6ÈMES RENCONTRES DE STATISTIQUE

SCIENCE DES DONNÉES

9-10 DÉCEMBRE 2021

Faculté Sciences & Sciences de l'Ingénieur

Campus de Tohannic - VANNES

Université Bretagne Sud

Faculté sciences & sciences de l'ingénieur
Université Bretagne Sud

LMBA
Laboratoire de Mathématiques de Bretagne Atlantique
UMR6205 - CNRS/UBO/UBS

www.univ-ubs.fr

COMITÉS D'ORGANISATION ET SCIENTIFIQUE

Présidents :
Gilles Durrieu et **Ion Grama**

Arlette Antoni
Thierry Dhorne
Evans Gouno
Salim Lardjane
Audrey Poterie
François Septier
Jean-Marie Tricot

CONFÉRENCIERS ET INTERVENANTS INVITÉS

Paul Doukhan,
CY Cergy Paris Université

Jean-François Dupuy,
INSA Rennes

Gilles Durrieu,
UBS Vannes

Emmanuel Frenod,
UBS Vannes et Seed

Robin Genuer,
Université de Bordeaux

Antoine Girard,
Bordeaux

Kévin Jaunâtre,
IFPEN Paris

Salim Lardjane,
UBS Vannes

Federico Maddanu,
University of Rome et CY Cergy Paris Université

Audrey Poterie,
UBS Vannes

Lionel Truquet,
ENSAI Rennes

Jonathan Villain,
Université Gustave Eiffel, Lille

JEUDI 9 DÉCEMBRE 2021

10h15

Accueil des participants et café dans l'amphi Yves Coppens, Université Bretagne Sud, 573 Rue André Lwoff, 56000 Vannes.

10h45-11h

Introduction

11h

Lionel Truquet, ENSAI Campus de Ker Lann, Rennes
(45 mn + questions)

Titre :

Multivariate time series models for mixed data

Résumé :

We introduce a general approach which unifies some previous attempts for modeling the dynamic of multivariate time series or for regression analysis when the data are of mixed type (binary/count/continuous). Our approach is quite flexible since conditionally on past values, each coordinate at time t can have a distribution compatible with a standard univariate time series model such as GARCH, ARMA, INGARCH or logistic models whereas past values of the other coordinates play the role of exogenous covariates in the dynamic. The simultaneous dependence in the multivariate time series can be modeled with a copula. Additional exogenous covariates are also allowed in the dynamic. We first study some usual stability properties of these models and then show that autoregressive parameters can be consistently estimated equation-by-equation using a pseudo-maximum likelihood method, leading to a fast implementation even when the number of time series is large. Moreover, we prove consistency results when a parametric copula model is fitted to the time series and in the case of Gaussian copulas, we show that the likelihood estimator of the correlation matrix is strongly consistent. We carefully check all our assumptions for two prototypical examples : a GARCH/INGARCH model and logistic/log-linear INGARCH model. Our results are illustrated with numerical experiments as well as two real data sets.

12h-14h

Déjeuner au Tableau (https://www.au-tableau.com/)

14h-15h

Jean-François Dupuy ,*INSA, Rennes* (45 mn + questions)

Titre :

Données de comptage avec censure et données manquantes

Résumé :

Le modèle de régression de Poisson est très utilisé en assurance (modélisation du nombre de sinistres dans un portefeuille d'assurés), en surveillance épidémiologique (suivi de la mortalité pour une cause donnée)... où il permet de modéliser des observations se présentant sous la forme de comptages, ou dénombrements. Dans cet exposé, nous nous intéressons au cas où ces comptages peuvent être «censurés» (par exemple, si on modélise la consommation individuelle quotidienne de fruits et légumes, et que l'on relève cette consommation au travers d'un questionnaire contenant la modalité «Je consomme au moins 5 fruits et légumes par jour», la donnée recueillie est dite censurée à droite : on sait que le nombre de fruits et légumes consommés est au moins égal à 5). Nous considérons de plus le cas où l'information sur la censure n'est pas connue précisément (mauvais recueil des données, perte d'information...). Par exemple, on a recueilli le nombre «5» pour la consommation de fruits et légumes d'une personne interrogée, mais on ne sait pas si elle a consommé exactement 5 fruits et légumes ou s'il s'agit du nombre minimum consommés. Ne pas prendre en compte toutes ces incertitudes dans l'analyse statistique peut engendrer des biais et par suite, une mauvaise appréhension du phénomène étudié. Nous proposons donc une méthode d'estimation dans le modèle de Poisson, adaptée à ce cadre. Elle est basée sur l'imputation multiple. Nous en expliquons le principe, puis montrons qu'elle fournit des estimations de bonne qualité, en particulier, au moyen d'une étude de simulations.

15h-16h

Paul Doukhan, *CY Cergy Paris Université* (45 mn + questions)

Titre :

Le projet EcoDep

Résumé :

Ce projet financé par l'institut d'études avancé de l'université CY de Cergy Pontoise a pour objectif la mise en place d'outils mathématiques et statistiques permettant d'appréhender la dynamique des populations animales ou végétales au fil du temps. Précisément des évolutions au fil du temps nécessitent la clarification de modèles dynamiques. Tout d'abord des modèles de type modèles de vie ou de mort ou d'autres modèles probabilistes adhoc. Lorsque la modélisation des abondances (qui quantifie les volumes d'espèces au fil du temps) est nécessaire, on prendre en compte les attendus physiques de leur évolution. Une fois que de tels modèles sont validés théoriquement, des techniques de statistiques permettent leur validation empirique. D'autres quantités comme des indicateurs naturels de ces évolutions sont envisagées, par exemple la loi de Taylor. Des questions analogues se posent aussi dans ce cas. Le projet EcoDep accueille une trentaine de chercheurs venant de nombreux pays, USA, chili, Japón, Allemagne, UK, Allemagne, ainsi que de nombreux laboratoires francais, Rennes, vannes, Lyon, Paris, mais son noyau est localisé à Cergy, dans les laboratoires de mathématiques, physique, et économie. La variété des chercheurs et leurs compétences permet ainsi de considérer des points de vue variés. A titre d'exemple des chercheurs participant à ce projet on déjà questionné des interrogations terrifiantes comme le nombre de vie que peut supporter notre planète ou l'estimation de la fin de l'humanité dans les 25 prochaines années. Dans le souci d'unifier les objectifs des réunions régulières, des conférences, et un séminaire sont appuyés par des enseignements adaptés. L'exposé a pour but de montrer au travers du site de ce projet que nous avons les moyens de notre ambition. Les différentes tâches seront expliquées et les résultats déjà obtenus seront détaillés. Notre époque si chaotique a un besoin urgent d'actions de fonds dans le domaine vital de l'écologie, nous espérons que ce projet pourra donner un soutien à tous les efforts visant à notre simple survie.

16h-16h15

Pause café

16h15-17h

Témoignages d'anciens étudiants (Antoine Girard, *Data analyst freelance*, Kévin Jaunâtre, *IFPEN, Paris*, Jonathan Villain, *Université Gustave Eiffel, Lille*)

17h-18h

Conférence-débat sur le thème des Data Science

19h30

Dîner au Piano barge (http://www.pianobarge.com/Page/Accueil)

VENDREDI 10 DÉCEMBRE 2021

Ouverture de la journée

Exposés de chercheurs en statistique de l'UBS

8h30

Salim Lardjane, *LMBA UBS Vannes*

Titre :

L'intuition mathématique

Résumé :

Depuis les travaux séminaux de Poincaré et Hadamard sur la genèse de l'invention mathématique, l'intuition mathématique n'a cessé d'intéresser mathématiciens et psychologues. Dans cet exposé, je croiserai différentes approches historiques de celleci avec des développements récents sur la modélisation de l'esprit humain et le rôle de l'intuition dans les processus créatifs.

9h00

Victor Watson, *LMBA UBS Vannes et CEA-DAM île-de-France*

Titre :

Détection séquentielle d'une perturbation temporaire dans une série temporelle multivariée.

Résumé :

Cette présentation porte sur la détection de signaux faibles par un champ de capteurs dans un environnement bruité. Il s'agit là de trouver, grâces à des hypothèses sur les statistiques du signal et du bruit, une méthode récursive de détection. Une nouvelle méthode visant à rendre détectables les signaux temporaires sera introduite. Cette dernière est dérivée de la méthode de somme cumulée (CUSUM) qui permet de calculer de manière récursive le rapport de vraisemblance des données. L'adaptation au cas multivarié avec un critère adaptatif d'agrégation des statistiques de test locales et la prise en compte de l'exposition temporaire des capteurs issue de la propagation de l'évènement conduisent à l'introduction de la méthode TE-CUSUM pour Temporary-Event-CUSUM. La présentation montrera par une étude comparative, l'efficacité de la méthode sur des exemples réalistes de cas de faibles niveaux de pollution se propageant dans un champ de capteurs.

9h30

Audrey Poterie, *LMBA UBS Vannes*

Titre :

Approches par arbres de décision et forêts aléatoires pour données structurées.

Résumé :

L'apprentissage supervisé consiste à expliquer et/ou prédire une sortie y par des entrées x. Dans de nombreux problèmes, les entrées x ont une structure de groupes connue et/ou clairement identifiable. Le regroupement des variables peut être naturel ou bien défini dans le but de modéliser les relations entre les différentes variables. Par exemple, en biologie, lorsque l'on souhaite étudier la composition chimique d'un sérum à l'aide de la spectrométrie de masse, les variables explicatives, de nature fonctionnelle, peuvent être divisées en groupes représentants différentes parties de la courbe. Dans ce contexte, l'élaboration d'une règle de prédiction prenant en compte cette structure de groupes peut se révéler plus pertinente qu'un algorithme effectué sur les variables individuelles tant au niveau des performances prédictives que de l'interprétation. Des algorithmes supervisés construits sur des groupes de variables ont déjà été proposés. Un des plus connus est certainement le Group lasso. Dans cette présentation, je vous présenterai des méthodes d'arbres de décision et de forêts aléatoires qui ont été développées dans le cadre des variables explicatives ayant une structure de groupes. Je parlerai aussi brièvement de deux projets en cours portant sur le développement de nouveaux algorithmes de forêts aléatoires dans deux cadres particuliers.

10h00

Gilles Durrieu, *LMBA UBS Vannes*

Titre :

A nonparametric statistical procedure for the detection of marine pollution and global warming effects.

Résumé :

This talk is devoted to the estimation of the derivative of the regression function in fixed and random design nonparametric regression. We establish the almost sure convergence as well as the asymptotic normality of our estimates. We provide concentration inequalities which are useful for small sample sizes. We also illustrate our nonparametric estimation procedure on simulated data and real life data associated with sea shores water quality and global warming.

10h30

Emmanuel Frenod, *LMBA UBS Vannes et Seed*

Titre :

Intelligence Artificielle pour les Entreprises

Résumé :

Dans cet exposé, je donnerai des exemples de réalisations d'Intelligences Artificielles (IA) qui ont été développées pour les besoins d'entreprises. Plusieurs de ces IA font appel à du Couplage Modèle-Données. La modélisation mathématique et statistique permet d'embarquer de la connaissance issue des experts des entreprises pour qui elles sont développées. Puis de l'apprentissage amène à y intégrer à partir des historiques de données de la connaissance complémentaire. J'expliquerai des aspects scientifiques et techniques de ces constructions. J'aborderai également comment ces projets sont développés par les Data Scientists de See-d ; en particulier avec des collaborations avec des chercheurs du LMBA.

11h-12h

Robin Genuer, *Université de Bordeaux*

Titre :

Fréchet random forests for metric space valued regression with non euclidean predictors

Résumé :

Random forests are a statistical learning method widely used in many areas of scientific research essentially for its ability to learn complex relationships between input and output variables and also its capacity to handle high-dimensional data. However, current random forest approaches are not flexible enough to handle heterogeneous data such as curves, images and shapes. In this talk, we present Fréchet trees and Fréchet random forests, which allow to manage data for which input and output variables take values in general metric spaces. To this end, a new way of splitting the nodes of trees is introduced and the prediction procedures of trees and forests are generalized. Then, random forests out-of-bag error and variable importance scores are naturally adapted. The method is illustrated through several simulation scenarios on heterogeneous data combining longitudinal, image and scalar data. Finally, a real dataset from an HIV vaccine trial is analyzed with the proposed method.

12h-14h

Déjeuner au Tableau (https://www.au-tableau.com/)

14h-15h

Federico Maddanu, *University of Rome Tor Vergata, Italie et CY Cergy Paris Université*. Joint work with Tommaso Proietti, *University of Rome*.

Titre :

Modelling Persistent Cycles in Solar Activity

Résumé :

Solar activity at decadal time scales is characterized by persistent periodic patterns with global effects on the Earth's climate. The paper deals with the analysis and prediction of the revised monthly sunspots numbers, adopting a recently proposed time series model for long-range dependent cycles. Learning is based on maximum likelihood and optimal signal extraction filters are available for cycle estimation and prediction. The analysis suggests the presence of stationary long memory in the sunspots generating process. Moreover, our formulation provides a reliable method for solar cycles predictions, yielding forecasts of the oncoming 25th cycle. In particular, we claim a main peak in early 2024 with amplitude 114 and the ending of the cycle in early 2030.

15h-16h

Jonathan Villain, *Université Gustave Eiffel, Lille* (45 mn + questions)

Titre :

Apprentissage statistique et sécurité des communications sans fils

Résumé :

Les communications par ondes radiofréquences sont de plus en plus utilisées pour la flexibilité de connexion qu'elles apportent. En effet, qu'il s'agisse de réseaux satellites, de réseaux cellulaires, de réseaux Wi-Fi, de réseaux Bluetooth ou de différents protocoles avec des débits inférieurs tels que LoRaWan ou SigFox, les communications sont établies par ondes radio. Les technologies de communication sans fil continuent de progresser rapidement et leur utilisation est de plus en plus courante comme la surveillance des frontières, les applications de soins de santé, la surveillance de l'environnement, le renseignement domestique, etc. Cependant, des défis restent à relever en matière de couverture et de déploiement, d'évolutivité, de qualité de service, de taille, de puissance de calcul, d'efficacité énergétique et de sécurité. Les risques d'une mauvaise protection d'un réseau sans fil sont divers. Il est possible d'interagir avec les communications que ce soit pour brouiller la vidéosurveillance, interrompre les appels, récupérer des informations personnelles ou s'introduire dans un système sécurisé. Dans ce contexte, l'utilisation d'algorithme d'apprentissage est de plus en plus courante et a pour but d'automatiser les systèmes permettant de sécuriser les communications. Ils sont notamment utilisés dans le but de détecter des anomalies dans les communications et la localisation de source de communication illicite.

• CLÔTURE DES 6ÈMES RENCONTRES •