sparse PLS with group lasso: inside the dual.spls package

Louna Alsouki*[1,2], François Wahl[2,3], Laurent Duval[3], Clément Marteau[2], Rami El Haddad[1]

[1] Université Saint-Joseph de Beyrouth – Liban
[2] Institut Camille Jordan – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, F-6962 Villeurbanne Cedex, France
[3] IFPEN Energies Nouvelles, France

The problem of high dimensionality is usually tackled with two dimension reduction techniques: projection methods, like the PLS, or variable selection algorithms, like the lasso. Sparse PLS combines both. It aims at accuracy in predictions and in localization of important variables. We propose a generalized method: the dual sparse partial least squares (dual-SPLS) which relies on the PLS1 algorithm modified by different types of penalizations. The sparsity of its results shows

good localizations while maintaining accurate prediction performance. Thanks to its flexibility, it provides a strategy that deals with different complementary data. The method is implemented in the package dual.spls available in R. It includes additional functions, notably, for calibration splits and data simulations. In this paper, we focus on a brief presentation of the method with a demo on the use of the package. We also illustrate the application on the case where two sets of predictors are related to the same response.

Sélection de modèles linéaires emboîtés par règle d'arrêt prématuré

Samy Clementz[*1,2], Alain Celisse[1], Sylvain Arlot[2]

[1] Statistique, Analyse et Modélisation Multidisciplinaire – Université Paris 1 Panthéon-Sorbonne : EA4543 – France
[2] Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8628 – France

In many model selection problems, candidate estimators are constructed iteratively. This holds for example with Principal Component Analysis (PCA), in the denoising by projection framework, or in the ordered-variable selection framework. Classically, model selection only starts after the construction of all estimators has been completed, which is costly in terms of time and computing resources. For addressing this issue, we analyze an alternative selection method, known in the inverse problem literature as the discrepancy principle. This method stops early the iterative construction of the successive estimators. It is therefore suited to situations where time and computational resources are limited.

Régression non paramétrique par projection sur des bases à support non compact

Florian Dussap[*1]

[1] Mathématiques Appliquées Paris 5 – Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Centre National de la Recherche Scientifique : UMR8145, Université de Paris : UMR8145 – France

On etudie le problème d'estimation non parametrique d'une fonction de regression avec un design aleatoire sur $\mathbb{R}^p$ pour $p \geq 2$. On utilise pour cela un estimateur par projection calcule avec un critère de moindres carres. Notre contribution est de considerer des domaines d'estimation non compacts et d'etudier theoriquement le risque de l'estimateur pondere par la loi du design. On propose une procedure de selection de modèle dans laquelle la collection de modèles est aleatoire et prend en compte l'ecart entre la norme empirique et la norme associee à la loi du design. On demontre que l'estimateur resultant optimise automatiquement le compromis biais-variance pour

les deux normes.

Phénomène de Stein pour l'estimation multiple de moyennes en grande dimension

Jean-Baptiste Fermanian*[1]

[1] Laboratoire de Mathématiques d'Orsay – CNRS, Université Paris Sud, Université Paris Saclay – France

Nous proposons une amélioration de l'estimation multiple de moyennes, dont le but est l'estimation conjointe des moyennes de plusieurs distributions à partir d'échantillons séparés et indépendants. L'approche naïve consiste à estimer la moyenne empirique de chaque échantillon séparément, alors que la méthode que nous proposons exploite les similarités possibles entre les moyennes, sans qu'aucune information a priori ne soit connue. Tout d'abord, pour chaque échantillon, on détecte les moyennes similaires ou voisines à l'aide de tests multiples. Ensuite, chaque estimateur naïf est décalé vers la moyenne locale de ses voisins. Cette transformation fait écho à l'estimateur de James Stein où la moyenne empirique est décalée vers zéro. Ainsi, ici notre point de référence n'est pas zéro mais la moyenne des voisins de l'échantillon considéré. Bien que du biais soit ajouté, l'estimation est améliorée par la réduction de la variance. Cette amélioration peut être significative lorsque la dimension de l'espace est importante, démontrant un phénomène de "bénédiction de la dimension". Une application de cette approche est l'estimation multiple de kernel mean embeddings, qui joue un rôle important dans de nombreuses applications modernes. Les résultats théoriques sont vérifiés sur des données artificielles et réelles.

Estimation of high dimensional gamma convolutions through random projections

Oskar Laverny*[1]

[1] Institut Camille Jordan [Villeurbanne] – Université Claude Bernard Lyon 1 – France

La classe des convolutions généralisées de lois gammas est définie par une structure convolutionelle semi-paramétrique. La flexibilité de ses structures de dépendance, les possibilités marginales et la structure convolutionelle pratique de ces distributions en fait une classe intéressante pour le praticien. Cependant, l'estimation de ces distributions quand la dimension augmente est complexe et représente un défi.
Nous proposons une procédure d'estimation stochastique basée sur une approximation de l'erreur quadratique en base de Laguerre via des cumulants (shiftés), évaluée sur des projections aléatoires du jeu de donnée. À travers l'analyse de la perte via des outils venant des cubatures de Grassmanniens, de l'optimisation de mesure sparse et de flots de gradients en espace Wasserstein, nous démontrons la convergence de la descente de gradient stochastique vers un estimateur

consistent de la distribution de grande dimension.
Nous proposons plusieurs exemples en dimension base et en grande dimension.

Applications : biologie et santé

(Amphi 10 - 17h20-18h35)

Méthodes pour l'inférence post-clustering appliquées à l'expression génique

Benjamin Hivert [*1,2,3], Denis Agniel [4,5], Rodolphe Thiébaut [1,2,3,6], Boris Hejblum [1,2,3]

[1] Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France – Université de Bordeaux (Bordeaux, France) – France
[2] Inria SISTM team – Inria Bordeaux Sud Ouest – France
[3] Vaccine Research Institute, VRI – Hôpital Henri Mondor, Créteil F-94000, France – France
[4] Harvard Medical School – États-Unis
[5] RAND Corporation – États-Unis
[6] CHU Pellegrin – Groupe Hospitalier Pellegrin, Bordeaux F-33076, France – France

L'analyse des données d'expression génique est souvent organisée autour de deux étapes successives : i) une classification non supervisée utilisant l'ensemble des gènes pour regrouper les unités d'observations (patients, échantillons ou cellules) en sous-groupes distincts et homogènes ; puis ii) l'analyse différentielle se faisant à l'aide de tests d'hypothèse visant à identifier quels gènes, c'est-à-dire quelles variables, sont différentiellement exprimés entre ces sous-groupes. Cependant, cette approche utilisant les même données lors des deux étapes ne permet pas de garantir un bon contrôle de l'erreur de type I à l'étape ii).
Nous proposons deux méthodes d'inférence pour tenir compte de l'étape initiale de classification non supervisée lors de l'analyse différentielle et ainsi garantir un contrôle effectif de l'erreur de type I. La première méthode se base sur le concept d'inférence sélective tandis que la seconde repose sur une définition de la séparation de classes faisant uniquement intervenir les concepts d'unimodalité et de multimodalité. Nous avons évalué les performances des deux méthodes grâces à différentes simulations numériques, ainsi que dans une application sur un jeu de données réelles de faible dimension. Les méthodes proposées conduisent à des p-valeurs valides sous l'hypothèse nulle d'absence de différence entre les sous-groupes dans l'expression d'un gène sélectionné, indépendamment de la classification, tout en conservant une bonne puissance statistique.
En grande dimension, cette inflation de l'erreur de type I peut-être contre-balancée par la dilution du signal utilisé pour la classification, à condition que les variables soient indépendantes. En revanche, en présence de corrélation (comme c'est le cas en pratique pour l'expression génique),

des classes artificielles apparaissent alors que celles-ci ne sont pas séparables. Une adaptation des méthodes à ce contexte de grande dimension est donc nécessaire.

Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique

Charles-Elie Rabier*[1], Simona Grusea[2]

[1] IMAG – CNRS, Université de Montpellier – France
[2] Institut de Mathématiques de Toulouse UMR5219 – Institut National des Sciences Appliquées - Toulouse – France

La sélection génomique (GS) consiste à sélectionner des individus sur la base de prédictions génomiques effectuées à l'aide d'une grande densité de marqueurs. Une question d'importance en GS est de déterminer le nombre de marqueurs nécessaires pour une prédiction fiable. Pour ce faire, nous introduisons de nouveaux proxies pour la précision de la prédiction. Ces proxies sont appropriés dans le cadre d'une carte génétique discrète, où il est fréquent d'observer du déséquibre de liaison incomplet, i.e. la situation où les allèles à l'emplacement d'un gène et à l'emplacement d'un marqueur à proximité, diffèrent. De plus, nos proxies suggérés sont utiles pour concevoir des puces SNPs basées sur une densité
modérée de marqueurs. Nous analysons des données de riz des Philippines, et nous nous focalisons sur la date de floraison recueillie durant la saison sèche 2012. En utilisant différentes densités de marqueurs, nous montrons qu'au moins 1553 marqueurs sont nécessaires afin d'obtenir une prédiction fiable et de mettre en place la GS.
Déterminer le nombre optimal de marqueurs est crucial afin d'optimiser le programme de sélection.

Identification of prognostic and predictive biomarkers in high-dimensional data

Wencan Zhu*[1,2], Céline Lévy-Leduc[1], Nils Ternès[3]

[1] MIA-Paris – AgroParisTech, INRA - Université Paris-Saclay – France
[2] SANOFI R&D – Sanofi Aventis R&D [Chilly-Mazarin] – France
[3] Sanofi R&D – SANOFI Recherche – France

In clinical trials, identification of prognostic and predictive biomarkers is essential to precision medicine. Prognostic biomarkers can be useful for the prevention of the occurrence of the disease, and predictive biomarkers can be used to identify patients with potential benefit from the treatment. Previous researches were mainly focused on clinical characteristics, and the use of genomic data in such an area is hardly studied. We propose a new approach called PPLasso (Prognostic

Predictive Lasso) integrating prognostic and predictive effects into one statistical model as a variable selection problem in an ANCOVA (Analysis of Covariance) type model. PPLasso also takes into account the correlations between biomarkers that can alter the biomarker selection accuracy. Through numerical experiments, we show that PPLasso outperforms the traditional Lasso approach on both prognostic and predictive biomarker identification in various scenarios.

Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection

Tom Rohmer*[1], Anne Ricard[2,3], Ingrid David[1]

[1] Génétique Physiologie et Systèmes dÉlevage – Institut National Polytechnique (Toulouse), Université Fédérale Toulouse Midi-Pyrénées, École nationale supérieure agronomique de Toulouse [ENSAT], Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR1388 – INRAE Auzeville - 24, chemin de Borde-Rouge -Auzeville Tolosane31326 Castanet Tolosan, France
[2] Génétique Animale et Biologie Intégrative – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR1313 – Domaine de Vilvert F-78252 Jouy-en-Josas, France
[3] Institut Français du Cheval et de l'Equitation [Saumur] – IFCE – Terrefort, BP 207, F-49411 Saumur, France

Dans les modèles multi-caractères utilisés en génétique animal, les composantes de variances sont très souvent estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des caractères, bien qu'en pratique cette hypothèse n'est pas toujours réaliste. Nous avons simulé des populations mimant un schéma de sélection classique rencontré dans les élevages porcins et mesuré l'impact d'une distribution multivariée non Gaussienne sur les résidus (de part une copule non-Gaussienne) sur les estimations réalisées en pratique. Les résultats ont montré que lorsque les reproducteurs sont sélectionnés au hasard, nous n'observons aucun impact significatif sur les estimés, malgré l'hypothèse Gaussienne sous-jacente. Néanmoins, lorsque les reproducteurs sont sélectionnés de façon à améliorer les deux caractères d'intérêt par un processus de troncation basé sur les prédictions BLUP des valeurs génétiques, on observe des différences significatives avec les paramètres théoriques, en particulier avec des distributions bivariées asymétriques sur la partie résiduelle.

Évaluation de l'impact de scénarios de consommation de tabac sur la charge future de l'infarctus du myocarde en France jusqu'en 2035 : une approche basée sur un modèle illness-death

Johann Kuhn*[1], Yann Le Strat[1], Christophe Bonaldi[1], Clémence Grave[1], Valérie Olié[1], Pierre Joly[2,3]

[1] Santé publique France - French National Public Health Agency – France
[2] Univ. Bordeaux – Univ. Bordeaux – Isped, Centre INSERM U897-Epidemiologie-Biostatistique, F-33000 Bordeaux, France
[3] Bordeaux Population Health – Bordeaux Population Health U1219 Inserm - Université de Bordeaux – France

En France, l'infarctus du myocarde est une cause importante de morbidité, de recours aux soins, d'altération de la qualité de vie et de mortalité. Dans un premier travail, nous avons projeté plusieurs indicateurs de l'infarctus du myocarde (nombre de cas prévalents, prévalence, âge moyen des cas incidents) jusqu'en 2035, montrant une augmentation de la prévalence avec un quasidoublement des cas chez les hommes et les femmes en 2035 comparé à 2015 et une augmentation du nombre de cas chez les femmes relativement jeunes. Dans ce second travail, nous quantifions l'impact de la diminution du tabagisme en France sur des indicateurs épidémiologiques de l'infarctus. Nous modélisons le statut tabagique des individus et l'arrêt naturel du tabac, ce qui nous permet ensuite de créer des scénarios simulant des diminutions de consommation du tabac afin de comparer l'impact de ces différents scénarios et les résultats de nos premières projections ne tenant pas compte d'une modification de consommation du tabac.

Planification – Plans d'expérience

(Amphi 11 - 17h20-18h35)

Simulation d'événements rares par échantillonnage préférentiel adaptatif pour des processus de Markov déterministes par morceaux

Guillaume Chennetier*[1,2], Josselin Garnier[2], Anne Dutfoy[1], Hassane Chraïbi[1]

[1] EDF R&D dept. Périclès – EDF – France
[2] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France

On souhaite estimer la probabilité de défaillance de systèmes industriels multi-composants hautement fiables modélisés par des processus de Markov déterministes par morceaux. Pour réduire le nombre de simulations nécessaires par rapport à une méthode de Monte-Carlo standard, nous proposons une méthode d'échantillonnage préférentiel adaptatif par entropie croisée. La paramétrisation de cette méthode, délicate pour des systèmes de grande dimension, repose sur la notion fiabiliste de "chemin minimal".

Estimation de courbes de fragilité sismique par planification séquentielle d'expériences.

Clément Gauchy*[1], Josselin Garnier[2], Cyril Feau[1]

[1] Universite Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques – CEA, CNRS, Université Paris-Saclay, CEA Saclay 91191 Gif sur Yvette France – France
[2] École Polytechnique – CMAP, École Polytechnique – France

Les études probabilistes de sûreté sismiques consistent à évaluer les probabilités de défaillance de structures mécaniques soumises à des excitations sismiques. Ces études nécessitent l'estimation de courbes de fragilité sismique, qui sont la probabilité de défaillance de la structure conditionnellement à une mesure d'intensité du signal sismique. Cependant, leur estimation requiert de nombreuses expériences numériques qui peuvent être très coûteuses en temps de calcul, ce qui rend l'estimation par une méthode Monte Carlo inappropriée. Nous proposons dans ce papier de construire un algorithme de planification séquentielle d'expériences en supposant un a priori de processus Gaussien sur la réponse du code de calcul mécanique.

Design of Experiment for Bayesian transferred model

Loïc Iapteff*[1,2], Julien Jacques[3], Benoit Celse[1], Victor Lameiras Franco Da Costa[1]

[1] IFP Energies nouvelles (IFPEN) – IFP Energies nouvelles, Rond-point de l'échangeur de Solaize, 69360 Solaize, France – France
[2] Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon 2 – France
[3] Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon II : EA3083, Université Claude Bernard - Lyon I (UCBL) – Université Lumière Lyon 2 5 avenue Pierre Mendès-France 69676 Bron Cedex, France

Le groupe IFP commercialise des catalyseurs et doit s'engager sur leur performance. Il est donc nécessaire de disposer de modèles prédictifs fiables pour chaque nouvelle génération de catalyseurs. Ces modèles sont construits à partir de données expérimentales très couteuses. Afin d'optimiser les coûts, notre ambition est de réduire le nombre d'expérimentations nécessaires pour estimer un modèle associé à un nouveau type de catalyseur. De précédents travaux ont montré qu'une approche de transfert Bayésien permettait d'améliorer la qualité des modèles lorsque le nombre d'observations est réduit. Dans cet article, les plans d'expériences sont étudiés afin de déterminer comment sélectionner ce nombre réduit d'observations permettant d'obtenir les meilleurs modèles par transfert Bayésien.
Cet article montre que l'algorithme Kennard and Stone peut, sous certaines conditions, offrir de meilleurs résultats que des plans optimaux.

Données expérimentales à bas coût pour la prévision de la distance parcourue par une avalanche de pierres

Frederique Leblanc*[1]

[1] Laboratoire Jean Kunzmann (LJK) – CNRS : UMR5224 – France

Les chutes de bloc rocheux constituent un risque important dans les zones montagneuses. La trajectoire de chaque bloc est aléatoire et dépend de sa forme, de la topographie, et de nombreux autres facteurs. Les résultats présentés ici qui portent sur la distance de propagation maximale de l'un des blocs sont obtenus à partir de données acquises sur des essais, très peu coûteux, réalisés en laboratoire sur un modèle réduit et dans des environnements caractérisés par des facteurs contrôlés.

Plans d'expériences pour mélanges avec une distribution de Dirichlet

Astrid Jourdan*[1]

[1] Equipes Traitement de l'Information et Systèmes – Centre National de la Recherche Scientifique : UMR8051, CY Cergy Paris Université – France

Le support de la distribution de Dirichlet est un simplexe tel que les coordonnées appartiennent à (0,1) et leur somme égale 1. Cette loi est donc appropriée pour définir des plans d'expérience pour mélanges. En fonction de ses paramètres, la distribution de Dirichlet permet d'obtenir une distribution des points du plan, symétrique ou asymétrique, uniforme ou concentrée à l'intérieur du domaine expérimental. Le cas de la distribution uniforme est largement utilisé pour les plans d'expérience sur les mélanges. L'uniformité des points de plan est généralement évaluée à l'aide d'un critère de discrépance. Dans ce travail, nous proposons un nouveau critère. Nous utilisons la divergence de Kullback-Leibler pour mesurer l'écart entre la distribution empirique des points du plan et une distribution de Dirichlet. Son estimation est faite, soit avec une méthode plug-in en remplaçant la fonction de densité par une estimation par noyau, soit par plus proches voisins.

Prix ENSAI

(Amphi 10 - 18h35-19h00)

Modélisation de la courbe de saturation en O2 de l'hémoglobine chez des patients en réanimation

Enora Alaoui*[1], Pierre Barbe[1], Felix Lucas[1], Victoria Mas[1]

[1] ENSAI – Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Modélisation de la courbe de saturation en O2 de l'hémoglobine chez des patients en réanimation.

19h00 – 20h30 Coktail de Bienvenue

---

14 juin 2022

Programme

9h00 – 10h00 PLENIERE : Francoise Berthoud **55**

10h00 – 10h10 Pause **55**

10h10 – 12h10 . . . . . **55**

SFdS - Environnement 55

Mathis Chagneux 56

Florian Lasgorceux [et al.] 56

Maximilien Servajean [et al.] 57

SFdS - Enseignement 57

Manon Andre 58

Nefeli Paparisteidi [et al.] 58

Roger Beecham 59

Sylvie Viguier-Pla [et al.] 59

Antoine Rolland 60

Application : santé & société 60

Clément Benoist [et al.] 60

Victoria Cornelius 61

Léonie Courcoul [et al.] 61

Deltreil Guillaume [et al.] 62

Pan Zhao [et al.] 63

Données Fonctionnelles 63

Gaëlle Chagny [et al.] 63

Eddy Ella Mintsa 64

Julien Ah-Pine [et al.] 64

Jean Steve Tamo Tchomgui [et al.] 65

Rémi Servien [et al.] 65

Statistique mathématique 66

Bastien Batardière [et al.] 66

Ayoub Belhadji 67

Christian Derquenne 67

Myrto Limnios [et al.] 67

Slimane Makhlouf [et al.] 68

12h10 – 13h20 Déjeuner **68**

Déjeuner scientifique . . . 68
13h20 – 14h20 . . . **69**
PLENIERE ANNULÉE : Chloé-Agathe Azencott . . . 69
PLENIERE : Nicolas Papadakis . . . 70
14h20 – 14h30 Pause . . . **70**
14h30 – 16h00 . . . **70**
SFdS - Fiabilité . . . 71
Mitra Fouladirad [et al.] . . . 71
Nicola Esposito [et al.] . . . 71
Vlad Stefan Barbu [et al.] . . . 72
Jérôme Jacob [et al.] . . . 73
Baptiste Kerleguer [et al.] . . . 73
SFdS - Jeunes Statisticiens . . . 74
Agrégation experts . . . 75
Yvenn Amara-Ouali [et al.] . . . 75
Nhat Thien Pham [et al.] . . . 75
Camila Fernandez [et al.] . . . 76
Sothea Has . . . 76
Margaux Zaffran [et al.] . . . 77
Trungtin Nguyen [et al.] . . . 77
Apprentissage non supervisé . . . 78
Francesco Amato [et al.] . . . 78
Emmanuelle Claeys [et al.] . . . 78
Nicolas Jouvin [et al.] . . . 79
Giulia Marchello [et al.] . . . 80
Ariane Marandon [et al.] . . . 80
Transport optimal . . . 81
Marie Breeur [et al.] . . . 81
Lucas De Lara [et al.] . . . 82
Marion Jeamart [et al.] . . . 82
Kimia Nadjahi . . . 83
16h00 – 16h10 Pause . . . **83**
16h10 – 16h30 COMPUTO . . . **83**
16h30 – 16h40 Pause . . . **84**
16h40 – 18h00 AG SFdS . . . **84**

9h00 – 10h00

PLENIERE : Françoise Berthoud

(Amphi 9 - 9h00-10h00)

Numérique et transition écologique : amis ou ennemis ?

Françoise Berthoud

CNRS, Grenoble

J'aborderai dans cet exposé la matérialité du monde numérique et examinerai quelques pistes proposées pour réduire son empreinte afin de rester dans les limites planétaires : économie circulaire, découplage, utilisation du numérique pour réduire les émissions de CO2 d'autres secteurs, efficacité, sobriété numérique etc. Je passerai par un petit voyage virtuel aux pays de la "naissance" et de la "mort" de ces bijoux technologiques, pays où les réalités dépassent l'imaginaire de nos dystopies.

10h00 – 10h10 Pause

10h10 – 12h10

SFdS - Environnement

(Amphi 7 - 10h20-11h50)

Comptage de macrodéchets en zone riparienne par apprentissage profond et modèle d'états

Mathis Chagneux*[1]

[1] Télécom Paris – Dpt Image-Data-Signal, LTCI, Telecom Paris, Institut Polytechnique de Paris, France – France

Les déchets sont une cause connue de dégradation des environnements marins et la plupart d'entre eux voyagent dans les rivières avant d'atteindre les océans. Dans cet article, nous présentons un nouvel algorithme pour aider à la surveillance des déchets le long des cours d'eau. Alors que plusieurs tentatives ont été faites pour quantifier les déchets en utilisant la détection neuronale d'objets dans des photographies d'objets flottants, nous nous attaquons à la tâche plus difficile du comptage direct dans des vidéos en utilisant des caméras embarquées sur des bateaux. Nous nous appuyons sur le suivi d'objets multiples (MOT), mais nous nous concentrons sur la diminution comptages faux et/ou redondants. Notre système ne nécessite qu'une supervision au niveau de l'image et effectue un filtrage bayésien via un modèle d'état basé sur le flux optique. Nous présentons un nouvel jeu de données ouvert d'images recueillies via une campagne de crowdsourcing et utilisées pour entraîner un détecteur d'objets basé sur la prédiction d'une carte de scores où chaque pixel donne la probabilité d'être au centre d'un objet. Des séquences vidéo réalistes assemblées par des experts en surveillance de la pollution marine sont annotées et fournies pour évaluation. Les améliorations de la qualité du comptage sont démontrées en comparaison à des systèmes construits à partir de traqueurs multi-objets partageant les mêmes capacités de détection. Une décomposition précise des erreurs met en évidence les défis restants.

Bivariate Log-Gaussian Cox Process Model for Presence-only Data

Florian Lasgorceux*[1], Julien Papaïx[1], Yoann Bunz[2], Damien Combrisson[2], Thomas Opitz[1]

[1] Biostatistique et Processus Spatiaux – Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) – France
[2] Parc national des Ecrins – Parc national des Ecrins – France

Les gestionnaires d'espaces naturels ont de plus en plus recours à des bases de données de comptage d'espèce alimentées de manière opportuniste (*données de présence seule*). Ces données offrent une meilleure couverture spatio-temporelle mais sont affectées de multiples biais

d'acquisition du fait de l'absence de protocole d'observation. Ceci se traduit en particulier par un effort d'échantillonnage hétérogène et inconnu. De plus, la mise en évidence des interactions entre espèces à partir de ce type de données n'est pas aisée puisque la co-occurrence d'observations ne présage pas forcément d'une influence d'une espèce sur l'autre. Dans ce contexte, nous nous sommes intéressés à la modélisation jointe de la distribution d'espèces et à l'inférence d'un tel modèle à partir de données opportunistes. Nous développons, dans un premier temps, un modèle de processus ponctuel bivarié (pour deux espèces), à savoir un processus de Cox log-gaussien, dont les intensités intègrent des covariables environnementales et des effets aléatoires spatiaux, dont un commun aux deux variables. L'identifiabilité des paramètres de ce modèle est étudiée par simulations. Dans un deuxième temps, un effort d'échantillonnage hétérogène en espace est ajouté au modèle et nous discutons des conséquences du choix de modélisation de ce dernier. Les simulations et l'estimation des paramètres du modèle sont réalisées dans un cadre bayésien via la méthode INLA-SPDE. L'estimation est explorée pour différentes configurations des paramètres liés à l'abondance et la variabilité spatiale de chaque espèce, à l'interaction entre les deux espèces et à la structure de l'effort d'échantillonnage. Le cadre de simulation est adapté au contexte du Parc National des Ecrins pour lequel des données opportunistes des communautés aviaires sont disponibles.

When strong annotators lead to bad model choices

Maximilien Servajean*, Waleed Ragheb[1], Valentin Leveau, Alexis Joly

[1] PhD Researcher – LIRMM – France

Les contextes participatifs tels que les sciences citoyennes impliquent souvent des procédures d'annotation où les labels d'un jeu de données sont produits de manière collaborative. De telles procédures ont été étudiées dans la littérature et leurs performances mesurées au travers de métriques telles que l'accuracy. Cependant, peu de travaux ont connecté la qualité des annotations aux performances des modèles de machine learning entraînés sur ces dernières. Dans cette étude, nous montrons que l'accuracy n'est pas nécessairement un bon critère. En particulier, nous montrons que sous la présence de bruit, même des labels obtenus via le classifieur optimal de Bayes ne permettent pas de garantir que le modèle obtenu minimise bien l'erreur souhaitée.

SFdS - Enseignement

(Amphi 8 - 10h20-12h10)

Analyse des tendances de la qualité de l'air

Manon Andre*[1]

[1] Université Grenoble Alpes – Département STID IUT – France

Au cours de ma seconde année de DUT STID, j'ai eu l'opportunité d'effectuer mon stage au sein de l'Institut des Géosciences de l'Environnement, sur le campus de Saint-Martin-d'Hères. L'objectif de ce stage était d'observer les tendances des concentrations massiques des éléments chimiques composant les particules fines et de les quantifier afin de comparer un site urbain et un site agricole, mais également de comparer les différentes tailles de particules au sein d'un même site.
Les données ont été nettoyées puis agrégées par mois avant de tester la présence ou non de tendance puis en déduire la quantification de celle-ci. L'idée est d'étudier les tendances en confrontant deux méthodes statistiques ayant des hypothèses différentes, l'une se basant sur la moyenne et l'autre sur la médiane. Pour cela, l'étude a été réalisée avec le langage python.
Il est apparu que la concentration massique des éléments des particules fines suivent majoritairement une diminution significative sur les deux typologies de sites (urbain et rural). Ces résultats indiquent premièrement que la mise en place de politiques publiques larges échelles ont un impact sur le trafic routier notamment avec le carbone élémentaire (EC), le cuivre (Cu) et le zinc (Zn). Cette constatation est confirmée par une diminution également présente sur les fractions grossière et fine du site rural, qui est éloigné des sources locales. Deuxièmement, on observe que sur Grenoble les politiques locales semblent fonctionner avec une accélération de la baisse de carbone élémentaire (EC) sur les cinq dernières années.

Hybridation Numérique : la réussite des étudiants à EPITECH

Nefeli Paparisteidi[2,1], Aurélien Pellet*[1], Sonico Samedy[1]

[2] Ecole Doctorale Frontiere de l'Innovation en Recherche et Education – Université Paris Cité – France
[1] Laboratoire MNSHS – Epitech – France

Cette étude vise à révéler l'expérience d'apprentissage de ceux qui participent à EPITECH, un établissement d'enseignement supérieur orienté vers le numérique et l'informatique pendant les années universitaires 2018/2019, 2019/2020 et 2020/2021, où l'apprentissage en ligne est devenu une nécessité en raison de la pandémie COVID-19. A travers cette étude, nous souhaitons examiner la performance et l'adaptation des étudiants à l'apprentissage hybride en comparant le revenu de leurs parents. Les résultats de l'étude montrent que les étudiants sont capables de s'adapter à l'apprentissage hybride indépendamment de leurs revenus familiaux. Les résultats de l'étude contribueront à une meilleure compréhension de l'impact de l'e-learning sur l'adaptation et la performance des étudiants dans l'enseignement supérieur.
Mots-clés : Enseignement supérieur, Hybridation numérique, Réussite

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy

Roger Beecham*[1]

[1] School of Geography, University of Leeds – Royaume-Uni

Concerns around the replicability of published scientific findings has prompted much introspection into the way in which scientific knowledge is produced. To address issues of data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that research findings should only be claimed through pre-registered confirmatory data analyses. Pre-registration studies are, though, somewhat inimical to the more informal research environments typical of modern applied data analysis ('Data Science'). In this talk I enumerate some of these challenges and demonstrate, through an analysis of road crash data in the UK, how nascent visualization techniques can be used to navigate and inject statistical rigour into contemporary data analysis environments.

Comprendre l'échec des étudiants en vue d'un enseignement différencié

Sylvie Viguier-Pla*[1], Mouna Kamel[2]

[1] LAboratoire de Mathématiques et PhySique – Université de Perpignan Via Domitia : EA4217 – France
[2] Institut de Recherche en Informatique de Toulouse – Université Paul Sabatier (Toulouse, France) – France

Le niveau de scolarisation étant en constante progression depuis les années 1900, l'accès aux études est devenu plus large et leur durée plus longue, notamment par le biais de l'Université qui offre l'opportunité à tous les étudiants d'accéder à au moins une formation. Cette évolution est accompagnée, notamment au cours des dernières décennies, d'une importante augmentation des échecs, des abandons et des réorientations, plus spécifiquement chez les étudiants de 1ère année du premier cycle universitaire. Différentes études ont identifié les éléments pouvant être à l'origine de ces taux d'échec, ceci à l'échelle d'une université, d'un pays, voire au niveau international. L'étude que nous présentons aujourd'hui s'inscrit dans cette démarche, à la différence qu'elle vise à promouvoir l'enseignement différencié en se focalisant sur l'identification des facteurs d'échec pour une formation précise, en un lieu donné. Nous avons analysé la formation STID d'IUT de Carcassonne. Les critères identifiés comme facteurs sont le type de bac, la CSP du parent 1, le sexe et le fait d'être néo-bachelier ou pas. L'échec est aussi lié à la motivation. La méthodologie mise en œuvre est reproductible pour toute formation.

Transposition didactique et enseignement de la statistique

Antoine Rolland *[1]

[1] Université Lumière Lyon 2 – IUT Lumière Lyon – France

La plupart des travaux sur la didactique de la statistique dans l'enseignement supérieur se focalisent sur l'enjeu de présenter la statistique à des étudiants a priori réfractaires. Au contraire, nous interrogeons les choix pédagogiques effectués par les enseignants dans les filières de formation en statistique. A travers l'exemple de la régression linéaire, et en mobilisant le cadre de la transposition didactique, nous analysons ces choix par l'étude de documents de cours dans six formations de différents niveaux. Cette étude met en évidence d'une part l'existence d'un " savoir savant " partagé, et d'une autre part un " savoir à enseigner " témoin d'une tension entre théorie généraliste et application pratique, tension propre à la science statistique.

Application : santé & société

(Amphi 9 - 10h20-12h10)

Application des réseaux de neurones aux données longitudinales de santé: cas du suivi de la fonction rénale aux patients transplantés rénaux

Clément Benoist *[1], Marc Labriffe [2,3], Anders Asberg [4], Pierre Marquet [5,6], Jean-Baptiste Woillard [5,6]

[1] Service de Pharmacologie, toxicologie et pharmacovigilance [CHU Limoges] – CHU Limoges – France
[2] Service de Pharmacologie, toxicologie et pharmacovigilance [CHU Limoges] – CHU Limoges, Pharmacologie et transplantation, INSERM 1248, Université de Limoges, Limoges, France – France
[3] INSERM 1248, Université de Limoges, Limoges, France – INSERM 1248, Université de Limoges, Limoges, France – France
[4] Departement of Pharmacy, Faculty of Mathematics and Natural Sciences, Oslo, Norway – Norvège
[5] Service de pharmacologie, Toxicologie et pharmacovigilance, CHU de Limoges,France – Service de pharmacologie, Toxicologie et pharmacovigilence, CHU de Limoges,France – France
[6] Pharmacologie & Transplantation, INSERM 1248, Université de Limoges, Limoges, France – Pharmacologie – France

Chez les patients transplantés rénaux, le fonctionnement du greffon est régulièrement suivi par l'estimation du débit de filtration glomérulaire (DFG). Plus cette valeur est basse, plus elle indique un défaut de fonctionnement du greffon. Savoir prédire le DFG à 6 mois ou 1 an permettrait d'anticiper une détérioration du DFG afin de proposer une prise en charge précoce adaptée et éviter un retour en dialyse. La surestimation du DFG ayant pour conséquence l'absence d'alerte du clinicien et d'un suivi plus rapproché pouvant entraîner une augmentation du risque de perte du greffon à moyen terme. Les données de suivi post-transplantation utilisées étaient sous forme de données longitudinales ou de séries temporelles courtes (3819 patients). Nous avons modélisé le DFG à 12 mois pour chaque patient grâce à un Long Short Term Memory (LSTM). Une fonction d'erreur adaptée à la problématique médicale a été développée permettant de donner pénaliser plus fortement les erreurs associées à des DFG faibles. La médiane de l'erreur absolue était 5.2 mL/min/1.73 m^2, compatible avec une utilisation en clinique de l'algorithme. Le 99e percentile de l'erreur était 25.4 mL/min/1.73 m^2.

Data visualisation of adverse events in randomised clinical trials

Victoria Cornelius *[1]

[1] Imperial College London – Royaume-Uni

Methods to analyse efficacy outcomes in randomised controlled trials (RCTs) are well established. With substantial improvements in statistical software it's now trivial to undertake advanced statistical modeling and present this data well. Despite this progress, the analysis and presentation of adverse events (AEs) in trial publications has seen very little progress.
AE data is particularly difficult to analyse due to its multi-faceted nature. One of these features is the large number of AE outcomes that get recorded in a trial. There is a lack of guidance on what and how to visually display complex AE data.
We previously undertook a methodology review to identify statistical methods specifically developed to analyze AE data. (Phillips et al 2020) This current paper examines two visual analysis methods identified to be suitable for any adverse events collected during a trial, and one new approach proposed by the authors for pre-specified harm outcomes that is particularly valuable for multi-arm studies We explore their value using data from a COVID-19 treatment trial and COVID-19 vaccination trial.

Impact de la variabilité de la pression artérielle sur le risque d'AVC

Léonie Courcoul *[1], Antoine Barbieri [1], Christophe Tzourio [1], Hélène Jacqmin-Gadda [1]

[1] Université de Bordeaux, ISPED, Inserm BPH U1219, F-33000, Bordeaux – Université de Bordeaux, ISPED, Inserm BPH U1219 – France

Etant donné l'incidence des accidents vasculaires cérébraux (AVC) et leurs mauvais pronos-

tics, leur prévention est un enjeu majeur de santé publique. Il est maintenant bien démontré qu'un niveau élevé de pression artérielle est un facteur de risque d'AVC, mais un nombre croissant d'études suggère que la variabilité de la pression artérielle pourrait également être un facteur de risque indépendant d'AVC. Cependant, ces études souffrent souvent de faiblesses méthodologiques importantes. L'objectif de ce travail était de développer un modèle conjoint à variance hétérogène pour les mesures répétées d'un marqueur longitudinal et le risque d'événements compétitifs afin d'étudier l'association entre la variabilité de la pression artérielle et le risque d'AVC en tenant compte du risque compétitif de décès. Ce modèle conjoint combine un modèle mixte incluant un effet aléatoire spécifique au sujet pour la variance résiduelle et un modèle à risque proportionnel cause-spécifique pour les risques compétitifs. Les risques peuvent dépendre simultanément de la variance résiduelle spécifique au sujet et de la valeur et de la pente courantes du marqueur. Le modèle a été estimé sur les données de l'essai clinique PROGRESS pour la prévention de la récurrence des accidents vasculaires cérébraux qui inclut 6105 sujets suivis pendant 5 ans avec 12 temps de mesure de la pression artérielle. Nous avons constaté que le risque de récidive d'AVC augmentait avec la valeur courante de la pression artérielle mais n'était pas associé à la variabilité intra-individuelle ou à la pente de la pression artérielle. Les capacités prédictives des modèles avec différentes structures de dépendance ont été comparées en utilisant la vraisemblance conditionnelle du temps d'événement schant le marqueur, le Brier Score et l'aire sous la courbe ROC en utilisant des estimateurs tenant compte de la censure à droite et des risques compétitifs.

Utilisation des matrices emplois expositions pour les troubles musculo-squelettiques. Application aux douleurs sévères du genou

Deltreil Guillaume *[1], Patrick Tardivel [2], Alexis Descatha [3], Mikael Escobar-Bach [1], Piotr Graczyk [1]

[1] LAREMA, Université d'Angers, France – Université d'Angers – France
[2] IMB, Université de Bourgogne, France – Université de Bourgogne-Franche-Comté – France
[3] Univ Angers Inserm, CHU Angers, Hofstra/ Northwell, France-USA – Univ Angers, Institut National de la Santé et de la Recherche Médicale - INSERM, CHU Angers – France

Les troubles musculo-squelettiques (TMS) sont la première cause de maladie professionnelle, avec les pathologies du dos, des membres supérieurs et des genoux. Ils sont liés à plusieurs facteurs de risques professionnels dont les postures contraignantes et le port de charges lourdes qui surviennent tout au long de la vie professionnelle. Durant notre présentation, nous étudierons les effets de deux variables (le temps total individuel d'exposition et la moyenne individuelle d'exposition) sur la probabilité de développer un TMS.

Optimal Individualized Treatment Regime of Plasma Transfusion for Severe Trauma

Pan Zhao*[1], Nicolas Gatulle[2], Julie Josse[1], Antoine Chambaz[3]

[1] Inria Montpellier – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[2] Sorbonne University, GRC 29, AP-HP, DMU DREAM, Department of Anaesthesiology and Critical Care, Pitié-Salpêtrière Hospital – Sorbonne University, GRC 29, AP-HP, DMU DREAM, Department of Anaesthesiology and Critical Care, Pitié-Salpêtrière Hospital – France
[3] MAP5, Université de Paris – CNRS : UMR8145 – France

An individualized treatment regime (ITR) is a decision rule that assigns a treatment, among the available options, to a patient based on the patient's characteristics. Hemorrhagic shock is the second leading cause of death in severe trauma patients, and considered as the leading cause of preventable mortality in hospitals. We conduct an analysis of TraumaBase® data, and develop a new super-learning based method to learn the optimal ITR of plasma transfusion for severe trauma patients. In addition, we tackle the issue of missing data.

Données Fonctionnelles

(Amphi 10 - 10h20-12h10)

Estimation adaptative dans le modèle linéaire fonctionnel à sortie fonctionnelle

Gaëlle Chagny*[1], Anouar Meynaoui[2], Angelina Roche[3]

[1] Laboratoire de Mathématiques Raphaël Salem – Université de Rouen Normandie – France
[2] Laboratoire de Mathématiques Raphaël Salem – Université de Rouen Normandie, Centre National de la Recherche Scientifique : UMR6085 – France
[3] CEntre de REcherches en MAthématiques de la DEcision – Université Paris-Dauphine, Centre National de la Recherche Scientifique : UMR7534 – France

On considère un modèle de régression linéaire fonctionnel à sortie fonctionnelle : les variables explicative et réponse sont des variables aléatoires "fonctionnelles", à valeurs dans un espace de

Hilbert. On s'intéresse à la question de l'estimation non-paramétrique de l'opérateur intégral reliant ces deux variables, à partir d'un échantillon. Une collection d'estimateurs par projection sur la base de l'ACP empirique de la covariable est construite. On obtient une décomposition biais-variance pour un risque quadratique moyen de prédiction. Une procédure de sélection de modèle (minimisation de contraste pénalisé) permet ensuite un choix automatique du meilleur estimateur dans la collection. Celui-ci satisfait une inégalité de type oracle, et atteint des vitesses de convergences minimax sur des espaces de régularité de type ellipsoïde : la borne supérieure du risque de prédiction
correspond à la borne inférieure, que nous calculons aussi. Ces résultats théoriques sont illustrés par des applications à des jeux de données simulées et réelles.

Procédure de classification plug-in pour des données fonctionnelles

Eddy Ella Mintsa*[1]

[1] Laboratoire LAMA UMR 8050 – Université Gustave Eiffel – France

Les récents progrès de la technologie moderne ont généré des données étiquetées regardées comme des réalisations d'une fonction aléatoire. Ce travail porte sur un problème de classification multiclasses pour des données fonctionnelles modélisées par une équation différentielle stochastique. Peu de travaux étudient le cas où les données fonctionnelles sont modélisées par des processus de diffusion, c'est pourquoi, la construction de procédures de classification adaptées à ce type de modèle est un enjeu majeur. Nous nous concentrons sur des processus de diffusion homogènes en temps avec des coefficients de dérive et de diffusion inconnus. L'objectif est de proposer une procédure de classification de type plug-in implémentable basée sur une estimation non-paramétrique à partir d'un échantillon d'apprentissage, des fonctions de dérive et du coefficient de diffusion par la minimisation d'un contraste des moindres carrés sur une base de fonctions B-splines. Nous établissons ensuite la consistance du classifieur obtenu.
Mots-clés. Classification supervisée, processus de diffusion, estimation non-paramétrique, classifieur plug-in

Fusion tardive en analyse de données fonctionnelles élastique

Julien Ah-Pine*[1,2,3], Noé Lebreton

[1] Laboratoire de Mathématiques Blaise Pascal (LMBP) – Université Clermont Auvergne, CNRS : UMR6620 – France
[2] Laboratoire (ERIC) – Université Lumière -Lyon 2 : EA3083, Université Claude Bernard - Lyon I : EA3083 – France
[3] Centre d'Études et de Recherches sur le Développement International – Centre National de la Recherche Scientifique : UMR6587, Université Clermont Auvergne : UMR6587 – France

Dans cette communication nous comparons la fusion précoce et la fusion tardive en classification supervisée de données fonctionnelles lorsqu'une séparation des variations d'amplitude et de phase est pertinente. Le cadre méthodologique utilisé est l'analyse de données fonctionnelles élastique qui passe par un alignement des fonctions basé sur la distance de Fisher-Rao. Ensuite, afin de considérer conjointement les variations d'amplitude et de phase, la fusion précoce concatène les fonctions alignées et les fonctions de déformation temporelle dans une fonction composite avant réduction de dimension et apprentissage du modèle de classification. Cette méthode est proposée dans la littérature récente. *A contrario*, nous examinons une fusion tardive dans laquelle, deux réductions de dimension et apprentissages de classifieurs sont appliquées indépendamment sur les deux types de fonctions d'amplitude et de phase, et ce sont les estimations des prédictions des deux modèles qui sont fusionnées par un opérateur d'agrégation.

Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP

Jean Steve Tamo Tchomgui[*1,2], Julien Jacques[3], Stéphane Chrétien[2], Guillaume Fraysse[4], Vincent Barriac[4]

[1] Orange Innovation – Orange Labs Networks – France
[2] Univ Lyon, Univ Lyon 2, ERIC – Université Lumière - Lyon II – France
[3] Univ Lyon, Univ Lyon 2, ERIC – Université Lumière - Lyon 2, Université Lumière - Lyon 2 – France
[4] Orange Innovation – Orange Labs Networks – France

Ce travail propose une étude comparative entre les techniques récentes de l'analyse des données fonctionnelles et celles des signatures pour la résolution d'un problème de prédiction de la Qualité d'Expérience (QoE), une mesure qui reflète la perception de l'utilisateur final d'un service de télécommunications. La QoE ainsi que les facteurs pouvant l'influencer étant mesurés à haute fréquence, le problème que nous considérons trouve donc sa formulation naturelle dans le contexte de la régression linéaire fonctionnelle, où la variable à prédire et les variables explicatives sont toutes fonctionnelles. Notre contribution principale est celle de montrer que l'utilisation des signatures permet de résumer de façon efficace l'information contenue dans les variables explicatives et de produire des prédictions souvent meilleures que celles obtenues par des méthodes classiques. **Mots-clés.** Données fonctionnelles, Modèles de régression fonctionnels, Signatures.

Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires

Rémi Servien[*1], Nathalie Vialaneix[2]

[1] INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France – Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UR0050 – France

[2] Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France – Institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) : UR875 – France

Nous nous intéressons ici au problème de sélection de variables dans un cadre de régression fonctionnelle. Le but est de sélectionner des points de mesure consécutifs afin de déterminer les intervalles importants dans la prédiction de la variable cible. Pour cela nous nous basons sur les forêts aléatoires et évaluons des variantes possibles pour trois étapes de l'approche générale que nous proposons (création de groupes, définition des résumés, sélection) que nous comparons sur des données simulées et réelles.

Statistique mathématique

(Amphi 11 - 10h20-12h10)

Multivariate Poisson Lognormal model: optimisation, inference and application to high dimensional data

Bastien Batardière*[1], Julien Chiquet[1], Joon Kwon[1], Laure Sansonnet[1]

[1] Mathématiques et Informatique Appliquées – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0518 – France

In Single-cell analysis, it is usual to deal with high-dimensional non-Gaussian observations. A typical example is the raw single-cell transcriptomic count distribution, which aims to understand the co-variations of each gene's activity (number of times the gene is expressed) in each cell. While the Gaussian setting provides a canonical way of modeling such dependencies with continuous data, it does not apply to count data. The Poisson lognormal model (PLN) does. Moreover, efficient dimension reduction techniques need to be performed to handle high-dimensional data. The Poisson lognormal PCA (PLN-PCA) model address this additional issue. However, inference is tricky since it deals with intractable integrals. We focus on two different inference approaches: a variational algorithm maximizing an Evidence Lower BOund (ELBO) and a Monte-Carlo approach. No theoretical guarantee exists for the error made by the ELBO maximization. We study this error thanks to the estimation of the log-likelihood and approve the validity of the approximation. The algorithmic complexities of both approaches evolve linearly with the number of genes and cells for the PCA part. We also study the difference between techniques with dimension reduction (PLN-PCA) and those without (PLN).

Les approximations à noyau à base de processus ponctuels déterminantaux

Ayoub Belhadji*[1]

[1] ENS de Lyon – Univ Lyon, ENS de Lyon, CNRS, Inria, UCB Lyon 1, IXXI, LIP, 69342 Lyon – France

On étudie l'approximation d'une fonction, qui vit dans un espace à noyau, par un mélange finie de translatées de noyau quand les noeuds de l'approximation suivent la distribution d'un processus ponctuel determinantal (DPP) adapté au noyau. On montre que le taux de convergence est presque optimal et dépend principalement des valeurs propres de l'opérateur d'intégration correspondant. En particulier, cette analyse unifiée permet d'avoir un nouveau regard sur l'utilisation des DPPs dans le domaine de l'intégration numérique.

Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple

Christian Derquenne*[1]

[1] EDF Labs – OSIRIS – France

Cet article se place dans le cadre de la multicolinéarité entre les prédicteurs au sein d'un modèle linéaire de régression multiple. Le phénomène de multicolinéarité peut entrainer des incohérences sur les coefficients de régression et des oublis de prédicteurs, cela peut par conséquent poser des problèmes d'interprétation qui peuvent entrainer de mauvaises décisions. Le critère proposé revient à régulariser la matrice des corrélations simples entre les variables candidates à l'explication de manière à respecter conjointement la cohérence des signes entre les coefficients de régression multiple et de corrélation simple, ainsi que l'ordre de liaison entre les prédicteurs et la réponse, et les niveaux de significativité, tout en préservant la qualité de reconstitution des données à l'aide du modèle estimé.

High-dimensional Two-sample Rank Statistics as Scalar Performance Criteria

Myrto Limnios*[1], Stephan Clémençon[2,3], Nicolas Vayatis[1]

[1] CB - Centre Borelli - UMR 9010 – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR9010, Ecole Normale Supérieure Paris-Saclay, Université de Paris – France
[2] Laboratoire Traitement et Communication de l'Information [Paris] (LTCI) – Télécom ParisTech, CNRS :

UMR5141 – CNRS LTCI Télécom ParisTech 46 rue Barrault F-75634 Paris Cedex 13, France
[3] Telecom Paris – Télécom ParisTech – France

Nous proposons une généralisation de statistiques de rang à deux échantillons en grande dimension. Cette approche s'inspire de modèles issus de l'apprentissage d'ordonnancement. Nous détaillons comment ces collections de statistiques peuvent être utilisées comme critères scalaires de performance et, en particulier, pour les modèles d'ordonnancement biparti.

Clustering de modalités en régression logistique

Slimane Makhlouf*[1], François-Xavier Jollois[2], Avner Bar-Hen[3]

[1] CEDRIC. Méthodes statistiques de data-mining et apprentissage – Centre d'études et de recherche en informatique et communications – France
[2] Paris Descartes – Université Paris V - Paris Descartes – France
[3] Conservatoire National des Arts et Métiers [CNAM] – Conservatoire National des Arts et Métiers (CNAM), Conservatoire National des Arts et Métiers [CNAM] – France

Ce travail s'intéresse à la réduction de dimension de données catégorielles par regroupement de modalités, dans le cadre d'une régression logistique. L'inflation du nombre de modalités augmente rapidement le coût de calcul et dégrade fortement les capacités prédictives des modèles. La réduction de dimensionalité peut s'effectuer par sélection et/ou par regroupement de variables (clustering). L'interprétation des groupes ainsi constitués peut apporter des informations supplémentaires sur la structure des données étudiées. De nombreux travaux de recherche se concentrent sur la sélection ou le regroupement de variables quantitatives. Nous présentons dans cet article nos travaux sur le regroupement de modalités à l'intérieur de variables catégorielles et tout particulièrement en grande dimension. Nos résultats sont appliqués sur un jeu d'enchères en ligne.

12h10 – 13h20 Déjeuner

Déjeuner scientifique - Jeunes Statisticiens

Statistique et écologie

Le déjeuner scientifique revient cette année sur le thème "Statistique et écologie" avec une discussion autour de la question "Être chercheur.se en statistique en pleine crise climatique : comment le vivre ?".

Ainsi, nous accueillerons Romain Couillet (Université Grenoble Alpes), Julie Delon (Université Paris Cité) et Nicolas Papadakis (Institut de Mathématiques de Bordeaux) le mardi 14 juin de 12h à 13h15.

Cet évènement est soumis à inscription via ce formulaire.

Lien du formulaire : https://framaforms.org/dejeuner-scientifique-jds-2022-lyon-1653462811.

13h20 – 14h20

PLENIERE ANNULÉE : Chloé-Agathe Azencott

(Amphi 7 - 13h20-14h20)

Structured feature selection in high-dimensional genomic data

Chloé-Agathe Azencott

Cbio Mines Paris Tech

Many problems in genomics require the ability to identify relevant features in data sets containing many more orders of magnitude than samples. One such example is genome-wide association studies (GWAS), in which hundreds of thousands of variants are measured for orders of magnitude fewer samples. Even for the most classical approaches, where one tests for the association of each variant with the phenotype, these studies are severely underpowered. Accounting for multiple effects of several variants is even more challenging. This talk will describe several approaches that alleviate this difficulty by incorporating prior knowledge (such as biological networks, population membership, or linkage disequilibrium) as structure on the data.

PLENIERE : Nicolas Papadakis

(Amphi 10 - 13h20-14h20)

Gradient Step Denoiser for convergent Plug-and-Play

Nicolas Papadakis

CNRS Bordeaux

In image sciences, Plug-and-Play methods constitute a class of iterative algorithms for solving Bayesian inverse problems where regularization is performed by an off-the-shelf denoiser. Although Plug-and-Play methods can lead to tremendous visual performance for various image problems, the few existing convergence guarantees are based on unrealistic (or suboptimal) hypotheses on the denoiser, or limited to strongly convex data terms. In this work, we propose a new type of Plug-and-Play methods, based on half-quadratic splitting, for which the denoiser is realized as a gradient descent step on a functional parameterized by a deep neural network. Exploiting convergence results for proximal gradient descent algorithms in the non-convex setting, we show that the proposed Plug-and-Play algorithm is a convergent iterative scheme that targets stationary points of an explicit global functional. Besides, experiments show that it is possible to learn such a deep denoiser while not compromising the performance in comparison to other state-of-the-art deep denoisers used in Plug-and-Play schemes. We apply our proximal gradient algorithm to various ill-posed inverse problems, e.g. deblurring, super-resolution and inpainting. For all these applications, numerical results empirically confirm the convergence results. Experiments also show that this new algorithm reaches state-of-the-art performance, both quantitatively and qualitatively.

14h20 – 14h30 Pause

14h30 – 16h00

SFdS - Fiabilité

(Amphi 7 - 14h30-16h00)

Degradation Model Selection Using Depth Functions

Mitra Fouladirad[1], Diego Rodolpho Tomassi[2], Arefe Asadi*[3]

[1] Centrale Marseille – École Centrale Marseille, Marseille, France, Aix Marseille Université (Aix-en-Provence), Université de Technologie de Troyes – France
[2] Biofortis – Biofortis – France
[3] Université de Technologie de Troyes – Université de Technologie de Troyes – France

For lifetime prediction or maintenance planning of complex systems, degradation modeling is essential. The reason is that for highly reliable systems that failure times are difficult to observe, degradation measurements often provide more information than failure time to improve system reliability (1). According to Lehmann (2), the stochastic-process-based model shows great flexibility in describing the failure mechanisms caused by degradation.
The aim of degradation modeling is to select a model from a set of competing models capturing the features of the underlying degradation process. An efficient statistical tool is able to discard irrelevant models.
The concept of statistical depth could be considered as a statistical tool for the model suggestion. Tukey (3) introduced a data depth to extend the notion of a median to multi-variate random variables. Considering functional data, the data which are recorded densely over time with one observed function per subject (4), the notion of depth has been extended by (5, 6, 7). An alternative point of view based on the graphic representation of curves is proposed in (8).
A depth function reflects the centrality of the observation to a statistical population (9). The models that show high values of depth function are compared based on different statistical criteria, namely Failure Time Distribution, and the best model is selected to predict failure time for the system under study.

On misspecification of a Gamma with an inverse Gaussian process and its effects in terms of prognostic and related decision processes

Nicola Esposito*, Bruno Castanier[1], Massimiliano Giorgio[2]

[1] Université d'Angers/Laris – Université Nantes Angers Le Mans – France
[2] Università di Napoli Federico II – Italie

The gamma and inverse Gaussian processes are two common choices of models for increasing degradation phenomena in reliability field (see Van Nortwijk (2009) and Ye and Chen (2014)). These models have similar features. In fact, both have independent increments and are especially suitable to describe degradation phenomena that take place gradually over time in a sequence of tiny increments, such as crack growth or corrosion. Factually, there are many experimental situations where these models show comparable fitting ability. Hence, based on all these reasons, in the current literature, they are often treated as equivalent to each other. Nonetheless, this is not true. This circumstance makes the misspecification issue a problem of concern.
In fact, in this paper we focus on this misspecification issue and on the effect of a misspecification on the estimates of the remaining useful lifetime and on the long-run average maintenance cost rate, computed under a maintenance policy recently proposed in the literature. Model selection is performed by using the Akaike Information Criterion (AIC). It is assumed that a misspecification occurs when the AIC leads to select the wrong model.
This study extends Esposito *et al.* (2021) and Esposito *et al.* (2022), where it is examined the misspecification of a gamma process with an inverse Gaussian, by considering the symmetric case where the true model is the inverse Gaussian process.
The new results give evidence and confirm that, under both settings, when the model used to analyze the available data is selected by using the Akaike information criterion, the (possible) misspecification of one of the considered competing model with the other one produces a negligible impact both in terms of remaining useful life estimation and in terms of maintenance costs. Conversely, it was observed that the consequences could be severe if the model is selected without adopting a formal inferential procedure.

Sequential Interval Reliablity for semi-Markov systems

Vlad Stefan Barbu*[1], Guglielmo D'amico[2], Thomas Gkelsinis[1]

[1] Université de Rouen Normandie – Laboratoire de mathématiques Raphaël Salem – France
[2] University G. D'Annunzio of Chieti-Pescara – Italie

The present work, based on Barbu et al. (2021), is dealing with reliability modeling for multi-state systems with time dependence. We propose a generalization of reliability indexes such as reliability, interval reliability and availability for homogeneous semi-Markov repairable systems in discrete time. That measure, called sequential interval reliability, is the probability that the system works in a sequence of non-overlapping time intervals.

Building wind loading estimation under uncertain inflow parameters by means of lattice Boltzmann simulation.

Jérôme Jacob*[1], Lucie Merlier[2], Pierre Sagaut[1]

[1] Laboratoire de Mécanique, Modélisation et Procédés Propres – Centre National de la Recherche Scientifique : UMR7340 / UMR6181, Ecole Centrale de Marseille, Aix Marseille Université – France
[2] Centre dÉnergétique et de Thermique de Lyon – Université Claude Bernard Lyon 1, Institut National des Sciences Appliquées de Lyon, Centre National de la Recherche Scientifique : UMR5008 – France

Large eddy simulation is becoming lately an interesting tool for the simulation of flows inside complex urban areas for the estimation of building wind loading. To ensure the reliability of the predicted wind loading it becomes more and more important to analyze its response to inflow uncertainties. The present study combines the use of a lattice Boltzmann method(1) (LBM) for the simulation of building wind loading and the anchored ANOVA POD/Kriging(2) (c-APK) method for the uncertainty quantification in order to efficiently predict the response surfaces of pressure coefficients and aerodynamic forces on the different facades of a high rise building. For that purpose, numerical simulation have been carried out for the case of a simplified building located in the Shinjuku area in Tokyo using the same configuration described in Jacob and Sagaut(3) considering several inflow wind velocity and direction around a reference configuration considered as the deterministic reference. The obtained results highlighted differences between the physical quantities obtained from the reference sample and the ones obtained from the c-APK analysis. From these results it is shown that the variability of integrated forces on building walls in terms of magnitude and orientation is more important on the highest half of the building which is the area where these forces are the more important. The inflow wind direction is also highlighted to be the main parameter influencing wind loading in the top half of the building whereas inflow wind magnitude is of main importance in the bottom half, mainly in the built area. All the obtained results show the importance of uncertainty quantification to improve the understanding of wind loading sensitivity to the model setting in order to improve the reliability of the simulated wind loading in the framework of industrial applications.
(1) – J. Jacob, O. Malaspinas and P. Sagaut, A new hybrid recursive regularized Bhatnagar-Gros-Krook collision model for lattice Boltzmann method-based large eddy simulation. Journal of Turbulence, 2018
(2) – L. Margheri and P. Sagaut, A hybrid anchored-ANOVA POD/Kriging method for uncertainty quantification in unsteady high fidelity CFD simulations. Journal of Computational Physics, 2016
(3) – J. Jacob and P. Sagaut, Wind comfort assessment by means of large eddy simulation with lattice Boltzmann method in full scale city area. Building and Environment, 2018

Méta-modélisation multi-fidélité par processus Gaussien en utilisant une base d'ondelettes

Baptiste Kerleguer*[2,1], Claire Cannamela[2], Josselin Garnier[3]

[2] DAM Île-de-France – Commissariat à l'énergie atomique et aux énergies alternatives : DAM/DIF – France
[1] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[3] École Polytechnique – CMAP, École Polytechnique – France

La régression par processus gaussien est largement utilisée pour émuler la sortie d'un code coûteux. Nous nous intéressons à des codes multi-fidélité, qui peuvent être exécutés avec différents niveaux de précision et de coût et dont les sorties sont de grande dimension. Ici nous considérons des sorties sous la forme de séries temporelles. Nous étendons le modèle de co-krigeage auto-regressif proposé par Kennedy-O'Hagan aux sorties de grande dimension. Pour cela nous réalisons la régression sur les coefficients d'ondelettes de la sortie à l'aide d'une fonction de covariance spéciale. Nous montrons que le modèle permet de prédire efficacement en terme d'erreur de prédiction, mais aussi de quantification d'incertitudes.

SFdS - Jeunes Statisticiens

(Amphi 8 - 14h30-16h00)

La santé mentale des jeunes chercheurs et chercheuses

Cette année, notre session spéciale portera sur la santé mentale des jeunes chercheurs et chercheuses. A cette occasion, nous espérons ouvrir la discussion sur ce sujet encore trop souvent tabou dans le monde de la recherche.

Lors de cet évènement, un créneau consacré à la parole et au témoignage de celles et ceux qui sont actuellement doctorant.e.s, stagiaires, ou jeunes docteur.e.s. Afin de recueillir cette parole, anonymement ou non, nous avons rédigé ce bref questionnaire que nous vous proposons de remplir (que vous comptiez participer ou non aux JdS). Votre témoignage pourrait être précieux pour la réussite de cet évènement, ainsi que pour ceux qui l'entendront.

Il s'agit d'une occasion que nous espérons importante de partager vos expériences, sans jugements et sans conséquences néfastes sur votre vie professionnelle. Nous espérons ainsi que certains vécus, aujourd'hui invisibilisés, puissent sortir du tabou, afin d'éviter une perpétuation des mêmes problèmes dans le futur.

Lien du questionnaire : https://framaforms.org/session-jeunes-jds-2022-la-sante-mentale-des-jeunes-chercheurs-et-chercheuses-1653321969

Agrégation d'experts

(Amphi 9 - 14h30-16h00)

Modélisation de la charge et de l'occupation des véhicules électriques

Yvenn Amara-Ouali*[1], Bachir Hamrouche[2], Yannig Goude[2], Pascal Massart[1], Jean-Michel Poggi[1]

[1] Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8628 – France
[2] EDF Labs – Electricité de France - EDF – France

Le développement des véhicules électriques est un levier majeur vers un transport bas carbone. Il s'accompagne d'un nombre croissant d'infrastructures de recharge qui peuvent être utilisées comme actifs flexibles de gestion du réseau. Pour permettre cette recharge intelligente (smart-charging), une prévision journalière efficace des comportements de charge est nécessaire. L'objectif de nos travaux est d'évaluer la performance de modèles de prévision des courbes de charge et d'occupation des points de charge sur 8 jeux de données ouverts. Nous étudions deux approches de modélisation : directe et bottom-up. L'approche directe consiste à prévoir la courbe de charge agrégée (resp. l'occupation des points de charge) d'une zone/station. L'approche bottom-up consiste à modéliser les sessions de recharge individuelles pour ensuite les agréger. Cette dernière est essentielle pour implémenter des stratégies de smart-charging. Nous montrons que les approches directes sont généralement plus performantes que les approches bottom-up. Le meilleur modèle peut néanmoins être amélioré en mélangeant les prévisions des approches directes et bottom-up à l'aide une stratégie d'agrégation adaptative.

Functional mixture-of-experts for classification

Nhat Thien Pham*[1,2], Faicel Chamroukhi[1,2]

[1] Université de Caen Normandie – Université de Caen Normandie – France
[2] Laboratoire de Mathématiques Nicolas Oresme – Centre National de la Recherche Scientifique : UMR6139, Université de Caen Normandie – France

We develop a mixtures-of-experts (ME) approach to the multiclass classification where the predictors are univariate functions. It consists of a ME model in which both the gating network and the experts network are constructed upon multinomial logistic activation functions with functional inputs. We perform a regularized maximum likelihood estimation in which the coefficient functions enjoy interpretable sparsity constraints on targeted derivatives. We develop an EM-Lasso like algorithm to compute the regularized MLE and evaluate the proposed approach on simulated and real data.

Ensemble Methods for Data Analytics

Camila Fernandez[*1,2], Chung Shue Chen[2], Pierre Gaillard[3], Alonso Silva[4]

[1] Sorbonne – Université Paris-Sorbonne - Paris IV – France
[2] Nokia Bell Labs – Nokia Bell Labs France – France
[3] Inria – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[4] Safran – SAFRAN Group, Safran Group – France

Time-to-event analysis is a branch of statistics that have been in many researchers' interest due to its wide application field, such as predictive maintenance, healthcare, customer churn prediction, among others. In this communication we present a simple way to aggregate time-to-event analysis methods that outperform on average the performance of every single predictor.

Consensual aggregation on randomly projected high-dimensional features of predictions for regression

Sothea Has[*1]

[1] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France

In this talk, we present a study of a kernel-based consensual aggregation on randomly projected high-dimensional features of predictions for regression. The aggregation scheme is composed of two steps: the high-dimensional features of predictions given by a large number of regression estimators are randomly projected into a smaller subspace using Johnson-Lindenstrauss Lemma (J-L) in the first step, then a kernel-based consensual aggregation is implemented in the second step. We theoretically show that the performance of the aggregation scheme is close to the one of the aggregation method implemented on the original high-dimensional features, with high probability. Then, we numerically show that the consensual aggregation method upholds its performance on highly correlated features of predictions given by different types of regression estimators, plainly constructed without model selection or cross-validation. The efficiency of the aggregation scheme is illustrated on several types of synthetic and real datasets.

Prédictions Conformelles Adaptatives pour les Séries Temporelles

Margaux Zaffran*[1,2,3], Aymeric Dieuleveut[4], Olivier Féron[1,5], Yannig Goude[6], Julie Josse[2,7]

[1] EDF – EDF Recherche et Développement – France
[2] Inria – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[3] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[4] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique – France
[5] Laboratoire de Finance des Marchés dÉnergie – Université Paris Dauphine-PSL, crest, EDF R&D – France
[6] EDF Labs – Electricité de France - EDF – France
[7] Institut Desbrest de santé publique – Institut National de la Santé et de la Recherche Médicale : UA11, Université de Montpellier – France

La quantification d'incertitudes des modèles prédictifs est cruciale dans les problèmes de prise de décision. La prédiction conformelle est une approche générale et théoriquement solide. Cependant, elle nécessite des données échangeables, excluant les séries temporelles. Dans cet exposé, nous soutenons que l'Inférence Conformelle Adaptative (Gibbs & Candès, 2021), développée pour les séries temporelles avec des changements de distribution, est une bonne procédure pour les séries temporelles avec une dépendance générale. Nous analysons théoriquement l'impact du taux d'apprentissage sur son efficacité dans le cas échangeable et auto-régressif. Nous proposons une méthode sans paramètre, AgACI, reposant de manière adaptative sur ACI basée sur l'agrégation d'experts en ligne. Nous menons une comparaison avec des méthodes de la littérature conformelle sur de nombreuses simulations. Les résultats numériques plaident pour l'utilisation d'ACI pour les séries temporelles. Finalement, nous appliquons ces méthodes à la prévision des prix de l'électricité, où AgACI fournit des intervalles de prédiction efficaces.

Model selection by penalization in mixture of experts models with a non-asymptotic approach

Trungtin Nguyen*[1], Faicel Chamroukhi[2], Hien Duy Nguyen[3], Florence Forbes[1]

[1] STATIFY team – Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France – France
[2] Laboratoire de Mathématiques Nicolas Oresme – Normandie Univ, UNICAEN, CNRS, LMNO, Caen, France – France
[3] School of Mathematics and Physics, University of Queensland, St. Lucia, Australia – Australie

Cette étude est consacrée au problème de la sélection de modèles parmi une collection de modèles de mélanges d'experts avec experts gaussiens et fonctions d'activations gaussiennes normalisées, caractérisés par le nombre de composantes du mélange et la complexité des experts moyens, dans

un cadre d'estimation par maximum de vraisemblance pénalisée. En particulier, nous établissons des limites de risque non asymptotiques qui prennent la forme d'inégalités oracles faibles, sous une condition de limite inférieure pour la pénalité. Leur bon comportement empirique est ensuite démontré en simulation et sur des données réelles.

Apprentissage non supervisé

(Amphi 10 - 14h30-16h00)

Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians

Francesco Amato*[1], Julien Jacques[2]

[1] Entrepôts, Représentation et Ingénierie des Connaissances – Université Lumière - Lyon 2 : EA3083, Université Claude Bernard Lyon 1 – France
[2] Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon II : EA3083, Université Claude Bernard - Lyon I (UCBL) – Université Lumière Lyon 2 5 avenue Pierre Mendès-France 69676 Bron Cedex, France

In social sciences or medicine, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model to perform temporal clustering on such data. The model relies on mixture of matrix-variate normal distributions, accounting for the within and between time-dependence structures simultaneously. A MC-EM algorithm for the model estimation is used. Applications on synthetic and real data are presented.

Process Mining : une nouvelle approche de l'évaluation des clusters à l'aide d'un Quantitative Model Checking

Emmanuelle Claeys*[1], Pierre Cry[2], Benoit Barbot[2], Paolo Ballarini[3]

[1] IRIT – Institut de recherche en informatique de Toulouse - IRIT – France
[2] lacl – LABORATOIRE D'ALGORITHMIQUE, COMPLEXITÉ ET LOGIQUE (LACL) – France

[3] central – Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centrale Supélec – France

Le *Process Mining* regroupe différentes méthodes pour construire un modèle formel à partir d'un ensemble de traces émises par un système réel. Dans ce cadre, les traces sont données sous la forme d'un *event log,* c'est-à-dire un jeu de données enregistrant un ensemble d'événements temporels. L'un des objectifs finaux du *Process mining* est de produire une représentation (sous forme de réseaux de Petri) et d'extrapoler des temps de parcours dans différentes situations. Malheureusement, la qualité des techniques classiques d'extraction de réseaux de Petri dépend fortement l'ensemble de données, pour lequel des parcours possibles trop grands ou trop complexes rendent les réseaux de Petri produits inutilisables pour faire des prédictions. Par conséquent, il est parfois nécessaire de discriminer ou de regrouper cet *event log* afin de générer des processus plus lisibles et exploitables pour l'utilisateur. Ces groupes de parcours similaires sont appelés *clusters*. De nombreux algorithmes permettant de *clusteriser* les parcours sont apparus ces dernières années, cependant, il est difficile d'évaluer la qualité des clusters obtenus, le choix des métriques existantes étant arbitraire. Nous proposons dans cet article une nouvelle façon d'évaluer la qualité du clustering en convertissant ces clusters en réseaux de Petri et en utilisant un outil de *model checking* statistique basé sur une logique temporelle HASL. Nous expérimentons notre technique dans un benchmark classique de process mining avec des résultats concluants.

Model-Based Clustering by Greedy Maximization of the Integrated Classification Likelihood

Nicolas Jouvin*[1], Etienne Côme[2], Pierre Latouche[3], Charles Bouveyron[4]

[1] Mathématiques et Informatique Appliquées – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0518 – France
[2] Université Gustave Eiffet / Département Composants et systèmes (COSYS) / Génie des Réseaux de Transport Terrestres et Informatique Avancée (UGE / COSYS - GRETTIA) – IFSTTAR, PRES Université Paris-Est – France
[3] Mathématiques Appliquées Paris 5 – Centre National de la Recherche Scientifique : UMR8145, Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Université Paris Descartes - Paris 5 : UMR8145 – France
[4] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France. – Université Côte d'Azur (UCA) – France

Nous présentons une nouvelle approche pour le clustering probabiliste. S'appuyant sur une maximisation directe de la vraisemblance classifiante intégrée par rapport à la partition, elle permet de faire conjointement clustering et sélection du nombre de groupes. Ce problème d'optimisation hautement combinatoire est résolu grâce à un algorithme génétique, et une étape finale de clustering hiérarchique permet d'explorer les résultats à différentes échelles ainsi que d'extraire un ordre sur les clusters, utile pour la visualisation. Applicable à une grande variété de modèles, cette méthode convient au clustering de différents types de données, ainsi qu'a des données hétérogènes. La présentation s'appuiera notamment sur le paquet R **greed** qui implémente la méthode pour les principaux modèles utilisés dans la littérature.

A Dynamic Latent Block Model for Co-clustering of Zero-Inflated Count Data Streams

Giulia Marchello*[1], Marco Corneli[2], Charles Bouveyron[3,4]

[1] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France – Université Côte d'Azur, Institut National de Recherche en Informatique et en Automatique, Centre Sophia-Antipolis Méditerranée – France
[2] Université Côte d'Azur, Inria, Maison de la Modélisation des Simulations et des Interactions (MSI), Maasai team, Nice, France. – Université Côte d'Azur, Institut National de Recherche en Informatique et en Automatique, Centre Sophia-Antipolis Méditerranée – France
[3] Laboratoire J.-A. Dieudonné, UMR CNRS 7531, Université Côte d'Azur – CNRS : UMR7531 – France
[4] Equipe Maasai, Inria Sophia Antipolis – Institut National de Recherche en Informatique et en Automatique – France

The simultaneous clustering of observations and features of data sets (known as co-clustering) has recently emerged as a central machine learning application to summarize large data sets. This work introduces a novel latent block model for the dynamic co-clustering of count data streams with high sparsity. To properly model this type of data, we assume that the observations follow a time and block dependent mixture of zero-inflated Poisson distributions. To model and detect abrupt changes in the dynamics of both clusters memberships and data sparsity, the mixing and sparsity proportions are modeled through systems of ordinary differential equations. The model inference relies on an original variational procedure whose maximization step trains recurrent neural networks in order to solve the dynamical systems. Numerical experiments on simulated data sets demonstrate the effectiveness of the proposed methodology.

False clustering rate control in mixture models

Ariane Marandon*, Etienne Roquain[1], Tabea Rebafka[2], Nataliya Sokolovska[3]

[1] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France
[2] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France
[3] University Paris 6, INSERM 1166, ICAN – Univeristy Paris 6 – France

The clustering task consists in delivering labels to the members of a sample. However, for most data sets, some individuals are ambiguous and intrinsically difficult to attribute to one or another cluster. Moreover, in practical applications, misclassifying individuals is potentially disastrous. To overcome this difficulty, the idea followed here is to classify only a part of the sample in order

to maintain a low misclassification rate. This approach is well known in the supervised setting, and referred to as classification with an abstention option. The purpose of this paper is to revisit this approach in an unsupervised mixture-model framework. The problem is formalized in terms of controlling the false clustering rate (FCR) below a prescribed level α, while maximizing the number of clustered items. New procedures are introduced and their behavior is shown to be close to the optimal one by establishing theoretical results and conducting numerical experiments. An application to breast cancer data illustrates the benefits of the new approach from a practical viewpoint.

Transport optimal

(Amphi 11 - 14h30-16h00)

Transport optimal pour l'appariement automatique de données métabolomiques non-ciblées

Marie Breeur *[1], George Stepaniants[2], Philippe Rigollet[2], Vivian Viallon[1]

[1] Centre International de Recherche contre le Cancer - International Agency for Research on Cancer – Organisation Mondiale de la Santé / World Health Organization Office – France
[2] MIT Mathematics Department – États-Unis

Le profilage métabolomique non ciblé permet de mesurer une large gamme de métabolites présents dans un échantillon biologique. Cependant, l'analyse et l'interprétation de ces données sont parfois difficiles, notamment en raison du faible nombre d'échantillons par étude, d'où la nécessité de pouvoir combiner ou méta-analyser les données issues de plusieurs études afin de gagner en pouvoir statistique. Or, un métabolite mesuré dans le cadre d'une approche non-ciblée n'est pas immédiatement identifiable et est défini uniquement par son rapport masse sur charge (m/z) et son temps de rétention (RT). En outre, les ensembles de données obtenus dans des conditions non identiques sont soumis à des variations de m/z et à des altérations significatives du RT. Par conséquent, identifier les métabolites communs à deux études différentes est particulièrement difficile et constitue un obstacle majeur à la mise en commun de ces données.
Nous présentons ici une méthode non supervisée destinée à détecter et apparier les métabolites communs à deux études métabolomiques non ciblées en utilisant leurs m/z et RT et leurs intensités de signal. Notre approche repose sur la distance de Gromov-Wasserstein (GW), une extension du transport optimal conçue pour coupler des ensembles en se basant sur leur structure. Nous avons ajouté une contrainte visant à restreindre ce couplage aux paires ayant des m/z proches et avons estimé la déviation des RT afin de ne retenir que les paires présentant des RT compatibles.

Nous avons ensuite évalué empiriquement les performances de notre méthode sur des données simulées et réelles, en la comparant à une autre approche récente prenant en compte les mêmes informations (m/z, RT et intensités mesurées pour les différents échantillons). Les résultats obtenus sont prometteurs et doivent être répliqués sur données réelles. Si ces bonnes performances se confirment, notre méthode pourrait avoir de nombreuses applications en métabolomique en permettant la comparaison de protocoles d'acquisition, la combinaison ou méta-analyse de données issues de différentes études, ce qui permettrait à la métabolomique non-ciblée d'être utilisée à son plein potentiel, par exemple en épidémiologie.

GAN Estimation of Lipschitz Optimal Transport

Lucas De Lara*[1], Alberto Gonzalez-Sanz[1], Louis Bethune[2]

[1] IMT TOULOUSE – Université Paul Sabatier - Toulouse III – France
[2] IRIT – Université Paul Sabatier-Toulouse III - UPS – France

This paper introduces the first statistically consistent estimator of the optimal transport map between two probability distributions, based on neural networks. Building on theoretical and practical advances in the field of Lipschitz neural networks, we define a Lipschitz-constrained generative adversarial network penalized by the quadratic transportation cost. Then, we demonstrate that, under regularity assumptions, the obtained generator converges uniformly to the optimal transport map as the sample size increases to infinity. Furthermore, we show through a number of numerical experiments that the learnt mapping has promising performances. In contrast to previous work tackling either statistical guarantees or practicality, we provide an expressive and feasible estimator which paves way for optimal transport applications where the asymptotic behaviour must be certified.

JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport

Marion Jeamart[1], Renan Bernard[1], Nicolas Courty[1], Chloé Friguet*[1], Valerie Gares[2]

[1] Univ. Bretagne-Sud – Université Bretagne Sud, IRISA, Vannes, France – France
[2] Univ Rennes INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes – INSA Rennes : UMR6625 – France

L'adaptation de domaine est un champ de l'apprentissage par transfert où les données d'entrainement (source) du modèle et celles utilisées pour le test (cible) proviennent de deux domaines dont les distributions sous-jacentes sont différentes, et il convient d'adapter le modèle appris pour qu'il puisse être utilisé sur les données cibles avec de bonnes performances. On présente ici un algorithme permettant de traiter l'adaptation de domaine dans le cas d'espaces source et cible hétérogènes car représentés par des espaces de caractéristiques différents. La méthode

développée utilise le transport optimal pour coupler les distributions des deux domaines et son implémentation est illustrée sur des données de référence.

La Distance de Sliced-Wasserstein pour l'Apprentissage à Grande Échelle

Kimia Nadjahi*[1]

[1] Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France

Les algorithmes actuels d'inférence statistique et d'apprentissage utilisent des distances définies sur l'espace des distributions de probabilité pour comparer des jeux de données. La distance de Wasserstein, issue du problème de transport optimal, est un choix intéressant mais souffre de limites computationnelle et statistique à grande échelle. Plusieurs métriques alternatives ont alors été proposées, notamment la distance de Sliced-Wasserstein (SW), de plus en plus utilisée en pratique en raison de son coût de calcul avantageux. Cependant, peu de travaux avaient analysé SW d'un point de vue théorique, ce qui limite fortement l'étude des méthodes basés sur cette distance. Nous proposons alors d'étudier les propriétés théoriques de SW afin de mieux comprendre son impact dans les problématiques modernes d'apprentissage automatique, notamment la modélisation générative. En particulier, nous prouvons un ensemble de résultats qui montrent la robustesse de SW à la dimension des données, ce qui est un avantage considérable par rapport à la distance de Wasserstein en grande dimension. Néanmoins, SW est couramment estimée en pratique avec une simple méthode de Monte Carlo, qui induit une erreur d'approximation susceptible de détériorer la performance des algorithmes basés sur SW. Pour éviter ce problème, nous introduisons une nouvelle approximation déterministe de SW en s'appuyant sur des résultats de concentration de la mesure. Nous prouvons que l'estimateur obtenu présente des garanties non asymptotiques sous une condition de dépendance faible, tout en étant plus rapide à calculer que la technique de Monte Carlo. Nos contributions sont illustrées sur des problèmes avec données synthétiques ainsi que des applications d'apprentissage profond pour la génération d'images.

16h00 – 16h10 Pause

16h10 – 16h30 COMPUTO

16h30 – 16h40 Pause

16h40 – 18h00 AG SFdS

---

15 juin 2022

Programme

9h00 – 10h00 PLENIERE : Conférence Le Cam – Markus Reiss . . . **87**

10h00 – 10h10 Pause . . . **87**

10h10 – 11h50 . . . **87**

SFdS - Società Italiana di Statistica . . . 88

Andrea Bucci [et al.] . . . 88

Cira Perna [et al.] . . . 88

Donatella Vicari . . . 89

SFdS - ENBIS . . . 89

Euan Enticott . . . 89

Tran Vi-Vi Élodie Perrin [et al.] . . . 90

David Obst [et al.] . . . 91

Amandine Pierrot [et al.] . . . 91

SFdS - Sport . . . 92

Julien Schipman [et al.] . . . 92

Audrey Difernand . . . 92

Quentin De Larochelambert [et al.] . . . 93

Léo Gerville-Réache . . . 94

Sébastien Déjean [et al.] . . . 94

Thibaut Ledanois . . . 95

Statistique mathématique . . . 95

Smail Adjabi . . . 95

Subhasish Basak [et al.] . . . 96

Florian Combes [et al.] . . . 96

Salima Helali [et al.] . . . 97

Hugo Senetaire [et al.] . . . 97

Patrick Tardivel [et al.] . . . 98

11h50 – 12h00 Pause . . . **98**

12h00 – 13h00 . . . **98**

PLENIERE : Li-Chun Zhang . . . 98

PLENIERE : Emilie Kaufmann . . . 99

13h00 – 14h00 Déjeuner . . . **100**

14h00 – 19h00 Programme Social **100**

19h00 – 21h00 Soirée festive **100**

9h00 – 10h00

PLENIERE : Markus Reiss

(Amphi 9 - 9h00-10h00)

Rank detection for time-varying covariance matrices and how Le Cam Theory may help

Markus Reiß

Humboldt-Universität zu Berlin

Rank detection for covariance matrices is one of the fundamental inference problems in statistics. Here we focus on the case of a time-varying instaneous (or spot) covariance matrix $S(t)$ of a continuous-time process $X(t)$. The data are given by high-frequency observations of X on $[0,T]$, possibly corrupted by noise. We ask for testing the hypothesis $\mathbb{H}_1 : \int_0^T \lambda_{r+1}(S(t))\mathrm{d}t \geq v_n$ i.e. that the mean $(r+1)$st eigenvalue is larger than some signal detection rate v_n, tending to zero with sample size n. This problem can be embedded in the classical nonparametric signal detection framework, but it has many unexpected features. For instance, the optimal detection rate depends on a regularity assumption on $S(t)$ under the null, not the alternative and a possible spectral gap leads to significantly better detection rates. We show how an asymptotically equivalent nonparametric Gaussian white noise model may help to understand the structure of the statistical problem and to come up with efficient and implementable methods. The rank detection is illustrated with applications to intraday data from government bonds.

Based on joint work with Lars Winkelmann, Markus Bibinger, Nikolaus Hautsch, Peter Malec

10h00 – 10h10 Pause

10h10 – 11h50

SFdS – Società Italiana di Statistica

(Amphi 7 - 10h10-11h40)

Analysing spatio-temporal patterns in the association of confirmed deaths by COVID-19 and intensive care admissions

Andrea Bucci*[1], Luigi Ippoliti[1], Pasquale Valentini[1]

[1] G. d'Annunzio University Chieti-Pescara – Italie

Understanding whether and to what extent the intensive care capacity strain for COVID-19 affects the death rate is critical to preparing authorities for further potential pandemic waves. As responses from governments to the COVID-19 pandemic have been varied across countries and through time, interest may also focus on finding groups of Regions with similar behaviour in terms of that association. To this end, we propose a flexible Bayesian nonparametric approach, based on a mixture of Gaussian processes coupled with the Dirichlet process, capable of clustering time series in terms of the time-varying parameter relative to this association, and allowing spatial correlation, measurement errors and predictions. We evaluate the proposed methodology on the weekly counts of deaths and new admissions to intensive care recorded at the NUTS-2 regional level for several European countries.

Sieve bootstrap based on extreme learning machines in time series analysis

Cira Perna*[1], Michele La Rocca[1]

[1] Department of Economics and Statistics, University of Salerno – Italie

The aim of the talk is to propose and discuss a sieve bootstrap scheme based on Extreme Learning Machines for non linear time series. The use of Extreme Learning Machines in the resampling scheme can dramatically reduce the computational burden of the bootstrap procedure, with

performances comparable to the NN-Sieve bootstrap and computing time similar to the AR-Sieve bootstrap. Moreover, the proposed procedure is fully nonparametric in its spirit and retains the conceptual simplicity of the residual bootstrap. A Monte Carlo simulation experiment has been implemented, in order to evaluate the performance of the proposed procedure and to compare it with the NN-Sieve bootstrap.

Advances in clustering asymmetric proximity data

Donatella Vicari *[1]

[1] Dipartimento di Scienze Statistiche, Sapienza, Università di Roma – Italie

Résumé. L'ensemble à classer est caractérisé par une matrice de proximité carrée asymétrique telle qu'une matrice d'échange. Quelques modèles de classification non hiérarchiques sont présentés, basés sur la décomposition de cette matrice dans sa partie symétrique et antisymétrique.
Abstract. Some non-hierarchical clustering models for asymmetric proximity data are presented which rely on the decomposition into symmetric and skew-symmetric components.

SFdS - ENBIS

(Amphi 8 - 10h10-11h40)

Scalable additive stacking models for electricity demand forecasting

Euan Enticott *[1]

[1] University of Bristol – Royaume-Uni

This submission is for a special session.
Short-term electricity demand forecasts are regularly used to inform decisions in grid management. However, the increasing reliance on renewable production will create a new challenge in demand forecasting. Renewable energy sources are less centralised and often dependent on external factors such as weather. To limit the need for large-scale and expensive storage infrastructure, smart grid management systems can be employed. These systems will require reliable demand

forecasts at lower levels of aggregation, possibly down to the individual household level. At this level, demand is characterised by a low signal-to-noise ratio, with frequent abrupt changepoints in demand dynamics. The challenges posed by forecasting at a low level of aggregation motivate the use of an ensemble approach that can incorporate information from several models and across households. The idea of additive stacking for probabilistic forecasting was proposed by Capezza (2020). However, the number of models that could be used was limited as the number of unknown parameters in the stacking model scales linearly with the number of experts.
In this talk, we will discuss more scalable solutions for additive stacking. Specifically, we explore the idea of adding structure to the weights to reduce the number of parameters required. This provides modeling advantages by reducing both computational complexity and the risk of overfitting. The area of application is not limited to demand forecasting, but extend to any setting where ensemble modeling is used.

Metamodeling and sensitivity analysis for models with spatial output. Application to coastal flooding models.

Tran Vi-Vi Élodie Perrin*[1], Olivier Roustant[2], Jérémy Rohmer[3], Jean-Philippe Naulin[4]

[1] Office national d'études et de recherches aérospatiales (Toulouse) – ONERA – France
[2] Institut National des Sciences Appliquées de Toulouse (INSA) – Institut National des Sciences Appliquées - Toulouse – France
[3] Bureau de Recherches Géologiques et Minières (BRGM) – Ministère de l'enseignement supérieur, de la recherche et de l'innovation, Ministère de la Transition Ecologique, ministère de l'Economie, des Finances et de la Relance – France
[4] Caisse Centrale de Réassurance (CCR) – Caisse Centrale de Réassurance – France

Motivé par l'évaluation des risques de submersions marines, on considère les modèles hydrodynamiques numériques développés par le BRGM et la CCR. La sortie de ces simulateurs est une carte d'inondation. L'objectif est de réaliser une analyse de sensibilité (AS) afin de mesurer et de hiérarchiser l'influence des paramètres d'entrée sur la sortie. Afin de réduire le temps de calcul des modèles et la dimension de la sortie spatiale, on propose d'utiliser l'ACP fonctionnelle (ACPF). La sortie est décomposée dans une base de fonctions, adaptée pour traiter les variations locales, telle que les ondelettes ou les B-splines. Une ACP avec une métrique ad-hoc est appliquée aux coefficients les plus importants, selon un critère d'énergie après orthonormalisation de la base, ou directement sur la base originale avec une approche de régression pénalisée. Des méta-modèles (comme le krigeage) sont construits sur les premières composantes principales, sur lesquels peut être réalisée l'AS. Comme résultat complémentaire, une formule analytique est obtenue pour les indices de sensibilité basés sur la variance, généralisant celle connue pour des bases orthonormées. L'ensemble des travaux a été appliqué à un cas d'inondation côtière. Les processus d'inondation ont été simulés avec les modèles numériques du BRGM et de la CCR.

Textual Data for Electricity Demand Forecasting

David Obst [1,2], Badih Ghattas [2], Sandra Claudel [1], Jairo Cugliari [*3,4], Yannig Goude [5], Georges Oppenheim [6]

[1] EDF Labs – EDF Recherche et Développement – France
[2] Institut de Mathématiques de Marseille – Aix Marseille Université : UMR7373 – France
[3] Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon 2 : EA3083 – Université Lumière Lyon 2, 5 avenue Pierre Mendès-France 69676 Bron Cedex, France
[4] Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon 2 : EA3083, Université Claude Bernard Lyon 1 – Université Lumière Lyon 2, 5 avenue Pierre Mendès-France 69676 Bron Cedex, France
[5] EDF Labs – Electricité de France - EDF – France
[6] Paris Saclay – Université de Paris-Sud Orsay – France

Les modèles traditionnels de prévision de l'électricité à moyen terme s'appuient sur des informations calendaires et météorologiques telles que la température et la vitesse du vent pour obtenir des performances élevées. Cependant, le fait de dépendre de ces variables présente des inconvenients, car elles peuvent ne pas être suffisamment informatives en cas de conditions météorologiques extrêmes, ce qui entraîne des erreurs de prevision plus importantes, et les données météorologiques historiques sont souvent disponibles avec retard ou à un prix élevé. Bien qu'omniprésentes, les sources d'information textuelles sont rarement incluses dans les algorithmes de prédiction des séries temporelles, malgré les informations pertinentes qu'elles peuvent contenir. Dans ce travail, nous proposons d'exploiter les rapports météorologiques librement accessibles pour les problèmes de prediction de la demande d'électricité et des séries temporelles météorologiques. Nos experiences sur des données de charge françaises et britanniques montrent que les sources textuelles considérées permettent d'améliorer la précision globale du modèle de référence, en particulier par temps froid. De plus, nous appliquons notre approche au problème de l'imputation des valeurs manquantes dans les séries temporelles météorologiques, et nous montrons que notre approche textuelle bat les méthodes standard. De plus, l'influence des mots sur les prédictions des séries temporelles peut être interprétée pour les schémas d'encodage considérés du texte, ce qui conduit à une plus grande confiance dans nos résultats.

On forecasting bounded variables with latent stochastic bounds

Amandine Pierrot [*1], Pierre Pinson

[1] Technical University of Denmark – Danemark

En lien avec l'application à la prévision éolienne, nous nous intéressons à la prévision de variables continues et bornées quand les bornes de l'intervalle varient au cours du temps ou en fonction de variables exogènes sans pouvoir être observées. Out of the application to wind power forecasting, we are interested in forecasting bounded continuous variables when the bounds may vary over

time or depending on exogenous variables while not being observed.

SFdS - Sport

(Amphi 9 - 10h10-11h40)

L'origine de la déficience visuelle est-elle un facteur de performance ? Analyse des para-nageurs et para-athlètes de niveau international

Julien Schipman*[1], Bryan Le Toquin, Quentin De Larochelambert, Guillaume Sauliere, Stephanie Ducombe, Jean-François Toussaint

[1] IRMES – Institut national du sport, de l'expertise et de la performance – France

L'objectif de cette étude était d'analyser l'effet des déficiences visuelles congénitales et acquises sur les performances internationales en natation et athlétisme Paralympiques issues de l'ensemble des compétitions entre 2009 et 2019, soit 20 689 performances récoltées. L'origine de la déficience a été recueillie sur le site du Comité International Paralympique (CIP) et catégorisée en deux groupes (déficience congénitale et acquise). Dans les classes sportives des déficients visuels (11-12-13), le niveau de performance et la relation âge/performance ont été étudiés en fonction de l'origine de la déficience. Dans les classes 11 et 12, les performances maximales ont été atteintes plus tôt par les nageurs et nageuses atteints d'une déficience congénitale que par ceux et celles atteints d'une déficience acquise ($p < 0.05$). Aucune différence n'était présente dans la classe 13 ni dans aucune classe en para-athlétisme ($p > 0.05$). Un niveau de performance similaire a été observé entre les deux disciplines sportives pour chaque classe ($p > 0.05$). Cette étude a démontré que l'origine de la déficience peut influencer le parcours de performance chez les nageurs déficients visuels.

Relative age effect among French swimmers

Audrey Difernand*[1]

[1] Institut de recherche biomédicale et d'épidémiologie du sport – Institut National du Sport, de l'Expertise et de la Performance, Université de Paris : URP7329 – France

Résumé. L'objectif de cette étude était de révéler la présence de l'effet de l'âge relatif parmi les nageurs Français et de proposer une méthode de rééquilibrage afin de mieux apprécier le potentiel de l'athlète en fonction de sa catégorie et de sa discipline. 62 610 nageurs entre 10 et 16 ans sur la discipline du 100m nage libre en bassin de 50m sont considérés pour cette étude. Parmi eux, moins d'un nageur sur cinq entre 13 et 16 ans est né dans le dernier trimestre de l'année. Pour éviter l'abandon ou la perte de vue des nageurs, nous avons mis en place une méthode de rééquilibrage basée sur la performance du nageur, son âge exact au moment de la compétition et le coefficient de régression entre la performance et l'âge au sein de la catégorie considérée. Après application de la méthode, nous avons remarqué aucune différence significative entre les performances rééquilibrées et les performances réalisées a posteriori excepté pour la catégorie des 13 ans.
Mots-clés. Age relatif, natation, détection, méthodes de rééquilibrage

Abstract. The aim of this study was to highlight the presence of the relative age effect among French swimmers and to propose a method of rebalancing in order to better appreciate the potential of each athlete according to his category and discipline. 62 610 males' swimmers between the ages of 10 and 16 in the 100m freestyle in a 50m pool are considered for this study. Less than one in five swimmers aged 13-16 was born in the last quarter of the year. We implemented a rebalancing method based on the swimmer's performance, his exact age at the time of the competition and the regression coefficient between performance and age within the category considered. After applying the method, we found no significant differences between the rebalanced performances and the performances achieved except for the 13 years old category.
Keywords. Relative age, swimming, talent identification, rebalancing method

Parametric and non-parametric modeling by a Makovian multi-state model of the evolution of performance in French alpine skiing

Quentin De Larochelambert [*3,2,1], Audrey Difernand [4], Juliana Antero [4], Adrien Sedeaud [5], Jean-François Toussaint [5], Nicolas Coulmy [6], Pierre-Yves Louis [7]

[3] Institut de recherche biomédicale et d'épidémiologie du sport – Institut National du Sport, de l'Expertise et de la Performance, Université de Paris : URP7329 – France
[2] Département Sportif et Scientifique – Fédération Française de Ski – France
[1] Institut de mathématiques de Bourgogne – Université de Bourgogne-Franche-Comté – France
[4] Institut de recherche biomédicale et d'épidémiologie du sport – Institut National du Sport, de l'Expertise et de la Performance, Université de Paris : URP7329 – France
[5] IRMES – Institut National du Sport, de l'Expertise et de la Performance – France
[6] Fédération Française de ski (FFS) – Fédération Français de ski – 50 Rue des Marquisats 74000 Annecy, France
[7] AgroSup Dijon – Université de Bourgogne Franche-Comté (UBFC), AgroSup Dijon, UMR PAM A 02.102 – France

We estimate the level of future performance in alpine skiing by modeling the probabilities of transition between performance levels categorized into performance lanes, through a time-

inhomogeneous Markov chain as well as nonparametric and parametric multi-state models. We show that the chances of reaching the best level of performance in adulthood are very slightly higher for an individual in the first level of performance at age 12 than an individual in the last level of performance (5% difference). This difference increases with age (16% at 15 years).

Reparlons de notes en patinage

Léo Gerville-Réache*[1]

[1] Université de Bordeaux – France

Les Jeux olympiques d'hiver de 2022 étant terminés, cette communication vous propose de replonger dans l'une des épreuves qui a permis à la France de capitaliser une médaille d'or de plus. Il s'agit de la danse sur glace. Comme d'habitude, c'est un jury composé de 9 juges qui, via un ensemble de notes, a classé les prestations des 23 couples en compétition. En retirant toujours la meilleure et la moins bonne note de chaque évaluation, il nous a semblé opportun de rappeler et réutiliser les possibilités statistiques qu'offre le modèle de Gauss-Markov.

Modélisation de la probabilité de réussite d'une passe au football

Sébastien Déjean*[1], Javier Lopez Sanchez, Philippe Saint-Pierre[2]

[1] Institut de Mathématiques, Université de Toulouse et CNRS – Centre National de la Recherche Scientifique – UMR 5219, F-31062 Toulouse, France
[2] Institut de Mathématiques, Université de Toulouse et CNRS – Centre National de la Recherche Scientifique – France

Depuis quelques années, le sport offre un environnement très dynamique pour l'analyse statistique de données. Que ce soit pour accompagner le développement de systèmes de paris sportifs, pour agrémenter les diffusions télévisées de rencontres sportives ou en vue d'optimiser les performances d'un athlète ou d'une équipe, l'analyse de données est une aide précieuse dans de tels contextes. Pour le football, plusieurs études ont conduit à populariser la notion d'expected goal traduisant la probabilité de marquer un but pour un joueur (ou une équipe) dans une situation donnée. À ce jour, cependant, peu de travaux semblent se consacrer à la probabilité de réussite d'une passe pourtant élément majeur dans tout sport collectif. C'est avec l'objectif de contribuer à ce sujet que nous avons utilisé la régression logistique pour modéliser la probabilité de réussite d'une passe en fonction de différentes variables explicatives. Nos analyses montrent des résultats qui, loin de révolutionner le football (plus une passe est courte et vers l'arrière, plus il est probable de la réussir) mettent en évidence de nombreuses perspectives pour l'optimisation de la performance d'une équipe.

Estimation of race scenarios in standard triathlon according to level density

Thibaut Ledanois*[1]

[1] French Institute of Sport (INSEP), Research Department, Laboratory Sport, Expertise and Performance (EA7370) – Institut National du Sport, de l'Expertise et de la Performance – France

Le triathlon est un sport jeune dans les Jeux Olympiques, le mouvement stratégique et le rythme de la compétition sont différents selon les étapes, les saisons et le sexe. L'objectif de cette étude est de prédire les scénarios de course dans le triathlon standard (1,5 km de natation - 40 km de vélo - 10 km de course à pied) avec drafting au niveau international en utilisant des approches de méthode de clustering. Le point de départ était d'observer les différences de temps entre les athlètes et d'estimer le niveau de chaque triathlète par un modèle d'évaluation de Stephenson. Des analyses non supervisées, supervisées et statistiques sur les données historiques des résultats de compétition des hommes et des femmes de l'Union internationale de triathlon depuis 2009 à 2019 sont réalisées. Deux niveaux d'événement et trois modèles de course pour chaque sexe à la sortie de la section natation sont définis. Dans un premier temps, les résultats montrent une différence de niveau et de densité entre le niveau international (World Triathlon Series (WTS) / World Championship (WChamp), Jeux Olympiques (JO) et Test Event (TE) et le niveau continental (Continental Cup (ConCup)/ Continental Championships (ConChamp). Le niveau de la coupe du monde (WCup) présente un niveau très variable. Deuxièmement, nos résultats ont démontré une relation entre le niveau de densité du champ de départ des athlètes et la probabilité de tendre vers la nage groupée (indice de dispersion = 189.06, 176.47), la nage dispersée (indice de dispersion = 227.48, 204.42) et la nage avec échappée (indice de dispersion = 229.52, 219.89). En conclusion, cette étude fournit des informations aux athlètes et à l'encadrement avant la course en fonction de la liste de départ pour mieux préparer leur stratégie. Ainsi que d'avoir une analyse plus fine de l'objectif de performance fixé pour un athlète.

Statistique mathématique

(Amphi 10 - 10h10-11h40)

Estimateurs non paramétriques à noyau général birnbaum-saunders des densités multivariées à support non négatifs

Smail Adjabi*[1,2,3]

[1] Nabil ZOUGAB – Algérie
[2] Lynda HARFOUCHE – Algérie
[3] Yasmina ZIANE – Algérie

On présente l'estimateur à noyau Général Birnbaum-Saunders (GBS) de la fonction densité multivariée des données non négatives et ses propriétés statistiques (biais, variance et erreur quadratique moyenne intégrée). Pour réduire le biais de cet estimateur, on applique la technique de correction du biais notée MBC (Multiplicative Biais Correction) de Jones-Linton-Nielsen (1995). On obtient ainsi un nouvel estimateur et on donne ses propriétés statistiques. Pour étudier la qualité de l'estimateur obtenu, des études sont menées sur des données simulées à partir de densités cibles bivariées connues et sur des données réelles en utilisant le critère : biais quadratique intégré (ISB)

Integration of bounded monotone functions: Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods

Subhasish Basak*[1,2], Julien Bect[2], Emmanuel Vazquez[2]

[1] ANSES – Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail – France
[2] L2S – Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France. – France

In this article we revisit the problem of numerical integration for monotone bounded functions, with a focus on the class of nonsequential Monte Carlo methods. We first provide new a lower bound on the maximal L_p error of nonsequential algorithms, improving upon a theorem of Novak when $p > 1$. Then we concentrate on the case $p = 2$ and study the maximal error of two unbiased methods-namely, method based on the control variate technique, and the stratified sampling method.

Subsampling under constraint

Florian Combes*[1,2], Badih Ghattas[1], Ricardo Fraiman[3]

[1] Aix Marseille University, CNRS, I2M – France
[2] Renault Group, Customer usage, – Renault Group, Customer usage, – France
[3] Centro de Matemática, Universidad de la República – Uruguay

In our setting we have an input X in a general space, and an output $Y = f(X)$ where f is a very complicated function, whose computational cost for every new input is very high. We are given

two sets of observations of X, S_1 and S_2 of different sizes such that only $f(S_1)$ is available. We tackle the problem of selecting a subsample $S_3 \in S_2$ of smaller size on which to run the complex model f, and such that distribution of $f(S_3)$ is close to that of $f(S_1)$. We suggest two algorithms to solve this problem and show their efficiency using simulated datasets.

Recursive conditional distribution estimation using kernel method and Bernstein polynomials

Salima Helali*[1], Yousri Slaoui[2]

[1] Laboratoire Angevin de Recherche en Mathématiques - Angers – angers – France
[2] Laboratoire de Mathématiques et Applications – Université de Poitiers, Centre National de la Recherche Scientifique : UMR7348 – Téléport 2 - BP 30179 Boulevard Marie et Pierre Curie 86962 Futuroscope Chasseneuil Cedex, France

The estimation of the conditional distribution function is a fundamental issue in non parametric estimation. From this perspective, a recursive estimator using kernel method and Bernstein polynomials, is introduced. We derive the asymptotic properties of the proposed estimator such as its asymptotic bias, variance and mean squared error, which strongly depend on the choice of some parameters. The asymptotic normality of the estimator is also established. Under some conditions, the proposed recursive estimator will be very competitive to other estimators, in terms of estimation error. Eventually, the performance of the proposed estimator is explored through a few simulation examples.

Explainability as statistical inference

Hugo Senetaire*[1], Jes Frellsen[1], Pierre-Alexandre Mattei[2]

[1] Danish Technical University – Danemark
[2] Inria Sophia Antipolis – Inria Sophia Antipolis (MAASAI) – France

A wide variety of model explanation approaches have been proposed in recent years. Here, we are interested in the removal-based approach, which quantifies how much a prediction change when only a subset of features is shown to the model. More precisely, we look into amortized interpretability models, where a neural network is used as a selector to allow for fast interpretation at inference time. Traditional methods rely on separate training of the predictor and the selector models. Here, we propose a single probabilistic model with a single maximum likelihood based loss. This allows for fast and efficient training of the selector and predictor at the same time. Our proposed method is independent of the predictor network architecture and can be apply to any problem.

La Géométrie pour l'Unicité, la Parcimonie et le Regroupement des Estimateurs Pénalisés

Patrick Tardivel*[1], Ulrike Schneider[2]

[1] Institut de mathématiques de Bourgogne – Université de Bourgogne-Franche-Comté – France
[2] TU Wien – Autriche

Durant la présentation nous donnerons une condition nécessaire et suffisante pour l'unicité d'un estimateur des moindres carrés pénalisés dont le terme de pénalité est une norme polyèdrique. Nos résultats couvrent de nombreuses méthodes incluant les estimateurs OSCAR, SLOPE et LASSO ainsi que la méthode de poursuite de base qui est une extension du LASSO.
Cette condition d'unicité est géométrique et implique l'espace vectoriel engendré par les lignes de la matrice de planification ainsi que les faces de la boule unité de la norme duale coupées par cet espace.
Des résultats théoriques sur la parcimonie des estimateurs LASSO et poursuite de base sont déduit de cette condition via la caractérisation des vecteurs signes accessibles pour ces deux méthodes.

11h50 – 12h00 Pause

12h00 – 13h00

PLENIERE : Li-Chun Zhang

(Amphi 7 - 12h00-13h00)

Descriptive inference of big-data statistics

Li-Chun Zhang

Univ. Southampton

We consider descriptive inference where the targets of interest can in principle be observed in a 'perfect census', in contrast to analytic inference where such a 'perfect census' does not exist even conceptually. For instance, which ones among the residents of France (as a given population) are infected by a certain virus on a given day is a descriptive inference problem, whereas the 'true' regression relationship between a scaler response and a given set of explanatory variables is a problem of analytic inference.

A fundamental challenge for descriptive inference based on supervised (machine) learning is to 'extrapolate' the model learned from the available observations (as a sample) to the unobserved ones, without which the learning would have little use. No matter how learning is organised within the sample, one cannot ensure the adopted model is valid outside it unless the sample is selected from the population in some controlled manner.

We shall consider, in two situations particularly, how probability sampling (or design-based) methods can be combined with supervised learning, such that the validity of descriptive inference is ensured with respect to hypothetical repeated probability sampling, regardless the adopted model is 'true' or not. In the first situation, supervised learning is based on a probability sample such that one can obtain an estimator of the total errors from applying the learned model to the out-of-sample units, which is unbiased over repeated sampling. In the second situation, we assume that the adopted model is learned from a very large 'convenience sample' (or big data), such that it is necessarily misspecified to some extent for the out-of-sample units, due to problems of incomplete coverage, imperfect measurement or informative selection, and the bias of any resulting statistic overwhelms the associated variance. A so-called auditing sample can now be used to provide accuracy measures that are valid over repeated sampling and unaffected by the failure of the assumptions underlying the big-data statistics themselves.

PLENIERE : Emilie Kaufmann

(Amphi 9 - 12h00-13h00)

Algorithmes de bandits non-paramétriques: optimalité et robustesse

Emilie Kaufmann

CNRS CRIStAL Lille

Dans un modèle de bandit, un agent sélectionne de manière séquentielle des "bras", qui sont des lois de probabilité initialement inconnues de l'agent, dans le but de maximiser la somme des échantillons obtenus, qui sont vus comme des récompenses. Les algorithmes de bandits les plus populaires sont basés sur la construction d'intervalles de confiance ou l'échantillonnage d'une loi a posteriori, mais ne peuvent atteindre des performances optimales qu'un ayant une connaissance a priori sur la famille de distributions des bras. Dans cet exposé nous allons présenter des approches alternatives basées sur du ré-échantillonnage de l'historique de chaque bras. De tels algorithmes peuvent s'avérer plus robustes en deux sens. Nous verrons qu'ils peuvent être optimaux pour plusieurs classes de distributions, et être aisément adaptés à des situations où le critère de performance n'est pas lié à la récompense moyenne de l'agent, mais prend en compte une mesure de risque.

13h00 – 14h00 Déjeuner

14h00 – 19h00 Programme Social

19h00 – 21h00 Soirée festive

La soirée du mercredi 15 juin sera l'occasion de nous retrouver pour le traditionnel dîner de gala. Cette année, nous vous proposons une nouvelle organisation, avec une soirée qui se déroulera dans 5 restaurants disséminés dans la ville:

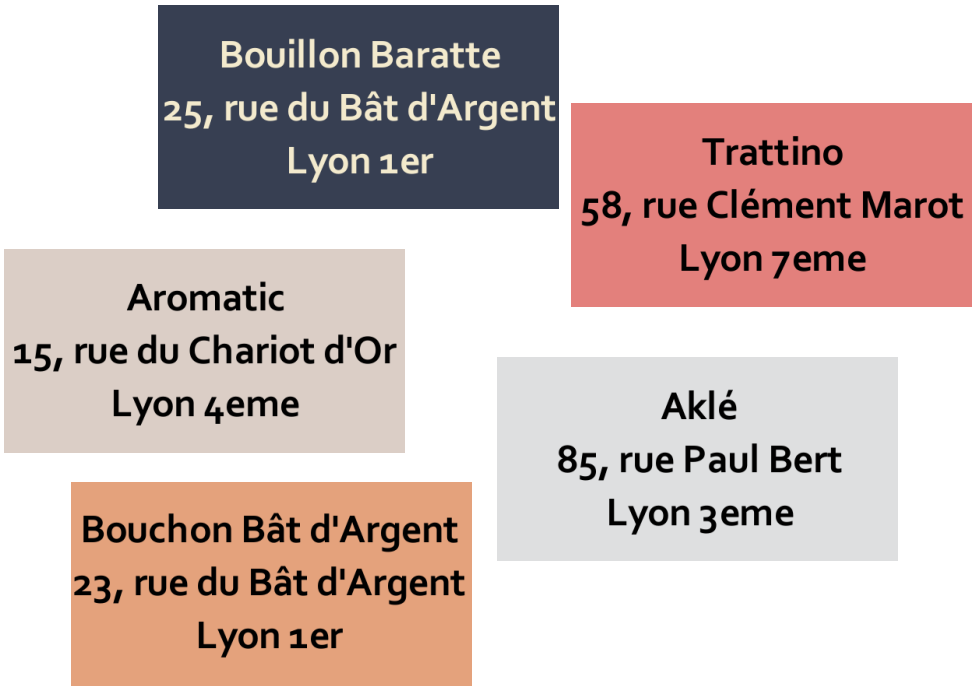
- **Akle**, 85 Rue Paul Bert, 69003 Lyon
- **l'Aromatic**, 15 Rue du chariot d'Or, 69004 Lyon
- **Bouillon Baratte**, 25 Rue du Bat d'argent, 69001 Lyon
- **Bouchon Bat d'argent**, 23 Rue du Bat d'argent, 69001 Lyon

- **Trattino**, 58 Rue Clément Marot, 69007 Lyon

Vous avez choisi entre plusieurs ambiances, et recevrez un ticket nominatif indiquant le restaurant qui vous attendra. Le nom associé au ticket ne sera qu'indicatif (pour avoir un suivi des tickets distribués). Les participants qui le souhaitent pourront s'échanger leur ticket, sous réserve de trouver un(e) collègue avec qui échanger.

Vous serez attendus dès 19h pour l'apéritif !

Excellente soirée !



16 juin 2022

Programme

9h00 – 10h00 . . . **105**

PLENIERE : Samory Kpotufe . . . 105
PLENIERE : Pierre Chainais . . . 106

10h00 – 10h20 Pause . . . **106**

10h20 – 11h40 . . . **106**

SFdS - Statistique mathématique . . . 106
Victor-Emmanuel Brunel . . . 107
Stanislav Nagy . . . 107
SFdS - BFA . . . 107
Marie Kratz [et al.] . . . 108
Etienne Marceau . . . 108
Gilles Stupfler [et al.] . . . 109
Olivier Wintenberger [et al.] . . . 109
Applications . . . 110
Jean-Baptiste Aubin [et al.] . . . 110
Guillaume Flament . . . 111
Camille Garcin [et al.] . . . 111
Valerie Gares [et al.] . . . 112
Wojciech Reise . . . 112
Extrêmes & Prediction . . . 113
Gloria Buritica [et al.] . . . 113
Marine Demangeot [et al.] . . . 113
Thibault Modeste [et al.] . . . 114
Romain Pic [et al.] . . . 115
Nicklas Werge . . . 115
Survie . . . 116
Ariane Cwiling [et al.] . . . 116
Audrey Lavenu [et al.] . . . 117
Eliz Peyraud . . . 118
Thanh Huan Vo [et al.] . . . 118

11h40 – 12h00 Pause . . . **119**

12h00 – 13h10 . . . **119**

SFdS - AMIES . . . 119
Van Tuan Nguyen . . . 119
Véronique Maume-Deschamps . . . 120
Esso-Ridah Bleza [et al.] . . . 120
Djibril Sarr . . . 120
ML - Graphes - Réseaux . . . 121
Dingge Liang [et al.] . . . 121
Annabelle Beaudoin [et al.] . . . 122
Hugo Schmutz [et al.] . . . 122
Louis Ohl [et al.] . . . 123
Méta-modèles – Analyse de sensibilité . . . 124
Naoufal Acharki [et al.] . . . 124
Clément Bénesse . . . 125
Gabriel Sarazin [et al.] . . . 125
Bénédicte Colnet [et al.] . . . 126
Oumar Baldé [et al.] . . . 126
ML & Regret . . . 127
Pierre Laforgue [et al.] . . . 127
Antoine Barrier [et al.] . . . 128
Aymen Al Marjani . . . 128
Léo Grill . . . 128
Carl Remlinger . . . 129
Statistique mathématique . . . 129
Clément Lalanne [et al.] . . . 130
Oumaima Ben Mrad [et al.] . . . 130
Marouane Il Idrissi [et al.] . . . 131
Thi Thanh Yen Nguyen [et al.] . . . 131
Steven Golovkine [et al.] . . . 132
Sylvie Viguier-Pla [et al.] . . . 132
13h30 – 14h30 Déjeuner . . . 133
14h30 – 15h30 . . . 133
PLENIERE : Frédéric Chazal . . . 133
PLENIERE : Gersende Fort . . . 133
15h30 – 15h50 Pause . . . 134
15h50 – 17h20 . . . 134
Séries Temporelles . . . 134
Guy Mélard [et al.] . . . 135
Raphael Mignot [et al.] . . . 135
William Todo [et al.] . . . 136
ML et Extrêmes . . . 136
Siham Alaoui Belghiti [et al.] . . . 137
Ibrahim Bouzalmat [et al.] . . . 137
Tanguy Lefort [et al.] . . . 138
Mira Rahal [et al.] . . . 138

Sara Armaut 139
Eric Adjakossa 139
Méthodes Génératives - ML - Deep 139
Teddy Ardouin [et al.] 140
Pierre Marion 140
Antoine Salmona [et al.] 141
Arsen Sultanov [et al.] 141
Pierre Wolinski [et al.] 142
Processus - Statistique mathématique 142
Francois Bachoc 142
Hadrien Lorenzo [et al.] 143
Cécile Mercadier 143
Amine Ounajim [et al.] 144
Ousmane Sacko 144
Théo Moins [et al.] 145
Statistique mathématique – Valeurs Manquantes 145
Alexis Ayme [et al.] 145
Christophe Crambes [et al.] 146
Firas Ibrahim [et al.] 146
Aude Sportisse [et al.] 147
Herbert Susmann 147
17h20 – 17h30 Pause 148
17h30 – 18h00 Clôture 148
18h00 – 18h30 Pause 148
18h30 – 19h30 148
Rencontre Jeunes statisticiens 148
Café de la Statistique 149

9h00 – 10h00

PLENIERE : Samory Kpotufe

(Amphi 7 - 9h00-10h00)

Adaptivity in Domain Adaptation and Friends

Samory Kpotufe

Columbia University

Domain adaptation, transfer, multitask, meta, few-shots, representation, or lifelong learning ... these are all important recent directions in ML that all touch at the core of what we might mean by 'AI'. As these directions all concern learning in heterogeneous and ever-changing environments, they all share a central question: what information a data distribution may have about another, critically, in the context of a given estimation problem, e.g., classification, regression, bandits, etc.

Our understanding of these problems is still rather fledgeling. We plan to present both some recent positive results and also some negative ones. On one hand, recent measures of discrepancy between distributions, fine-tuned to given estimation problems (classification, bandits, etc) offer a more optimistic picture than existing probability metrics (e.g. Wasserstein, TV) or divergences (KL, Renyi, etc) in terms of provable rates. On the other hand, when considering seemingly simple extensions to choices between multiple datasets (as in multitask), or multiple prediction models (as in Structural Risk Minimization), it turns out that minimax oracle rates are not always adaptively achievable, i.e., using just the available data without side information.

The talk will be based on joint work with collaborators over the last few years, namely, G. Martinet, S. Hanneke, J. Suk.

PLENIERE : Pierre Chainais

(Amphi 10 - 9h00-10h00)

Echantillonner efficacement grâce à l'approximation de distributions

Pierre Chainais

ECLille CRIStAL Lille

Les méthodes bayésiennes pour les problèmes inverses en traitement du signal et des images ont l'avantage de donner accès à la distribution a posteriori des paramètres à estimer. Ainsi, on accède non seulement à une solution au problème, mais aussi à des intervalles de crédibilité précieux. Par exemple, en astrophysique ou en médecine, il n'existe en général pas de vérité terrain. Fournir des prédictions assorties d'intervalles de confiance est essentiel : la lecture de l'image reconstruite se fait avec un niveau de confiance contrôlé. Néanmoins, les méthodes de Monte Carlo utilisées pour ces simulations de lois a posteriori sont réputées gourmandes en temps de calcul et limitées quant au passage à l'échelle en grande dimension ou pour une grand nombre de paramètres à estimer. Nous présenterons une famille d'approches appelées « augmentation de données asymptotiquement exacte » (AXDA). Cette approche, inspirée du splitting en optimisation, permet de construire de façon systématique une distribution approchée moins coûteuse à échantillonner que la loi cible du modèle initial, dans le cadre d'un compromis efficacité numérique/qualité de l'approximation. Ces méthodes ouvrent la voie à de nombreuses déclinaisons que nous évoquerons et illustrerons par des applications à la résolution de problèmes inverses.

10h00 – 10h20 Pause

10h20 – 11h40

SFdS - Statistique mathématique

(Amphi 7 - 10h20-11h40)

Profondeur de Tukey: Ensembles de niveau empiriques et théoriques

Victor-Emmanuel Brunel*[1]

[1] Centre de Recherche en Économie et STatistique (CREST) – ENSAE – France

La profondeur de Tukey est une notion statistique qui suscite beaucoup d'intérêt en statistique multivariée, car elle permet d'ordonner les données en généralisant, d'une certaine manière, la notion de quantile. Etant donnés une loi de probabilité et des données i.i.d. suivant cette loi, nous étudions la convergence des ensembles de niveau de la profondeur empirique vers les ensembles de niveau théoriques. Ces ensembles de niveau peuvent être interprétés comme des quantiles multivariés. Sous des hypothèses raisonnables, nous montrons la concentration des ensembles de niveau empiriques à la vitesse paramétrique, en utilisant des outils de géométrie convexe, de processus empiriques et de programmation linéaire semi-infinie.

Statistical depth functions: Characterization and reconstruction problems

Stanislav Nagy*[1]

[1] Department of Probability and Mathematical Statistics, Charles University – République tchèque

The depth is a concept of nonparametric statistics that generalizes ranks, orderings, and quantiles to multivariate data. Just as for quantiles, it is desired that a proper depth function fully characterizes probability measures, and that a measure is possible to be reconstructed efficiently from its depth. We consider two important depth functions from the literature - the halfspace and the simplicial depth - and explore the theory regarding their characterization and reconstruction properties.

SFdS - BFA

(Amphi 8 - 10h20-11h40)

Multi-normex for evaluating the distribution of aggregated heavy tailed risks

Marie Kratz*[1], Evgeny Prokopenko[2]

[1] ESSEC Business School – CREAR – France
[2] Sobolev Institute of Mathematics – Novosibirsk, Russie

We build a sharp approximation of the whole distribution of the sum of iid heavy-tailed random vectors, combining mean and extreme behaviors. It extends the so-called 'normex' approach from a univariate to a multivariate framework. We propose two possible multi-normex distributions, named d-Normex and MRV-Normex. Both rely on the Gaussian distribution for describing the mean behavior, via the CLT, while the difference between the two versions comes from using the exact distribution or the EV theorem for the maximum. The main theorems provide the rate of convergence for each version of the multi-normex distributions towards the distribution of the sum, assuming second order regular variation property for the norm of the parent random vector when considering the MRV-normex case. Numerical illustrations and comparisons are proposed with various dependence structures on the parent random vector, using QQ-plots based on geometrical quantiles.

Lundberg-Aumann-Serrano dangerousness risk index

Etienne Marceau*[1]

[1] École d'actuariat, Université Laval, Québec (Québec) – Canada

Consider an entity exposed to a risk whose temporal evolution is modeled using a random process in discrete time. For example, in actuarial mathematics, this entity is an insurance company and the risk corresponds to the overall net losses of a portfolio of this company. Effective management requires appropriate risk measurement with respect to exposure to that risk. A risk measure aims to achieve one or more of the following objectives: (1) describe the risk of an event as precisely as possible; (2) communicate the level of risk; (3) making comparisons between risks; (4) identify the most serious risk; (5) reflect the evolution of risk over time; (6) measure/evaluate the effectiveness of strategies to reduce/prevent risk; (7) recommend and make decisions based on a given level of risk. In the classical risk model (in actuarial mathematics; proposed by De Finetti), the risk process is a random walk with negative drift and with independent and identically distributed increments. First, we introduce the Lundberg-Aumann-Serrano risk index as a functional of the risk process in discrete time and taking values in positive reals. The Lundberg-Aumann-Serrano

risk index is defined as the multiplicative inverse of the Lundberg adjustment coefficient. This coefficient is found in several applications of applied probabilities. Via this risk index, we aim to quantify and analyze the risk exposure of the entity in question from a perspective that is at the border of the theory of risk measures and the theory of ruin. It is demonstrated that the Lundberg-Aumann-Serrano risk index satisfies desirable properties for a risk measure with respect to the mentioned objectives, such as monotonicity, homogeneity, convexity and sub-additivity. We also prove that the risk index has the properties of being consistent with respect to orders in stochastic and convex increasing dominance for the processes. We discuss the application of the Lundberg-Aumann-Serrano risk index for risk processes defined from random walks whose increments can be dependent. We demonstrate the impact of this temporal dependence on the Lundberg-Aumann-Serrano index. We use Euler's principle to address the issue of calculating the contributions of model components to the Lundberg-Aumann-Serrano index. We conclude the presentation with numerical examples to illustrate the applicability of the Lundberg-Aumann-Serrano risk index.

Optimal pooling and distributed inference for the tail index and extreme quantiles

Gilles Stupfler*[1], Abdelaati Daouia[2], Simone A. Padoan[3]

[1] Centre de Recherche en Économie et STatistique (CREST) – ENSAI, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France
[2] Toulouse School of Economics – Université Toulouse I (UT1) Capitole, Université Toulouse I [UT1] Capitole – France
[3] Bocconi Institute for Data Science and Analytics – Italie

Ce travail, motivé par des problématiques d'estimation de risque provenant de plusieurs sources distinctes, se concentre sur des méthodes de pooling (ou groupement) pour l'estimation de paramètres extrêmes d'une loi à queue lourde. On construit et étudie une classe d'estimateurs de Hill groupés de l'indice de valeurs extrêmes, et une classe d'estimateurs de Weissman groupés calculés par moyenne géométrique. On montre la convergence de ces estimateurs lorsque les distributions au sein des échantillons peuvent présenter de l'hétérogénéité, et en présence de dépendance entre les échantillons. On obtient des estimateurs optimaux du point de vue de la variance ou de l'erreur quadratique, dont on étudie les propriétés asymptotiques en comparaison à celles de l'estimateur de Hill sur l'échantillon fusionné dans un cadre d'inférence distribuée. On considère diverses extensions, comme le cas où le nombre d'échantillons augmente, ou la présence de covariables et/ou de dépendance temporelle. Deux applications sur des données de pluie et d'assurance illustrent le comportement des estimateurs en pratique.

Clustering parcimonieux pour extrêmes multivariés

Olivier Wintenberger*[1], Nicolas Meyer[2]

[1] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France
[2] IMAG – Université Montpellier II - Sciences et Techniques du Languedoc – France

Résumé.} Identifier les directions dans lesquelles des événements exceptionnels apparaissent est un des problèmes majeurs de la théorie multivariée des valeurs extrêmes. D'un point de vue théorique, l'information concernant de tels événements est contenue dans la mesure spectrale, qui apparaît comme la limite de la composante angulaire de vecteurs aléatoires à variation régulière. Estimer cette mesure s'avère être un point délicat, notamment en grande dimension. Dans cette présentation, nous introduisons une méthode de réduction de la dimension fondée sur la projection euclidienne sur le simplexe. Cette projection a été étudiée dans le cadre des valeurs extrêmes par Meyer et Wintenberger (2021) qui ont établi plusieurs résultats théoriques. La présentation s'attachera à exposer une approche statistique fondée sur de la sélection de modèle qui permet d'identifier les groupes de coordonnées susceptibles d'être extrêmes simultanément. Cette approche donne lieu à un algorithme appelé MUSCLE pour MUltivariate Sparse Clustering for Extremes.

Applications

(Amphi 9 - 10h20-11h40)

Votes par évaluation avec des fonctions de profondeur

Jean-Baptiste Aubin[*1,2], Irène Gannaz[1], Samuela Leoni[1], Antoine Rolland[3]

[1] Institut Camille Jordan – Ecole Centrale de Lyon, Université de Lyon, Université Claude Bernard Lyon 1, Institut National des Sciences Appliquées de Lyon, Institut National des Sciences Appliquées : UMR5208, Université Jean Monnet [Saint-Etienne], Centre National de la Recherche Scientifique – France
[2] Déchets Eaux Environnement Pollutions – Institut National des Sciences Appliquées de Lyon : EA7429, Université de Lyon, Institut National des Sciences Appliquées – France
[3] Laboratoire ERIC – France

Cet article propose un cadre unifié pour les procédures de vote basées sur une évaluation des candidats. L'idée est de représenter les notes de chaque électeur sur d candidats comme un point dans $\mathbb{R}^d$ et de définir le vainqueur du vote en utilisant le point le plus profond du nuage de points ainsi formé. Le point le plus profond est obtenu par la maximisation d'une fonction de profondeur. Il est démontré que les principales procédures par évaluation (jugement majoritaire, vote par approbation...) rentrent dans ce cadre. De plus, les propriétés des procédures obtenues étudiées sont l'universalité, l'unanimité, la neutralité, la monotonie et l'indépendance aux alternatives non pertinentes (IIA).

Impact of the energy transition on long-term factor productivity

Guillaume Flament *[1]

[1] CREST – ENSAI, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Dans cet article, nous nous appuyons sur le modèlde DICE où nous introduisons une nouvelle quantité, communément appelée exergy. Les données montrent que la variation de l'exergy primaire est un bon prédicteur de la variation de la productivité des facteurs. Cette observation nous permet de concevoir de nouveaux scénarios économiques sous condition de respect des accords de Paris. Ces scénarios sont particulièrement intéressants pour les futurs tests de résistance climatique organisés par le régulateur. De plus, notre modèle est capable de conciler des projections très différentes, d'une part, si l'exergy augmente, alors nous obtenons des scénarios similaires à ceux de Nordhaus, à l'inverse, si elle diminue, alors nous obtenons des scénarios similaires à ceux de Meadows.

Pl@ntNet-300K: une base de données d'images de plantes avec une ambiguïté entre classes importante et une distribution à longue traîne

Camille Garcin *, Maximilien Servajean, Alexis Joly [1], Joseph Salmon [2]

[1] INRIA (INRIA) – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – Montpellier, France
[2] Institut Montpelliérain Alexander Grothendieck (IMAG) – CNRS, Université de Montpellier – France

Cet article présente un nouveau jeu de données d'images à forte ambiguïté intrinsèque et à distribution à longue traîne, construit à partir de la base de données de l'observatoire citoyen Pl@ntNet. Il est composé de 306146 images de plantes couvrant 1081 espèces. Nous soulignons deux caractéristiques particulières du jeu de données, inhérentes à la manière dont les images sont acquises et à la diversité intrinsèque de la morphologie des plantes :
(i) le jeu de données présente un fort déséquilibre entre classes, c'est-à-dire que quelques espèces seulement représentent la plupart des images, et,
(ii) de nombreuses espèces sont visuellement similaires, rendant l'identification difficile même pour un œil expert.

Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score

Valerie Gares*[1], Guillaume Chauvet[2], David Hajage[3]

[1] Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France – INSA Rennes, Univ Rennes, IRMAR – UMR CNRS 6625 – France
[2] Univ Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France – Ecole Nationale de la Statistique et de l'Analyse de l'Information [Bruz], Univ Rennes, IRMAR – UMR CNRS 6625 – France
[3] Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié-Salpêtrière, Département de Santé Publique, Centre de Pharmacoépidémiologie, Paris, France – Université Paris-Sorbonne - Paris IV, Centre de Recherche Inserm, Institut Pierre Louis d'Epidémiologie et de Santé Publique, CHU Pitié-Salpêtrière [AP-HP] : Département de Santé Publique, Centre de Pharmacoépidémiologie – France

Les méthodes utilisant le score de propension sont largement utilisées dans les études observationnelles pour évaluer les effets marginaux du traitement. Le score de propension généralisée (GPS) est une extension du score de propension, historiquement développé dans le cas des expositions binaires, pour des expositions quantitatives ou continues. Nous proposons des estimateurs de variance pour les estimateurs de l'effet du traitement sur des variables d'intéret continues. Les fonctions dose-réponse (DRF) ont été estimées en pondérant par l'inverse du GPS ou en utilisant la stratification. Les estimateurs de variance ont été évalués à l'aide de simulations Monte Carlo. Malgré l'utilisation de poids stabilisés, la variabilité de l'estimateur pondéré du DRF était particulièrement élevée, et aucun des estimateurs de la variance n'a pu capturer adéquatement cette variabilité, en particulier lorsque la proportion de la variabilité de l'exposition quantitative expliquée par les covariables était importante. L'estimateur stratifié était plus stable et les estimateurs de la variance plus efficaces pour saisir la variabilité empirique des paramètres du DRF. L'estimateur de la variance linéarisé groupé et l'estimateur groupé basé sur le modèle avaient tendance à surestimer la variance, tandis que l'estimateur bootstrap, qui prend intrinsèquement en compte l'étape d'estimation du GPS, a donné des estimations de la variance corrects.

Topological period counting method for reparametrized functions

Wojciech Reise*[2,1]

[2] LMO – Université de Paris-Sud Orsay – France
[1] INRIA-Saclay – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France

On considère un signal constitué de plusieurs périodes d'une fonction périodique. On observe une reparamétrisation bruitée de ce signal et on cherche à retrouver le nombre de périodes observées. C'est une version plus facile du problème d'estimation de phase, traitée souvent quand la fonction est simple ou connue. On s'intéresse ici à des estimateurs basés sur la quantification de la forme du signal et au cas où la fonction reparamétrisée est inconnue. On étudie l'homologie persistante

du signal qui synthétise l'évolution de ses sous–niveaux et on propose des estimateurs du nombre de périodes. Ceux-ci ne nécessitent pas la connaissance de cette fonction et sont adaptés à une classe de signaux génériques.

Extrêmes & Prediction

(Amphi 10 - 10h20-11h40)

Extremal cluster inference for heavy-tailed time series

Gloria Buritica *[1], Thomas Mikosch [2], Olivier Wintenberger [1]

[1] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France
[2] Univerisity of Copenhagen – Danemark

We study regularly varying time series. In this setting, extremes can appear as consecutive large observations. We refer to extreme episodes with large lp-norm, as p-clusters. To evaluate the clustering effect, we consider functionals acting on p-clusters.
A summary statistic that we can recover in this manner is the extremal index. Its reciprocate is interpreted as the mean simultaneous large values in one cluster. We review the p-cluster inference setting based on extremal lp–blocks. We compare two estimators of the extremal index, based on extremal lα,l∞–blocks, respectively, where $\alpha > 0$ is the tail index of the series. Numerical experiments from the auto-regressive model of order one show the approach with p=α is more robust to handle time dependencies than the p=∞ strategy.

Étude de la dépendance spatiale dans les extrêmes : un point de vue géostatistique

Marine Demangeot *[1], Emilie Chautru [2], Anne Sabourin [3]

[1] Sorbonne Université, LPSM – Sorbonne Universités, UPMC, CNRS – France
[2] Mines ParisTech, centre de Géosciences, équipe Géostatistique – MINES ParisTech, PSL Research University – France
[3] Télécom Paris, centre IDS, équipe S2A – Télécom Paris – France

La fonction coefficient extrémal est une mesure bivariée de dépendance spatiale pour les processus stationnaires max-stables, cf. Schlather et Tawn (2003). Elle est généralement estimée à partir de données temporelles, lorsque le phénomène spatial étudié est observé plusieurs fois dans le temps (e.g. des pluies ou des températures extrêmes, de fortes concentrations de polluants dans l'air). Parfois, nous n'avons pas accès à de telles données : seulement un ou quelques enregistrements sont disponibles. C'est le cas, par exemple, des études sur l'estimation des ressources minières ou sur l'évaluation de la pollution des sols et plus généralement de toute recherche dont l'objet d'étude varie très peu au cours du temps ou pour lequel le coût d'échantillonnage est trop élevé. Ce cas de figure est très peu abordé par la communauté des extrêmes. Au contraire, c'est un cadre d'analyse auquel la Géostatistique s'intéresse particulièrement. Un des outils fondamentaux de cette discipline est le variogramme qui est aussi une mesure bivariée de dépendance spatiale. En considérant le variogramme des indicatrices, au dessus d'un certain seuil, d'un processus stationnaire max-stable, nous proposons un nouvel estimateur non paramétrique de la fonction coefficient extrémal basé sur l'estimateur de type Nadaraya-Watson de ce variogramme. Ce dernier a notamment été étudié par Garcia-Soidan et al. (2004) et Garcia-Soidan (2007) ; à partir de leurs travaux, nous établissons les propriétés asymptotiques de notre estimateur lorsqu'il est calculé à partir d'un unique jeu de données spatialisées. En particulier, sous certains conditions, nous montrons qu'il est consistant et asymptotiquement normal. Ces résultats sont illustrés sur des simulations numériques puis sur des données de précipitations maximales. Nous comparons également les performances de notre estimateur avec celle de l'estimateur basé sur le F-madogramme proposé par Cooley et al. (2006).

M. Schlather and J.A. Tawn, A dependence measure for multivariate and spatial extreme values: Properties and inference, Biometrika, 2003

J.-P. Chilès and P. Delfiner, Geostatistics: Modeling Spatial Uncertainty, Wiley series in probability and statistics, second edition, 2012

P.H. Garcia-Soidan, M. Febrero-Bande and W. Gonzalez-Manteiga, Nonparametric kernel estimation of an isotropic variogram, Journal of Statistical Planning and Inference, 2004

P.H. Garcia-Soidan, Asymptotic normality of the Nadaraya-Watson semivariogram estimators, TEST, 2007

D. Cooley, P. Naveau and P._~Poncet Variograms for spatial max-stable random fields, pages 370-390, Springer New-York, 2006

Maximum Mean Discrepancy invariant par translation et convergence en loi

Thibault Modeste*[1], Clément Dombry[2], Anne-Laure Fougères[1]

[1] Université Claude Bernard – Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-6962 Villeurbanne Cedex, France – France

[2] Université Bourgogne Franche-Comté – Laboratoire de Mathématiques de Besançon, Laboratoire de Mathématiques de Besançon – France

In this talk, we will focus on the relations between RKHS and probability measures on Rd. We will be interested in kernels with a translation invariant Maximum Mean Discrepancy (MMD) which allows unbounded kernels. First, we will look at the links between these MMD and weak

convergence. Then we will use the fact that the kernels k are not necessarily bounded to consider a stronger distance, the W1 Wasserstein distance.

Vitesse de Convergence Minimax pour la Régression Distributionnelle basée sur le Continuous Ranked Probability Score

Romain Pic*[1], Clément Dombry[2], Philippe Naveau[3], Maxime Taillardat[4]

[1] Laboratoire de Mathématiques de Besançon – CNRS : UMR6623, UBFC, Univ Bourgogne Franche-Comte, F- 25000 Besancon, France – France
[2] Université de Franche-Comté – Laboratoire de Mathématiques de Besançon – France
[3] Laboratoire des Sciences du Climat et de l'Environnement [Gif-sur-Yvette] – Université de Versailles Saint-Quentin-en-Yvelines : UMR8212, Commissariat à l'énergie atomique et aux énergies alternatives : DRF/LSCE, Université Paris-Saclay, Institut National des Sciences de l'Univers : UMR8212, Centre National de la Recherche Scientifique : UMR8212 – France
[4] Météo France – CNRS : UMR3589 – France

La régression distributionnelle répond à un besoin fondamental de l'analyse statistique : permettre de faire des prévisions tout en quantifiant leur incertitude. Cette approche surmonte les limites de la régression classique qui estime uniquement l'espérance conditionnellement aux covariables en fournissant un estimateur de l'intégralité de la loi conditionnelle. Cette méthodologie, dite de prédiction probabiliste, est largement adoptée dans de nombreux domaines tels que la météorologie et la production d'énergie, mais ses aspects théoriques restent peu développés. Par analogie avec la théorie classique de l'apprentissage statistique, nous définissons un cadre où le prédicteur est une loi de probabilité, dite loi prédictive, et où la fonction de perte est donnée par un score strictement propre au sens de Gneiting et Raftery (2007). Le prédicteur de Bayes coïncide alors avec la loi conditionnelle. Dans le cas du CRPS, nous étudions ensuite la vitesse minimax de convergence et montrons en particulier que l'algorithme des k plus proches voisins atteint le taux minimax optimal en dimension supérieure ou égale à 2 et que les méthodes à noyau réalise ce taux optimal en dimension quelconque.

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Nicklas Werge*[1]

[1] Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Université : UMR8001 – France

Motivated by the high-frequency data streams continuously generated, real-time learning is becoming increasingly important. These data streams should be processed sequentially with the property that the data stream may change over time. In this streaming setting, we propose tech-

niques for minimizing convex objective functions through unbiased estimates of their gradients, commonly referred to as stochastic approximation problems. Our methods rely on stochastic gradient algorithms because of their applicability and computational advantages. The reasoning includes iterate averaging that guarantees optimal statistical efficiency under classical conditions. Our non-asymptotic analysis shows accelerated convergence by selecting the learning rate according to the expected data streams. We show that the average estimate converges optimally and robustly for any data stream rate. In addition, noise reduction can be achieved by processing the data in a specific pattern, which is advantageous for large-scale machine learning problems. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed algorithms.

Survie

(Amphi 11 - 10h20-11h40)

Machine learning for survival data prediction: Application of the super learner on pseudo-observations

Ariane Cwiling*[1], Olivier Bouaziz[1], Vittorio Perduca[1]

[1] MAP5 – Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Centre National de la Recherche Scientifique : UMR8145, Université de Paris : UMR8145 – France

La moyenne restreinte du temps de survie ("restricted mean survival time" ou RMST) est aisément interprétable, ce qui en fait un objet d'étude intéressant en analyse de survie. Sa prédiction par rapport aux caractéristiques d'un patient peut être très utile dans le domaine de la santé. Cependant peu de méthodes traitent de cette question en analyse de survie. Un article récent de Zhao (2021) propose d'appliquer un réseau de neurones profonds sur des pseudo-observations. Ces dernières peuvent être décrites comme une transformation des temps censurés en données pouvant être gérées comme non censurées. Dans ce travail, nous proposons une nouvelle méthode de prédiction pour le RMST basée sur les pseudo-observations et combinée avec le super learner, un algorithme de prédiction qui propose une combinaison pondérée optimale de différents algorithmes d'apprentissage. Nous avons évalué notre modèle à l'aide de simulations approfondies. Because of its simple interpretation, the restricted mean survival time (RMST) is an interesting quantity of interest in survival analysis. Its prediction with regard to the attributes of a patient can be of great interest in health care. However few survival methods exist in practice for this purpose. To achieve this goal, a recent article applied a deep neural network on pseudo-observations, the latter being a transformation of the incomplete observed times into data that can be handled as uncensored (Zhao, 2021). In this work, we propose a new prediction model

for the RMST based on pseudo-observations combined with the super learner, a prediction algorithm which fits a weighted combination of candidate learners. We evaluated our model through extensive simulations.

Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation.

Audrey Lavenu [*3,2,1], Juliette Murris [4,5,6,7], Alexis Mareau [4], Timothé Rouzé [4], Magalie Fromont [9,8], Valerie Gares [9,10], Sandrine Katsahian [4,5,7,11]

[3] Institut de Recherche Mathématique de Rennes – Agrocampus Ouest, Universite de Rennes 1, Université de Rennes 2, École normale supérieure - Rennes, Centre National de la Recherche Scientifique : UMR6625, Institut National des Sciences Appliquées - Rennes – France
[2] Université Rennes 1 – université Rennes 1 – France
[1] Centre dÍnvestigation Clinique [Rennes] – Universite de Rennes 1, Hôpital Pontchaillou, Institut National de la Santé et de la Recherche Médicale : CIC1414 – France
[4] CIC - HEGP – Institut National de la Santé et de la Recherche Médicale : CIC1418, Université de Paris, Hôpital Européen Georges Pompidou [APHP], Assistance publique - Hôpitaux de Paris (AP-HP) – France
[5] Health data- and model- driven Knowledge Acquisition – Inria de Paris, Centre de Recherche des Cordeliers – France
[6] PIERRE FABRE – PIERRE FABRE – France
[7] Université de Paris, Sorbonne Université – France
[9] Institut de Recherche Mathématique de Rennes – Agrocampus Ouest, Universite de Rennes 1, Université de Rennes 2, École normale supérieure - Rennes, Centre National de la Recherche Scientifique : UMR6625, Institut National des Sciences Appliquées - Rennes – France
[8] Université Rennes 2 – Université Rennes 2 - Haute Bretagne, Place du recteur Henri Le Moal, CS 24307, 35043, Rennes cedex - France – France
[10] INSA Rennes – INSA - Institut National des Sciences Appliquées – France
[11] Assistance publique - Hôpitaux de Paris (AP-HP) – France

Avec l'augmentation du nombre de données sur les patients dans les domaines de l'imagerie médicale ou encore de la génomique, les méthodes d'analyses classiques sont souvent inadéquates dans les cas où il y a moins d'observations que de variables. L'objectif de notre travail est d'étudier différents critères de performance et leur estimation de la méthode Cox Boost pour analyser des données de survie en grande dimension sur petits échantillons. Nous nous intéressons à la prédiction et la discrimination des variables pronostiques, mais aussi à la "tunabilité" du modèle Cox Boost (gain par optimisation des hyperparamètres).
Nous simulons les temps de survie avec une loi exponentielle et les temps de censure avec une loi uniforme. Pour fixer le taux de censure à un taux prédéfini, nous montrons comment calculer le paramètre de la distribution de censure.
Avec un schéma de simulation faisant varier: la taille d'effet des covariables, la taille de l'échantillon, et le taux de variables actives, nous comparons différents critères de performance de la méthode (C de Harrell et mesure d'importance de variable) estimés par validation croisée à 2 et 5 blocs,

avec trois méthodes de choix des hyperparamètres. Nous montrons la difficulté d'optimiser les hyperparamètres pour de petits échantillons, et les lacunes des mesures d'importance des variables à détecter les variables simulées actives, même quand la performance du modèle en termes de prédiction est correcte.

Modèle de Cox avec des données hétérogènes

Eliz Peyraud *[1]

[1] Laboratoire ERIC – Université Lumière - Lyon II – France

Dans un contexte médical, l'analyse de survie se fait essentiellement à l'aide de modèles à risques proportionnels de Cox. Le présent papier cherche à construire un tel modèle afin de prédire la probabilité de survie d'un patient à un temps t après un acte chirurgical lourd. Nous mettons en évidence la nécessité de prendre en compte l'hétérogénéité des patients lors de la modélisation et expliquons comment nous pouvons prendre en compte des variables explicatives de natures variées.

Cox regression with linked data

Thanh Huan Vo *[1,2], Guillaume Chauvet [3], André Happe [4], Emmanuel Oger [4], Stéphane Paquelet [1], Valerie Gares [2]

[1] Artificial Intelligence – Institute of Research and Technology b<>com – France
[2] IRMAR – INSA Rennes – France
[3] IRMAR – ENSAI, Ecole Nationale de la Statistique et de l'Analyse de l'Information – France
[4] EA 7449 REPERES – Univ Rennes, EHESP, REPERES (Pharmacology and health services research) - EA 7449, F-35000 Rennes, France – France

Les méthodes d'appariement sont couramment utilisées pour combiner des variables provenant de plusieurs bases de données, mais associées à la même entité. Cependant, l'appariement est rarement parfait, et ne pas prendre en compte les erreurs afférentes peut conduire à des estimations biaisées. Plusieurs méthodes ont été proposées pour traiter ce problème dans le cas du modèle linéaire généralisé, mais à notre connaissance le modèle de Cox n'a pas été beaucoup étudié dans la littérature alors qu'il soit l'un des modèles statistiques les plus cités. Dans cet article, nous proposons une équation estimante ajustée pour tenir compte des erreurs dues à l'appariement. Cette méthode est utilisable pour un utilisateur secondaire de données appariées, quand l'appariement a été réalisé par un opérateur tiers et quand la personne réalisant l'analyse n'a pas d'information sur les variables d'appariement. Les résultats de nos simulations montrent que la méthode proposée permet de réduire significativement le biais d'estimation pour les paramètres du modèle de Cox causé par les erreurs d'appariement.

11h40 – 12h00 Pause

12h00 – 13h10

SFdS - AMIES

(Amphi 7 - 12h00-13h10)

Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations

Van Tuan Nguyen*[1]

[1] University de Paris – LPSM – France

This paper introduces a prognostic method called lights to deal with the problem of joint modeling of longitudinal data and censored durations, where a large number of both longitudinal and time-independent features are available. In the literature,standard joint models are either of type shared random-effect or joint latent class ones; where the association structure between the longitudinal and the time-to-event submodels takes respectively the form of either shared association features learned from the longitudinal processes and included as potential risk factor in the survival model, or latent classes modeling population heterogeneity. We pick modeling ideas from both worlds and use appropriate penalties during inference for being able to learn from a high-dimensional context. The statistical performance of the method is examined on an extensive Monte Carlo simulation study, and finally illustrated on a publicly available dataset. Our proposed method significantly outperforms the state-of-the-art joint models regarding risk prediction in terms of C-index in a so-called real-time prediction paradigm, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant features being relevant from a practical perspective. Thus, we propose a

powerfull tool with the ability of automatically determining significant prognostic longitudinal features, which is of increasing importance in many areas: for instance personalized medicine, or churn prediction in a customer profile and activity monitoring setting, to name but a few.

L'agence pour les mathématiques en interaction avec l'entreprise et la société (AMIES)

Véronique Maume-Deschamps*[1,2]

[1] Institut Camille Jordan [Villeurbanne] – Université Claude Bernard Lyon 1 – France
[2] AMIES – INSMI (CNRS), University of Grenoble Alpes (UGA) – France

Comment AMIES agit au niveau national pour développer les relations avec les entreprises ?

Prédiction des niveaux de risque pollinique à partir de données historiques multi-sources

Esso-Ridah Bleza*[1,2], Valerie Monbet[2], Pierre-François Marteau[1]

[1] IRISA – Université de Bretagne Sud [UBS] – France
[2] IRMAR – Université de Rennes I – France

Dans la littérature scientifique, de nombreuses études montrent que les conditions météorologiques ont un impact sur l'émission, la dispersion et la suspension des pollens dans l'air menaçant la santé des millions de personnes en France. L'objectif est d'étudier la capacité à prédire à 3 jours (J+3) les niveaux de risques de présence de pollens dans l'air sur un territoire donné (en France Métropolitaine) avec des techniques d'apprentissage statistique exploitant des données historiques, et les paramètres météorologiques jusqu'au jour (J). Nous nous sommes intéressés à la prévision de risque pour 3 familles de pollens qui font partie des espèces les plus allergisantes (ambroisie, cupressacées et graminées). L'agrégation de modèles de régression logistiques par un classifieur de type Forêt aléatoire a permis de prédire les niveaux du risque pollinique ('nul', 'faible', 'moyen' et 'fort') avec des performances de l'ordre de 75% à 90% d'AUC et 70% de précision et de rappel, avec quelques confusions pour les niveaux faible et moyen.

Deep Calibration of Interest Rates Models

Djibril Sarr*[1,2]

[1] Université Sorbonne Paris Nord – Laboratoire Analyse, Géométrie et Applications, LAGA, CNRS,

UMR 7539, F-93430, Villetaneuse, France. – France
[2] FBH Associés – FBH Associés – France

For any financial institution it is a necessity to be able to apprehend the behaviour of interest rates. Despite the use of Deep Learning that is growing very fastly, for many reasons (expertise, ease of use, ...) classic rates models such as CIR, or the Gaussian family are still being used widely. We propose to calibrate the five parameters of the G2++ model using Neural Networks. To achieve that, we construct synthetic data sets of parameters drawn uniformly from a reference set of parameters calibrated from the market. From those parameters, we compute Zero-Coupon and Forward rates and their covariances and correlations. Our first model is a Fully Connected Neural network and uses only covariances and correlations. We show that covariances are more suited to the problem than correlations. The second model is a Convolutional Neural Network using only Zero-Coupon rates with no transformation. The methods we propose have performed better than classic calibration methods on a comparison data set, with a global error, 2 times lower and a calibration in less than a second, compared to 164 seconds for the classic method.

ML - Graphes - Réseaux

(Amphi 8 - 12h00-13h10)

The deep latent position model for node partitioning in graphs

Dingge Liang*[1], Marco Corneli[1], Charles Bouveyron[1], Pierre Latouche[2]

[1] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France. – Université Côte d'Azur (UCA) – France
[2] Mathématiques Appliquées Paris 5 – Institut National des Sciences Mathématiques et de leurs Interactions : UMR8145, Centre National de la Recherche Scientifique : UMR8145, Université de Paris : UMR8145 – France

With the significant increase of interactions between individuals in different domains, the clustering of vertex in graphs has become a fundamental approach for analysing large and complex networks. We propose here the deep latent position model (DeepLPM), an end-to-end clustering approach which combines the widely used latent position model (LPM) for network analysis with a graph convolutional network (GCN) encoding strategy. Thus, DeepLPM can automatically assign each node to its group without using any additional algorithms and better preserves the network topology. To demonstrate its effectiveness in performing clustering task, an application

on the ecclesiastical network in Merovingian Gaul is conducted. Moreover, we inspect the capability of DeepLPM to automatically select the number of clusters, without using external model selection tools.

Etude de la Pénalisation GraphNet en Analyse de Données Multiblocs

Annabelle Beaudoin *[1], Natalia Pietrosemoli [1], Cathy Philippe [2], Hervé Abdi [3], Vincent Guillemot [1]

[1] Hub de Bioinformatique et Biostatistique – - : InstitutPasteur, - – France
[2] Unité Baobab – Service NEUROSPIN : DRF/JOLIOT/NEUROSPIN – France
[3] University of Texas at Dallas [Richardson] – États-Unis

L'intégration de données multiblocs est maintenant incontournable pour analyser des données complexes allant, par exemple, des données multi-omiques, aux données d'imagerie génétique. Par ailleurs, les bases de données biologiques de référence contiennent maintenant une information très riche qu'il convient d'intégrer dans de telles analyses.
Nous proposons d'explorer la méthode netSGCCA qui permet l'intégration de réseaux dans le cadre de l'Analyse des Corrélations Canonique Généralisée pénalisée à l'aide d'une pénalité GraphNet. Plus particulièrement, nous souhaitons mettre en lumière un des désavantages de cette pénalité, qui est d'introduire des composantes "haute-fréquence".
L'exemple que nous étudions est issu d'une étude clinique sur la Spondylarthrite ankylosante et comprend trois blocs : deux blocs de données d'expression, et un bloc de données cliniques. Le réseau de référence que nous utilisons est extrait de la base de données STRING-DB. Nous montrons sur cet exemple un moyen de ne conserver que les éléments "basse fréquence" induits par l'introduction de la pénalité GraphNet.

Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing

Hugo Schmutz *[1,2], Olivier Humbert [3], Pierre-Alexandre Mattei [4]

[1] Laboratoire Jean Alexandre Dieudonné – Université Nice Sophia Antipolis, Centre National de la Recherche Scientifique – France
[2] UMR E4320 – Université Côte d'Azur, Université Nice Sophia Antipolis (... - 2019), COMUE Université Côte d'Azur (2015 - 2019), Commissariat à l'énergie atomique et aux énergies alternatives : DRF/JOLIOT, COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019) – Faculté de Médecine, 28 avenue de Valombrose 06107 Nice Cedex 2, France
[3] UMR E4320 – Université Côte d'Azur, Université Nice Sophia Antipolis (... - 2019), COMUE Uni-

versité Côte d'Azur (2015 - 2019), Commissariat à l'énergie atomique et aux énergies alternatives : DRF/JOLIOT, COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019), COMUE Université Côte d'Azur (2015 - 2019) – Faculté de Médecine, 28 avenue de Valombrose 06107 Nice Cedex 2, France
[4] Inria Sophia Antipolis – Inria Sophia Antipolis (MAASAI) – France

L'apprentissage semi-supervisé (SSL) offre un moyen efficace d'exploiter les données non étiquetées pour améliorer les performances d'un modèle. Néanmoins, la plupart des méthodes présentent l'inconvénient commun de manquer de securité. Par sécurité, nous entendons la qualité de ne pas dégrader un modèle entièrement supervisé lors de l'inclusion de données non étiquetées. Notre idée a été de remarquer que l'estimateur du risque que la plupart des méthodes SSL discriminatives minimisent est biaisé, même asymptotiquement. Ce biais rend ces techniques non fiables sans un jeu de validation approprié, mais nous proposons un moyen simple de supprimer ce biais. Notre approche de débiasage est simple à mettre en œuvre et applicable à la plupart des méthodes SSL profondes. Nous fournissons des garanties théoriques simples sur la sécurité de ces méthodes modifiées, sans avoir à s'appuyer sur les hypothèses fortes sur la distribution des données que le SSL exige habituellement. Nous montrons expérimentalement que le débiaisage peut rivaliser avec les techniques SSL profondes classiques dans divers contextes classiques et qu'il donne même de bons résultats lorsque le SSL traditionnel échoue.

Generalised Mutual Information Maximisation for Deep Clustering

Louis Ohl[*1,2], Pierre-Alexandre Mattei[3,2], Warith Harchaoui[4], Frederic Precioso[2,5], Charles Bouveyron[2,6]

[1] Laboratoire d'Informatique Signaux et Systèmes de Sophia-Antpolis – France
[2] Equipe Maasai, Inria Sophia Antipolis – Institut National de Recherche en Informatique et en Automatique – France
[3] Laboratoire J.-A. Dieudonné, UMR CNRS 7531, Université Côte d'Azur – Université Nice Sophia Antipolis, Centre National de la Recherche Scientifique, 3iA Côte d'Azur – France
[4] DERAISON.ai – Chercheur indépendant – France
[5] Laboratoire d'Informatique Signaux et Systèmes de Sophia-Antpolis – France
[6] Laboratoire J.-A. Dieudonné, UMR CNRS 7531, Université Côte d'Azur – CNRS : UMR7531 – France

Au cours des dernières années, le problème de partionnement des données a été abordé avec succès par les méthodes d'apprentissage profond. Cependant, la plupart de ces méthodes font intervenir la maximisation de l'information mutuelle comme objectif d'entrainement et font appel à des techniques de régularisations sophistiquées pour corriger son instabilité plutôt que de chercher à l'améliorer. Nous prenons un autre point de vue sur l'information mutuelle et la considérons comme une distance moyenne entre les distributions des classes. Nous étendons sa définition pour inclure d'autres distances que celle de Kullback-Leibler entre les distributions: l'information mutuelle généralisée (GEMINI). Nos expériences soulignent qu'un choix judicieux de GEMINI peu améliorer le partitionnement lorsque l'information mutuelle est utilisée comme objectif tout en étant un outil efficace pour trouver un bon nombre de partitions.

Méta-modèles – Analyse de sensibilité

(Amphi 9 - 12h00-13h30)

Heterogeneous Treatment Effects Estimation: When Machine Learning meets multiple treatments regime

Naoufal Acharki*[1], Josselin Garnier[2], Antoine Bertoncello, Lugo Ramiro

[1] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[2] École Polytechnique – CMAP, École Polytechnique – France

Résumé. Dans beaucoup de domaines scientifiques et techniques, l'inférence de l'effet d'un traitement et l'exploration de son hétérogénéité sont déterminantes pour l'optimisation et la prise de décision. De nombreux méta-algorithmes ont été développés pour estimer la fonction d'effet moyen conditionnel du traitement (CATE) pour un traitement binaire, leur principal avantage est de ne pas restreindre l'estimation à une méthode d'apprentissage supervisée particulière. Dans ce travail, nous étudions le régime de traitement multiple sous le modèle causal de Rubin et nous nous focalisons sur l'estimation des effets hétérogènes de traitement. Nous généralisons les meta-algorithmes pour l'estimation de la fonction CATE pour chaque niveau possible de traitement. Nous évaluons la qualité de chaque méta-algorithme sur des données observationnelles en utilisant un jeu de données Semi-synthétique et nous soulignons en particulier les performances du X-learner.
Mots clés. Apprentissage Automatique; Inférence Causale; Traitement multiple; Effets hétérogènes.

Abstract. In many scientific and engineering domains, inferring the effect of treatment and exploring its heterogeneity is crucial for optimization and decision making. Several meta-algorithms have been developed to estimate the Conditional Average Treatment Effect (CATE) function in the binary setting, with the main advantage of not restraining the estimation to a specific supervised learning method. In this work, we investigate the multiple treatment regime under Rubin Causal Model and we focus on estimating heterogeneous treatment effects. We generalize Meta-learning algorithms to estimate the CATE for each treatment value. Using semi-synthetic simulation datasets, we assess the quality of each meta-learner in observational data and we highlight in particular the performances of the X-learner
Keywords. Machine Learning; Causal Inference; Multiple Treatments; Heterogeneous Effects.

Risk bounds for Sobol'-based Global Sensitivity Analysis Indices when using metamodels.

Clément Bénesse*[1]

[1] Institut de Mathématiques de Toulouse – Université Paul Sabatier - Toulouse III – France

Les métamodèles sont de plus en plus utilisé comme un moyen d'obtenir à moindre coût une approximation d'algorithmes complexes et gourmands, dans le but d'en extraire le maximum d'informations sur les caractéristiques du système. Parmi celles-ci, les indices que l'on retrouve en Analyse de Sensibilité Globale (GSA) permettent de connaître l'influence des variables d'entrée sur la sortie d'une boîte noire. Nous nous intéressons ici aux déviations d'une classe de ces indices lors de l'utilisation d'un métamodèle afin de pouvoir les borner.

Test d'indépendance basé sur les indices HSIC-ANOVA d'ordre total

Gabriel Sarazin*[1], Amandine Marrel[1,2], Sébastien Da Veiga[3], Vincent Chabridon[4]

[1] Commissariat à l'énergie atomique et aux énergies alternatives – Centre de recherche du Commissariat à l'Energie Atomique - CEA Cadarache (Saint Paul-lez-Durance, France) – France
[2] Institut de Mathématiques de Toulouse – Institut de Mathématiques de Toulouse – France
[3] Safran Tech - Modeling and Simulation – SAFRAN (FRANCE) – France
[4] EDF R&D – EDF Recherche et Développement – France

L'apprentissage statistique dans le cas de données simulées par un code de calcul industriel, aussi appelé "métamodélisation", est une tâche dont la difficulté de mise en oeuvre croît avec la dimension du problème et le manque de données d'apprentissage. Une analyse de sensibilité préliminaire peut venir en soutien de la construction du métamodèle pour éliminer les variables les moins pertinentes et trier les variables restantes par ordre d'influence sur la sortie. Pour mener une analyse de sensibilité, l'approche historique de Sobol' offre un cadre conceptuel confortable qui est articulé autour de la décomposition de la variance de la sortie. Toutefois, l'estimation précise des indices associés n'est plus possible si l'échantillon d'apprentissage est de petite taille. Pour contourner cette difficulté, il est désormais fréquent d'utiliser une mesure de sensibilité basée sur le critère d'indépendance de Hilbert-Schmidt (notée HSIC). Elle est appliquée à chaque couple entrée-sortie, et permet ainsi de définir la collection des indices HSIC. Leur interprétation est généralement moins intuitive que celle des indices de Sobol car leur construction repose sur la théorie des espaces de Hilbert à noyaux reproduisants. Face à ce constat, les indices HSIC-ANOVA ont été récemment introduits et permettent une séparation des effets principaux et des interactions, à l'instar de la décomposition de Hoeffding dans le formalisme des indices de Sobol'. Cette avancée a été obtenue au prix d'une hypothèse d'indépendance mutuelle des entrées et sous réserve de l'utilisation de noyaux spécifiques, comme les noyaux de Sobolev. Dans ce travail, on commence par montrer que tout noyau de Sobolev est caractéristique, c'est-à-dire

que la nullité d'un indice HSIC-ANOVA est équivalente à une situation d'indépendance au sein du couple formé par l'entrée considérée et la sortie. Dans un second temps, il est montré qu'un test d'indépendance peut être construit pour l'indice HSIC-ANOVA d'ordre total en s'inspirant de ce qui est fait pour l'indice HSIC traditionnel. Enfin, une étude numérique révèle empiriquement que le nouveau test d'indépendance est au moins aussi puissant que celui basé sur l'indice HSIC traditionnel, ce qui offre des perspectives intéressantes pour améliorer le processus de sélection des variables à impliquer dans la construction d'un métamodèle.

Sélection de variables lors de la généralisation d'un effet causal

Bénédicte Colnet*[1], Julie Josse, Scornet Erwan[2], Gaël Varoquaux[3,4]

[1] Inria Saclay - Ile de France – Institut National de Recherche en Informatique et en Automatique – France
[2] Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[3] Parietal – INRIA – France
[4] Unicog – Inserm : U992 – France

Les essais contrôlés randomisés (ECR) sont considérés comme l'étalon-or pour conclure à l'effet causal d'une intervention donnée. Pour autant, ils présentent des limites quand la population éligible à l'essai est significativement différente de la population cible.Disposer d'un échantillon de la population cible d'intérêt permet de généraliser l'effet causal.Ce processus nécessite des covariables dans les deux jeux de données, en particulier toute variable qui serait un modulateur de l'effet du traitement ou dont la distribution change d'une population à l'autre.Dans ce travail, nous proposons d'étudier l'impact des covariables manquantes ou partiellement manquantes sur l'estimation, en quantifiant le biais sous une hypothèse semi-linéaire. Notre travail complète également les preuves de consistance pour les trois estimateurs usuels basés sur la repondération (IPSW), la modélisation de la réponse (G-formula), ou la combinaison des deux dans des approches dites doublement robustes (AIPSW). Ce résultat fournit une analyse de sensibilité pour les praticiens. Inversement, nous avons également étudié l'impact de l'ajout d'un trop grand nombre de covariables dans l'estimation, en particulier sur la variance asymptotique.

Calage conditionnel bayésien d'un modèle numérique

Oumar Baldé*[1], Guillaume Damblin[1], Loic Giraldi[2], Antoine Bouloré[3], Amandine Marrel[4,3]

[1] Centre CEA de Saclay DES/ISAS/DM2S/STMF/LGLS – CEA – France
[2] CEA Centre de Cadarache DES/IRESNE/DEC/SESC/LSC – DES – France
[3] CEA, DES, IRESNE, DEC, SESC, LSC – Centre de recherche du Commissariat à l'Energie Atomique - CEA Cadarache (Saint Paul-lez-Durance, France) – France

[4] Institut de Mathématiques de Toulouse – Université Toulouse III - Paul Sabatier, Université Toulouse III - Paul Sabatier, Université Toulouse III- PaulSabatier – France

Notre travail porte plus précisément sur le calage bayésien conditionnel. Il s'agit de réaliser le calage d'un vecteur de paramètres d'entrée $\theta(\lambda) \in$ R p (p ≥ 1) fonction d'une covariable incertaine ($\lambda \in$ R D avec D ≥ 1), à partir des données expérimentales disponibles sous l'hypothèse que ces données observées ne dépendent pas de la valeur de λ. Pour cela, nous proposons un formalisme bayésien hiérarchique du problème, en supposant que le modèle est linéaire en $\theta(\lambda)$. L'approche est hiérarchique au sens où chaque composante de $\theta(\lambda)$ est supposée être la trajectoire indépendante d'un processus gaussien a priori dont les hyperparamètres sont estimés par maximisation de la vraisemblance marginale des données observées. À court terme, cette approche sera appliquée au calage des paramètres d'un modèle de gaz de fission en fonction de la conductivité ($\lambda \in$ R) du combustible nucléaire.

ML & Regret

(Amphi 10 - 12h00-13h30)

Multitask Online Mirror Descent

Pierre Laforgue*[1], Nicolò Cesa-Bianchi[1], Andrea Paudice[1,2], Massimiliano Pontil[2,3]

[1] Università degli Studi di Milano – Italie
[2] Istituto Italiano di Tecnologia – Italie
[3] University College, London – Royaume-Uni

Nous introduisons et analysons MT-OMD, une extension multitâche d'*Online Mirror Descent* (OMD) basée sur l'échange d'informations entre les différentes tâches. Nous prouvons que le regret de MT-OMD est de l'ordre de $\sqrt{1+\sigma^2(N-1)}\sqrt{T}$, où σ^2 est la variance des tâches vis-à-vis de la géométrie induite par le régulariseur, N est le nombre de tâches, et T est l'horizon. Dès que les tâches sont similaires, c'est-à-dire quand $\sigma^2 \leq 1$, notre méthode présente un regret inférieur à $\sqrt{NT}$ obtenu en appliquant OMD indépendamment sur chaque tâche. Nous prouvons enfin que l'amélioration permise par MT-OMD est optimale, et appuyons nos résultats théoriques par une analyse empirique.

Identification de Meilleur Bras à Budget Fixé

Antoine Barrier[*1,2], Aurélien Garivier[2], Gilles Stoltz[1]

[1] Laboratoire de Mathématiques d'Orsay – Université Paris-Saclay, Centre National de la Recherche Scientifique : UMR8628 – France
[2] Unité de Mathématiques Pures et Appliquées – Centre National de la Recherche Scientifique : UMR5669, École Normale Supérieure - Lyon – France

Le problème d'identification de meilleur bras consiste à trouver, parmi un ensemble de distributions, celle de plus grande espérance en observant de manière séquentielle des réalisations indépendantes de ces lois. Les applications (essais cliniques, systèmes de recommandation, ...) en font un cadre d'étude dense pour lequel plusieurs objectifs sont considérés. Dans le cadre du budget fixé, le nombre d'observations est limité et on s'intéresse aux propriétés d'algorithmes qui minimisent la probabilité d'identifier une mauvaise distribution (probabilité d'erreur). Dans ce travail on fait apparaître une notion de complexité à base de théorie de l'information, complétant les travaux préalables qui ne faisaient intervenir que les écarts entre les moyennes des distributions.

On Best Policy Identification with Instance-specific sample complexity

Aymen Al Marjani[*1]

[1] Unité de Mathématiques Pures et Appliquées – École Normale Supérieure - Lyon – France

Le problème de l'identification de la meilleure politique (BPI) a récemment suscité l'intérêt de la communauté de l'apprentissage par renforcement (RL). Formellement, un agent de RL interagit avec un processus de décision Markovien (PDM) et vise à apprendre une politique ϵ-optimale avec une grande confiance en utilisant le moins d'observations possible. Dans le cas particulier d'un PDM avec un facteur d'actualisation $\gamma = 0$ (ou de manière équivalente lorsque l'horizon $H = 1$), le problème se réduit à l'identification du meilleur bras (BAI) dans un bandit multi-armé. Inspirés par le succès des algorithmes du type Track-and-Stop dans ce cas particulier, nous proposons des algorithmes pour BPI avec des bornes spécifiques à l'instance sur leur complexité d'échantillon. Notamment, nos algorithmes sont optimaux pour l'instance dès lors qu'un oracle résolvant un certain problème d'optimisation est disponible.

Bayesian Actor for Deep Reinforcement Learning

Léo Grill[*2,1]

[2] Orange Labs [Lannion] – France Télécom – France
[1] SISMI Poitiers – Université de Poitiers, UMR 7348 du CNRS – France

L'apprentissage renforcé s'est beaucoup développé, et trouve de nombreuses applications. En parallèle les méthodes bayésiennes sont de plus en plus étudiées, notamment pour l'apprentissage profond. Ces méthodes permettent de palier à des problèmes tels que le surapprentissage et la transparence des modèles. Dans ce papier nous étudions comment un acteur bayésien peut s'appliquer aux méthodes d'acteur-critique de l'apprentissage par renforcement profond. Une étude numérique est présentée, elle montre l'intérêt de l'utilisation de ces méthodes dans l'apprentissage renforcé profond.

Apprentissage robuste pour la résolution d'EDPs, Applications aux marchés de l'énergie

Carl Remlinger *[1,2]

[1] EDF – EDF Lab, Saclay – France
[2] Université Gustave Eiffel – Laboratoire d'analyse et de mathématiques appliquées (LAMA) – France

Un modèle résolvant une famille d'équations aux dérivées partielles (EDPs) avec un seul apprentissage est proposé. Pour cela, nous considérons des réseaux d'opérateurs profonds afin d'approcher avec précision des opérateurs continus non linéaires. Nous voulons résoudre des EDPs lorsque l'environnement n'est pas stationnaire ou pour plusieurs conditions initiales à la fois. Notre modèle apprend la solution générale associée à chaque fonction paramètre simultanément. Mais, en fin de compte, nous voulons généraliser la résolution avec des modèles sous-jacents ou des conditions qui n'étaient pas présents lors de l'entraînement. Nous confirmons l'efficacité de la méthode avec plusieurs problèmes de gestion des risques.
Dans ce cas, re-calibrer un modèle de facteurs de risque ou ré-entraîner un modèle pour la résolution chaque fois que les conditions de marché changent est coûteux et insatisfaisant. Nous évaluons notre DeepOHedger pour la couverture d'options, incluant des modèles de volatilité locale et des options spread impliquées dans les marchés de l'énergie.

Statistique mathématique

(Amphi 11 - 12h00-13h30)

Estimation confidentielle de quantiles en présence d'atomes

Clément Lalanne*[1], Clément Gastaud, Nicolas Grislain, Aurélien Garivier, Rémi Gribonval

[1] LIP ENS Lyon – ENS Lyon, Sciences po Lyon, CNRS, Université Lyon 2, Université Jean Monnet – France

Nous abordons l'estimation sous confidentialité différentielle de plusieurs quantiles (MQ) d'un ensemble, un élément clé de l'analyse moderne des données. Nous appliquons le mécanisme récent de sensibilité inverse (SI) non lissé à ce problème spécifique et établissons que la méthode qui en résulte est étroitement liée à l'état de l'art actuel, l'algorithme JointExp, partageant en particulier la même complexité algorithmique et une efficacité similaire. Cependant, nous démontrons à la fois théoriquement et empiriquement que JointExp (non lissé) souffre d'un manque important de performance dans le cas de distributions à pics, avec un impact potentiellement catastrophique en présence d'atomes. Alors que sa version lissée permettrait de tirer parti des garanties de performance de IS, elle reste un défi ouvert à mettre en œuvre. Pour résoudre ce problème, nous proposons une méthode simple et numériquement efficace appelée Heuristically Smoothed JointExp (HSJointExp), qui est dotée de garanties de performance pour une large classe de distributions et permet d'obtenir des résultats significativement meilleurs sur des ensembles de données problématiques.

Généralisation et estimation de la Q-distribution gaussienne

Oumaima Ben Mrad*[1], Afif Masmoudi[2], Yousri Slaoui[1]

[1] Laboratoire de Mathématiques et applications – Université de Poitiers, UMR 7348 du CNRS – France
[2] Laboratoire de Probabilités et Statistique – Tunisie

Résumé : Notre objectif est de compléter le travail de Diaz et Pariguan (2009) en introduisant la distribution q-Gaussienne Nq(0,σ^2) centrée et de variance σ^2. De plus, les notions d'entropie de Shannon et de quantile seront généralisées. Nous mettons en évidence la relation de convergence de Nq(0,σ^2) vers les distributions Uniforme sur (-σ, σ) et Gaussienne sur la droite réelle lorsque q tend vers 0 et 1 respectivement. En outre, pour chaque valeur de q on obtient une q-Gaussienne donc un nouveau paramètre qui s'ajoute avec la moyenne et la variance ce qui implique plus de flexibilité. Enfin, une estimation des paramètres q et σ^2 sera effectuée en suivant l'algorithme E-M.
Mots-clés : q-calcul, distribution q-Gaussienne, q-Entropie, q-Quantile.

Abstract: The aim of this paper is to complet the work of Diaz and Pariguan (2009) by constructing the centered q-Gaussian distribution Nq(0,σ^2) with variance σ^2. In addition, the notion of entropy of Shannon and quantile will be generalized. However, we highlight the relation of convergence of Nq(0,σ^2) to the Uniform distribution on (-σ, σ) and to the Gaussian distribution on the real line when q tends to 0 and 1 respectively. Furthermore, we obtain a q-Gaussian for each value of q, implying that a new parameter is added to the mean and variance, implying

more flexibility. Finally, an estimation of the parameters q and σ^2 will be achieved by following the E-M algorithm.
Keywords : q-Calculus, q-Gaussian distribution, q-Entropy, q-Quantile.

Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale

Marouane Il Idrissi [*3,2,1], Nicolas Bousquet [2,3,4], Fabrice Gamboa [1], Bertrand Iooss [2,3,5], Jean-Michel Loubes [1]

[3] Saclay Industrial Lab for Artificial Intelligence Research – THALES, TOTAL FINA ELF, EDF – France
[2] EDF R&D PRISME - Performance, Risque Industriel, Surveillance pour la Maintenance et l'Exploitation – EDF, EDF Recherche et Développement – France
[1] IMT - Institut de Mathématiques de Toulouse UMR5219 – Institut National des Sciences Appliquées - Toulouse, Institut National des Sciences Appliquées, université Toulouse 1 Capitole, Université Toulouse - Jean Jaurès, Université Toulouse III - Paul Sabatier, Université Fédérale Toulouse Midi-Pyrénées, Centre National de la Recherche Scientifique : UMR5219 – France
[4] Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France
[5] Institut de Mathématiques de Toulouse (IMT) – PRES Université de Toulouse, CNRS : UMR5219 – UPS IMT, F-31062 Toulouse Cedex 9, France INSA, F-31077 Toulouse, France UT1, F-31042 Toulouse, France UT2, F-31058 Toulouse, France, France

Motivés par des applications en analyse de robustesse pour la quantification d'incertitude et l'apprentissage statistique, nous nous intéressons au problème de projection d'une mesure de probabilité univariée donnée. L'ensemble de projection est composé de mesures de probabilité respectant certaines contraintes sur leur quantiles. Ce travail explore le cas de la distance de Wasserstein pour des mesures de probabilité portées par $\mathbb{R}$. Ce problème particulier de projection dans des espaces de mesures sous contraintes quantile se réécrit comme un problème de projection dans $L^2((0,1))$ avec contraintes de monotonie et d'interpolation. Ce dernier problème se résout analytiquement. Nous présentons également une régularisation de la solution en nous restreignant aux fonctions polynomiales par morceaux. Finalement, nous illustrons ces résultats sur des données réelles.

Sparse regression to anticipate the declarations of natural disaster in France

Thi Thanh Yen Nguyen [*1], Antoine Chambaz [2], Ecoto Geoffrey [3]

[1] MAP5 – Université Paris Cité : UMRCNRS 8145 – France

[2] MAP5, Université de Paris – CNRS : UMR8145 – France
[3] Caisse Centrale de Réassurance – Caisse Centrale de Réassurance, R – France

Drought events are the second most expensive natural disasters within the legal framework of the natural disasters compensation scheme in France. We develop a new methodology to anticipate which cities will benefit from a declaration of natural disaster, a key step of the national compensation scheme. The methodology hinges on optimal transport theory and an inertial proximal algorithm for nonconvex optimization.
Keywords. Optimal transport, proximal algorithm, Sinkhorn algorithm, natural disasters

Estimation adaptative optimale de fonctions moyennes et covariances irrégulières

Steven Golovkine*[1], Valentin Patilea[2], Nicolas Klutchnikoff[3]

[1] University of Limerick – Irlande
[2] Centre de Recherche en Économie et Statistique (CREST) – INSEE, École Nationale de la Statistique et de l'Administration Économique – France
[3] Université Rennes 2 – Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes – France

Nous proposons des estimateurs non-paramétriques pour les fonctions moyenne et covariance de données fonctionnelles. Les courbes aléatoires ne sont pas nécessairement différentiables, de régularité inconnue et mesurées avec erreur sur un ensemble de points discret tirés aléatoirement. La définition de nos estimateurs non-paramétriques dépend de la régularité locale du processus stochastique générant les données. D'abord, nous proposons un estimateur simple de cette régularité locale utilisant les informations intra- et inter-courbes. Ensuite, l'approche "smoothing first, then estimate" est utilisée pour l'estimation des fonctions moyenne et covariance. Ces nouveaux estimateurs non-paramétriques atteignent des vitesses de convergence optimales.

Réduction d'une fonction aléatoire cyclostationnaire définie sur R

Sylvie Viguier-Pla*[1], Alain Boudou[2]

[1] LAboratoire de Mathématiques et PhySique – Université de Perpignan Via Domitia : EA4217 – France
[2] Institut de Mathématiques de Toulouse UMR5219 – Université Toulouse III - Paul Sabatier, Université Toulouse III- PaulSabatier – France

Considérant une fonction aléatoire cyclostationnaire, nous proposons une transformation pour en déduire une série multidimensionnelle stationnaire, dont on peut éventuellement faire une analyse en composantes principales. Nous définissons ce qu'est une série résumé de plus faible dimension, et nous en étudions une particulière.

13h30 – 14h30 Déjeuner

14h30 – 15h30

PLENIERE : Frédéric Chazal

(Amphi 7 - 14h30-15h30)

Quelques propriétés statistiques des descripteurs topologiques des données

Frédéric Chazal

INRIA Univ Paris Saclay

L'Analyse Topologique des Données (TDA) est un domaine récent qui connait un succès croissant depuis quelques années. Il vise à comprendre, analyser et exploiter la structure topologique et géométrique de données complexes. Avec l'émergence de la théorie de la persistance homologique, la géométrie et la topologie ont fourni des outils nouveaux et efficaces pour aborder ces questions. Dans cet exposé, nous introduirons quelques outils permettant de construire des descripteurs robustes de la topologie des données. Nous nous intéresserons en particulier à leurs propriétés statistiques et nous illustrerons, sur quelques exemples concret, l'intérêt des approches topologiques pour l'analyse des données et l'apprentissage statistique.

PLENIERE : Gersende Fort

(Amphi 10 - 14h30-15h30)

Algorithmes Majoration-Minoration stochastiques pour l'Apprentissage Statistique grande échelle

Gersende Fort

CNRS IMT Toulouse

En Apprentissage Statistique, notamment pour la minimisation de fonctions de risque, on s'intéresse à la minimisation de fonctions de type 'sommes finies' i.e. s'exprimant comme la somme d'un grand nombre de termes, eux-mêmes pouvant ne pas avoir d'expressions explicites. Il est alors nécessaire de définir des procédures d'optimisation stochastiques, capables de réduire le coût computationnel lié à la gestion de grands ensembles d'apprentissage, mais aussi d'intégrer des approximations consistantes de quantités incalculables.

Cet exposé sera consacré aux procédures d'optimisation de type Majoration-Minoration (MM). Les algorithmes de gradient stochastique et leurs extensions proximales, ou encore les algorithmes Expectation-Maximization pour l'apprentissage dans les modèles à données latentes, sont des exemples de procédures MM très populaires en Statistique.

Cet exposé présentera de nouveaux algorithmes d'approximations stochastiques accélérés pour répondre à des problèmes d'apprentissage statistique dans le contexte usuel de grands ensembles d'apprentissage, mais aussi celui de l'apprentissage en ligne et de l'apprentissage fédéré. Des éléments d'analyse de convergence et d'analyse de complexité seront aussi discutés.

15h30 – 15h50 Pause

15h50 – 17h20

Séries Temporelles

(Amphi 7 - 15h50-17h20)

Une nouvelle méthode pour les propriétés asymptotiques de modèles à seuil autoexcités

Guy Mélard*[1], Marcella Niglio[2]

[1] Université libre de Bruxelles, Faculté SBS-EM, ECARES – Belgique
[2] Università degli Studi di Salerno – Italie

Une nouvelle méthode pour obtenir les propriétés asymptotiques des estimateurs de modèles à seuil autoexcités autorégressifs (self-excited threshold autoregressive, SETAR) est introduite.
Les modèles à seuil sont des modèles non linéaires pour une série temporelle $x(t)$, $t = 1, ..., n$, avec k régimes, où le régime dépend de la valeur d'une variable $y(t - d)$, $d > 0$, appelée variable de seuil, par rapport à une (quand $k = 2$) ou à plusieurs (quand $k > 2$) valeur(s) de seuil. Comme pour la plupart des modèles non linéaires, la méthode usuelle pour établir les propriétés asymptotiques de tels modèles consiste à obtenir une solution stationnaire et ergodique de l'équation du modèle et à utiliser l'ergodicité pour prouver la consistance et la normalité asymptotique d'un estimateur des paramètres du modèle.
Une nouvelle méthode pour établir ces propriétés asymptotiques a été récemment proposée par les auteurs quand la variable de seuil $y(t - d)$ est exogène et indépendante des innovations du modèle. La méthode est basée sur la théorie asymptotique pour des modèles tdARMA scalaires ou vectoriels, où les coefficients ne sont pas constants mais sont des fonctions déterministes du temps et d'un petit nombre de paramètres. La méthode est donc valable bien au-delà de modèles autorégressifs à seuil (TAR), comme des modèles TARMA et leurs généralisations multivariées, mais toujours en supposant une variable de seuil exogène.
Les modèles SETAR, c'est-à-dire quand $y(t - d)$ est identique à $x(t - d)$, par exemple, ont des coefficients aléatoires de sorte que la théorie n'est plus directement applicable. Néanmoins, il est possible d'adapter les résultats fondamentaux d'Alj et al. (2017) au cas de coefficients aléatoires tels qu'ils apparaissent dans les modèles SETAR ainsi que dans leurs généralisations SETARMA scalaires ou vectorielles. Le seul problème avec cette nouvelle méthode est que l'existence et la non-singularité de la matrice d'information doivent être supposées ou prouvées.

Barycentres de séries temporelles : une nouvelle approche basée sur la méthode de la signature

Raphael Mignot*[1], Konstantin Usevich[2], Marianne Clausel[3], Georges Oppenheim[4], Laure Coutin[5], Antoine Lejay[6,7]

[1] Institut Élie Cartan de Lorraine – Université de Lorraine, Centre National de la Recherche Scientifique : UMR7502 – France
[2] Centre de Recherche en Automatique de Nancy – Université de Lorraine, Centre National de la Re-

cherche Scientifique : UMR7039 – France
[3] Université de Lorraine – Institut Elie Cartan de Lorraine (IECL) – France
[4] Université Paris-Est – Université Paris Est (UPE) – France
[5] Institut de Mathématiques de Toulouse – Université Paul Sabatier - Toulouse 3 – France
[6] Institut Élie Cartan de Lorraine (IECL) – Université de Lorraine, Centre National de la Recherche Scientifique : UMR7502 – Université de Lorraine, Boulevard des Aiguillettes BP 70239 54506 Vandoeuvre-les-Nancy CedexIle du Saulcy - 57 045 Metz Cedex 01, France
[7] Tosca, Inria Nancy Grand Est / Institut Elie Cartan de Lorraine – INRIA – IECL, campus scientifique, BP 70239, 54506 Vandoeuvre-les-Nancy, France

Résumé. La méthode de la signature a été largement utilisée pour l'analyse des séries temporelles. Cette approche a prouvé son efficacité pour de nombreuses applications en apprentissage statistique. La définition d'une notion de barycentre dans l'espace des signatures est un premier pas prometteur permettant de développer de nouvelles extensions de l'analyse en composantes principales (ACP) ou de l'algorithme des k-moyennes aux séries temporelles.

Abstract. The signature method is a widely used method for time series analysis. This novel approach has shown to perform well in various machine learning contexts. Designing a notion of barycenter in the signature space would allow the use of the signature transform in ubiquitous strategies in data science such as the Principal Component Analysis (PCA) for data compression or k-means for clustering.

Réduction de dimension pour les séries temporelles avec des auto-encodeurs variationnels

William Todo*[1], Jean-Michel Loubes[1], Béatrice Laurent[1], Merwann Selmani

[1] Institut de Mathématiques de Toulouse – Université Toulouse III - Paul Sabatier, Université Toulouse III - Paul Sabatier, Université Toulouse III- PaulSabatier – France

Dans cette présentation, nous explorons des techniques de réduction de dimensions sur des séries temporelles univariées et multivariées. En particulier, une comparaison détaillée entre les auto-encodeurs variationnels et la décomposition en ondelettes pour la réduction de dimension. On montre que les auto-encodeurs variationnels sont performants pour réduire la dimension de données de grandes dimensions comme les ECG. Ses comparaisons sont faites sur des jeux de données réels et disponibles publiquement qui présentent une grande variabilité. On utilise l'erreur de reconstruction comme métrique de performance. Enfin nous montrerons la robustesse de ces modèles avec des données bruitées que ce soit pour l'entrainement ou l'inférence. Ces tests permettant de simuler des problèmes souvent rencontré avec des séries temporelles et les VAE sont robustes à ces perturbations.

ML et Extrêmes

(Amphi 8 - 15h50-17h20)

In-store traffic flow forecasting using high dimensional time series methods

Siham Alaoui Belghiti*[1], Badih Ghattas[2]

[1] I2M AMU, Act LOCALA – France
[2] I2M AMU – France

Predicting the flow of visits to stores is of great importance to businesses since it can help to optimize and streamline their operations by adapting to expected traffic volumes. In this article, we frame the problem of predicting traffic flow as a high dimensional time series prediction problem. We pre-process the data in order to enable us to utilize state of the art high dimensional time series algorithms to predict future time points. We then analyze and compare in detail, several state-of-the-art forecasting methods.
We found that a high dimensional approach utilising deep learning and learning both global and local structure works the best in our case.

Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte

Ibrahim Bouzalmat*[1], Benoîte De Saporta[2], Solym Manou-Abi[3]

[1] IMAG, Univ Montpellier, CNRS, Montpellier – CNRS, Université de Montpellier – France
[2] IMAG, Univ Montpellier, CNRS, Montpellier, France – IMAG – France
[3] IMAG, Univ Montpellier, CNRS, CUFR Mayotte – CNRS, Université de Montpellier – France

L'objectif de ce travail est de modéliser la propagation de la fièvre typhoïde à Mayotte à partir d'un jeu de données d'hospitalisations fourni par l'Agence Régionale de Santé. Nous utilisons un processus de naissance et mort linéaire avec immigration comptabilisant les personnes infectées par la maladie. L'objectif est alors d'estimer les taux de contamination de personne à personne, contamination par l'environnement et guérison à partir des données. L'originalité de notre approche et la diculté du problème provient de deux sources. D'une part, les observations ne sont pas disponibles en temps continu mais seulement à des dates fixes (hospitalisations journalières), et d'autre part à ces dates le nombre total de personnes infectées n'est pas observé directement, seuls les nouveaux cas depuis la date précédente sont comptabilisés. Pour traiter ces spécicités, nous obtenons d'abord une expression explicite des lois de transition à pas xe pour construire des

estimateurs de nos paramètres basés sur les fréquences des transitions pour le nombre d'infectés. Ensuite nous nous plaçons dans le cadre des chaînes de Markov cachées et adaptons l'algorithme de Baum-Welch à notre cas.

Crowdsourcing label noise simulation on image classification tasks

Tanguy Lefort*[1], Benjamin Charlier[2], Joseph Salmon[1], Alexis Joly[3]

[1] Institut Montpelliérain Alexander Grothendieck (IMAG) – CNRS, Université de Montpellier – France
[2] Institut Montpelliérain Alexander Grothendieck (IMAG) – Université de Montpellier, Centre National de la Recherche Scientifique : UMR5149 – UMR CNRS 5149 - Université Montpellier 2, Case courrier 051, 34095 Montpellier cedex 5 - France, France
[3] INRIA (INRIA) – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – Montpellier, France

It is common to collect labelled datasets using crowdsourcing. Yet, labels quality depends deeply on the task difficulty and on the workers abilities. With such datasets, the lack of ground truth makes it hard to assess the quality of annotations. There are few open-access crowdsourced datasets, and even fewer that provide both heterogeneous tasks in difficulty and all workers answers before the aggregation. We propose a new crowdsourcing simulation framework with quality control. This allows us to evaluate different empirical learning strategies empirically from the obtained labels. Our goal is to separate different sources of noise: workers that do not provide any information on the true label against poorly performing workers, useful on easy tasks.

Evaluating the Impact of Parenthood on Career Progression in STEM jobs, across Gender

Mira Rahal*[1], Adeline Samson[2], Elise Arnaud[2], Isabel Torres[3]

[1] Laboratoire d'Ecole d'Economie de Paris – Paris School of Economics - CNRS – France
[2] Laboratoire Jean Kuntzman – Université Grenoble Alpes – France
[3] Mothers in Science – Mothers in Science – France

Nous avons étudier l'impact de la parentalité sur les carrières scientifiques de Science, Technology, Engineering, Mathematics et Médecine (STEMM) à partir des réponses à une enquête réalisée par Mothers in Science en France en 2020. Dans les secteurs académiques, nous avons trouvé que les femmes connaissent souvent une baisse de taux de publications après être devenues parents, ce qui est souvent moins le cas pour les pères. En utilisant un modèle à effets mixtes, nous avons montré que les femmes travaillant dans le secteur des sciences physiques connaissent une baisse significative de leur productivité après être devenues parents. Nous avons également mis en évidence un effet inégal de la parentalité sur la réputation professionnelle selon le sexe, à

travers un modèle de régression logistique ordinale.

Profondeur à noyau multivariée et mesure de risque extrême associée

Sara Armaut*[1]

[1] Laboratoire Jean Alexandre Dieudonné – Université de Nice Sophia-Antipolis – France

De nos jours, dans pratiquement tous les domaines tels que la finance, la médecine, l'écologie, l'industrie..., il est impossible d'éviter des risques ! Par exemple, en finance, le "risque" signifie souvent la possibilité de perdre de l'argent. En hydrologie, le risque peut par ailleurs représenter la quantité d'eau dépassant le niveau de remplissage maximum d'un barrage. La Conditional Covariate Tail Expectation (ou CCTE) est une mesure de risque qui quantifie un coût moyen associé à d $>$ 0 facteurs de risque non nécessairement homogènes. Dans notre cadre d'étude, la zone de risque est représentée par un ensemble de niveau inférieur associé à une fonction de profondeur statistique multivariée. Nous proposons un estimateur consistant de la CCTE pour des niveaux extrêmes : cet estimateur fait intervenir une estimation de l'ensemble de niveau associé à la profondeur en question via une méthode plug-in. Dans le modèle gaussien, on obtient une vitesse de convergence de la CCTE basée sur la profondeur dite à noyau.

Kalman Recursions Aggregated Online

Eric Adjakossa*[1]

[1] UMR 518 MIA-Paris, F-75005 Paris – AgroParisTech, INRA - Université Paris-Saclay – France

Dans ce travail, nous améliorons la qualité de la prédiction par agrégation d'experts en utilisant les propriétés sous-jacentes des modèles qui fournissent ces experts. Nous nous limitons au cas où les prédictions d'experts sont issues de récursions de Kalman par ajustement de modèles espace-état. En utilisant des poids exponentiels, nous avons construit différents algorithmes d'agrégation de récursions de Kalman en ligne (KAO) qui compétissent avec le meilleur expert ou la meilleure combinaison convexe des experts de façon adaptative ou non. Nous améliorons les résultats existants de la littérature sur l'agrégation d'experts lorsque les experts sont des récursions de Kalman en utilisant leurs propriétés de second ordre. Nous appliquons notre approche aux récursions de Kalman et l'étendons au contexte général d'experts en ajustant un modèle espace-état aux erreurs d'experts fournis.

Méthodes Génératives - ML - Deep

(Amphi 9 - 15h50-17h20)

Etude du comportement de l'algorithme DbAS pour l'optimisation de propriétés de séquences nucléiques

Teddy Ardouin *[2,1], Laurent Drazek[1], Pierre Mahé[1], Adeline Samson[2]

[2] Laboratoire Jean Kuntzmann – Institut National de Recherche en Informatique et en Automatique, Centre National de la Recherche Scientifique : UMR5224, Université Grenoble Alpes, Institut polytechnique de Grenoble - Grenoble Institute of Technology – France
[1] Biomérieux – BIOMERIEUX – France

Cette communication porte sur l'optimisation des propriétés d'une séquence de nucléotides résolue sur un espace discret. Dans cette perspective, nous étudions les performances de DbAS (Design by Adaptive Sampling), algorithme prometteur pour l'optimisation de propriétés de protéines tel que la conformation 3D ou la fluorescence. L'objectif est de générer un ensemble de séquences à la fois varié et possédant les meilleures valeurs d'une fonction f à optimiser. Un Auto-encodeur Variationnel (VAE) est utilisé comme modèle génératif associé à DbAS et deux types de fonctions f, unimodale et multimodale, sont étudiées. Bien que l'algorithme fonctionne bien pour des séquences nucléiques, l'augmentation de la taille des séquences et l'optimisation d'une fonction multimodale dégradent les performances. Ces limitations proviennent à la fois du modèle génératif et de l'objectif de DbAS qui ne pénalise pas le manque de diversité.

Limite d'échelle pour les réseaux de neurones résiduels

Pierre Marion *[1]

[1] Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Université : UMR8001, Centre National de la Recherche Scientifique : UMR8001, Université de Paris : UMR8001 – France

Les réseaux de neurones résiduels, introduits par He et al. (2016), ont permis des avancées majeures dans plusieurs domaines de recherche en apprentissage automatique, en permettant l'entraînement de réseaux avec plusieurs milliers de couches. Néanmoins, l'entraînement de ces réseaux reste complexe et instable, en particulier à cause d'un mauvais conditionnement du réseau lorsque la profondeur augmente. Nous nous intéresserons dans cet exposé à l'étude du comportement de ce type de réseaux lorsqu'on ajoute un facteur d'échelle qui dépend de la profondeur, dans la limite où la profondeur L est grande. Nous montrerons en particulier qu'à l'initialisation, le facteur d'échelle doit évoluer comme $L^{-1/2}$ pour obtenir une limite non

dégénérée. Nous illustrerons ce résultat par des expériences simples. Nous ferons également le lien avec les modèles de réseaux de neurones en temps continu, et en particulier les modèles d'équations différentielles ordinaires neuronales introduites par Chen et al. (2018).

Can Push-forward Generative Models Fit Multimodal Distributions?

Antoine Salmona[*1,2], Julie Delon[2], Agnès Desolneux[1], Valentin De Bortoli[3]

[1] Centre Borelli – CNRS : UMR9010, ENS Paris-Saclay – France
[2] MAP5 – CNRS : UMR8145, Université de Paris – France
[3] DI/ENS Ulm – CNRS : UMR8548, Ecole Normale Supérieure de Paris - ENS Paris – France

Les modèles génératifs sont aujourd'hui l'un des sujets de recherche les plus populaires en apprentissage automatique, notamment grâce à leur impressionnante capacité à générer des images synthétiques réalistes. Cependant, il reste souvent difficile de savoir si ces modèles s'approchent correctement de la distribution sous-jacente des données ou s'ils génèrent uniquement des échantillons qui semblent similaires aux données. Dans ce travail, nous nous concentrons sur la classe particulière des modèles génératifs *push-forward*, qui inclut les *Variational Autoencoders*, les *Generative Adversarial Networks* et les *Normalizing Flows*. Nous montrons que ces modèles doivent avoir de grandes constantes de Lipschitz afin de bien approcher les distributions multimodales. Ainsi, il existe pour ces modèles une compétition entre la stabilité de leur apprentissage et leur capacité à générer des distributions multimodales. Generative models are today one of the most popular research topic in machine learning thanks to their impressive ability to generate realistic synthetic images. However, it remains unclear whether these models approach correctly the underlying data-distribution or only generate samples that seem similar to the data. In this work, we focus on the particular class of the push-forward generative models, which includes Variational Autoencoders, Generative Adversarial Networks and Normalizing Flows. We show that these models must have large Lipschitz constants in order to approximate multimodal distributions. Thus, there is for these models a contradiction between their training stability and their ability to generate multimodal distributions.

Generating new crystal structures with statistical methods

Arsen Sultanov[*1], Jean-Claude Crivello[1], Tabea Rebafka[2], Nataliya Sokolovska[3]

[1] Univ Paris Est Creteil, CNRS, ICMPE – Université Paris Est, ICMPE (UMR 7182), CNRS – France
[2] LPSM, Sorbonne Université, Université de Paris & CNRS – Probability, statistics and modeling lab – France
[3] Sorbonne University, INSERM, NutriOmics – Sorbonne Universités UPMC Univ Paris 06, Inserm, CNRS – France

Dans cette étude, nous proposons une méthode basée sur un modèle probabiliste de *denoising diffusion* pour générer des matériaux hypothétiques définis par leurs structures cristallines. Nous expliquons le modèle statistique sous-jacent, comment il peut être appliqué à la génération de structure cristalline et pourquoi cette approche est bien justifiée à ce problème.

Comment imposer des pré-activations gaussiennes dans un réseau de neurones ?

Pierre Wolinski*[1], Julyan Arbel[2]

[1] Inria Grenoble - Rhône-Alpes – Institut National de Recherche en Informatique et en Automatique, Université Grenoble Alpes – France
[2] MISTIS (INRIA Grenoble Rhône-Alpes / LJK Laboratoire Jean Kuntzmann) – Laboratoire Jean Kuntzmann, INRIA – France

Le but de ce travail est de proposer un moyen de modifier la loi d'initialisation des poids d'un réseau de neurones ainsi que sa fonction d'activation, de façon à assurer que toutes les pré-activations soient gaussiennes. Nous proposons une famille de couples initialisation/activation, où les fonctions d'activation couvrent un continuum allant des fonctions bornées de type Heaviside ou tanh, jusqu'à la fonction identité.
Ce travail est motivé par la contradiction entre des travaux existants quant au caractère gaussien des pré-activations : d'un côté, les travaux dans la lignée des Neural Tangent Kernels et de l'Edge of Chaos en font abondamment usage ; de l'autre, des résultats théoriques et expérimentaux mettent à mal cette hypothèse.
La famille de couples initialisation/activation que nous proposons permettra d'avancer sur la question qui traverse des travaux actuels : est-il souhaitable d'avoir des pré-activations gaussiennes dans un réseau de neurones ?

Processus - Statistique mathématique

(Amphi 10 - 15h50-17h20)

Vitesses de contraction du posterior pour les processus gaussiens profonds contraints en classification et estimation de densité

Francois Bachoc*[1]

[1] Institut de Mathématiques de Toulouse – Centre national de la recherche scientifique - CNRS (France), Université Paul Sabatier - Toulouse III – France

Nous fournissons des vitesses de contraction du posterior pour les processus gaussiens contraints profonds en estimation non paramétrique de densité et en classification. Les contraintes sont des bornes sur les valeurs et dérivées des processus gaussiens dans les couches de la structure de composition. Les vitesses de contraction sont d'abord données dans un cadre général, sous la forme d'une nouvelle fonction de concentration que l'on introduit et qui prend les contraintes en compte. Ensuite, le cadre général est appliqué au mouvement Brownien intégré, au processus de Riemann Liouville et au processus de Matérn, avec des classes de fonctions standard. Dans chacun des exemples, on retrouve des vitesses minimax classiques.

Sélection de variables en régression SIR par seuillage doux ou seuillage dur de la matrice d'intérêt

Hadrien Lorenzo[*2,1], Jerôme Saracco[2,1]

[2] ASTRAL – INRIA – France
[1] Institut de Mathématiques de Bordeaux – CNRS : UMR5251, Institut polytechnique de Bordeaux, Université de Bordeaux (Bordeaux, France) – France

La régression inverse par tranches (Sliced Inverse Regression, SIR, en anglais) a ceci d'intéressant qu'elle permet la construction d'indices, comme combinaisons linéaires de la variable explicative multidimensionnelle, les plus associés (en un certain sens) à la variable réponse étudiée. Ce type de méthode est fort utile dans des contextes non linéaires. La problématique de la sélection de variables dans ce cadre est importante. Nous présentons dans cette communication deux approches fondées sur le seuillage doux ou le seuillage dur de la matrice d'intérêt de la méthode SIR. Nous indiquons comment sélectionner l'hyper-paramètre du seuillage considéré et nous présentons les performances numériques de ces nouvelles méthodologies obtenues dans le cadre de simulations.

Liens entre les décompositions de Hoeffding–Sobol et de Möbius

Cécile Mercadier[*1]

[1] Institut Camille Jordan [Villeurbanne] – Ecole Centrale de Lyon, Université Claude Bernard Lyon 1, Université Jean Monnet [Saint-Etienne], Institut National des Sciences Appliquées de Lyon, Centre National de la Recherche Scientifique : UMR5208 – France

Cet exposé présente les résultats obtenus dans *Linking the Hoeffding–Sobol and Möbius formulas through a decomposition of Kuo, Sloan, Wasilkowski, and Woźniakowski,* publié récemment

dans *Statistics & Probability Letters*. Il s'agit de comprendre les liens entre les décompositions fonctionnelles de Hoeffding–Sobol et de Möbius. Toutes deux sont utilisées en statistique et permettent d'écrire une fonction comme la somme de termes de complexité croissante. Nous montrons qu'elles sont étroitement liées à celle obtenue en 2010 par Kuo, Sloan, Wasilkowski et Wozniakowski. Afin de les retrouver dans un énoncé commun, nous avons dû généraliser le résultat de Kuo et coll. (*Math. Comp.*, 79 (2010), 953–966).

Mélange de modèles d'analyse factorielle longitudinale pour l'analyse de données longitudinales multivariées

Amine Ounajim*[1], Yousri Slaoui[1], Pierre-Yves Louis[2], Denis Frasca[3], Philippe Rigoard[4]

[1] Laboratoire de Mathématiques et applications – Université de Poitiers, UMR 7348 du CNRS – France
[2] AgroSup Dijon – Université de Bourgogne Franche-Comté (UBFC), AgroSup Dijon, UMR PAM A 02.102 – France
[3] INSERM UMR-1246 – Université de Nantes – France
[4] PRISMATICS Lab (Predictive Research In Spine/neurostimulation Management and Thoracic Innovation in Cardiac Surgery) – CHU Poitiers – France

Afin d'étudier l'évolution de plusieurs indicateurs observés parmi des individus, il est important de se concentrer sur les tendances longitudinales parmi les variables latentes en utilisant une modélisation conjointe basée sur les structures de covariance entre ces indicateurs observés. Cependant, ce type de données peut représenter une hétérogénéité dans le temps et selon les groupes d'individus. Dans cet article, nous proposons une extension de l'analyse factorielle classique dans laquelle la non-invariance des groupes est prise en compte en utilisant un modèle de mélange.
Nous commençons par définir le mélange du modèle d'analyse factorielle longitudinale et ses paramètres. Ensuite, nous proposons un algorithme EM pour estimer le modèle. Nous proposons également un critère d'information bayésien pour identifier le nombre optimal de composantes du mélange. L'initialisation, la convergence et le temps de calcul seront discutés. Nous montrons l'applicabilité du modèle en utilisant des données simulées et réelles.

Estimation adaptative dans un modèle de régression-convolution en base d'Hermite

Ousmane Sacko*[1]

[1] Mathématiques Appliquées Paris 5 – Université Paris Descartes - Paris 5 : UMR8145 – France

Dans ce travail, on considère le modèle de régression-convolution suivant: $y(x_k) = f \star g(x_k) + e\_k, x_k = kT/n, k = -n, \ldots, n-1$, T fixé; g est connue et f est la fonction inconnue que l'on

cherche à estimer. Les erreurs $(e_k)_{-n\leq k\leq n-1}$ sont i.i.d., centrées de variance finie et connue. Nous proposons un estimateur par projection en exploitant les propriétés de la base d'Hermite. Une borne du risque est prouvée: sous des conditions de régularité, nous obtenons une vitesse de convergence. Une procédure adaptative pour choisir la dimension pertinente est aussi proposée en s'inspirant de la méthode de Goldenshluger & Lepski et nous prouvons que l'estimateur résultant satisfait une inégalité oracle pour e sous-Gaussien. Enfin, nous illustrons la procédure par des graphes.

A local version of R-hat for MCMC convergence diagnostic

Théo Moins*[1], Julyan Arbel[1], Anne Dutfoy[2], Stéphane Girard[1]

[1] Statify – Laboratoire Jean Kuntzmann, INRIA, Université Grenoble Alpes, Institut polytechnique de Grenoble (Grenoble INP) – France
[2] EDF R&D dept. Périclès – EDF – France

Diagnostiquer la convergence des chaînes de Markov est un enjeu crucial pour les méthodes de Monte Carlo par chaîne de Markov. Parmi les méthodes les plus populaires, le diagnostique de Gelman-Rubin, communément appelé R-hat, est un indicateur qui permet de vérifier la convergence en se basant sur une comparaison des variances inter- et intra-chaîne. Nous proposons ici une version localisée R-hat(x) qui se concentre sur un quantile x de la distribution, et analysons certaines propriétés de la valeur théorique associée R(x). Ceci conduit à proposer un nouvel indicateur qui permet à la fois de localiser la convergence de la chaîne en différents quantiles de la distribution, et en même temps de traiter certains problèmes de convergence non détectés par d'autres versions de R-hat.

Statistique mathématique – Valeurs Manquantes

(Amphi 11 - 15h50-17h20)

Vitesse de convergence pour une prediction linéaire en présence de données manquantes

Alexis Ayme*[1], Claire Boyer[1], Aymeric Dieuleveut[2], Erwan Scornet[2]

[1] LPSM – Sorbonne Universités, UPMC, CNRS – France
[2] CMAP – Polytechnique - X – France

Les valeurs manquantes apparaissent dans la plupart des bases de données. Leur nature même empêche généralement d'exécuter les algorithmes classiques d'apprentissage supervisé. Dans cet article, nous étudions les modèles linéaires largement étudiés, mais qui en présence de valeurs manquantes, s'avère être une tâche difficile. En effet, cela nécessite finalement de résoudre un nombre de tâches d'apprentissage, exponentiel par rapport au nombre de variables, ce qui rend les prédictions impossibles sur de vraies bases de données. Nous établissons une première borne non-asymptotique sur l'excès de risque d'un estimateur de type moindres carrés, qui croît exponentiellement avec la dimension. Nous proposons ensuite un nouvel algorithme et une borne qui s'adapte à la distribution des valeurs manquantes et qui, de plus, s'avère être optimale au sens minimax. Des expériences numériques mettent en évidence les avantages de notre méthode par rapport aux algorithmes fréquemment utilisés pour la prédiction avec valeurs manquantes.

Imputation multiple dans le modèle linéaire fonctionnel avec une covariable partiellement observée et des valeurs manquantes dans la réponse

Christophe Crambes[1], Chayma Daayeb*[1], Ali Gannoun[1], Yousri Henchiri[2]

[1] Institut Montpelliérain Alexander Grothendieck – Université de Montpellier, Centre National de la Recherche Scientifique : UMR5149 – France
[2] Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur [Tunis] – Tunisie

Les données manquantes sont courantes dans les études de météorologie, de biomécanique, d'économie, de médecine et de sciences sociales et posent souvent un sérieux défi dans l'analyse des données. Les méthodes d'imputation multiple (IM) sont des outils populaires et naturels pour traiter les données manquantes, qui offrent une méthode réalisable sur le plan informatique pour étudier un large spectre de problèmes en présence de données manquantes, en remplaçant chaque valeur manquante par un ensemble de valeurs raisonnables qui représentent une incertitude sur les valeurs à imputer.

Usefulness of surrogate outcomes on the regression model and the PaO2 prediction

Firas Ibrahim*[1], Mustapha Rachdi[2], Jacques Demongeot[2]

[1] LMBA – Université de Bretagne Sud [UBS] – France
[2] AGEIS – , Universitey of Grenoble Alpes (UGA) – France

Dans ce travail, on traite du problème de l'estimation de la régression d'une variable de réponse scalaire, présentant des données manquantes mais avec des données de substitution, expliquée par une variable aléatoire fonctionnelle. Nous construisons un estimateur de l'opérateur de régression en utilisant, en plus des valeurs disponibles (vraies) de la variable réponse, les valeurs des données de substitution. Nous établissons ensuite des propriétés asymptotiques de l'estimateur construit en termes de convergences presque complète et en moyenne quadratique. Notons que les résultats obtenus ici, dans le cadre fonctionnel, généralisent une partie des résultats obtenus dans le cadre de dimension finie. Enfin, une illustration sur l'applicabilité de nos résultats sur des données simulées et une application à des données réelles ont été réalisées (prédiction de PaO2). Nous avons ainsi démontré la supériorité de notre estimateur sur les estimateurs classiques, lorsque nous manquons de données complètes.

Apprentissage semi-supervisé avec réponses manquantes informatives

Aude Sportisse*[1], Charles Bouveyron[2], Pierre-Alexandre Mattei[3]

[1] Inria, 3Ia – Inria Sophia Antipolis (MAASAI), 3IA, Université Côte d'Azur (UCA) – France
[2] Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France. – Université Côte d'Azur (UCA) – France
[3] Laboratoire Jean Alexandre Dieudonné – Université Nice Sophia Antipolis, Centre National de la Recherche Scientifique, 3iA Côte d'Azur – France

Pour prédire un phénomène, des données parfaites seraient constituées de plusieurs variables explicatives et d'une variable cible ; bien sûr, toutes seraient observées. En pratique, bien que la quantité de données disponibles soit souvent énorme, l'étiquetage des données est coûteux et prend du temps. L'apprentissage semi-supervisé vise à exploiter des données étiquetées et non étiquetées pour entraîner des modèles prédictifs. Dans ce travail, le non-étiquetage des données est considéré comme un problème de données manquantes. Nous nous concentrons sur le cas MNAR (Missing Not At Random), lorsque le manque des étiquettes dépend de leurs valeurs elles-mêmes. Par exemple, cela se produit lorsque les gens sont plus enclins à étiqueter les images de certaines classes qui sont faciles à reconnaître. Dans ce cadre, nous proposons d'estimer le mécanisme des données manquantes et nous proposons des estimateurs pondérés utilisant cette estimation qui tiennent compte de toutes les données.

Estimation ciblée pour les modèles structurels marginaux

Herbert Susmann[1,2], Antoine Chambaz[2]

[1] University of Massachusetts Amherst

[2] MAP5, Université de Paris CNRS : UMR8145

Deux des tâches principales de l'inférence causale sont de définir et d'estimer l'effet d'un traitement sur une issue d'intérêt. Formalement, ces effets sont definis comme un sommaire fonctionnel de la loi générant les données, et sont dénommés paramètres d'intérêt. L'estimation du paramètre d'intérêt peut être difficile, surtout lorsqu'il est de grande dimension. Les Modèles Structurels Marginaux (MSM) permettent de résumer tels paramètres en utilisant un modèle de travail spécifié par l'utilisateur. Nous déterminons la borne d'efficacité semi-parametrique pour l'estimation des paramètres des MSM dans un cadre général. Ensuite, nous présentons un estimateur fondé sur le principe de la minimisation ciblée de pertes qui atteint cette borne. Nos resultats peuvent être facilement adaptés à des structures de donné et paramètres d'intérêt spécifiques.

17h20 – 17h30 Pause

17h30 – 18h00 Clôture

(Amphi 7 - 17h30-18h00)

18h00 – 18h30 Pause

18h30 – 21h30

Rencontre Jeunes statisticiens

Rencontres Jeunes Statisticiens et Conférenciers invités... et le quizz !

C'est aussi le grand retour des rencontres en présentiel ! Venez les rencontrer autour d'un apéro et du fameux quizz du groupe Jeunes, le jeudi 16 juin à partir de 18h, dans un lieu encore à préciser.

Pour nous permettre d'organiser au mieux ces rencontres, merci de bien vouloir vous inscrire en suivant ce lien : https://framaforms.org/inscription-aux-rencontres-jeunes-statisticiennes-et-conferencieres-invitees-1654079720

Café de la Statistique

Café Toï Toï le Zinc, 17-19 rue Marcel Dutartre, 69100 Villeurbanne,
ou visioconférence. Sur inscription à statcafe@sfds.asso.fr.

19h00 – 21h30 (entrée recommandée dès 18h30)

La transition énergétique : quels choix en France et en Europe ?

Cédric Philibert

Analyste énergie et climats – Chercheur associé à l'IFRI- Ancien responsable des énergies renouvelables à l'Agence Internationale de l'Energie

Les prévisions issues des données et des modèles climatiques présentées par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) font largement consensus dans la communauté scientifique et, en principe, entre les Etats qui se fixent un objectif global de neutralité carbone à l'horizon 2050. Cependant les débats en France et en Europe, dans le contexte d'une baisse rapide des coûts des renouvelables (éolien, solaire, biomasse) et d'un refus de la dépendance à l'égard du gaz et du pétrole russes, manifestent des oppositions et des incertitudes sur les voies à suivre et les moyens nécessaires en ce qui concerne notamment :

- les parts souhaitables et soutenables des énergies renouvelables, du nucléaire et des énergies carbonées...
- les besoins à satisfaire et les économies possibles dans les usages des énergies.

Au vu des données récentes sur les coûts et la faisabilité des options, telles qu'on les trouve présentées en particulier dans les derniers rapports de l'Agence internationale de l'énergie (AIE)

et du Réseau de transport de l'électricité (RTE), notre Café de la Statistique du 16 juin, introduit par Cédric Philibert, ancien responsable de l'AIE et expert en matière d'énergies renouvelables, sera l'occasion d'éclairer ces questions essentielles. Un accent sera mis sur quelques problèmes statistiques du monde de l'énergie, à travers par exemple les notions d'énergies primaire et finale, et les répartitions d'émissions directes et indirectes des secteurs de consommation finale.

Pour participer à cette soirée débat, **inscrivez-vous** en adressant un courriel à statcafe@sfds.asso.fr en précisant votre choix présentiel ou visioconférence.
Pour permettre à tous ceux qui le souhaitent de suivre la conférence à distance, un lien de connexion vous sera adressé pour accéder à la conférence.
Pour permettre à tous ceux qui veulent avoir le temps de commander boissons, tartines ou planches avant le début de l'exposé, nous vous recommandons d'arriver vers 18h30.

Liste des auteurs

Abdi, Hervé, 122
Acharki, Naoufal, 124
Adjabi, Smail, 95
Adjakossa, Eric, 139
Agniel, Denis, 46
Ah-Pine, Julien, 64
Ahmed, Mohamed-Salem, 14
Al Marjani, Aymen, 128
Alamichel, Louise, 17
Alaoui Belghiti, Siham, 137
Alaoui, Enora, 52
Allard, Denis, 13
Allassonnière, Stéphanie, 14, 39
Allouche, Michael, 25
Alquier, Pierre, 43
Alsouki, Louna, 43
Amara-Ouali, Yvenn, 75
Amato, Francesco, 78
Amovin-Assagba, Messan Martial, 21
Andre, Manon, 58
Antero, Juliana, 93
Arbel, Julyan, 16, 17, 142, 145
Ardouin, Teddy, 140
Arlot, Sylvain, 44
Armaut, Sara, 139
Arnaud, Elise, 138
Asadi, Arefe, 71
Asberg, Anders, 60
Aubin, Jean-Baptiste, 110
Ayme, Alexis, 145
Azencott, Chloé-Agathe, 69
Azzaoui, Nourddine, 19

Bachoc, Francois, 142
Baldé, Oumar, 126
Ballarini, Paolo, 78
Bar-Hen, Avner, 68
Baragatti, Meili, 18
Barbe, Pierre, 52
Barbieri, Antoine, 61
Barbillon, Pierre, 38
Barbot, Benoit, 78
Barbu, Vlad Stefan, 72
Barriac, Vincent, 65
Barrier, Antoine, 128
Basak, Subhasish, 96
Batardière, Bastien, 66
Beaudoin, Annabelle, 122
Bect, Julien, 96
Beecham, Roger, 12, 59
Belhadji, Ayoub, 67
Bellet, Aurélien, 37
Ben Mrad, Oumaima, 130
Benoist, Clément, 60
Bernard, Renan, 82
Bernier, Jacques, 20
Berthoud, Françoise , 55
Bertoncello, Antoine, 124
Bertrand, Quentin, 29
Bethune, Louis, 82
Biernacki, Christophe, 19
Bleza, Esso-Ridah, 120
Bobbia, benjamin, 25
Bode, Nikolai, 17
Bonaldi, Christophe, 48
Bouaziz, Olivier, 116
Boudou, Alain, 132
Bouhadjera, Feriel, 18
Boulin, Alexis, 26
Bouloré, Antoine, 126
Bousquet, Nicolas, 131
Boutet, Jerome, 20
Boutin, Rémi, 39
Bouveyron, Charles, 39, 79, 80, 121, 123, 147
Bouzalmat, Ibrahim, 137
Boyer, Claire, 145
Brault, Vincent, 20
Breeur, Marie, 81
Brunel, Victor-Emmanuel, 107
Bry, Xavier, 28
Bucci, Andrea, 88

Bunce, Catey, 10
Bunz, Yoann, 56
Buritica, Gloria, 113
Bystrova, Daria, 17
Bénesse, Clément, 125

Cannamela, Claire, 73
Carrière, Mathieu, 35
Castanier, Bruno, 71
Celisse, Alain, 44
Celse, Benoit, 50
Cesa-Bianchi, Nicolò, 127
Chabridon, Vincent, 125
Chagneux, Mathis, 56
Chagny, Gaëlle, 63
Chainais, Pierre, 106
Chambaz, Antoine, 63, 131, 147
Chamroukhi, Faicel, 75, 77
Charlier, Benjamin, 138
Chautru, Emilie, 113
Chauvet, Guillaume, 112, 118
Chavent, Marie, 22
Chazal, Frédéric , 133
Chen, Chung Shue, 76
Chennetier, Guillaume, 49
Cheysson, Felix, 30
Chiquet, Julien, 66
Chraïbi, Hassane, 49
Chrétien, Stéphane, 65
Claeys, Emmanuelle, 78
Claudel, Sandra, 91
Clausel, Marianne, 135
Clementz, Samy, 44
Clémençon, Stephan, 67
Colizza, Vittoria, 33
Colnet, Bénédicte, 126
Combes, Florian, 96
Combrisson, Damien, 56
Corneli, Marco, 80, 121
Cornelius, Victoria, 13, 61
Coulmy, Nicolas, 93
Courcoul, Léonie, 61
Courty, Nicolas, 82
Coutin, Laure, 135
Crambes, Christophe, 146
Crepey, Pascal, 33
Crivello, Jean-Claude, 141
Cry, Pierre, 78
Cugliari, Jairo, 91
Cwiling, Ariane, 116
Côme, Etienne, 79

D'Amico, Guglielmo, 72
Da Veiga, Sébastien, 125
Daayeb, Chayma, 146
Dabo-Niang, Sophie, 14
Damblin, Guillaume, 126
Daouia, Abdelaati, 109
David, Ingrid, 48
De Bortoli, Valentin, 141
De Lara, Lucas, 82
De Larochelambert, Quentin, 92, 93
De Saporta, Benoîte, 137
De Vilmarest, Joseph, 41
Delattre, Maud, 15
Delon, Julie, 141
Deltreil, Guillaume, 62
Demangeot, Marine, 113
Demongeot, Jacques, 146
Denis, Christophe, 32
Derquenne, Christian, 67
Descatha, Alexis, 62
Desolneux, Agnès, 141
Di Bernardino, Elena, 26
Dieuleveut, Aymeric, 77, 145
Difernand, Audrey, 92, 93
Digne, Julie, 35
Dombry, Clément, 15, 25, 114, 115
Drazek, Laurent, 140
Du Roy De Chaumaray, Marie, 27
Ducombe, Stephanie, 92
Duprey, Corentin, 19
Durrleman, Stanley, 39
Dussap, Florian, 44
Dutfoy, Anne, 49, 145
Duval, Laurent, 43
Duveau, Catherine, 23
Déjean, Sébastien, 94
Döhler, Sebastian, 30

El Haddad, Rami, 43
El Harfaoui, Echarif, 41
Ella Mintsa, Eddy, 64
Enticott, Euan, 89
Erwan, Scornet, 126
Escobar-Bach, Mikael, 62
Esposito, Nicola, 71

Favre, Julien, 19
Feau, Cyril, 50

Fermanian, Jean-Baptiste, 45
Fernandez, Camila, 76
Fischer, Aurélie, 22
Flament, Guillaume, 111
Forbes, Florence, 16, 77
Fort, Gersende , 134
Fougères, Anne-Laure, 114
Fouladirad, Mitra, 71
Fraiman, Ricardo, 96
Frasca, Denis, 144
Fraysse, Guillaume, 65
Frellsen, Jes, 97
Frevent, Camille, 14
Friguet, Chloé, 82
Fromont, Magalie, 117
Féron, Olivier, 77

Gaillard, Pierre, 76
Galant, Julie, 19
Gamboa, Fabrice, 131
Gannaz, Irène, 21, 110
Gannoun, Ali, 146
Garcin, Camille, 111
Garcin, Matthieu, 26
Gares, Valerie, 82, 112, 117, 118
Garivier, Aurélien, 128, 130
Garnier, Josselin, 49, 50, 73, 124
Gastaud, Clément, 130
Gatulle, Nicolas, 63
Gauchy, Clément, 50
Genin, Michaël, 14
Geoffrey, Ecoto, 131
Gerville-Réache, Léo, 94
Ghattas, Badih, 91, 96, 137
Gibaud, Julien, 28
Giorgio, Massimiliano, 71
Giraldi, Loic, 126
Girard, Stéphane, 25, 145
Giulini, Ilaria, 22
Gkelsinis, Thomas, 72
Gobet, Emmanuel, 25
Goga, Camelia, 15
Golovkine, Steven, 132
Gonzalez-Sanz, Alberto, 82
Goude, Yannig, 75, 77, 91
Graczyk, Piotr, 62
Grave, Clémence, 48
Gribonval, Rémi, 130
Grill, Léo, 128
Grislain, Nicolas, 130
Grollemund, Paul-Marie, 18
Grusea, Simona, 47
Guillemot, Vincent, 122
Guillin, Arnaud, 19

Hajage, David, 112
Hallin, Marc, 9
Hamrouche, Bachir, 75
Happe, André, 118
Harchaoui, Warith, 123
Harel, Michel, 41
Has, Sothea, 76
Hebiri, Mohamed, 32
Hejblum, Boris, 46
Helali, Salima, 97
Henchiri, Yousri, 146
Hilgert, Nadine, 18
Hivert, Benjamin, 46
Humbert, Olivier, 122

Iapteff, Loïc, 50
Ibrahim, Firas, 146
Il Idrissi, Marouane, 131
Iooss, Bertrand, 131
Ippoliti, Luigi, 88

Jacob, Jérôme, 73
Jacqmin-Gadda, Hélène, 61
Jacques, Julien, 21, 50, 65, 78
Jannot, Anne-Sophie, 14
Jeamart, Marion, 82
Jollois, François-Xavier, 68
Jolly, Caroline, 20
Joly, Alexis, 57, 111, 138
Joly, Pierre, 48
Josse, Julie, 63, 77, 126
Jourdan, Astrid, 51
Jouvin, Nicolas, 79

Kamel, Mouna, 59
Katsahian, Sandrine, 117
Kaufmann, Emilie , 99
Kerleguer, Baptiste, 73
Khalfi, Abderaouf, 42
Klopfenstein, Quentin, 29
Klutchnikoff, Nicolas, 132
Kon Kam King, Guillaume, 15, 17
Kpotufe, Samory , 105
Kratz, Marie, 108
Kuhn, Johann, 48
Kwon, Joon, 66

La Rocca, Michele, 88
Labriffe, Marc, 60
Labyt, Etienne, 20
Lacroix, Perrine, 28
Laforgue, Pierre, 127
Lalanne, Clément, 130
Laloe, Thomas, 26
Lambert, Raphaël, 20
Lameiras Franco Da Costa, Victor, 50
Lasalle, Etienne, 40
Lasgorceux, Florian, 56
Latouche, Pierre, 39, 79, 121
Laurent, Béatrice, 136
Lavenu, Audrey, 117
Laverny, Oskar, 45
Le Minh, Tâm, 40
Le Strat, Yann, 48
Le Toquin, Bryan, 92
Leblanc, Frederique, 51
Lebreton, Noé, 64
Ledanois, Thibaut, 95
Lefort, Tanguy, 138
Lejay, Antoine, 135
Leoni, Samuela, 110
Leveau, Valentin, 57
Liang, Dingge, 121
Limnios, Myrto, 67
Lopez Sanchez, Javier, 94
Lorenzo, Hadrien, 143
Loubes, Jean-Michel, 131, 136
Louis, Pierre-Yves, 93, 144
Lu, Yunjiao, 20
Lucas, Felix, 52
Lévy-Leduc, Céline, 47

Mahé, Pierre, 140
Makhlouf, Slimane, 68
Manou-Abi, Solym, 137
Mantoux, Clément, 39
Marandon, Ariane, 80
Marbac, Matthieu, 27
Marbac-Lourdelle, Matthieu, 19
Marceau, Etienne, 108
Marchello, Giulia, 80
Mareau, Alexis, 117
Marie, Nicolas, 43
Marion, Pierre, 140
Marquet, Pierre, 60
Marrel, Amandine, 125, 126
Marteau, Clément, 43
Marteau, Pierre-François, 120
Mas, Victoria, 52
Masmoudi, Afif, 130
Massart, Pascal, 75
Massias, Mathurin, 29
Massonaud, Clement, 33
Mattei, Pierre-Alexandre, 97, 122, 123, 147
Maume-Deschamps, Véronique , 120
Meah, Iqraa, 30
Mercadier, Cécile, 143
Merlier, Lucie, 73
Meyer, Nicolas, 109
Meynaoui, Anouar, 63
Michel, Bertrand, 35
Mignot, Raphael, 135
Mikosch, Thomas, 113
Modeste, Thibault, 114
Mohdeb, Zaher, 31
Moins, Théo, 145
Mokkadem, Abdelkader, 31
Monbet, Valerie, 120
Mougeot, Mathilde, 22
Mourer, Alex, 22
Murris, Juliette, 117
Mélard, Guy, 135

Nadjahi, Kimia, 83
Nagy, Stanislav, 107
Naulin, Jean-Philippe, 90
Naveau, Marion, 15
Naveau, Philippe, 115
Ngatchou-Wandji, Joseph, 41
Ngounou Bakam, Yves Ismaël, 31
Nguyen, Hien Duy, 16, 77
Nguyen, Thi Thanh Yen, 131
Nguyen, TrungTin, 16, 77
Nguyen, Van Tuan, 119
Niglio, Marcella, 135

Obst, David, 91
Oger, Emmanuel, 118
Ohl, Louis, 123
Olié, Valérie, 48
Olteanu, Madalina, 22
Opitz, Thomas, 56
Oppenheim, Georges, 91, 135
Ounajim, Amine, 144

Padoan, Simone A., 109
Pakzad, Cambyse, 26

Papadakis, Nicolas, 70
Paparisteidi, Nefeli, 58
Papaïx, Julien, 56
Paquelet, Stéphane, 118
Parent, Eric, 20
Patilea, Valentin, 132
Paudice, Andrea, 127
Pellet, Aurélien, 58
Perduca, Vittorio, 116
Perna, Cira, 88
Perrin, Tran Vi-vi Élodie, 90
Peyraud, Eliz, 118
Pham, Nhat Thien, 75
Philibert, Cédric, 149
Philippe, Cathy, 122
Pic, Romain, 115
Pierrot, Amandine, 91
Pietrosemoli, Natalia, 122
Pinson, Pierre, 91
Poggi, Jean-Michel, 75
Pommeret, Denys, 31
Pontil, Massimiliano, 127
Potier, Axel, 19
Precioso, Frederic, 123
Prieur, Clémentine, 36
Prokopenko, Evgeny, 108
Pruilh, Solange, 14
Pujol, Louis, 29

Quinton, Jean-Charles, 20

Rabier, Charles-Elie, 47
Rachdi, Mustapha, 146
Ragheb, Waleed, 57
Rahal, Mira, 138
Ramiro, Lugo, 124
Rebafka, Tabea, 23, 80, 141
Reise, Wojciech, 112
Reiss, Markus , 87
Rejeb, Sara, 23
Remlinger, Carl, 129
Ricard, Anne, 48
Rigoard, Philippe, 144
Rigollet, Philippe, 81
Roche, Angelina, 63
Rodolpho Tomassi, Diego, 71
Rohmer, Jérémy, 90
Rohmer, Tom, 48
Rolland, Antoine, 60, 110
Roquain, Etienne, 30, 80
Rosier, Amélie, 43
Rouanet, Anaïs, 32
Roustant, Olivier, 90
Roux, Jonathan, 33
Rouzé, Timothé, 117

Sabourin, Anne, 113
Sacko, Ousmane, 144
Sadoun, Mohamed Djemaa, 42
Sagaut, Pierre, 73
Saint-Pierre, Philippe, 94
Salmon, Joseph, 111, 138
Salmona, Antoine, 141
Samedy, Sonico, 58
Samson, Adeline, 138, 140
Sansonnet, Laure, 15, 66
Saracco, Jerôme, 143
Sarazin, Gabriel, 125
Sarr, Djibril , 120
Sauliere, Guillaume, 92
Schipman, Julien, 92
Schmutz, Hugo, 122
Schneider, Ulrike, 98
Scornet, Erwan, 145
Sedeaud, Adrien, 93
Selmani, Merwann, 136
Senetaire, Hugo, 97
Servajean, Maximilien, 57, 111
Servien, Rémi, 65
Silva, Alonso, 76
Slaoui, Yousri, 97, 130, 144
Smits, Nathalie, 18
Sokolovska, Nataliya, 80, 141
Sportisse, Aude, 147
Spychala, Cécile, 15
Stepaniants, George, 81
Stoltz, Gilles, 128
Stupfler, Gilles, 109
Sultanov, Arsen, 141
Susmann, Herbert, 147

Taillardat, Maxime, 115
Tamo Tchomgui, Jean Steve, 65
Tardivel, Patrick, 62, 98
Ternès, Nils, 47
Thiébaut, Rodolphe, 46
Thomas-Agnan, Christine, 24
Todo, William, 136
Torres, Isabel, 138
Toussaint, Jean-François, 92, 93

Trottier, Catherine, 28
Tzourio, Christophe, 61

Usevich, Konstantin, 135

Vaiter, Samuel, 35
Valentini, Pasquale, 88
Vandewalle, Vincent, 19
Varoquaux, Gaël, 126
Varron, Davit, 25
Vayatis, Nicolas, 67
Vazquez, Emmanuel, 96
Verbanck, Marie, 34
Vialaneix, Nathalie, 65
Viallon, Vivian, 81
Vicari, Donatella, 89
Viguier-Pla, Sylvie, 59, 132
Vo, Thanh Huan, 118

Wahl, François, 43
Werge, Nicklas, 115
Wintenberger, Olivier, 41, 109, 113
Woillard, Jean-Baptiste, 60
Wolinski, Pierre, 142

Yao, Anne-Françoise, 19

Zaffran, Margaux, 77
Zaouche, Mounia, 17
Zaoui, Ahmed, 32
Zhang, Li-Chun , 99
Zhao, Pan, 63
Zhu, Wencan, 47

Sponsors


LES MATHS
VECTEUR D'INNOVATION

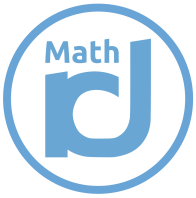

BIOMÉRIEUX


cea


ÉCOLE
CENTRALELYON


ENS DE LYON


eric


Math


Google


Inria


IXXI


ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES


LIMOS


Lip


Lyon 1


UNIVERSITÉ
LUMIÈRE
LYON 2


LABEX
MILYON
UNIVERSITÉ DE LYON


MINES
Saint-Étienne


La Région
Auvergne-Rhône-Alpes


UMPA
ENS DE LYON