Bienvenue !

Le comité d'organisation et le comité scientifique des 51^{es} Journées de Statistique sont heureux de vous accueillir sur le site de la Faculté des Sciences et Technologies à Vandoeuvre-lès-Nancy. Cette importante manifestation scientifique, congrès annuel de la Société Française de Statistique, est organisée cette année par l'Université de Lorraine en collaboration avec le CNRS et INRIA.

Ce livret contient le programme de l'événement, ainsi que les résumés des conférences plénières et des communications sélectionnées par le comité scientifique. Nous sommes reconnaissants envers tous ses membres : ils ont fait leur maximum pour proposer à tous les participants un programme riche et du meilleur niveau, mêlant la grande tradition statistique avec les thématiques suscitant l'actuel engouement de notre communauté mais aussi des attentes fortes de la société. Puissent les Journées de Statistique aider à leur nécessaire enrichissement mutuel ! Cette variété et cette exigence se reflètent autant dans les 17 conférences plénières que dans les 52 sessions au programme de cette édition.

La préparation de ce congrès a été portée par l'Institut Elie Cartan de Lorraine, laboratoire de rattachement de la plupart des membres du comité d'organisation. Elle a aussi bénéficié du soutien financier, humain et logistique de l'IUT Nancy-Charlemagne et se déroule dans les locaux de la Faculté des Sciences et Technologies. Par ailleurs, l'événement n'aurait pu avoir lieu sans le soutien de tous nos autres partenaires et sponsors, ni sans l'expertise de nos webmasters dont Vincent Brault qui a fait preuve d'une patience infinie. Nous adressons à tous nos vifs remerciements.

Nous vous souhaitons une semaine très fructueuse sur le plan scientifique et un séjour agréable dans l'agglomération nancéienne.

Aurélien Garivier	Angelo Efoévi Koudou
Président du comité scientifique	Président du comité d'organisation

Comité scientifique

Le comité scientifique des JdS 2019 est présidé par Aurélien Garivier (ENS Lyon). Il est composé de :
— Stéphane Boucheron (Université Paris VII)
— Marianne Clausel (Université de Lorraine)
— Nicolas Bousquet (Quantmetry & Sorbonne Université)
— Benjamin Guedj (Inria)
— Arnaud Guyader (Université Paris VI)
— Pierre Latouche (Université Paris I)
— Christophe Ley (Université de Ghent)
— Anne Philippe (Université de Nantes)
— Mélanie Prague (Inria)
— Joseph Salmon (Université de Montpellier)
— Nicolas Savy (Université Toulouse III)
— Christine Thomas-Agnan (Toulouse School of Economics, Université Toulouse I)
— Yves Tillé (Université de Neuchâtel)

Comité d'organisation

Président
Angelo Efoévi Koudou (Université de Lorraine)
Vice-présidente
Anne Gégout-Petit (Université de Lorraine / Inria)
Trésorière
Sophie Mézières (Université de Lorraine)
Trésorier adjoint
Pascal Wild (INRS)
Site Web
Bruno Scherrer (Inria)
Coralie Fritsch (Inria)
Communication
Sandie Ferrigno (Université de Lorraine)
Romain Azaïs (Inria)
Coralie Fritsch (Inria)
Matériel, logistique, activités sociales, subventions
Eliane Albuisson (Université de Lorraine/CHRU)
Nathalie Benito (Université de Lorraine)
Christelle Breuils (Université de Lorraine)
Madalina Deaconu (Inria)
Aurélie Gueudin (Université de Lorraine)
Clémence Karmann (Université de Lorraine)
Armand Maul (Université de Lorraine)
Jean-Marie Monnez (Université de Lorraine)
Pascal Moyal (Université de Lorraine)
Joseph Ngatchou-Wandji (Université de Lorraine)
Radu Stoica (Université de Lorraine)
Pierre Vallois (Université de Lorraine)

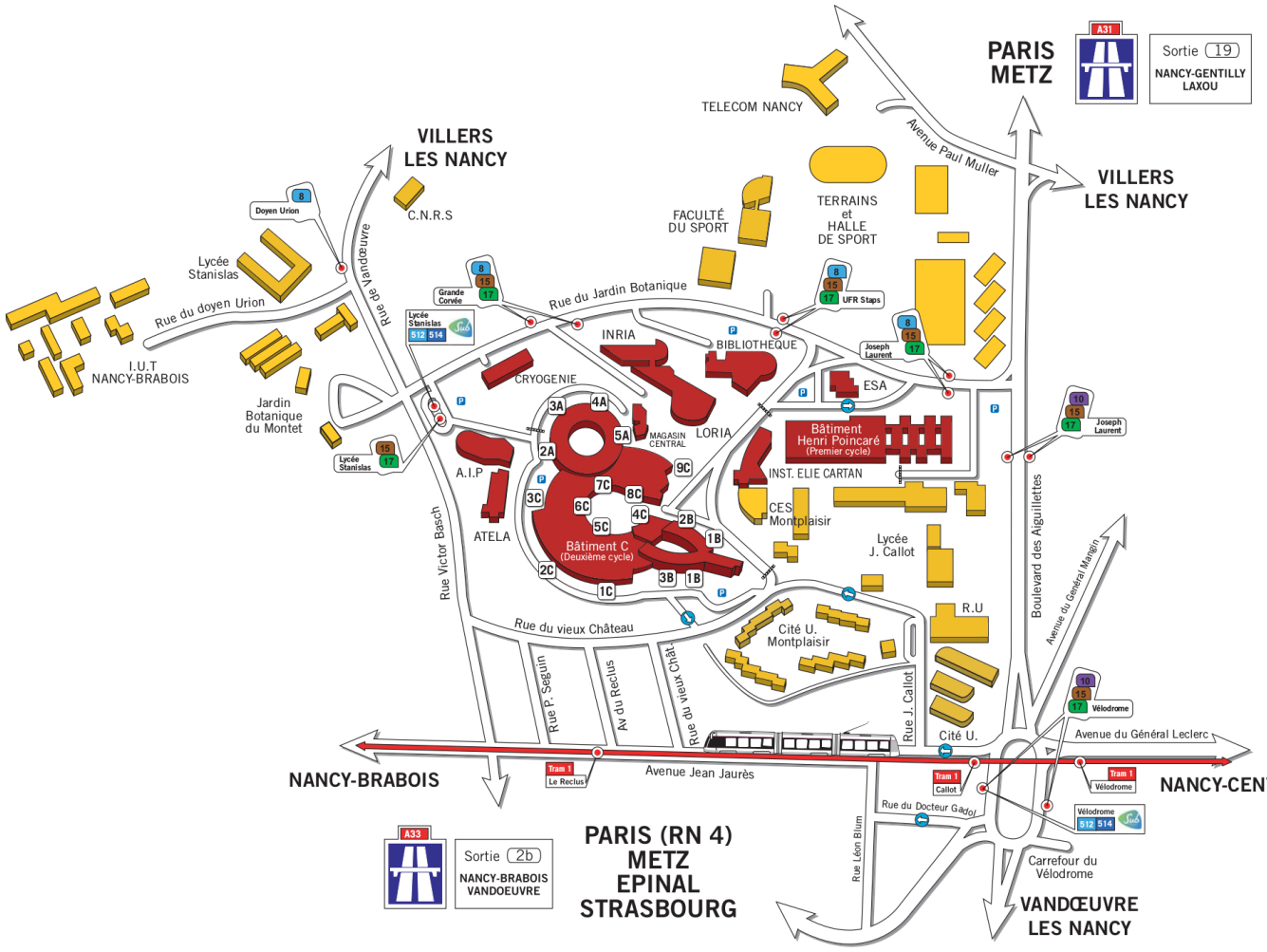
avec la complicité permanente de Vincent Brault, webmestre de la SFdS.

Table des matières

Informations pratiques

Pour les exposés : Les auteurs doivent avoir leur présentation en format pdf sur une clé USB. Merci de venir au cours de la pause café qui précède la session et/ou 10 minutes en avance dans la salle pour installer la présentation sur le PC de l'amphi. Chaque présentation dure 20 mn questions comprises, et c'est le modérateur qui définit le temps imparti entre présentation et questions.

Pour se repérer : Voici ci-dessous un plan du campus de la Faculté des Sciences et Technologies (F.S.T.). La conférence a lieu dans les amphis 11 à 16 du **Bâtiment Henri Poincaré (Premier cycle)**. Les repas sont servis au Restaurant Universitaire (**R.U.**) en échange des tickets qui vous ont été remis à l'accueil. Deux salles (salle 40 et 'Wifi') sont en accès libre.



Connexion Wi-Fi : Accès 'invité' au réseau wifi de l'Université (SSID : Université de Lorraine)
Login : jds2019
Mot de passe : iecl!2019

Evénements satellites

Lundi 3 juin

Réunion relations internationales

Lieu : salle 40 (même étage que les amphis, couloir de droite).
Horaire : 13h-14h

Conseil de la SFdS

Lieu : Salle Döblin, IECL
Horaire : 13h-14h

Mardi 4 juin

Groupe jeunes

Lieu : salle 40
Horaire : 13h-14h30

Jeudi 6 juin

Déjeuner scientifique 1, Causalité

Lieu : salle 40
Horaire : 13h15-14h30

Déjeuner scientifique 2, Enseignement

Lieu : salle Wifi (à côté de salle 40)
Horaire : 13h15-14h30

Rencontres jeunes statisticiens

Lieu : Bar le Pinochio, place Saint-Epvre
Horaire : 18h30

Programme social

Cocktail de bienvenue (lundi 3 juin, 19h-21h)

Le cocktail de bienvenue a lieu à l'Hôtel de Ville situé place Stanislas (10-15 minutes à pied depuis la gare / arrêt de tram Cathédrale ou Point Central).

Activités (mercredi 5 juin, 14h-18h)

Pour les horaires et lieux de départ, voir le site web de la conférence.

1. **Nancy XVIIIème. Centre historique et place d'Alliance**
   La visite débutera par les places Stanislas, Carrière et d'Alliance dans leur écrin architectural XVIIIème. Puis, après un passage par « le Kiosque» et « le Rodin » du Parc de la Pépinière, le charme de la Vieille Ville, médiévale et renaissance, vous apparaîtra dans l'un ou l'autre des hauts lieux de l'histoire de la Lorraine : Palais Ducal, Hôtels particuliers, Porte de la Craffe.
   Contacts : Sandie Ferrigno (06 20 03 62 46), Aurélie Gueudin (06 86 49 00 89)
2. **Nancy XVIIIème. UNESCO et musée des Beaux-Arts**
   L'ensemble architectural XVIIIème inscrit depuis 1983 au patrimoine mondial de l'Unesco vous sera présenté de la place Stanislas à la Place d'Alliance en passant par la Place Carrière et le parc de la Pépinière. La visite se poursuivra par la présentation des peintres Lorrains du musée des Beaux-Arts et vous laissera la possibilité de découvrir vous-même, si vous le souhaitez, le restant des richesses du musée (en particulier les pièces de la collection DAUM).
   Contact : Sophie Mézières (06 07 36 01 81), Pascal Wild (06 26 06 08 70)
3. **Nancy, Art Nouveau et musée des Beaux Arts**
   De la place Maginot à la place Stanislas, vous découvrirez comment les progrès technologiques, l'excellence des arts-décoratifs et de l'artisanat ont été mis à profit par « l'école de Nancy ». Vous explorerez en particulier quelques lieux de la vie économique des années 1900 comme des magasins, des banques ou encore des brasseries. La visite se poursuivra par la présentation au musée des Beaux-Arts du peintre Emile Friant et des collections des verreries Daum.
   Contact : Zhanhao Liu (07 60 81 39 64)
4. **Nancy, Grand tour, Art Nouveau, Art déco et musée de l'école de Nancy**
   Nancy est internationalement connue pour être la capitale de l'Art nouveau en France. Elle l'est beaucoup moins pour son patrimoine Art Déco. Cette balade commencera avec la rue Félix Faure. Les maisons, remarquables pour l'ensemble de leurs façades, permettront de se faire une idée de la différence entre les deux styles. Puis le jardin du Musée de l'« école de Nancy » et une petite incursion dans le musée devraient vous surprendre. Avant que la « Villa Majorelle » ne vous impressionne, merveille de l'Art Nouveau dans un quartier entier très ... art déco ! Vous finirez avec d'autres très beaux ensembles architecturaux du quartier de la gare.
   Contact : Christiane Wild (06 22 86 29 93)
5. **Mine de Neuves-Maisons (Val de Fer) et musée de l'Histoire du Fer (Jarville)**
   La mine de Neuves-Maisons est située au cœur de la deuxième réserve mondiale de minerai de fer. Elle présente les conditions de travail dans des galeries datant des années 1870 à 1930 où l'on multipliait l'extraction du fer pour créer les chemins de fer partout à travers le monde, les structures en fer pour le bâtiment, les ouvrages d'art comme le pont Alexandre III à Paris, et la Tour Eiffel, phare de la technologie de l'époque. Nous proposons de vous faire découvrir la

mine et son environnement au travers d'une visite qui parcourt les anciennes galeries et permet de comprendre les conditions de l'exploitation d'une mine de fer en pénétrant au cœur des galeries dans les mêmes conditions que les travailleurs de l'époque. Les conditions de la mine (12 degrés humides toute l'année) demandent de prévoir un équipement chaud et des chaussures fermées. L'après-midi se poursuivra avec la visite du musée de l'Histoire du Fer, consacré à l'utilisation du fer et de ses dérivés, depuis le Moyen Âge et l'apparition du haut-fourneau jusqu'au XXème siècle. Le musée accueille également un espace dédié à Jean Prouvé, grand innovateur dans le domaine de l'architecture métallique.
Contact : Bruno Scherrer (06 98 73 73 47)

6. **Maison de la mirabelle – distillerie**
   La famille Grallet-Dupic vous accueille à la Maison de la Mirabelle pour vous présenter un parcours-découverte guidé de la Mirabelle de Lorraine. La distillerie est située à Rozelieure, entre Nancy et Epinal. Un parcours pour découvrir l'histoire, la vie et la conduite du mirabellier en Lorraine. La visite se termine par une dégustation !
   Contact : Coralie Fritsch (06 17 14 15 45)
7. **Randonnée sportive sur les hauts de Nancy**
   Pour ceux qui veulent se dégourdir les jambes, départ du lieu de la conférence, montée et pique-nique au parc de Brabois, traversée du plateau jusqu'en forêt de Haye, descente vers la ville par les sentiers des jardins ouvriers de Laxou. Retour au centre-ville. 2h30-3h de marche suivant le rythme du groupe. Prévoir chaussures adaptées à la marche et petit sac confortable.
   Contact : Anne Gégout-Petit (06 62 64 97 54)

Dîner de gala (mercredi 5 juin, 19h30-1h)

Le dîner de gala a lieu au Palais du Gouvernement situé *49, Place du Général de Gaulle, Nancy*, à 5 minutes à pied de la place Stanislas.

Lundi 3 juin

8h30-9h45 : Accueil des participants

9h45-10h10 : Ouverture des journées

10h10-11h10

Luc Devroye (Prix Laplace) (Amphi 11, retransmis en Amphi 12) :
Apprendre un waterzooï de distributions

Université McGill, Montréal

Mots clefs : apprentissage automatique, estimation d'une densité, modèle d'Ising, distribution gaussienne, distributions sur les graphes, minimax, variation totale, complexité.

Modération : Pascal MASSART

11h10-11h30 : Pause café

11h30-12h50

Valeurs extrêmes, événements rares (Amphi 11)

Modération : Anne-Laure FOUGERES

Prédiction d'évènements rares : une revue de la littérature

Mathieu BERTHE (Laboratoire de Mathématiques- UMR CNRS 6620)
Pierre DRUILHET (Laboratoire de Mathématiques- UMR CNRS 6620)
Stéphanie LEGER (Laboratoire de Mathématiques- UMR CNRS 6620)
Olivier BRACHET (IPA (Innovation, Performance, Analytics))
La prediction d'évènements rares est l'un des challenges actuels de la data science. Les jeux de données déséquilibrées d´egradent la performance des techniques de modélisation habituelles et rendent les outils de comparaison de modèles inefficaces et biaisés. Pour pallier ces problèmes, de nombreuses méthodes ont été d´eveloppées, que ce soit des méthodes dites de sampling (Oversampling, undersampling, SMOTE...), des améliorations et corrections de biais sur des méthodes existantes (régression logistique pour évènements rares) ou encore des méthodes dites ensemblistes (Bagging, boosting). Nous proposons de présenter, grâce à une revue de la littérature, les avantages et les inconvénients de chaque méthode. A l'aide de simulations, nous comparerons les méthodes ' afin de mettre enévidence les meilleures d'entre elles. Ceci nous permettra également de proposer des outils de comparaison plus adaptés aux jeux de données déséquilibrés, par rapport à ceux couramment utilisés.

Étude de l'erreur relative d'extrapolation associée à l'estimateur de Weissman pour les quantiles extrêmes

Clément ALBERT (Polytechnique)
Anne DUTFOY (EDF R&D)
Stéphane GIRARD (Inria)
Nous étudions le comportement asymptotique de l'erreur d'extrapolation (relative) associée à l'estimateur de Weissman, un estimateur semi-paramétrique des quantiles extrêmes dédié au domaine d'attraction de Fréchet. Des conditions sont alors fournies de telle sorte que l'erreur tende vers zéro quand la taille de l'échantillon augmente. Nous montrons que, dans le cas où la loi appartient au domaine d'attraction de Fréchet, sans surprise, l'erreur d'extrapolation relative tend vers zéro sous des conditions très faibles sur l'ordre du quantile. En revanche, de manière originale, nous montrons que l'erreur d'extrapolation tend vers zéro pour deux types de lois du domaine d'attraction de Gumbel sous des conditions raisonnables sur l'ordre du quantile. Mieux encore, des équivalents de l'erreur sont établis montrant que l'estimateur de Weissman mène à des erreurs d'extrapolation plus faibles que l'estimateur Exponential Tail pour certains types de loi du domaine d'attraction de Gumbel. Ces résultats sont illustrés numériquement.

Recursive estimator for extreme value index defined by stochastic approximation method

Fatma BEN KHADHER (Université de Monastir, Tunisie.)
Yousri SLAOUI (Université de Poitiers, France.)
The aim is to apply the stochastic approximation method to define a class of recursive kernel estimator of the conditional extreme value index. Then, we study the properties of this recursive estimator and compare them with the non-recursive kernel Hill's estimator. We show that using some optimal parameters, the proposed recursive estimator defined by the stochastic approximation algorithm, will be very competitive to the non-recursive kernel Hill's estimator. Finally, simulations are done to corroborate the obtained theoretical results.

Méthode de couplage en distance de Wasserstein pour la théorie des valeurs extrême.

Benjamin BOBBIA (laboratoire de mathématiques de Besançon, université de Franche-Comté)
Clément DOMBRY (laboratoire de mathématiques de Besançon, université de Franche-Comté)
Davit VARRON (laboratoire de mathématiques de Besançon, université de Franche-Comté)
Nous proposons une relecture de résultats classiques de la théorie des valeurs extrêmes, que nous étudions grâce aux outils que nous fournit la théorie du transport optimal. Dans ce cadre, nous pouvons voir la normalité des estimateurs comme une convergence de mesures dans un espace métrique muni de la distance de Wasserstein. Il s'agit d'une approche par couplage. Soient $(X_1, ..., X_n)$ et $(X_1^*, ..., X_n^*)$ deux

échantillons i.i.d, nous nous intéressons aux relations qui lient la distance de Wasserstein entre les mesures empiriques sur les échantillons et la distance de Wasserstein entre les lois qui ont généréces échantillons. Ce résultat nous permet de redémontrer la normalité du célèbre estimateur de Hill et de donner une vitesse de convergence.

Apprentissage statistique et statistique mathématique (Amphi 12)

Modération : Stéphane BOUCHERON

Bornes d'excès de risque asymptotiquement minimax pour l'estimation de densité et la régression logistique mal spécifiées

Jaouad MOURTADA (École polytechnique)
Stéphane GAIFFAS (Université Paris Diderot)
Erwan SCORNET (École polytechnique)
Nous introduisons un nouvelle procédure pour l'estimation de densité conditionnelle, qui satisfait une borne générale d'excès de risque prédictif pour la perte logarithmique. Cette borne reste valable dans le cas mal spécifié, et est d'ordre d/n dans de nombreux cas, où d est la dimension du modèle et n la taille de l'échantillon. En particulier, cette procédure est robuste au cas mal spécifié, contrairement à l'estimateur du maximum de vraisemblance qui y est sensible. Nous en déduisons une procédure minimax au premier ordre pour l'estimation de densité dans les familles exponentielles et la régression logistique mal spécifiées, avec un excès de risque asymptotique de $d/(2n) + o(1/n)$. Cette approche produit des estimateurs plus efficaces que celles utilisant des bornes sur le risque cumulées, et éliminent des facteurs logarithmiques superflus. Dans de nombreux cas (incluant la régression logistique), ces estimateurs sont calculables explicitement.

Apprentissage d'un classifieur minimax pour données discrètes

Lionel FILLATRE (Université Côte d'Azur, Laboratoire I3S)
Cyprien GILET (Université Côte d'Azur, Laboratoire I3S)
Susana BARBOSA (Université Côte d'Azur, Laboratoire IPMC)
L'apprentissage d'un classifieur supervisé lorsque les proportions par classe de la base d'apprentissage diffèrent de celles de la base de test, ou lorsque que les données sont non-balancées, peut augmenter le risque d'erreurs de classification pour de nouvelles observations. Nous nous intéressons ici au classifieur de Bayes non-naïf lorsque les variables prédictives sont discrètes ou discrétisées. Nous montrons que, sous ces conditions, le risque de Bayes, considéré comme une fonction des proportions des classes, est concave, non-différentiable et affine par morceaux. Nous proposons un algorithme de sous-gradient projeté permettant d'estimer les proportions qui maximisent ce risque de Bayes. Le classifieur minimax obtenu minimise le risque conditionnel maximum.

Arbres et forêts aléatoires de Fréchet

Louis CAPITAINE (Université de Bordeaux)
Robin GENUER (Université de Bordeaux)
Les forêts aléatoires sont une méthode d'apprentissage statistique très largement utilisée dans de très nombreux domaines de recherche scientifique tant pour sa capacité à décrire des relations complexes entre des variables explicatives et une variable réponse que pour sa capacité à traiter des données de très grande dimension. Cependant, avec l'emergence de nouvelles techniques d'acquisition de données, nous avons accès à des données de plus en plus complexes, des images, des formes, des données longitudinales,

des courbes et la méthode des forêts aléatoire n'est pas toujours adaptée à ces nouvelles entrées. Dans ce travail nous introduisons la notion de forêts aléatoires de Fréchet, qui permet d'apprendre des relations entre des variables de natures diverses dans des espaces métriques non ordonnés, et ce, même dans un cadre de grande dimension. Nous décrivons une nouvelle manière de découper les noeuds des arbres constituant notre forêt de Fréchet puis nous détaillons la procédure de prédiction pour une variable explicative à valeurs dans un espace non euclidien. Nous utilisons la structure des forêts afin de calculer une l'importance des variables constituant notre échantillon d'apprentissage. Nous terminons avec un exemple d'utilisation de nos Forêts de Fréchet pour la régression entre courbes à partir de leurs formes et nous donnons quelques simulations dans ce cadre.

Combinaison de la classification de variables et de la sélection de variables par forêts aléatoires

Robin GENUER (Inserm U-1219, ISPED, Université de Bordeaux, Inria Bordeaux Sud-Ouest)
Marie CHAVENT (Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Université de Bordeaux, Inria Bordeaux Sud-Ouest)
Jérôme SARACCO (Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Bordeaux INP, Inria Bordeaux Sud-Ouest)
Les approches standard pour aborder la classification supervisée en grande dimension font souvent intervenir une sélection de variables et/ou une réduction de la dimension. La méthodologie proposée dans ce travail combine la classification de variables et la sélection de variables. La classification hiérarchique des variables permet de construire des groupes de variables corrélées et résume chaque groupe par une variable synthétique. L'originalité est que les groupes de variables sont inconnus a priori. De plus, l'approche de classification traite à la fois des variables numériques et des variables catégorielles. Parmi toutes les partitions possibles, les variables synthétiques les plus pertinentes sont sélectionnées à l'aide d'une procédure utilisant des forêts aléatoires. Les performances numériques sont illustrées sur des ensembles de données simulées et réelles. La sélection de groupes de variables peut permettre d'améliorer les performances en prédiction et facilite l'interprétation des résultats.

Statistique des processus 1 (Amphi 13)

Modération : Céline LACAUX

Données fonctionnelles avec erreur hétéroscédastique

Steven GOLOVKINE (Renault)
Nicolas KLUTCHNIKOFF (Université Rennes2)
Valentin PATILEA (ENSAI)
Avec les récentes avancées technologiques, de plus en plus d'objets sont équipés de capteurs leur permettant, par exemple, de connaître la position d'autres objets dans son environnement. Ces capteurs fournissent un grand nombre de signaux pouvant être modélisés comme des données fonctionnelles multivariées entachées d'un bruit. Dans ce travail, nous supposons que ces données sont enregistrées avec un bruit hétéroscédastique d'échelle inconnue. Nous nous intéressons donc à l'estimation adaptatif du signal.

Application de la théorie des valeurs extrêmes et d'une approche par méta-modélisation adaptative à un problème d'optimisation sous contraintes probabilistes

Alexis COUSIN (IFPEN)
Josselin GARNIER (CMAP)

Martin GUITON (IFPEN)
Miguel MUNOZ ZUNIGA (IFPEN)
Nous nous intéressons à la résolution d'un problème d'optimisation sous contraintes probabilistes. La fonction déterministe à minimiser est linéaire et les contraintes s'expriment sous la forme de probabilités de dépassements de seuils. Ces probabilités doivent rester très faibles ce qui entraîne des efforts de calcul importants. Nous nous servirons de résultats de la théorie des valeurs extrêmes, de méta-modèles et de méthodes de Monte Carlo accélérées afin d'estimer ces probabilités avec précision, en temps raisonnable et avec une intégration pertinente dans la boucle d'optimisation. Nous appliquerons ces outils à un cas académique qui nous permettra de comparer cette approche à des méthodes existantes.

Processus ponctuels pour l'étude de l'apparition de fuites sur un réseau de distribution d'eau.

Nicolas DANTE (Universite de Lorraine, IECL)
Radu STOICA (Universite de Lorraine, IECL)
Bérengère SIXTA DUMOULIN (SEDIF)
Cet article présente les processus ponctuels sur réseaux linéiques pour la modélisation spatiale des fuites sur un réseau de distribution d'eau. Cette modélisation se fait sachant la position des capteurs de détection de fuites sur le réseau. Des statistiques exploratoires permettant la comparaison des observations avec la réalisation d'un processus poissonien stationnaire sont présentées. Ensuite un processus ponctuel de Poisson inhomogène est testé sur des données réelles. A la fin, des conclusions et des perspectives sont formulées.

Estimation d'exposants de Hurst dans un cadre stationnaire

Matthieu GARCIN (Pôle universitaire Léonard de Vinci)
Les propriétés de changement d'échelle d'une série temporelle peuvent être décrites par une statistique simple : l'exposant de Hurst. On lie aussi souvent la valeur de l'exposant de Hurst à la persistance de la série : si $H=1/2$, il n'y a pas d'autocorrélation, si $H>1/2$ la série est persistante et si $H<1/2$ elle est anti-persistante. Cependant l'interprétation de l'exposant de Hurst dépend fortement du modèle décrivant la dynamique. En particulier, le cas des modèles stationnaires est intéressant, notamment pour son application en finance. Nous présentons deux modèles fractales stationnaires dans lequel l'exposant de Hurst est impliqué : le processus d'Ornstein-Uhlenbeck fractionnaire et la transformée de Lamperti inverse du mouvement brownien fractionnaire. Nous exposons les spécificités de l'estimation de l'exposant de Hurst pour ces modèles, de même que les conséquences dans l'interprétation de ce qu'est un exposant de Hurst dans ce cas et de la manière dont une série stationnaire pourrait être prédite.

Luxembourg Society of Statistics (Amphi 14)

Modération : Nico WEYDERT

What makes Data Science different ? A discussion involving Statistics2.0 and Computational Sciences

Christophe LEY (Ghent University)
Data Science is today one of the main buzzwords, be it in industrial or academic settings. Machine learning, experimental design, data-driven modelling are all, undoubtedly, rising disciplines if one goes by the soaring number of research papers and patents appearing each year. The prospect of becoming a "Data Scientist" appeals to many. A discussion panel organised as part of the European Data Science

Conference 2016 in Luxembourg asked the question : "What makes Data Science different ?" In this talk, I give a personal and multi-facetted view on this question, from a statistics, machine learning and engineering perspective. In particular, I compare Data Science to Statistics and discuss the connection between Data Science and Computational Science.

Statistics on Cross-Border Workers in the Greater Region

Agnieszka WALCZAK (Luxembourg Institute of Socioeconomic Research)
Since 2010, the Luxembourg Institute of Socio-Economics Research (LISER) has been conducting, on behalf of the Central Bank of Luxembourg (BCL), the Household Finance and Consumption Survey among cross-border workers in Luxembourg. This survey collects information on the financial situation and behaviour of cross-border commuters, including their employment, access to banking and credit, housing decisions, education as well as consumption. It is the only source that collects detailed household level information on assets and liabilities of cross-border workers. The survey is important for Luxembourgish institutions and the community of researchers as Luxembourg's labour market is highly reliant on foreign workers from the Grande Région, which consists of Luxembourg, Wallonia and the German-speaking community in Belgium, Saarland and Rhineland-Palatinate (Germany) and Lorraine in France. This presentation will shed light on the HFCS Cross-border workers survey – its methodology and results from the three waves of the survey conducted so far.

Modélisation des dépendances dans les marchés

Simon PETITJEAN (Université du Luxembourg - LSF)
Jang SCHILTZ (Université du Luxembourg - LSF)
Capturer les dépendances des séries temporelles financières est un exercice complexe comportant de nombreuses difficultées techniques aussi bien du point de vue théorique que du point de vue pratique. Nous developpons une méthodologie qui permet de modéliser les portefeuilles financiers basée sur les arbres couvrant de poids minimal couplés avec des modèles économétriques, nous permettant de résoudre un grande partie de ces difficultées. Nous utilisons une tracking error équivalente à la distance euclidienne pour rapprocher les indices de marché d'itération en itération. Les risques financiers sont difficiles à gérer comme ils évoluent constemment et dépendent de multiples facteur comme la volatilité, la liquidité, les classes d'actifs etc. Pour gérer au mieux ces changements, nous développons une méthode récursive de validation de portefeuille qui révèle la vraie nature de l'évolution de la structure de risque des portefeuilles d'actifs financiers. Nous analysons plusieurs stratégies de gestion de portefeuille pour comprendre les dynamiques derrières les positions. Cela nous permet de valider l'analyse de portefeuille de nos stratégies. Nous effectuons une construction en copule de vigne, ce qui permet de séparer la modélisation de la dépendance de la modélisation des modèles marginaux. Pour créer un modèle précis et robuste, l'arbre à la base de la vigne doit être très précis et cohérent pour rapprocher les paires avec le meilleur potentiel pour la modélisation. Utiliser un arbre couvrant de poids minimum permet d'obtenir une solution capable de modéliser toutes les dépendances. En partant de l'analyse des portefeuilles, notre méthode nous permet de valider la valuation des actifs financiers sur une période spécifique.

12h50-14h10 : Repas

14h10-15h10

Stefano Favaro : Bayesian nonparametric disclosure risk assessment (Amphi 11)

University of Torino and Collegio Carlo Alberto

Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. When sample records are cross-classified according to categorical identification variables (key variables), any decision about release is supported by measures of disclosure risk, the most common being the number τ_1 of sample uniques cells that are also population uniques. We first make use of tools at the interface between Bayesian nonparametrics and the theory of exchangeable random partitions to develop a methodology that makes inference on τ_1 exact, computationally efficient and of easy implementation and reproducibility. Our approach relies on : i) a generalized Poisson-Dirichlet prior for modeling the random partition induced by the cross-classification of sample records ; ii) an empirical Bayes approach for estimating prior parameters in such a way to recognize a primary role to sample unique cells. These minimal model assumptions lead to an explicit, and simple, expression for the posterior distribution of τ_1, which allows to avoid the use of Markov chain Monte Carlo methods for posterior approximation. The proposed approach is tested on data from the U.S. 2000 census for the state of California and for the state of New York, revealing the same good experimental performance as recent Bayesian hierarchical semiparametric approaches that rely on modeling association among key variables at the cost of an increased computational effort.

Modération : Pierre LATOUCHE

Gallo Gueye et Anne Clémenceau : Les défis posés à la statistique officielle à l'heure du numérique (Amphi 14)

Modération : Serge ALLEGREZZA

15h10-16h30

Statistique computationnelle pour la segmentation (Amphi 11)

Modération : Guillem RIGAILL

Segmentation d'images spectrales à l'aide d'un algorithme hiérarchique avec contrainte spatiale

Agnès GRIMAUD (Lab. de Mathématiques de Versailles, UVSQ, CNRS, Université Paris-Saclay)
Gilles CELEUX (Inria Saclay-Ile-de-France)
Serge COHEN (IPANEMA CNRS ministère de la Culture UVSQ MNHN, USR3461, Université Paris-Saclay)
La caractérisation des matériaux anciens conduit à l'exploitation de données d'imagerie spectrale. La richesse d'information fournie par l'acquisition et l'exploitation d'un spectre complet est déterminante pour caractériser les détails des matériaux anciens, hétérogènes et d'une grande complexité. Ces images font souvent plusieurs GO, voire dizaine de GO et il importe d'avoir des algorithmes d'analyse performants. Nous proposons une procédure de classification hiérarchique des dissimilarités spectrales permettant de prendre en compte les proximités spatiales des pixels. Cette procédure est appliquée sur une image spectrale d'un fossile de poisson.

Assessment of various initialization strategies for the Expectation-Maximization algorithm for Hidden Semi-Markov Models with multiple categorical sequences

Brice OLIVIER (Université Grenoble Alpes)
Anne GUERIN-DUGUE (Gipsa-lab)
Jean-Baptiste DURAND (Grenoble INP)
In this study, we propose a method called sequence breaking framework to search high local maximum of the likelihood by providing starting values based on the observations for the Expectation-Maximization algorithm, for Hidden semi-Markov model parameters estimation. The method is shown to be efficient on several datasets with multiple categorical sequences.

Une méthode statistique pour détecter des ruptures multiples dans un arbre

Solène THEPAUT (Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay)
Guillem RIGAILL (Laboratoire de Mathématiques et Modélisation d'Evry, Université d'Evry Val d'Essonne & Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Gif sur Yvette, France)
Nous considérons le problème de détection de ruptures multiples dans la moyenne des nœuds d'un arbre. Ce problème est motivé par des applications en écologie où des mesures de diversité sont faites en n points d'un réseau de rivières. Ce réseau de rivières est représenté par un arbre. L'objectif est d'identifier des sous-arbres où les fluctuations d'abondance d'une espèce sont synchrones. Nous proposons d'inférer la position des ruptures et le signal par minimisation d'un risque empirique pénalisé. Nous dérivons une pénalité adaptée au problème et contrôlant le risque au travers d'une inégalité oracle non-asymptotique. Nous proposons deux algorithmes de programmation dynamique élagués pour retrouver la segmentation optimisant ce critère. Nous montrons empiriquement que leur complexité est en moyenne de $O(n^2)$ ou moins avec n le nombre de nœuds de l'arbre. Nous avons testé le comportement de notre approche sur des simulations et sur notre jeu de données écologique.

gfpop : Un package R pour la détection de ruptures contrainte par un graphe

Vincent RUNGE (Laboratoire de Mathématiques et Modélisation d'Évry, Université d'Évry Val d'Essonne)
Toby HOCKING (Northern Arizona University. School of Informatics, Computing and Cyber Systems)
Guillem RIGAILL (Institute of Plant Sciences Paris-Saclay, INRA)
T. D. Hocking et al. (2017) ont proposé un algorithme pour l'inférence de modèles de détection de ruptures avec des contraintes entre les paramètres des segments successifs. L'algorithme est exact au

sens où il optimise un risque pénalisé. Leur implémentation traite le cas particulier d'une contrainte haut-bas pour une perte Poisson. Nous avons développé un package R implémentant l'algorithme de manière générique traitant plusieurs pertes (avec leur équivalent robuste) et de nombreuses contraintes décrites par un graphe. Nous illustrerons d'abord les performances de notre algorithme dans le cas de la régression isotonique. Nous mettons en évidence l'intérêt d'utiliser des pertes robustes - récemment suggéré par Bach (2018) et Fearnhead and Rigaill (2018) - et celui de pénaliser le nombre de ruptures quand il y a effectivement des changements abrupts dans la moyenne du signal. Nous présentons enfin plus généralement les potentialités du package avec des graphes de contraintes plus exotiques.

Statistiques mathématiques : transport optimal (Amphi 12)

Modération : Adrien SAUMARD

Ordonner $\mathbb{R}^d$, $d \geq 2$: fonctions de répartition, fonctions quantiles et transports de mesures

Marc HALLIN (ECARES et Département de Mathématique, Université libre de Bruxelles)
Contrairement à la droite réelle, $\mathbb{R}^d$, pour $d \geq 2$, n'est pas ordonné de façon canonique. Une conséquence de cette absence d'un ordre canonique est que des concepts aussi fondamentaux que ceux de fonction de répartition ou de fonction quantile ne sont pas davantage définis de façon canonique. La définition usuelle, fondée sur les odres marginaux, ne jouit d'aucune des propriétés désirables pour une fonction de répartition ; en particulier, son inverse fournit une notion de quantile sans grande signification. Nous montrons comment une caractérisation de type transport de mesures permet d'étendre ces notions à $\mathbb{R}^d$, construisant ainsi un ordre spécifique à chaque loi (dans la population), induit par les observations (dans l'échantillon). A la différence des nombeuses propositions faites dans la littérature, les concepts ainsi obtenus possèdent toutes les propriétés qui font des rangs et des quantiles univariés des outils fondamentaux pour l'analyse des données aussi bien que pour l'inférence.

Using Efron-Stein's inequality for Optimal Transport Central Limit Theorem and Fairness issues in Machine Learning

Jean-Michel LOUBES (IMT)
Nous utilisons l'inégalité de Efron-Stein pour prouver des résultats sur des théorèmes de limite centrale pour la distance de Wasserstein. Nous montrerons quelques applications pour détecter un usage déloyal d'un algorithme d'apprentissage. Ces résultats sont tirés de del Barrio et Loubes (2019) et del Barrio et al. (2018).

Econométrie, finance (Amphi 13)

Modération : Pierre VALLOIS

Tests de rupture dans les modèles CHARN

Marwa LTAIFA (Doctorante)
Nous étudions un test du rapport de vraisemblance permettant de détecter les ruptures discrètes dans la moyenne des modèles CHARN. Nous montrons que dans le cas où les paramètres du modèle sont connus, le test est asymptotiquement optimal et nous donnons une forme explicite de sa puissance asymptotique. Dans le cas où les paramètres sont inconnus, en les remplaçant par des estimateurs appropriés, le test reste optimal et une expression explicite de la puissance locale est donnée. Une étude de simulation permettra d'évaluer les performances de la méthode.

Analyse d'événements Cyber grâce à un arbre de régression construit selon une vraissemblance de pareto généralisée et application à la tarification pour l'assurance des risques Cyber

Sébastien FARKAS (Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM)
Nous proposons une méthodologie d'analyse des bases de données de sinistres Cyber qui considère l'hétérogénéité des données pour proposer une calibration des tarifs et des conséquences d'un évènement extrême liés à un portefeuille de polices d'assurances Cyber. Nous appliquons cette méthodologie à une base publique constituée par l'association "Privacy Rights Clearinghouse" qui répertorie des événements de failles de données concernant des citoyens américains. Nous apportons une attention particulière aux valeurs extrêmes et analysons l'hétérogénité des données en réalisant une classification hiérarchique inspirée des arbres de régression par maximum de vraissemblance. Nous étayons cette modélisation de la sévérité par une modélisation de la fréquence des sinistres pour construire un modèle de tarification simple applicable à l'assurance du risque Cyber.

Autour d'estimateurs du Maximum de Vraisemblance explicites pour le modèle linéaire généralisé dans le cas de covariables catégorielles

Tom ROHMER (INRIA paris-île-de-France)
Christophe DUTANG (CEREMADE, Paris-Dauphine)
Alexandre BROUSTE (Le Mans Université)
Dans cette contribution, en reprenant l'approche des modèles linéaires généralisés, nous étudions la modélisation de variable d'intérêt, conditionnellement à des covariables catégorielles. D'une première part, nous déterminons une forme explicite général pour l'estimateur du maximum de vraisemblance (MLE) lorsque les covariables sont catégorielles ; en particulier l'existence ainsi que le comportement non-asymptotique du MLE y sont discutés. Nous allons baser nos illustrations sur deux lois d'extrême, pour lesquelles nous avons pu déterminer la loi des estimateurs des paramètres du modèle GLM et proposer des estimations débiaisées. Nous avons illustrons nos méthodes sur des données de sinistre en assurance.

Estimation de la variance asymptotique de l'estimateur des moindres carrés des modèles FARIMA faibles

Yacouba BOUBACAR MAINASSARA (Laboratoire de Mathématiques de Besançon)
Youssef ESSTAFA (Laboratoire de Mathématiques de Besançon)
Bruno SAUSSEREAU (Laboratoire de Mathématiques de Besançon)
Dans ce travail, nous considérons le problème de l'estimation de la matrice de variance asymptotique de l'estimateur des moindres carrés des modèles FARIMA (pour Fractionally AutoRegressive Integrated Moving-Average) dans le cas où les erreurs sont supposées non-corrélées mais non nécessairement indépendantes ni même des différences de martingales. Nous proposons un estimateur convergent de cette matrice de variance asymptotique et nous illustrons ensuite les résultats théoriques obtenus par des simulations.

Statistiques publiques (Amphi 14)

Modération : Nico WEYDERT

Déterminants de la confiance dans la statistique publique. Vers une perspective européenne ?

Serge ALLEGREZZA (STATEC)
La papier (et la présenation) expose les résultats d'une étude économétrique sur les déterminants de la confiance dans la statistique publique . En exploitant deux enquêtes représentatives (2015, 2017)dans le cas du Luxembourg, le papier explore les déterminants (catégories socio- économiques, éducation, participation à des enquêtes, utilisation de statistique, indépendance politique) sur la confiance dans les chiffres et l'institut que les produit . Ce travail rejoint les travaux de Chiche et Chanvril (2016) réalisés à Sciences Po. Il a vocation à s'étendre à d'autres pays de l'UE dans une optique de comparabilité. La confiance dans la statistique est essentielle pour permettre un débat public et des choix collectifs éclairés à l'ère du 'fake news' et de la désiformation. Dr Serge Allegrezza, directeur général du STATEC

Les effets du mode de collecte des données sur la mesure de l'emploi : une comparaison entre le web et le téléphone

Guillaume OSIER (STATEC)
Johann NEUMAYR (STATEC)
Joachim SCHORK (LIS et STATEC)
Cesare RIILLO (STATEC Research)
Les questionnaires par Internet sont de plus en plus utilisés en complément des modes de collecte des données plus traditionnels comme le face-à-face ou le téléphone. Le recours à des enquêtes « multimodes » pose cependant la question de savoir si l'utilisation du web conduit à introduire des biais de mesure qui sont spécifiques à ce mode de collecte. Notre hypothèse est que la magnitude de ce biais dépend fortement de la variable étudiée. En utilisant les données de l'enquête luxembourgeoise sur les forces de travail, qui associe un questionnaire web et un questionnaire par téléphone, nous étudions les différences entre ces deux modes de collecte à partir de données objectives (par ex. le statut dans l'emploi) et subjectives (par ex., la satisfaction avec l'emploi ou avec le salaire). Afin d'isoler le biais induit par le mode de collecte, nous appliquons une méthode d'appariement des données (Coarsened Exact Matching). Les variables qui sont à la base de cet appariement ont été déterminées en utilisant des algorithmes de sélection automatique mais également des éléments théoriques présents dans la littérature. Cette étude indique que les variables objectives ne sont pas affectées par le mode de collecte, tandis que les répondants par web ont tendance à déclarer un niveau de satisfaction inférieur à celui des répondants par téléphone. On voit ainsi que l'effet de mode dépend fortement de la nature de l'enquête et de la variable qui est étudiée.

Répondre aux enquêtes, PC vs. mobile : Les évaluations de la vie des personnes sont-elles comparables ?

Francesco SARRACINO (STATEC)
Cesare RIILLO (STATEC)
Malgorzata MIKUCKA (MZES)
La littérature sur les enquêtes en mode mixte a longtemps étudié si les modes d'enquête en face à face, par téléphone et en ligne permettaient de collecter des données fiables. On en sait beaucoup moins sur le biais potentiel associé à l'utilisation de différents dispositifs multimédia pour répondre aux enquêtes

en ligne. Nous comparons les mesures subjectives de bien-être collectées sur le Web via PC et téléphones portables pour vérifier si le dispositif utilisé affecte les évaluations de bien-être des personnes. Nous utilisons des données uniques, représentatives de la population Luxembourgeoise, qui contiennent cinq mesures du bien-être subjectif collectées en 2017. L'utilisation d'un multinomial logit avec Coarsened Exact Matching indique que le dispositif multimédia utilisé a une incidence sur les scores de satisfaction dans la vie. Sur une échelle de 1 à 5, où les scores les plus élevés représentent une plus grande satisfaction, les répondants utilisant un téléphone mobile sont plus susceptibles de choisir la catégorie de bien-être la plus élevée et moins susceptibles de choisir la quatrième catégorie. Nous n'observons aucune différence statistique entre les trois catégories restantes. Nous testons la robustesse de nos résultats en utilisant trois indicateurs indirects du bien-être subjectif. Les résultats indiquent que les outils multimédia n'induisent aucune différence statistiquement significative dans le bien-être subjectif. Nous discutons des conséquences potentielles de nos résultats sur l'inférence statistique.

Enseignement et Histoire de la Statistique (Amphi 15)

Modération : Antoine ROLLAND

Transition between education and profession : Experiences of statisticians

Layla GUYOT (Texas State University)
Les compétences en statistique sont de plus en plus recherchées, et ce quelque soit le domaine d'application. Bien que l'offre d'emploi pour les statisticiens soit en augmentation, il est parfois difficile de recruter les jeunes diplômés principalement parce que les statisticiens développent un savoir et des compétences spécifiques à leur milieu professionnel (Pfannkuch & Wild, 2000). Afin de reconnaître comment ce savoir et ces compétences spécifiques sont développées, des statisticiens se sont engagés dans une réflexion sur leur expérience. Les résultats révèlent d'importantes compétences en statistique requises en milieu professionnel, reconnues par les professionnels eux-mêmes. Reconnaître ces compétences permettra d'offrir une expérience pratique aux étudiants de statistiques et futurs statisticiens.

Représentations sociales de la statistique chez des étudiants de psychologie, mises en évidence à travers les réseaux d'associations

Jean-Marie MARION (Université Catholique de l'Ouest – Faculté des Sciences – BP 10808 - 49008 ANGERS CEDEX 01)
Véronique DUBREIL-FREMONT (Université Catholique de l'Ouest – Faculté des Sciences Humaines et Sociales – BP 10808 - 49008 ANGERS CEDEX 01)
Alain BIHAN-POUDEC (Université Catholique de l'Ouest – Faculté d'Éducation – BP 10808 - 49008 ANGERS CEDEX 01)
Noëlle ZENDRERA (Université Catholique de l'Ouest – Faculté des Sciences Humaines et Sociales – BP 10808 - 49008 ANGERS CEDEX 01)
En tant qu'enseignants en statistique dans l'enseignement supérieur, nous nous intéressons aux représentations sociales qu'ont nos étudiants à propos de cette discipline. Une première série d'études à base de questionnaires nous a permis de cerner les représentations sociales de la statistique auprès d'étudiants en sciences humaines et sociales (Bihan-Poudec, 2013), ainsi que leur évolution (Bihan-Poudec et Marion, 2014). Plus récemment, nous avons poursuivi nos travaux chez des étudiants en psychologie, à l'aide de la technique des réseaux d'associations (de Rosa, 1995) afin d'explorer notamment la dimension affective de l'attitude des étudiants vis-à-vis de la statistique (Dubreil-Frémont, Marion et Bihan-Poudec, 2018). Par ailleurs, comme son nom l'indique, cette technique permet aux participants d'associer eux-mêmes les mots, et aux expérimentateurs de ne plus se contenter de simples co-occurrences.

Sur le nombre et l'indépendance des jurés du comité de salut public

Ingrid ROCHEL (Université de Bordeaux)
Léo GERVILLE-REACHE (Univ. Bordeaux, CNRS, IMS, UMR 5218)
Le Comité de Salut Public a rendu plus de 4000 jugements entre le 6 avril 1793 et le 27 juillet 1794. Parmi ces jugements, un millier a été rendu en Gironde. En étudiant la nature de ces jugements et les profils des jugés selon les conventions, c'est la question de la probabilité des erreurs de jugements, chère aux mathématiciens de l'époque, qui est mise à rude épreuve. L'hypothèse d'indépendance des jurés et son implication sur la pluralité des voix restent purement théoriques.

À l'occasion du cent cinquantenaire de deux graphiques emblématiques de Charles-Joseph Minard (1781-1870)

Antoine de FALGUEROLLES (Enseignant-chercheur retraité)
Charles-Joseph Minard (1781-1870) publiait il y a 150 ans deux graphiques statistiques emblématiques : une carte figurative des pertes successives en hommes de l'armée qu'Hannibal conduisit d'Espagne en Italie en traversant les Gaules (selon Polybe) et celle de l'armée napoléonienne dans la campagne de Russie 1812-1813. Génération spontanée ou inspirée ? Emmanuel de Las Cases (1766-1842) avait antérieurement représenté ces thèmes dans son Atlas en matérialisant le parcours de ces armées par un "ruban colorié porté sur la carte". La mise en correspondance de leur travaux fait ressortir chez Las Cases une créativité artistique certaine. Mais c'est la formalisation statistique de ces événements qui, alliée à une bonne intuition des règles de la sémiologie de l'image, donne aux graphiques de Minard leur "éloquence brutale".

16h30-16h50 : Pause café

16h50-18h10

Statistique d'enquête (Amphi 11)

Modération : Yves TILLE

Paris sportifs au football : l'intérêt des expected goals

Paul STEFFEN (Univ. Bordeaux, CNRS, IMS, UMR 5218)
Léo GERVILLE-REACHE (Univ. Bordeaux, CNRS, IMS, UMR 5218)
Nicolas BISOFFI (Betclic)
Pour l'établissement des cotes du résultat d'un match de football, l'estimation des probabilités des issues possibles 1/N/2 est fondamentale. Les nombreux modèles statistiques qui sont utilisés se servent du résultat réel. Dans cette communication nous proposons d'utiliser les « issues espérées » construites à partir des buts espérés (expected goals, xG). En utilisant le modèle polytomique ordonné, nous comparons, sur plusieurs années de Ligue 1, les résultats obtenus selon l'utilisation des issues réelles ou des issues espérées.

Sélection des logements pour les enquêtes auprès des ménages dans les grandes villes : tirage équilibré versus tirage systématique en présence de non-réponse.

Lionel DELTA (Insee)
L'Insee a récemment renouvelé son échantillon-maître, c'est-à-dire l'échantillon de zones, appelées unités primaires, au sein desquelles seront tirés pendant dix ans les logements interrogés dans le cadre de différentes enquêtes de la statistique publique. Ces différentes zones ont été tirées selon un plan de sondage équilibré spatialement réparti permettant des gains de précision assez important par rapport à d'autres méthodes envisageables. Après ce premier degré de tirage, il était donc envisagé que le second degré de tirage, celui correspondant à la sélection des logements, soit également équilibré lorsqu'il concerne un nombre significatif de logements au sein d'une même unité primaire, généralement les grandes villes. En se basant sur des simulations de tirages à partir de différents plans de sondage y compris la simulation de phénomènes de non-réponse, nous comparons les performances en matière de précision de différents tirages équilibrés à des tirages systématiques avec tri préalable de la base de sondage selon les mêmes variables auxiliaires. Les contraintes liées à la taille des échantillons de logements au sein de chacune des unités primaires ainsi que l'existence de non-répondants sont de nature à limiter voire inverser les gains de précisions attendus, en particulier lorsqu'on effectue un calage sur marges. Cette étude permettait donc de justifier le choix final de conserver un tirage systématique au second degré et pose d'autres questions comme celle de la justification de la préférence pour cette option même en l'absence de non-réponse.

L'année 2019 pour le projet Nautile (Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes) et pour l'échantillon de l'Enquête Emploi en Continu

Thomas SAUVAGET (Insee)
Ludovic VINCENT (Insee)
Pierre-Arnaud PENDOLI (Insee)
Thomas MERLY-ALPA (Insee)
Sébastien FAIVRE (Insee)
Actuellement et pour encore quelques mois, les enquêtes ménages de l'Insee sont pour la plupart tirées dans le recensement de la population (RP), et réalisées en face-à-face par des enquêteurs. Pour cela, l'Insee a mis en place une méthodologie consistant à sélectionner un échantillon de zones (appelé Échantillon-Maître (EM)), puis, au sein de chaque zone sélectionnée, à échantillonner les logements interrogés. En 2019, les travaux de renouvellement décennal de l'EM, démarrés en 2016, arrivent à leur terme avec l'entrée en production de l'application Nautile (Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes) d'une part, et du nouvel échantillon de l'Enquête-Emploi en Continu (EEC) d'autre part. Ces deux renouvellements d'échantillon étaient nécessaires. En effet, côté EM, des études avaient montré un déséquilibre croissant des groupes de rotations du RP au cours du temps, dégradant la qualité de tout EM issu de la méthode actuelle. Par ailleurs, une nouvelle source, Fidéli (Fichier démographique des logements et des individus), a ouvert de nouvelles opportunités. Ce fichier, mis à jour chaque année à partir des fichiers fiscaux, présente les propriétés d'une bonne base de sondage : exhaustivité, unicité, fraîcheur des données. L'utilisation de Fidéli à la place du RP permet de diminuer l'étendue des zones d'enquêtes et d'améliorer leur plan de sondage. Elle rend également possible le tirage d'individus. Parallèlement, l'échantillon de l'Enquête Emploi en Continu (EEC) arrivait lui aussi à terme, si bien qu'un tirage coordonné avec l'EM a pu être mis en place. Au premier semestre 2019, des travaux complémentaires sur la constitution des grappes de logements à interroger ont permis de finaliser la méthode de mise à jour de l'échantillon qui sera employée au cours des 9 années à venir. Cette communication présente les derniers développements concernant l'entrée en production de Nautile et du nouvel échantillon de l'EEC au second semestre 2019.

Imputation équilibrée pour la non-réponse en fromage suisse

Yves TILLE (Université de Neuchâtel)
Audrey-Anne VALLEE (Université de Neuchatel)
La non-réponse en fromage suisse ou la non-réponse non monotone regroupe les cas où toutes les variables d'une enquête contiennent des valeurs manquantes sans schéma particulier. L'imputation des valeurs manquantes permet de réduire le biais et la variabilité causés par la non-réponse. Il est difficile de préserver

les distributions et les relations entre les variables lors de l'imputation dans le cas de non-réponse en fromage suisse. Dans cette présentation, l'imputation équilibrée par les K plus proches voisins (Hasler & Tillé, 2016) est développée pour le cas de la non-réponse en fromage suisse. Il s'agit d'une méthode d'imputation par donneurs qui est aléatoire et construite pour répondre à plusieurs exigences. D'abord, un non-répondant peut être imputé par des donneurs qui sont proches de lui. Les distances sont calculées avec les valeurs disponibles des variables. Ensuite, toutes les valeurs manquantes d'un non-répondant sont imputées par le même donneur choisi aléatoirement. Enfin, les donneurs sont sélectionnés de façon à ce que si on imputait les valeurs observées des non-répondants aussi, les estimations des totaux imputés et les totaux connus devraient être les mêmes. Pour imputer en respectant de telles contraintes, une matrice de probabilités d'imputation est construite à l'aide de méthodes de calage. Les donneurs sont ensuite choisis avec ces probabilités et des méthodes d'échantillonnage équilibré.

Apprentissage statistique et applications (Amphi 12)

Modération : Julien JACQUES

Apprentissage statistique sur modèles météorologiques pour l'éolien

Aurélie FISCHER (LPSM, Université Paris Diderot)
Cet exposé est consacré au problème dit de 'réduction d'échelle' d'une quantité météorologique à partir des sorties de modèles numériques de prévisions météorologiques. Il est essentiel dans le domaine de l'énergie éolienne de disposer de prévisions précises de la vitesse du vent de surface aux emplacements des parcs éoliens. Ces prévisions sont aussi cruciales pour la prévention des dommages liés à des vents violents. Nous étudions les performances de méthodes d'apprentissage statistique pour la reconstruction et la prévision de la vitesse du vent à partir des sorties du modèle du Centre européen de prévisions météorologiques à moyen terme (ECMWF), en utilisant tout d'abord les données de vitesse du vent locales fournies par la plateforme d'observation du Site Instrumental de Recherche par Télédétection Atmosphérique (SIRTA), puis des données observées à l'échelle de la France.

Le modèle des blocs latents, une méthode régularisée pour la classification en grande dimension

Christine KERIBIN (INRIA Saclay Ile de France ; Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France)
Christophe BIERNACKI (Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille)
Les modèles de mélange sont connus pour être un outil efficace de classification non supervisée quand la dimension des observations est faible, mais échouent en grande dimension à cause d'un manque de parcimonie. Certaines tentatives pour prendre en compte la redondance ou la pertinence des variables se heurtent à des problèmes de complexité explosive. Nous recommandons d'utiliser le modèle des blocs latents, un modèle probabiliste de classification croisée simultanée des individus et des variables, pour classifier des individus en grande dimension. Nous illustrons de façon empirique le compromis biais-variance de la stratégie de classification croisée dans des scénarii en grande dimension comportant des caractéristiques de redondance et de non pertinence et nous montrons son effet régularisateur sur la classification simple.

Étude de la variabilité inter-individuelle de données de connectivités intrinsèques : détection de réseaux instables et de sous-populations dans un tableau tridimensionnel

Loïc LABACHE (Groupe d'Imagerie Neurofonctionnelle, CEA & IMN, UMR 5293, 146 rue Léo Saignat, 33000 Bordeaux, France & CQFD team, Inria Bordeaux Sud Ouest & IMB, UMR 5251, ENSC - Bordeaux INP, 109 Avenue Roul, 33400 Talence, France)
Marc JOLIOT (Groupe d'Imagerie Neurofonctionnelle, CEA & IMN, UMR 5293, 146 rue Léo Saignat, 33000 Bordeaux, France)
Jérôme SARACCO (CQFD team, Inria Bordeaux Sud Ouest & IMB, UMR 5251, ENSC - Bordeaux INP, 109 Avenue Roul, 33400 Talence, France)
Nathalie TZOURIO-MAZOYER (Groupe d'Imagerie Neurofonctionnelle, CEA & IMN, UMR 5293, 146 rue Léo Saignat, 33000 Bordeaux, France)
Dans cette communication, nous proposons deux méthodologies permettant de mieux comprendre la variabilité inter-individuelle de données cérébrales d'Imagerie par Résonance Magnétique (IRM) fonctionnelle. Il s'agit de quantifier si le dendrogramme "moyen" est bien représentatif de la population initiale et d'identifier ses éventuelles sources d'instabilité. La première méthode permet d'identifier les réseaux pouvant conduire à des partitions instables du dendrogramme "moyen". La seconde approche permet d'identifier des sous-populations homogènes de sujets pour lesquelles leurs dendrogrammes "moyen" associés est plus stable que celui de la population initiale. Ces deux méthodes seront illustrées sur des données simulées à partir de données cérébrales de connectivités intrinsèques obtenues en IRM fonctionnelle.

Sur l'estimation du tau de Kendall conditionnel à l'aide de méthodes de classification

Alexis DERUMIGNY (CREST-ENSAE)
Jean-David FERMANIAN (CREST-ENSAE)
Nous montrons ici comment le problème d'estimation du tau de Kendall conditionnel peut être réécrit comme un problème de classification. Le tau de Kendall conditionnel est un paramètre de dépendance conditionnel qui est caractéristique d'une paire de variables aléatoires. Le but est de prédire si la paire est concordante (valeur de 1) ou discordante (valeur de -1) conditionnellement à un vecteur de covariables. Nous prouvons la consistance et la normalité asymptotique d'une famille d'estimateurs basés sur un maximum de vraisemblance approché, comportant comme cas particuliers l'équivalent des régressions logit et probit dans notre cadre. Nous détaillons des algorithmes spécifiques, adaptant les techniques usuelles d'apprentissage automatique comme les plus proches voisins, les arbres de décision, les forêts aléatoires et les réseaux de neurones dans le contexte de l'estimation du tau de Kendall conditionnel. Des simulations détaillent leurs propriétés à distance finie. Finalement, tous ces estimateurs sont appliqués à une base de données d'indices boursiers européens.

Statistique séquentielle (Amphi 13)

Modération : Gérard BIAU

Algorithmes de bandits pour le pilotage de la consommation électrique

Margaux BREGERE (EDF R&D, Université Paris-Sud, Inria)
Gilles STOLTZ (Université Paris-Sud)
Pierre GAILLARD (Inria)
Yannig GOUDE (EDF R&D)

L'électricité ne pouvant être stockée, l'équilibre entre la production et la consommation doit en permanence être maintenu. Nous proposons d'appliquer la théorie des bandits contextuels pour piloter, à l'aide d'incitations tarifaires, la demande électrique. Plus précisément, une consommation moyenne cible est fixée à chaque instant et la consommation moyenne est modélisée comme une fonction des prix envoyés et de variables contextuelles (température, heure, jour etc.). La performance des stratégies est mesurée en pertes quadratiques à travers un critère de regret. Inspiré des stratégies standards pour les bandits contextuels (LinUCB - [5, 2]), notre algorithme permet de borner ce regret en $T^{2/3}$ (aux termes poly-logarithmiques près). Des simulations sur des données publiques de UK Power Networks, dans lesquels des incitations tarifaires ont été proposées, montrent que notre stratégie influence efficacement la consommation électrique des usagers.

Gestion des logs dans les problèmes de bandits contextuels

Emmanuelle CLAEYS (IRMA)
Pierre GANCARSKI (ICUBE)
Myriam MAUMY-BERTRAND (IRMA)
Récemment, de nouvelles méthodes prometteuses optimisent les tests A/B en utilisant l'allocation dynamique. Elles permettent d'obtenir un résultat plus rapide pour déterminer quelle variation est la meilleure, ce qui permet à l'utilisateur d'économiser des coûts. Cependant, l'allocation dynamique par les méthodes traditionnelles reste contraignante sur le type de données utilisées. Dans cet article, nous présentons une nouvelle méthode qui permet d'intégrer des séries temporelles (logs) pour améliorer l'allocation dynamique dans le cadre d'un A/B test. Cet article fournit des résultats numériques sur des données d'essai réelles, pour démontrer l'amélioration apportée par la méthode par rapport aux méthodes traditionnelles.

Détection statistique de rupture dans le cadre online.

Nassim SAHKI (Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France)
Anne GEGOUT-PETIT (Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France)
Sophie WANTZ-MEZIERES (Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France)
Nous introduisons la version online de la statistique de CUSUM basée sur un test séquentiel du rapport de vraisemblance, que nous remplaçons par une fonction de score dans le cas non-paramétrique. La détection de rupture est basée sur une règle d'arrêt et la sélection d'un seuil de détection. Dans notre travail, nous proposons un seuil de détection instantanée dépendant du temps (dynamique), et des nouvelles règles d'arrêt dans le but de contrôler les paramètres de détection donnés par le taux de fausse alarme (FAR), le temps moyen entre fausses alarmes (MTBFA) et ainsi que le délai moyen de détection (ADD). Finalement, nous présentons des résultats de simulation par l'estimation des paramètres de détection.

Approche statistique de la causalité (Amphi 14)

Modération : Anne GEGOUT-PETIT

Causal mediation analysis in presence of multiple mediators uncausally related

Allan JEROLON (Université Paris Descartes)
Laura BAGLIETTO (Université de Pise)
Etienne BIRMELE (Université Paris Descartes)
Vittorio PERDUCA (Université Paris Descartes)
Flora ALARCON (Université Paris Descartes)

L'analyse de médiation vise à démêler les effets d'un traitement sur une variable de sortie par le biais de mécanismes de causalité alternatifs et est devenue une pratique courante dans les applications biomédicales et en sciences sociales. Le cadre causal basé sur les contrefactuels est actuellement l'approche standard de la médiation, avec d'importants progrès méthodologiques introduits dans la littérature au cours de la dernière décennie, en particulier pour la médiation simple, c'est-à-dire avec un médiateur à la fois. Parmi une variété d'approches alternatives, K. Imai et al. ont montré des résultats théoriques et développé un package R pour traiter la médiation simple ainsi que la médiation multiple impliquant plusieurs médiateurs indépendants conditionnellement au traitement et aux covariables. Cette approche ne permet pas de considérer la situation souvent rencontrée dans laquelle une cause commune non observée induit une corrélation fallacieuse entre les médiateurs. Dans ce contexte, que nous qualifions de médiation avec des médiateurs liés de manière non-causale, nous montrons que, sous de nouvelles hypothèses appropriées, les effets naturels directs et indirects sont identifiables de manière non paramétrique. Ces résultats sont rapidement traduits en estimateurs non biaisés utilisant le même algorithme quasi-bayésien mis au point par Imai et al que nous avons adaptés au cas multiple. Nous validons notre méthode par une étude de simulation originale. A titre d'illustration, nous appliquons notre méthode sur un ensemble de données réelles d'une grande cohorte afin d'évaluer l'effet du traitement hormonal sur le risque de cancer du sein par l'intermédiaire de trois médiateurs, à savoir les zones mammaires denses, les zones mammaires non denses et l'indice de masse corporelle.

Estimation double robuste d'effet du traitement avec facteurs confondants incomplets

Imke MAYER (CAMS, EHESS)
Julie JOSSE (CMAP, X-Paris-Saclay)
Jean-Pierre NADAL (CAMS, EHESS)
Stefan WAGER (Stanford Graduate School of Business)
Tobias GAUSS (Traumabase Group)
Jean-Denis MOYER (Traumabase Group)
Dans la recherche en santé et en sciences sociales, les études observationnelles prospectives sont fréquentes, relativement faciles à mettre en place (contrairement aux études expérimentales d'essais randomisés qui sont parfois même impossible à réaliser) et peuvent permettre différents types d'analyses postérieures telles que des inférences causales. L'estimation de l'effet moyen du traitement (average treatment effect en anglais, ATE), par exemple, est possible grâce à l'utilisation de scores de propension qui permettent de corriger les biais d'affectation du traitement dus à de la confusion, i.e. la présence de facteurs liés à la fois à l'affectation du traitement et à la variable d'intérêt. Cependant, un problème majeur des grandes études observationnelles est leur complexité et leur caractère souvent incomplet : les covariables sont souvent prises à différents niveaux et stades, elles peuvent être hétérogènes – catégorielles, discrètes, continues – et contiennent presque inévitablement des valeurs manquantes. Le problème des valeurs manquantes dans l'inférence causale a longtemps été ignoré et n'a regagné l'attention que récemment en raison des impacts non négligeables en termes de puissance et de biais induits par des analyses de cas complètes et des modèles d'imputation mal spécifiés. Nous discutons des conditions dans lesquelles une inférence causale peut être possible malgré la présence de valeurs manquantes dans les facteurs confondants, nous comparons différentes méthodes proposées dans le passé pour traiter les valeurs confondantes manquantes et proposons deux estimateurs ATE double robustes qui rendent directement compte des valeurs manquantes. Nous évaluons la performance de nos estimateurs sur une base de données prospective considérable contenant des informations détaillées sur environ 20 000 patients poly-traumatisés graves en France. À l'aide des estimateurs d'ATE proposés et de cette base de données, nous étudions l'effet sur la mortalité de l'administration de l'acide tranexamique aux patients présentant un choc hémorragique.

Leveraging contact network information in studies of contagion processes

Mélanie PRAGUE (Inria SISTM Team, Inserm U1219, Université de Bordeaux)
Patrick STAPLES (Harvard TH School of Public health, Boston, USA)
Victor DEGRUTTOLA (Harvard TH School of Public health, Boston, USA)
Jukka-Pekka ONNELA (Harvard TH School of Public health, Boston, USA)
Evaluate the effect of an exposition or intervention against a contagion pro- cess is a central question in the development of new methods for epidemic control or information propagation. We are interested

in contagion processes operating on a con- tact network, transmission can only occur through ties that connect contaminated and non-contaminated individuals. We investigate the use of contact network features as both confounders and efficiency covariates in exposure effect estimation. Using doubly-robust augmented generalised estimating equations (GEE), we estimate how correction of bias and gains in efficiency depend on the network structure and characteristics of the contagion agent. We apply this approach to estimate the effects of various exposures on the spread of a microfinance program in a collection of villages in Karnataka, India.

Modélisation PINAR(p) et prévision du nombre d'admissions hospitalier

Mohamed djemaà SADOUN (Operational Research Department, USTHB, Centre for Research in Applied Economic for Development (CREAD))
Cette communication propose une modélisation autorégressive à valeurs entières d'ordre arbitraire à coefficients périodique PINAR(p), dans le but d'analyser le nombre d'arrivées des services d'urgence hospitalières causées par des maladies ayant un comportement saisonnier. Deux méthodes d'estimation des paramètres du modèles seront proposées, à savoir : la méthode des moindres carrées conditionnelles (MCC) et la méthode du maximum de vraisemblance conditionnelle (MVC). De plus la fonction de prédiction du modèle sera donnée en utilisant une certaine représentation de l'espérance conditionnelle. Les performances des estimateurs obtenus seront montrés via une étude de simulation intensive. Une application sur données réelles sera réalisée pour modéliser le nombre mensuel d'admissions hospitalier causée par la grippe.

Environnement (Amphi 15, fin à 18h30)

Modération : Liliane BEL

Modélisation spatio-temporelle et précipitations extrêmes

Jean-Noel BACRO (IMAG, Université de Montpellier, CNRS, Montpellier, France)
Carlo GAETAN (DAIS, Università Ca' Foscari di Venezia, Venice, Italy)
Thomas OPITZ (BioSP, INRA, Avignon, France)
Gwladys TOULEMONDE (IMAG, Université de Montpellier, CNRS, Montpellier et INRIA, Project-team LEMON, France)
La modélisation statistique de données spatio-temporelles s'appuie aujourd'hui fortement sur des approches dites hiérarchiques. Ces dernières offrent en effet un cadre théorique relativement simple permettant une décomposition naturelle et exploitable des dépendances complexes inhérentes à ce type de données. Après avoir donné quelques éléments fondamentaux sur la modélisation spatio-temporelle, la présentation s'orientera sur des problématiques spécifiques aux données extrêmes. Un modèle spatio-temporel pour les dépassements de seuils élevés sera proposé et une application sur des précipitations dans le sud de la France sera présentée. L'intérêt et les limites de ce type de modélisation pour la simulation de scénarii de précipitations seront discutés.

Spatial analysis of heterogeneous precipitation data, application to urban hydrology

Marie BOUTIGNY (LMBA, Eau du Ponant)
Pierre AILLIOT (LMBA)
Benoît SAUSSOL (LMBA)
Antoine SINQUIN (Eau du Ponant)
In urban areas, where an important part of the sewerage system is combined, waste water dumping

can occur during rainy weather. Water collection and storage systems are generally designed using hydrological models which describe the functioning of the sewerage systems. Weather conditions such as evapotranspiration and most of all precipitation are very important forcing factors for such models. An usual approach to take into account the variability of weather conditions consists in forcing the hydrological model with up to 5 years of meteorological data, considered as 'representative' and observed in a site close to the system location. This permits to estimate the statistical distribution of water dumping with reasonable computational cost. In order to assess the sensitivity of the method to the choice of the 'representative' meteorological conditions, we propose to develop a stochastic weather generator tor (Wilks & Wilby (1999), Ailliot et al. (2015)) to simulate a high number (e.g. several centuries) of realistic spatio-temporal meteorological series. These artificial weather conditions can then be used as input to the hydrological model. This will allow us to estimate its sensitivity to different things, such as the length of the input time series or rainfall spatialization, and finally determine the forcing weather conditions to be used in studies aiming at designing sewage systems.

Modèle POT non-stationnaire pour l'analyse des températures et des précipitations extrêmes au Burkina Faso.

Béwentaoré SAWADOGO (UMR MIA-Paris, INRA, AgroParisTech, Université Paris Saclay, 75005, Paris, France)
Liliane BEL (UMR MIA-Paris, INRA, AgroParisTech, Université Paris Saclay, 75005, Paris, France)
Diakarya BARRO (LANIBIO, UFR-SEG, Université Ouaga II, 12 BP417 Ouagadougou 12, Burkina Faso)
Dans cette étude, les tendances des températures et des précipitations extrêmes au Burkina Faso sont évaluées en utilisant la théorie statistique des valeurs extrêmes, en particulier l'approche des dépassements de seuil dans un contexte non-stationnaire. La méthode de Parey et al. (2014) est utilisée pour décrire les tendances dans les paramètres de la distribution des excès au dessus d'un seuil fixé et l'intensité du processus de Poisson non-stationnaire. A partir de l'extrapolation des tendances identifiées, les niveaux de retour non-stationnaires et leur intervalle de confiance sont également calculés.

Scenarios of hydrometeorological variables based on auxiliary data for water stress retrieval in central Tunisia

Nesrine FARHANI (Université de Carthage / Institut National Agronomique de Tunisie/ LR17AGR01-GREEN-TEAM, Tunis, Tunisie)
Julie CARREAU (HydroSciences Montpellier (HSM), CNRS/IRD/UM1/UM2, Place Eugène Bataillon, 34095 Montpellier, France)
Gilles BOULET (Centre d'Etudes Spatiales de la Biosphère, Université de Toulouse, CNRS, CNES, IRD, UPS, Toulouse, France 10)
Zeineb KASSOUK (Université de Carthage / Institut National Agronomique de Tunisie/ LR17AGR01-GREEN-TEAM, Tunis, Tunisie)
Bernard MOUGENOT (Centre d'Etudes Spatiales de la Biosphère, Université de Toulouse, CNRS, CNES, IRD, UPS, Toulouse, France 10)
Michel LE PAGE (Centre d'Etudes Spatiales de la Biosphère, Université de Toulouse, CNRS, CNES, IRD, UPS, Toulouse, France 10)
Zohra LILI CHABAANE (Université de Carthage / Institut National Agronomique de Tunisie/ LR17AGR01-GREEN-TEAM, Tunis, Tunisie)
Rim ZITOUNA (INRGREF-LRVENC, Carthage University, BP 10 El Menzah IV, 1004 Tunis, Tunisia)

Water scarcity and the inter-annual variability of water resources in semi arid areas are limiting factors for agricultural production. Characterization of plant water use, generally determined by estimating evapotranspiration, is needed to better manage water resources. Furthermore, water stress derived from remote sensing data in the thermal infrared domain is particularly informative for monitoring agrosystem health and adjusting irrigation requirements. Evapotranspiration and water stress must be monitored at hourly to daily scales. Both can be simulated by a dual source energy balance model from climatic observations (air temperature, relative air humidity, global radiation and wind speed) and satellite information (NDVI, LAI, albedo and surface temperature). However, it may occur that the available climatic

observations are insufficient to account for the spatial and temporal variability of the area of interest due to the sparsity of gauged networks, the lack of long observation periods and the presence of numerous gaps. We aim to adapt a kind of stochastic weather generator that relies on low resolution ERA reanalysis data to provide spatio-temporal scenarios of multiple climatic variables. This approach serves to perform imputation of missing data and to drive simulation in order to extend the climatic time series in the past. In addition to ERA reanalysis data, other covariates are used to introduce inter-variable dependence, seasonal and diurnal cycles together with geographical information. We compare this method with other classic temporal extension method, bias correction which exploit also ERA reanalysis. This approach can be employed to account for the difference in spatial resolution between reanalysis and observations from the gauged network. By working on anomalies of diurnal cycles, existing bias correction methods can be applied at sub-daily temporal resolution. We consider a univariate and a multivariate bias correction method, CDFt and MBCn respectively, to assess to added-value of correcting simultaneously the climatic variables. The statistical approaches are applied to generate climatic scenarios in the Kairouan area in central Tunisia which is subject to semi-arid climate. The different scenarios are evaluated and compared in terms of their ability to reproduce several features of the climatic observations and also in terms of their ability to reproduce the simulation of evapotranspiration and water stress with a dual source energy balance model.

19h-21h : Cocktail de bienvenue à l'Hôtel de Ville

Mardi 4 juin

9h-10h

Charles Bouveyron : Bayesian sparsity for statistical learning in high dimensions (Amphi 11)

Université Côte d'Azur & INRIA

Although the ongoing digital revolution in fields such as chemometrics, genomics or personalized medicine gives hope for considerable progress in these areas, it also provides more and more high-dimensional data to analyze and interpret. A common usual task in those fields is discriminant analysis, which however may suffer from the high dimensionality of the data. The recent advances, through subspace classification or variable selection methods, allowed to reach either excellent classification performances or useful visualizations and interpretations. Obviously, it is of great interest to have both excellent classification accuracies and a meaningful variable selection for interpretation. This work addresses this issue by introducing a subspace discriminant analysis method which performs a class-specific variable selection through Bayesian sparsity. The resulting classification methodology is called sparse high-dimensional discriminant analysis (sHDDA). Contrary to most sparse methods which are based on the Lasso, sHDDA relies on a Bayesian modeling of the sparsity pattern and avoids the painstaking and sensitive cross-validation of the sparsity level. The main features of sHDDA are illustrated on simulated and real-world data. In particular, we propose an exemplar application to cancer characterization based on medical imaging using radiomic feature extraction is in particular proposed.

Modération : Marianne CLAUSEL

Jean Opsomer : Survey Estimators Under Partial Ordering (Amphi 14)

Westat

In many large-scale surveys, estimates are often produced for large numbers of domains. This can lead to small sample sizes, with resulting unreliable domain-level estimates. When a priori qualitative constraints between domain means can be specified, it is natural to attempt to ensure that the estimates likewise satisfy the constraints, with the goal of improving the precision of the estimates and their acceptability by data users. We describe constrained estimation methods for domains under partial ordering. The asymptotic properties of the methods are obtained within a classical design-based inferential framework. We illustrate the usefulness of the methodology on data from the U.S. National Survey of College Graduates.

Modération : Yves TILLE

10h-11h20

Statistique mathématique (Amphi 11)

Modération : Antoine GODICHON-BAGGIONI

M-estimation inference for partially linear single-index models : an empirical likelihood approach

Matthieu MARBAC (Ensai - Crest)
Valentin PATILEA (Ensai - Crest)

Partially linear single-index models represent a versatile tool to capture the relationship between response variables and possibly high-dimensional covariate vectors. The approximation of the response is given by the sum of a linear term and of a nonparametric link function of a second linear combination of covariates, usually called the index. This approximation is defined with respect to a loss function which characterizes a feature of the conditional law of the response given the covariates. We consider a general family of loss functions and investigate the corresponding partially linear single-index regression models. Except for imposing some moments to be finite, the conditional law of the error term is allowed to be general. For the inference, we adopt the empirical likelihood (EL) approach based on a class of moment conditions in which we plug-in estimates of the nuisance link function. We show the asymptotic pivotality of the likelihood ratio under weak high-level conditions. A simple data-driven choice of the tuning parameter for the estimation of the link function is proposed.

Normalité asymptotique des statistiques de tests des indices relatifs de dispersion et de variation

Aboubacar Y. TOURE (Université Bourgogne Franche-Comté)
Simplice DOSSOU-GBETE (Université de Pau et des Pays de l'Adour)
Célestin C. KOKONENDJI (Université Bourgogne Franche-Comté)
A partir des indices de dispersion relatifs aux lois de Poisson et binomiale pour les données de comptage et, récemment, de l'indice de variation exponentielle pour les données continues positives, nous introduisons d'abord la définition unifiée de l'indice de variabilité relative à une famille exponentielle naturelle positive à travers sa fonction variance. Ensuite, nous montrons la normalité asymptotique des statistiques de tests correspondantes et donnons des exemples applicables. Des études de simulations ont mis en évidence de bons comportements de ces statistiques de tests asymptotiques. Des remarques finales sont faites avec de possibles extensions.

On the Asymptotic Normality between Sample Quantiles and Dispersion Estimators

Marcel BRAUTIGAM (ESSEC Business School (CREAR), Sorbonne University (LPSM), LabEx MME-DII,)
Marie KRATZ (ESSEC Business School (CREAR))
In this study, we derive the joint asymptotic distributions of functionals of sample quantiles and functionals of measure of dispersion estimators (the sample variance and the sample mean absolute deviation).

Modèles de régression pour les géométriques Poisson-Tweedie des données ultra-dispersées de comptage

Rahma ABID (Doctorante)
Célestin C. KOKONENDJI (Profeseur)
Afif MASMOUDI (Professeur)
Une nouvelle classe de mélange Poisson-exponentiel-Tweedie (PET) est introduite dans le cadre des modèles linéaires généralisés pour l'analyse des données de comptage ultra-surdispersées. Le modèle proposé est équivalent aux modèles exponentiels-Poisson-Tweedie issus des sommes géométriques de variables Poisson-Tweedie. A cet égard, les modèles PET englobent les versions géométriques des modèles Hermite, de Neyman Type A, de Polya-Aeppli, de négative binomiale et de Poisson inverse Gaussienne. La géométrique décalée à zéro est considérée comme la distribution de référence. Des propriétés, dans les modèles PET, des nouveaux indices relatifs des phénomènes de dispersion et de zéro-inflation sont alors établies. Les modèles de régression correspondants sont ajustés par l'approche de quasi-vraisemblance. Les performances de ces modèles sont illustrées sur des données réelles dans les domaines de la fiabilité et de l'assurance.

Apprentissage statistique : méthodes à noyaux (Amphi 12)

Modération : Erwan SCORNET

De l'importance de la fonction de poids dans le noyau des sous-arbres

Florian INGELS (INRIA, équipe MOSAIC, RDP, ENS de Lyon)
Romain AZAIS (INRIA, équipe MOSAIC, RDP, ENS de Lyon)
Les méthodes à noyaux sont une des approches permettant l'apprentissage à partir de données arborescentes. Parmi elles, nous nous intéressons au noyau des sous-arbres, qui présente l'avantage combinatoire

de pouvoir énumérer l'ensemble des objets utilisés pour estimer la similitude entre deux arbres. Nous introduisons le concept de réduction DAG, puis de recompression DAG, qui mènent à un algorithme efficace pour calculer le noyau. La possibilité d'aborder une base de données sous un format fortement compressé nous permet d'introduire une nouvelle fonction de poids, qui parvient à capturer de l'information dans des données où les poids proposés dans la littérature n'y arrivent pas.

Smoothed discrepancy principle as early stopping rule in RKHS

Yaroslav AVERYANOV (Inria Lille-Nord Europe)
Alain CELISSE (Inria Lille-Nord Europe)
In this paper we work on the estimation of a regression function that belongs to a reproducing kernel Hilbert space (RKHS). We describe spectral filter framework for our estimator that allows us to deal with several iterative algorithms : gradient descent, Tikhonov regularization etc. The main goal of the paper is to propose a new early stopping rule by introducing smoothing parameter for empirical risk of the estimator in order to improve the previous results on discrepancy principle. Theoretical justifications as well as simulations experiments for the proposed rule are provided.

Estimation plug-in d'ensembles de niveau de la fonction de régression

Dang DAU (étudiant Ecole Polytechnique)
Thomas LALOE (Université Nice Côte d'Azur)
Rémi SERVIEN (INRA Toulouse)
Dans cette communication, nous étudions le comportement asymptotique d'un estimateur des ensembles de niveau de la régression. Nois choisissons une erreur donnée par le volume de la différence symétrique. Une vitesse exacte est obtenue pour un niveau fixé mais également dans le cas où le niveau est inconnu et doit être estimé.

Agrégation d'hold outs

Guillaume MAILLARD (Université Paris Sud)
Sylvain ARLOT (Université Paris Sud)
Matthieu LERASLE (Université Paris Sud)
La validation croisée est souvent utilisée pour sélectionner une règle d'apprentissage parmis une famille, souvent paramètrée (sélection d'hyperparamètres). L'article que nous présentons étudie une méthode voisine, appelée agrégation d'hold-out (Agghoo), qui mélange validation croisée et agrégation ; des liens peuvent aussi être établis avec le bagging. Nous obtenons les premières garanties théoriques sur Agghoo, ce qui assure que l'on peut l'utiliser sans risque : au pire, les performances d'Agghoo sont celles du hold-out, à constante près. Pour le hold-out, des inégalités oracle étaient connues dans le cas des pertes bornées, comme en classification binaire. Cette approche semble pouvoir être étendue, sous de bonnes hypothèses, à la plupart des problèmes de minimisation de risque. Sous des hypothèses faibles, nous obtenons notamment une inégalité d'oracle concernant le choix du paramètre de pénalisation des SVM à perte Lipschitz. Dans toutes ces situations, Agghoo vérifie donc une inégalité d'oracle. Cependant, des simulations suggèrent que le comportement réel est souvent bien meilleur que ce que la théorie peut démontrer pour l'instant. En particlier, l'agrégation conduit à une amélioration significative que les bornes théoriques actuelles venant du hold-out sont incapables d'expliquer. En conséquence, l'agrégation d'hold-out semble donc bien être compétitive en pratique, lorsqu'on la compare à la validation croisée.

Statistique directionnelle (Amphi 13)

Modération : Pham NGOC

On projection-based tests of uniformity on the hypersphere

Eduardo GARCIA-PORTUGUES (Carlos III University of Madrid)
Paula NAVARRO-ESTEBAN (University of Cantabria)
Juan CUESTA-ALBERTOS (University of Cantabria)
We study a projection-based class of uniformity tests on the hypersphere using statistics that integrate, along all possible directions, the weighted quadratic discrepancy between the empirical cumulative distribution function of the projected data and the projected uniform distribution. The class is motivated by the proposal by Cuesta-Albertos et al. (2009), which builds a Kolmogorov-Smirnov test statistic on the projections of the data in a set of randomly-chosen directions. Simple expressions for several test statistics are obtained for the circle and sphere, and relatively tractable forms for higher dimensions. Despite their different origins, the proposed class and the well-studied Sobolev class of uniformity tests (see, e.g., Prentice (1978)) are shown to be related. Our new parametrization proves itself advantageous by allowing to derive new tests for hyperspherical data that neatly extend the circular tests by Watson, Ajne, and Rothman, and by introducing the first instance of an Anderson and Darling (1954)-like test in such context. The asymptotic distributions and the local optimality against certain alternatives of the new tests are obtained. A simulation study corroborates the theoretical findings and evidences that, for certain scenarios, the new tests are competitive against previous proposals. Finally, a real data example illustrates the usage of the new tests. References [1] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. J. Amer. Statist. Assoc., 49 :765–769. [2] Cuesta-Albertos, J. A., Cuevas, A. and Fraiman, R. (2009). On projection-based tests for directional and compositional data. Stat. Comput., 19(4) :367–380. [3] Prentice, M. J. (1978). On invariant tests of uniformity for directions and orientations. Ann. Statist., 6(1) :169–176.

Inférence sur la position sphérique dans des situations de grande concentration

Davy PAINDAVEINE (Université libre de Bruxelles)
Thomas VERDEBOUT (Université libre de Bruxelles)
Motivé par le fait que les données circulaires ou sphériques sont souvent très concentrées autour d'une position $\boldsymbol{\theta}$, nous considérons l'inférence sur $\boldsymbol{\theta}$ dans des régimes asymptotiques de grande concentration sous lesquels la probabilité de toute calotte sphérique centrée à $\boldsymbol{\theta}$ converge vers un lorsque la taille d'échantillon n diverge vers l'infini. Plutôt que de nous restreindre aux distributions de Fisher-von Mises-Langevin, nous considérons une classe semiparamétrique beaucoup plus large de distributions à symétrie rotationnelle indicée par un paramètre de position $\boldsymbol{\theta}$, un paramètre scalaire de concentration κ et une nuisance fonctionnelle f. Nous déterminons la classe des distributions pour laquelle le phènomène de grande concentration décrit ci-dessus se matérialise lorsque κ diverge vers l'infini. Pour de telles distributions, nous considérons alors l'inférence (estimation ponctuelle, estimation par zone de confiance, tests d'hypothèses) sur $\boldsymbol{\theta}$ dans des scénarios asymptotiques où κ_n diverge vers l'infini à une vitesse arbitraire avec n. Notre étude asymptotique révèle que, de façon intéressante, les procédures d'inférence optimales ont des taux de convergence qui dépendent de f. En ayant recours à une asymptotique à la Le Cam, nous montrons que la moyenne sphérique est, pour toute nuisance f, un estimateur paramétriquement super-efficace de $\boldsymbol{\theta}$ et que les tests de Watson et de Wald pour $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ jouissent de propriétés non-standards d'optimalité similaires. Nos résultats sont illustrés par des simulations. D'un point de vue technique, nos résultats asymptotiques requièrent des développements délicats de fonctionnelles à symétrie rotationnelle pour des grandes valeurs de l'argument de la nuisance fonctionnelle f.

Kernel circular density estimation with errors in variables

Agnese PANZERA (University of Florence)
Marco DI MARZIO (University of Chieti-Pescara)
Stefania FENSORE (University of Chieti-Pescara)
Charles C. TAYLOR (University of Leeds)
On considère le problème de l'estimation non paramétrique d'une densité circulaire à partir de données contaminées par des erreurs angulaires. On propose pour la tâche un estimateur à noyau dont les poids rappellent les noyaux de déconvolution. Une étude de simulation a été réalisée pour démontrer la performance de l'estimateur proposé.

Detecting the direction of high-dimensional spherical signals

Davy PAINDAVEINE (Université libre de Bruxelles)
Thomas VERDEBOUT (Université libre de Bruxelles)
We consider one of the most important problems in directional statistics, namely the spherical location testing problem, whose null is that the modal location of a Fisher-von Mises-Langevin (FvML) distribution on the p-dimensional unit sphere coincides with a given location . The underlying concentration parameter plays the role of a nuisance. We derive local asymptotic normality (LAN) results in a general high- dimensional framework where the dimension pn goes to innity, at an arbitrary rate, with the sample size n, and where the concentration behaves in a completely free way with the sample size, which covers a spectrum of problems ranging from arbitrarily easy to arbitrarily challenging ones. We identify seven asymptotic regimes, depending on the convergence/divergence properties of the concentration, that yield different limiting experiments and different contiguity rates. In each regime, we derive Le Cam optimal tests and we compute, from the Le Cam third lemma, asymptotic powers of the classical Watson test under contiguous alternatives. To obtain a full understanding of the non-null behavior of this test, we derive its local asymptotic powers in the broader, semiparametric, model of rotationally symmetric distributions. Monte Carlo studies show that finite-sample behaviours remarkably agree with our asymptotic results.

Epidémiologie I (Amphi 14)

Modération : Mélanie PRAGUE

Cartographie du risque appliquée aux fièvres d'origines inconnues à Djibouti

Mohamed ABDI KHAIRE (Université Clermont Auvergne, Laboratoire de Mathématiques, Clermont-Ferrand)
Hawa ADEN FARAH (Université de Djibouti, Département de Statistiques, République de Djibouti)
Anne-Françoise YAO (Université Clermont Auvergne, Laboratoire de Mathématiques, Clermont-Ferrand)

Le rapport présente une étude épidémiologique faite sur les fièvres d'origines inconnues à Djibouti. Ces dernières sont définies par une température corporelle supérieure à 38 degrés qui ne s'associe à aucune maladie transitoire ou auto-limitée. Le nombre de personnes atteintes par la fièvre d'origine inconnue n'a cessé d'augmenter durant les 10 dernières années. Nous nous intéressons à l'estimation du risque et à son évolution dans la capitale. Les risques (taux de mortalité standardisée) seront projetés sur une carte de Djibouti-ville puis nous les utiliserons pour l'exploration des données via des analyses statistiques.

Prédiction d'épidémies de grippe extrêmes

Maud THOMAS (LPSM, Sorbonne Université)
Holger ROOTZEN (Chalmers University of Technology)
Les épidémies de la grippe provoquent chaque année plus de 500,000 décès à l'échelle mondiale et présentent une forte morbidité, ce qui fait peser un fardeau supplémentaire sur des systèmes de santé déjà fragiles. Un des enjeux de la planification en santé publique est donc de prédire le risque de la survenue d'épidémies sévères ou extrêmes. Une épidémie est dite extrême lorsque le taux d'incidence dépasse un certain seuil (élevé). Notre objectif est de prédire l'occurence d'une épidemie extrême à court terme, par exemple dans les prochaines semaines. Pour cela, nous travaillons sur les taux d'incidence hebdomadaires de syndromes grippaux en France publiés par le réseau Sentinelles depuis 1985. La théorie des valeurs extrêmes a été développée pour prédire, à partir d'une série d'observations, la probabilité d'événements

plus extrêmes que ceux précédemment enregistrés. En nous basant sur l'approche proposée par Rootzén et al. (2018) et Kiriliouk et al. (2018), nous ajustons d'abord une loi de Pareto généralisée sur les premières semaines de l'épidémie. Puis, des estimations de la probabilité d'excès d'un seuil élevé sont obtenues à partir de la loi conditionnelle du modèle précédemment ajusté. Ces prédictions sont evaluées à partir de graphiques représentant les probabilités d'excès et de scores de Brier et comparées avec une régression logistique standard sur données simulées et réelles.

Etude sur le Climat et Lien avec les Fièvres à Djibouti

Mohamed ABDI KHAIRE (Université Clermont Auvergne, Laboratoire de Mathématiques, Clermont-Ferrand,)
Hawa ADEN FARAH (Département de Statistiques, Université de Djibouti, Djibouti,)
Anne-Françoise YAO (Université Clermont Auvergne, Laboratoire de Mathématiques, Clermont-Ferrand,)

Le monde connait actuellement un réchauffement climatique très important et plusieurs conséquences possibles de ces changements font l'objet d'un consensus scientifique. La République de Djibouti est classée dans la catégorie des pays semi-arides de par ses conditions climatiques sévères. Afin de comprendre le concept de réchauffement climatique, l'évolution de la température et de la pluviométrie de Djibouti-ville seront suivies sur une période de 55 ans allant de 1961 à 2016. L'objectif principal de l'article est la localisation des points de changements sur les tendances estimées pour chacune des séries et ainsi estimer un modèle pour les cas des fièvres en considérant les variables pluviométries et températures.

Régression bayésienne sur profils d'exposition : application en épidémiologie des rayonnements ionisants

Marion BELLONI (IRSN, PSE-SANTE/SESANE/LEPID)
Sophie BELLONI (IRSN, PSE-SANTE/SESANE/LEPID)
Chantal GUIHENNEUC (Université Paris Descartes, EA 7537)
En épidémiologie des rayonnements ionisants, les effets sanitaires des expositions professionnelles sont souvent étudiés séparemment pour chaque source de rayonnement. Or, les travailleurs sont exposés simultanément à plusieurs sources de rayonnements ionisants qui sont corrélées entre elles et également à certains agents chimiques et physiques. On s'intéresse dans ce travail au risque de décès par cancer du poumon dans la cohorte française des mineurs d'uranium. Ces mineurs sont exposés au radon, aux rayonnements gamma et aux poussières d'uranium, ainsi qu'à d'autres agents chimiques. Une méthode adaptée à la corrélation entre ces expositions est ici proposée. Si cette corrélation n'est pas prise en compte dans la régression multiple, on obtient des estimateurs instables, et donc non interprétables. On présente ici une approche hiérarchique bayésienne appelée 'régression bayésienne sur profils d'exposition' qui permet de traiter ce problème avec des variables explicatives continues et catégorielles. Cela consiste à regrouper les individus ayant des profils similaires, c'est-à-dire des caractéristiques proches, et d'estimer le risque associé à chaque groupe ainsi constitué. La répartition en groupes et l'estimation de risque se font conjointement sous le paradigme bayésien. L'inférence bayésienne du modèle proposé a été implémentée sous Python via un algorithme de type MCMC. Après un post-traitement des chaînes de Markov obtenues, l'identification et la caractérisation des groupes de mineurs d'uranium ayant des profils à haut risque et à bas risque de décès par cancer du poumon sont faites.

Jeunes statisticiens (Amphi 15)

Modération : Marie PERROT-DOCKES

Témoignage d'un jeune maître de conférences en Statistiques

Pierre GLOAGUEN (AgroParisTech)
La statistique est un domaine de recherche qui connaît une forte progression dans les universités et les instituts de recherche en France. Les th´ematiques liées à la statistique sont très variées, allant de l'application à la théorie mathématique. Nous aurons le témoignage de Pierre Gloaguen. Après un Master en Biostatistiques à Montpellier 2, et des études au sein de l'université de Rennes 1, Pierre a effectué un doctorat à l'Institut Français de recherche pour l'exploitation de la mer (IFREMER), sur l'utilisation de méthodes statistiques pour l'analyse et la modélisation du mouvement en écologie. A la suite de son doctorat, Pierre a effectué deux postdoctorat d'un an chacun. Le premier à l'école d'ingénieur AgroParisTech et le second à l'université Bretagne Sud à Vannes. Il a obtenu un poste de Maître de Conférences à AgroParisTech en septembre 2018. Il nous racontera son parcours, et décrira plus particulièrement les candidatures pour le concours Maître de Conférences.

Témoignage d'une jeune maître de conférences dans la recherche académique

Myriam TAMI (CentraleSupélec, Université Paris-Saclay, laboratoire MICS)
La statistique est un domaine de recherche très porteur, notamment par les défis que représentent l'analyse, la gestion et le traitement de données voir de gros volumes de données. Ainsi, des liens forts se créent entre les statistiques et l'intelligence artificielle. Par exemple, le machine learning est aujourd'hui une thématique de recherche incontournable. En interaction entre deux communautés, ce champ de recherche est très riche, comprenant de nombreux pôles : algorithmes, applications et théories mathématiques fines etc. Nous aurons le témoignage de Myriam Tami, qui après un master en mathématiques appliquées et recherche a effectué un doctorat à l'Institut Montpelliérain Alexander Grothendieck sur les modèles à équations structurelles avec facteurs latents. À la suite de son doctorat, Myriam a effectué un postdoctorat de 2 ans au Laboratoire d'Informatique de Grenoble (LIG) sur un projet de recherche en partenariat avec Total. L'objectif était le développement et l'implémentation d'une méthode d'apprentissage basée sur les arbres de décision permettant de gérer des données hétérogènes et incertaines. Dans le cas de la régression, cette nouvelle méthode d'ensemble s'est avérée compétitive avec des forêts aléatoires et ouvre de nouvelles perspectives de recherche. Depuis février 2019, Myriam est enseignante chercheuse en intelligence artificielle à CentraleSupélec au sein de l'Université Paris-Saclay. Elle nous présentera son parcours, ses expériences de recherche avec les industriels et abordera le processus des candidatures pour le concours Maître de Conférences.

11h20-11h50 : Pause café

11h50-12h50

Conférence Le Cam - Oleg Lepski : Estimation in the convolution structure density model (Amphi 11, retransmis en Amphi 12)

Université d'Aix Marseille

We study the problem of nonparametric estimation under L_p-loss, $1 \leq p < \infty$, in the framework of the convolution structure density model on R^d. This observation scheme is a generalization of three classical statistical models, namely density estimation under direct and indirect observations as well as partially contaminated observations. The original pointwise selection rule from a family of "kernel-type" estimators is proposed. For the selected estimator, we prove an L_p-norm oracle inequality and several of its consequences. Next, the problem of adaptive minimax estimation under L_p-loss over the scale of anisotropic Nikol'skii classes is addressed. We fully characterize the behavior of the minimax risk for different relationships between regularity parameters and norm indexes in the de nitions of the functional class and of the risk. We prove that the proposed selection rule leads to the construction of an optimally or nearly optimally (up to logarithmic factors) adaptive estimator.

Modération : Dominique PICARD

12h50-14h20 : Repas

14h20-15h40

Statistique bayésienne (Amphi 11)

Modération : Anne PHILIPPE

ABC within Gibbs

Grégoire CLARTE (Université Paris Dauphine)
Kerrie MENGERSEN (QUT)
Pierre PUDLO (Aix-Marseille Université)
Christian ROBERT (Université Paris Dauphine et Université de Warwick)
Robin RYDER (Université Paris Dauphine)

Julien STOEHR (Université Paris Dauphine)
Les méthodes ABC pâtissent des grandes dimensions. La grande dimension de l'espace des observations est traité par l'usage de statistiques résumées. Les grands espaces de paramètres, doivent être explorés plus efficacement, nous nous proposons ici d'adapter l'échantillonneur de Gibbs en utilisant des conditionnelles approchées par méthode ABC. La convergence de cet algorithme n'est pas évidente dans le cas général, on démontrera la convergence en variation totale par des méthodes de couplage. L'efficacité de cet échantillonneur est montrée sur plusieurs exemples.

Détection bayésienne d'outliers et ses applications en archélogie

Jean-Michel GALHARRET (Laboratoire LMJL, Univ NANTES)
Anne PHILIPPE (Laboratoire LMJL, Univ NANTES)
Norbert MERCIER (CNRS)
Nous nous intéressons à la détection d'outliers pour des questions de datation en archéologie. La méthode proposée est basée sur une extension du modèle d'événement proposé par Lanos et Philippe (2018). On exploite les hyperparamètres de ce modèle robuste pour identifier les outliers. Ces valeurs aberrantes sont alors supprimées de l'échantillon avant une ré-estimation du paramètre d'intérêt par une méthode non robuste. On applique cette procédure à la combinaison de dates et à l'estimation d'un âge par luminescence. Dans les deux cas nous montrons par des simulations qu'il est préférable d'exclure les outliers détectés plutôt que d'utiliser la méthode d'estimation robuste. Les résultats sont meilleurs en terme d'exactitude et de précision.

Sélection bayésienne de variables pour modèle linéaire à coefficients dynamiques

Benjamin HEUCLIN (Institut Montpelliérain Alexander Grothendieck, Université de Montpellier)
Marie DENIS (UMR AGAP, CIRAD)
Frédéric MORTIER (UPR Forêts et Sociétés, CIRAD)
Catherine TROTTIER (Université Paul Valéry Montpellier 3)
Comment l'architecture génétique des caractères quantitatifs évolue-t-elle au cours du temps ? La réponse à cette question est cruciale pour de nombreux domaines d'application tels que la génétique humaine et la sélection végétale ou animale. Au cours des dernières décennies, des techniques de génotypage à haut débit ont été utilisées pour mieux comprendre les liens entre l'information génétique et les caractères phénotypiques. Récemment, des méthodes de phénotypage à haut débit ont également été utilisées pour fournir de grandes quantités d'information à l'échelle phénotypique. En particulier, ces méthodes permettent de mesurer les caractères dans le temps, et ce, pour un grand nombre d'individus. La combinaison de ces deux informations peut donner des indications sur l'évolution de l'architecture génétique au cours du temps. Toutefois, ces données soulèvent de nouveaux défis statistiques liés, entre autres, à la dimension élevée, aux dépendances temporelles et aux effets variant dans le temps. Dans ce travail, nous proposons un modèle linéaire dynamique bayésien permettant, en une seule étape, l'identification des marqueurs génétiques impliqués dans la variabilité des caractères phénotypiques et l'estimation de leurs effets dynamiques. Cette approche combine des priors de type spike-and-slab pour la sélection de variables avec de l'interpolation de type P-spline pour l'estimation fonctionnelle des effets dynamiques.

MALIA/SSFAM : Apprentissage statistique - nouveaux défis (Amphi 12)

Modération : Stéphane CHRETIEN

Application du Transport Optimal en Fair Learning

Paula GORDALIZA (Institut de Mathématiques de Toulouse and IMUVA)
Eustasio DEL BARRIO (IMUVA)
Jean-Michel LOUBES (Institut de Mathématiques de Toulouse)
Nous fournissons un théorème de limite centrale pour la distance de Monge-Kantorovich entre deux distributions empiriques de tailles n et m, $\mathcal{W}_p(P_n, Q_m),\ p \geq 1$, pour les observations sur la droite réelle. Dans le cas $p > 1$, nos hypothèses sont précises en termes de moments et de finesse. Nous prouvons des résultats concernant le choix des constantes de centrage. Nous fournissons une estimateur consistant de la variance asymptotique qui permet de construire tests sur deux échantillons et des intervalles de confiance pour certifier la similarité entre deux distributions. Celles-ci sont ensuite utilisées pour évaluer un nouveau critère d'équité des ensembles de données dans la classification.

Signature, chemins rugueux et apprentissage statistique

Adeline FERMANIAN (Sorbonne Université)
Les applications modernes de la statistique et de l'apprentissage automatique ont mené à une explosion de données temporelles. On peut par exemple penser à la finance quantitative, aux enregistrements d'appareils médicaux ou à des trajectoires d'écriture manuscrite. De tels flux de données sont classiquement considérés comme des réalisations de processus stochastiques échantillonnés. Afin d'utiliser des algorithmes d'apprentissage classiques, il est nécessaire de représenter ces processus sous la forme de vecteurs de dimension finie. Nous présentons ici la transformation d'un flux de données multi-dimensionnel en sa signature, qui encode des propriétés géométriques du processus associé. La signature a été introduite dans les année 60 quand Chen (1958) a remarqué qu'un chemin peut être représenté par ses intégrales itérées, et a ensuite été au centre de la théorie des chemins rugueux de Lyons dans les année 90. La transformation en signature combinée avec un algorithme d'apprentissage a obtenu des résultats de pointe pour plusieurs applications, comme par exemple Yang (2016), ce qui soulève la question de ses propriétés statistiques. Nous allons donc présenter les principales propriétés de la signature puis étudier ses applications en apprentissage. Nous nous intéresserons en particulier à l'utilisation de la signature en régression, présentée dans Levin (2013). Compte tenu des résultats prometteurs dans la littérature, nous mènerons plusieurs tests empiriques sur les performances de cette transformation en comparaison d'autres algorithmes classiques.

Interprétabilité, forêts aléatoires, et ensemble de règles.

Clément BENARD (Safran Tech)
Sébastien DA VEIGA (Safran Tech)
Erwan SCORNET (CMAP)
Gérard BIAU (Sorbonne Université)
Les forêts aléatoires introduites par Breiman sont des algorithmes de régression et de classification parmi les plus performants. Cependant, le grand nombre d'opérations nécessaires pour effectuer une prédiction leur confère un aspect 'boîte noire'. A l'opposé les arbres de décisions ont une structure très simple mais instable, et une prédictivité limitée. Ces caractéristiques limitent fortement l'utilisation des arbres et des forêts pour certaines applications, par exemple l'analyse des processus de production dans l'industrie manufacturière. En effet, les décisions impactant des chaînes de production ont des conséquences lourdes, et ne peuvent reposer aveuglement sur des modélisations aléatoires. Les modèles se doivent d'être interprétables, c'est à dire à minima, simples, stables et prédictifs. Un troisième type de modèles, les ensembles de règles, présentent un compromis intéressant avec une structure simple et une capacité de prédiction comparable aux forêts, mais se caractérisent aussi par une certaine instabilité. Nous proposons un nouvel algorithme de classification d'ensemble de règles, extraites d'une forêt aléatoire. Pour les problèmes avec des interactions d'ordre faible, la méthode hérite d'une capacité de prédiction approchant celle des forêts, de la simplicité des arbres de décision, et d'une structure stabilisée. L'algorithme proposé présente à la fois des garanties théoriques asymptotiques, et de bonnes performances sur des données réelles.

Apprentissage supervisé avec données manquantes

Nicolas PROST (CMAP, Inria)
Julie JOSSE (CMAP)
Erwan SCORNET (CMAP)
Gaël VAROQUAUX (Inria)
Dans de nombreuses applications, les données sont affectées par des valeurs manquantes, qui perturbent l'analyse statistique. Une littérature abondante traite des données manquantes dans un cadre d'inférence, où l'objectif est d'estimer des paramètres et leurs variances à partir de tableaux incomplets. Ici, nous considérons un cadre d'apprentissage supervisé où l'objectif est de prédire au mieux une variable cible lorsque des données manquantes apparaissent à la fois dans le jeu d'apprentissage et de validation. Nous montrons la consistance de deux approches pour estimer la fonction de régression. Nous prouvons en particulier que l'imputation par la moyenne en amont de la phase d'apprentissage, technique très utilisée en pratique, est consistante lorsque les données manquantes ne sont pas informatives. Ceci contraste avec le contexte inférentiel où l'imputation par la moyenne est connue pour ses sérieux inconvénients en termes de déformation des lois marginales et jointe des données. Le fait qu'une approche si simple soit consistante a d'importantes conséquences en pratique. Ce résultat est valable asymptotiquement, pour un algorithme d'apprentissage dont le risque, en l'absence de données manquantes, tend vers zéro. Nous apportons des analyses supplémentaires sur les arbres de décision, car ils sont naturellement adaptés à la minimisation du risque empirique avec données manquantes. Cela est dû à leur capacité à prendre en compte la nature semi-discrète des variables avec données manquantes. Après avoir comparé théoriquement et empiriquement différentes stratégies pour prendre en compte les données manquantes dans la construction des arbres, nous recommandons d'utiliser la méthode 'missing incorporated in attribute' car elle peut gérer des données manquantes informatives et non informatives.

Séries chronologiques (Amphi 13)

Modération : Marianne CLAUSEL

Modèles tdVARMA$^{(n)}$ à coefficients dépendant du temps : propriétés des estimateurs

Guy MELARD (Université libre de Bruxelles, SBS-EM, Ecares, Bruxelles, Belgique)
Abdelkame ALJ (Université Moulay Ismail, FSJES, Meknès, Maroc)
Rajae AZRAK (Université Mohammed V - Rabat, Faculté des Sciences juridiques, économiques et sociales, Salé, Maroc)
Dans un article récent, les modèles VARMA à coefficients dépendant du temps t mais pas de la longueur n de la série, ou tdVARMA, ont été étudiés. Lors des journées de statistique de 2017, une nouvelle théorie asymptotique assez générale a été proposée pour l'estimation paramétrique des processus stochastiques à temps discret, non nécessairement stationnaires et ergodiques. Nous appliquons ici ces résultats aux modèles tdVARMA$^{(n)}$ où les coefficients peuvent aussi (mais ne doivent pas) dépendre de n, et aussi la matrice de covariance des erreurs. Contrairement à l'approche des processus localement stationnaires, les coefficients ne doivent pas dépendre de t/n et il n'est pas exigé qu'ils soient des fonctions lisses du temps. Les aspects numériques et pratiques de l'estimation sont également discutés. Enfin, nous traitons le modèle tdVMA$^{(n)}$ d'ordre 1 qui représente un cas particulier de tdVARMA$^{(n)}$, où nous vérifions que les hypothèses de la théorie sont satisfaites et pour lequel des résultats simulés sont présentés.

Processus autorégressif à bruits gaussiens stationnaires.

MARIUS SOLTANE (Laboratoire Manceau de Mathématiques.)
Le but de cet exposé est de s'intéresser dans un premier temps aux propriétés asymptotiques de l'estima-

teur du maximum de vraisemblance dans l'expérience statistique présentée dans le titre. Dans un second nous établissons la propriété LAN relative au ratio de vraisemblance de l'expérience considérée afin de pouvoir définir une notion d'efficacité asymptotique pour les estimateurs et construire une procédure optimale pour tester la significativité du paramètre autorégressif.

Etude comparative entre plusieurs tests de détection de la non linéarité dans les modèles autorégressifs d'ordre 1

Nabil AZOUAGH (Département Mathématiques, Faculté des Sciences ,UMP)
Said EL MELHAOUI (Faculté de Droit, Département d'Economie, UMP)
Dans ce travail, on s'intéresse aux tests capables de détecter la non linéarité dans les modèles autorégressifs d'ordre 1 (AR(1)). Nous proposons une étude comparative par simulation entre plusieurs tests sélectionnés pour ce problème. Ainsi, nous mettons en compétition les tests de Hinich (1982), Keenan (1985), Tsay (1986), Saikkonen and Luukkonen (1988) avec le test pseudo-Gaussien de Allal and El Melhaoui (2006) basé sur l'approche de Le Cam. De plus, nous appliquons ces tests sur la série temporelle classique de lynx du Canada, qui est bien connue comme une série non linéaire, afin d'examiner leurs pouvoir de détection de la non linéarité pour des données réelles.

Corrected LM tests for AR models with time-varying variance

Raja BEN HAJRIA (Faculté des Sciences de Monastir)
Dans le présent article, nous étendons l'approche de Breusch-Godfrey (Test des multiplicateurs de Lagrange (LM)) pour contrôler la qualité d'ajustement des modèles autorégressifs dans le cas des processus ayant une variance non conditionnelle non constante au cours du temps. Nous effectuons une comparaison asymptotique de la puissance de ces deux tests au sens de l'efficacité de Bahadur contre des alternatives linéaires.

Théorie des distributions : laquelle choisir en quelles circonstances ? (Amphi 14)

Modération : Christophe LEY

Comparison and classification of flexible distributions for multivariate skew and heavy-tailed data

Sladana BABIC
Christophe LEY
David VEREDAS
We present, compare and classify the most popular families of flexible multivariate distributions. Our classification is based on the tail behaviour (a single tail weight parameter or multiple tail weight parameters) and the type of symmetry (spherical, elliptical, central symmetry or asymmetry). We compare the flexible families both theoretically (comparing the relevant properties and distinctive features) and with a Monte Carlo study (comparing the fitting abilities in finite samples). Probability distributions are the building blocks of statistical modelling and inference. It is therefore of utmost importance to know which distribution to use in what circumstances, as wrong choices will inevitably entail a biased analysis. An example of such a situation is the 2008 financial crisis where financial institutions tended to use the multivariate Gaussian distribution for modelling the behaviour of their assets. The Gaussian distribution not accounting for extreme events led to an underestimation of risks with its known consequences. Financial

data have the peculiarity of being heavy-tailed and often they exhibit some form of skewness as negative events are usually more extreme than positive events. Moreover, portfolios tend to have hundreds of assets, hence financial data are often vast-dimensional or high-dimensional. Consequently, one needs multivariate distributions that are flexible, in the sense that they can incorporate skewness and heavy tails, and applicable in high dimensions. Also, their parameters should bear clear interpretations and parameter estimation ought to be feasible. These are the main characteristics that we expect from good multivariate distributions. The need for such versatile probability laws is also motivated by the increased computing power of our modern days. More and larger data sets from various domains get collected and require a high-quality analysis. Until the 1970s the multivariate normal distribution played a central role in multivariate analysis as well as in practice. However, the incapacity of the Gaussian distribution to accommodate for instance heavy tails led researchers to search for more general alternative distributions. A natural extension of the multivariate Gaussian distribution is the family of elliptical or elliptically symmetric distributions. Even though they retain the property of elliptical symmetry and hence cannot model skew data, this familiy of distributions allows for tails that are heavier and lighter than tails of the Gaussian distribution and thus are more flexible for data modelling. Numerous statistical procedures therefore have been built under the assumption of elliptical symmetry instead of multivariate normality. However, the symmetry assumption together with the fact that elliptical distributions are governed by a one-dimensional radial density and hence a single tail weight parameter have proved to be too restrictive for various data sets. Financial data obviously are one example. Meteorological data also often present skewness and heavy tails, a striking example being coastal floods where, besides the fact that the sea level has risen during the past century, land-rising respectively land-sinking can make an analysis complicated. For all these reasons, more flexible multivariate distributions than the elliptical ones are needed. Various distinct proposals exist in the literature, based on mixtures, skewing mechanism and copulas. The different research groups each work on their family of distributions. In the present paper, we fill this gap. We present the most popular flexible families of multivariate distributions and discuss their advantages and drawbacks. Our comparison goes crescendo in the sense that we start from the simplest families and every new section extends on the previous one. These extensions are based on the tail behavior, moving from a single tail weight parameter to multiple tail weight parameters, and/or the type of symmetry, moving from spherical, elliptical, central symmetry to skewness. By doing so we introduce a natural classification of these multivariate distributions that are of use to both theoreticians and practitioners.

Extreme Value Statistics for Specification of Design Aerodynamic Coefficient

Arnab SARKAR (Associate Professor, Department of Mechanical Engineering, IIT (BHU), Varanasi)
Gaurav GUGLIANI (Deaprtment of Mechanical Engineering, IIT (bHU), Varanasi)
It is seen from the statistics of natural hazards that the major parts of the world are affected by extreme winds. The statistics of damages clearly reveals that even today structures and structural components are not sufficiently wind hazard resistant. So the engineers are interested not only to forecast but also to design the structures properly. The wind load w is obtained as a function of the air density ρ, the wind speed v and the aerodynamic coefficient c. At least v and c have to be treated as random variables. So the specification of the design aerodynamic coefficient is a prerequisite for the specification of the design wind load. The specification of the design aerodynamic coefficient requires extreme value analysis of extreme aerodynamic coefficients. Extreme aerodynamic coefficients can be sampled from different runs of full scale and wind tunnel experiments. Then they can be fitted in a suitable extreme value distributions like Gumbel (type I), Fréchet(type II) or Reverse Weibull (type III) distribution. The different probability distributions for extreme aerodynamic coefficients can be compared and the best one can be found. In this way, the aerodynamic coefficient can be specified for the design working life of a structure.

Florence FORBES, Wraith DARREN

Robust mixture modelling using skewed multivariate distributions with variable amounts of tailweight

Florence FORBES
Wraith DARREN
The family of location and scale mixtures of Gaussians has the ability to generate a number of flexible

distributional forms. It nests as particular cases several important asymmetric distributions like the Generalised Hyperbolic distribution. The Generalised Hyperbolic distribution in turn nests many other well-known distributions such as the Normal Inverse Gaussian (NIG) whose practical relevance has been widely documented in the literature. In a multivariate setting, we propose to extend the standard location and scale mixture concept into a so called multiple scaled framework which has the advantage of allowing different tail and skewness behaviours in each dimension of the variable space with arbitrary correlation between dimensions. Estimation of the parameters is provided via an EM algorithm with a particular focus on NIG distributions. Inference is then extended to cover the case of mixtures of such multiple scaled distributions for application to clustering. Assessments on simulated and real data confirm the gain in degrees of freedom and flexibility in modelling data of varying tail behaviour and directional shape.

STID (Amphi 15)

Modération : Antoine ROLLAND

Réalisation d'interfaces spécifiques permettant l'exploitation d'un datawarehouse immobilier (pour la session spéciale STID)

Samuel GOUTIN
Dans le cadre de mon stage de fin de cursus de DUT STID, j'ai choisi de le réaliser dans la société Habitelem, une filiale du groupe Action Logement qui oeuvre dans le secteur du logement social dans les Pyrénées-Atlantiques (64). En 2017, le service informatique a mis en service un entrepôt de données destiné à alimenter des restitutions qui prennent la forme de tableaux de bord, pour permettre un meilleur suivi de l'activité de l'entreprise. Initialement, ma mission consistait à répondre aux nouvelles demandes en réalisant des restitutions avec le QlikSense (un logiciel de business intelligence). Ainsi, durant mon stage, j'ai pu en réaliser trois : une concernant la modélisation de l'évolution de la vacance locative dans le temps, une autre sur le clustering des demandeurs de logement et une dernière sur la prédiction des impayés des clients. Cependant, QlikSense ne permet pas de réaliser des traitements statistique d'une telle complexité. Alors comment répondre durablement à des besoins spécifiques mettant en oeuvre des méthodes statistique poussées ? Comme substitut de QlikSense, j'ai imaginé et conçu une solution de reporting : les analyses seraient effectuées avec R, mise en forme sur LaTeX puis contrôlées par une interface graphique permettant de paramétrer, générer et envoyer les rapports par e-mail aux concernés. Outre aider directement à orienter les décisions, mon travail a laissé entrevoir à l'entreprise une nouvelle manière d'utiliser ses données qui, jusqu'à maintenant, n'étaient que très faiblement exploitées.

De quelles ressources les enseignants du secondaire ont-ils besoin pour enseigner les probabilités et la statistique ?

Frédérique LETUE (STID Grenoble, LJK)
Maxime ARRAGON (LP ESSM Grenoble)
Guillaume CHEDALEUX (LP ESSM Grenoble)
Ludivine LEGRAND (LP ESSM Grenoble)
Juliette MAHE (LP ESSM Grenoble)
Le groupe 'enseignement de la statistique' de la SFdS a confié à un groupe de projet tutoré de Licence professionnelle 'Etudes statistiques, sondages et marketing' une enquête auprès des enseignants de mathématiques de l'enseignement secondaire pour mieux connaitre leurs besoins en ressources pour enseigner les probabilités et la statistique. Ces besoins en ressources concernent aussi bien des ressources pour consolider ou approfondir des concepts et méthodes, que des ressources pour les aider à enseigner

ces concepts aux élèves. L'objet de cet exposé est de présenter les résultats de cette enquête et d'en tirer des perspectives de création de ressources.

ONTOSTATS : un outil pour instrumentaliser les ressources éducatives en statistiques

Jean-marc MEUNIER (Université Paris 8)
Enseigner les statistiques suppose que les outils, conceptuels et matériels utilisés acquiert le statut d'instrument. C'est également le cas des ressources pédagogiques tant pour les étudiants que pour les enseignants. Une condition nécessaire est la structure des connaissances. Dans le projet Ontostats, nous avons utilisé l'ontologie STATO développée par l'Université d'Oxford pour indexer un dépôt de ressources dans une plateforme sous Omeka-S. Dans cet article, nous adoptons une perspective instrumentale pour en explorer les avantages, les perspectives, mais aussi les limites.

Plateformes et nouveaux outils pour la diffusion de la statistique

Vincent VANDEWALLE (Université de Lille)
A l'instar de nombreuses disciplines, l'apprentissage en ligne de la statistique connait un développement croissant ces dernières années. Des supports de plus en plus nombreux et variés sont disponibles : cours, vidéos, mais aussi plus récemment MOOC et plateformes dédiées. Ces nouveaux medias ouvrent la voie à un apprentissage interactif de la *data-science*. En parallèle, les logiciels statistiques, à l'image de R, évoluent. De nouvelles fonctionnalités sont développées pour permettre de compiler facilement texte, formules mathématiques et blocs de code. Les interfaces sont simplifiées et permettent désormais à l'apprenant d'interagir ou de mettre aisément en pratique une notion pédagogique précise. Dans cette présentation, nous passerons en revue ces nouveaux outils et nous étudierons les opportunités qu'ils offrent dans l'apprentissage de la statistique à un large public.

15h40-16h : Pause café

16h-17h

Alexandre Gramfort : What can statistics applied to neural signals tell us about the brain ? (Amphi 11)

INRIA

Understanding how the brain works in healthy and pathological conditions is considered as one of the major challenges for the 21st century. After the first electroencephalography (EEG) measurements in 1929, the 90's was the birth of modern functional brain imaging with the first functional MRI (fMRI) and full head magnetoencephalography (MEG) system. By offering noninvasively unique insights into the living brain, imaging has revolutionized in the last twenty years both clinical and cognitive neuroscience. After pioneering breakthroughs in physics and engineering, the field of neuroscience has to face new major computational and statistical challenges. The size of the datasets produced by publicly funded populations studies (Human Connectome Project in the USA, UK Biobank or Cam-CAN in the UK etc.) keeps increasing with now hundreds of terabytes of data made available for basic and translational research. The new high density neural electrode grids record signals over hundred of sensors at thousands of Hz which represent also large datasets of time-series which are overly complex to model and analyze : non-stationarity, high noise levels, heterogeneity of sensors, strong variability between individuals, lack of accurate models for the signals.
In this talk I will present some recent statistical machine learning contributions applied to electrophysiological data, and illustrate how optimization, statistics and advanced signal processing are used today to get the best of such challenging, and sometimes massive, data.

Modération : Joseph SALMON

Forrest Crawford : Causal inference under spillover and contagion - structural versus agnostic methods (Amphi 14)

Biostatistics, Yale University

Causal inference under spillover and contagion : structural versus agnostic methods Résumé : Two competing paradigms dominate statistical and econometric approaches to estimating the effects of interventions in interconnected/interacting groups under spillover or interference between experimental units. "Mechanistic" or "structural" models capture dynamic features of the process by which outcomes are generated, permitting inferences with real-world interpretations and detailed predictions. "Agnostic", "design-based", or "reduced form" approaches, often based on notions of randomization, refrain from specifying the full joint distribution of the data, and provide inferences that are robust to model mis-specification. Statisticians, economists, epidemiologists, and other scientists often disagree about which of these paradigms is superior for studies of interventions among potentially interacting individuals, with competing claims about model realism, bias, and credibility of inferences. In this presentation, I review methods for estimating the causal effect of an individualistic treatment under spillover, with special attention to the case of contagion, whereby units can transmit their outcome to others in a way that depends on their treatment. I define a formal structural model of contagion, and ask what causal features agnostic or reduced-form estimates recover. I exhibit analytically and by simulation the circumstances under which coefficients in a marginal regression model imply an effect whose direction is opposite that of the true individualistic treatment effect. Furthermore, I show that widely recommended randomization designs and estimators may provide misleading inferences about the direct effect of an intervention when outcomes are contagious. These ideas are illustrated in three empirical examples : transmission of tuberculosis, product adoption, and peer recruitment in social networks.

Modération : Mélanie PRAGUE

17h-18h : Assemblée générale de la SFdS

Mercredi 5 juin

9h-10h

Ghislaine Gayraud : Bayesian Quantile regression - an overview (Amphi 11)

LMAC, Université de Technologie de Compiègne

If the distribution of the response variable is highly skewed or in presence of outliers, it is well-known that traditional mean regression model may fail to describe interesting aspect of the relationship between the prediction and response variables. As an alternative to this classical linear regression, quantile regression (QR) is one of the most popular and useful regression technique. In this talk, I will give an overwiew of recent QR developments with a particular focus in the Bayesian framework.

Modération : Anne PHILIPPE

Julie Josse : On the Consistency of Supervised Learning with Missing Values (Amphi 14)

Ecole Polytechnique - INRIA

In many application settings, the data have missing features which make data analysis challenging. An abundant literature addresses missing data in an inferential framework : estimating parameters and their variance from incomplete tables. Here, we consider supervised-learning settings : predicting a target when missing values appear in both training and testing data. We show the consistency of two approaches in prediction. A striking result is that the widely-used method of imputing with the mean prior to learning is consistent when missing values are not informative. This contrasts with inferential settings where mean imputation is pointed at for distorting the distribution of the data. That such a simple approach can be consistent is important in practice. We analyze further decision trees. These can naturally tackle empirical risk minimization with missing values, due to their ability to handle the half-discrete nature of incomplete variables. After comparing theoretically and empirically different missing values strategies in trees, we recommend using the "missing incorporated in attribute" method as it can handle both non-informative and informative missing values.

Modération : Aurélie FISHER

10h-11h20

Statistique et sport 1 (Amphi 11)

Modération : Christophe LEY

Classification trees to define spatial performance indicators in basketball

Marica MANISERA (University of Brescia, Italy)
Paola ZUCCOLOTTO (University of Brescia, Italy)
Marco SANDRI (DMS StatLab, University of Brescia, Italy)
In this contribution we propose the definition of spatial performance indicators for basketball players and teams. The analysis of players' and teams' scoring probability in different areas on the court is very relevant in basketball analytics, because it enables coaches and experts to define better game strategies and training programmes. In this contribution we propose a method based on classification trees, which defines a partition of the court in rectangles with maximally different scoring probabilities. Shooting efficiency measures computed within the rectangles can be used to define spatial scoring performance indicators. Also, each analyzed team/player has its/his own partition, so comparisons can be easily made among different teams/players.

Classification trees to define spatial performance indicators in basketball

Marica MANISERA Paola ZUCCOLOTTO Marco SANDRI In this contribution we propose the definition of spatial performance indicators for basketball players and teams. The analysis of players' and teams' scoring probability in different areas on the court is very relevant in basketball analytics, because it enables coaches and experts to define better game strategies and training programmes. In this contribution we propose a method based on classification trees, which defines a partition of the court in rectangles with maximally different scoring probabilities. Shooting efficiency measures computed within the rectangles can be used to define spatial scoring performance indicators. Also, each analyzed team/player has its/his own partition, so comparisons can be easily made among different teams/players.

Ranking soccer teams on the basis of their current strength : A comparison of maximum likelihood approaches

Christophe LEY (Ghent University)
Hans VAN EETVELDE (Ghent University)
We present different strength-based statistical models that we use to model soccer match outcomes with the aim of producing a new ranking. The models are of four main types : Thurstone–Mosteller, Bradley–Terry, independent Poisson and bivariate Poisson, and their common aspect is that the parameters are estimated via weighted maximum likelihood, the weights being a time depreciation factor giving less weight to matches that are played a long time ago and eventually a match importance factor. Since our goal is to build a ranking reflecting the teams' current strengths, we compare the models on the basis of their predictive performance via the Rank Probability Score at the level of both domestic leagues and national teams.

Time-to-event analyses in sports injury research

Laurent MALISOUX (Sports Medicine Research Laboratory, Luxembourg Institute of Health)
La prévention des blessures sportives est importante pour les cliniciens, chercheurs, athlètes et la population active puisque toute mesure de prévention efficace peut contribuer non seulement à la poursuite de la participation à la compétition , mais aussi au maintient d'une vie active. Les premières étapes de toute démarche de prévention consistent en l'évaluation de l'étendue du problème et la description des mécanismes de blessures. Malheureusement, ces derniers sont encore insuffisamment compris. Des études prospectives avec des designs expérimentaux robustes et des approches statistiques avancées sont nécessaires pour mieux comprendre les mécanisme de survenue des blessures. L'objectif de cette présentation est d'expliquer pourquoi les analyses de survie apparaissent comme la méthode de choix pour la recherche sur la prévention des blessures sportives, et d'aborder les approches statistiques avancées qui devraient être considérées dans l'étude de l'étiologie des blessures. Les besoins et les contraintes des chercheurs en prévention des blessures sportives seront présentés afin de démontrer en quoi les analyses de survie semblent les mieux adaptées. Entre autres, elles permettent suffisamment de flexibilité aux chercheurs pour gérer les données censurées, les variables d'exposition qui varient dans le temps, les variables d'intérêt qui varient dans le temps, les événements récurrents, ainsi que les différents types de blessures qui entrent en compétition (risque compétitif). Ces modèles complexes nécessitent une étroite collaboration entre les chercheurs en prévention des blessures sportives et les statisticiens.

Problèmes inverses et parcimonie (Amphi 12)

Modération : Joseph SALMON

Linear Simplex Support Vector Regression

Quentin KLOPFENSTEIN (Université de Bourgogne)
Samuel VAITER (CNRS, Université de Bourgogne)
La déconvolution est une méthode utilisée dans la recherche contre le cancer afin d'obtenir la composition cellulaire d'un échantillon tumoral. Cette modélisation mathématique permet à partir de l'expression des gènes de la tumeur d'obtenir les proportions de cellules présentes au sein de la tumeur. L'estimation de ces quantités repose sur des modèles linéaires et notamment sur un modèle de regression à vecteurs de support. De par la nature de ce qui est estimé, l'estimateur obtenu dans le modèle doit avoir des coefficients positifs dont la somme est égale à 1. La méthode la plus utilisée aujourd'hui ne gère ces contraintes que dans un second temps, dans une étape de post-normalisation. Nous proposons donc un nouvel estimateur appelé Linear Simplex Support Vector Regression (LSSVR) qui prend en compte les contraintes liées à l'estimation de proportions directement dans le modèle. Nous étudions l'impact de ce changement dans la qualité de l'estimation et la stabilité de l'estimateur.

A sparsity regularization for functional linear discriminant analysis

Juhyun PARK (Lancaster University)
Jeongyoun AHN (University of Georgia)
Yongho JEON (Yonsei University)
For functional classification problem, it is well known that functional linear discriminant analysis can achieve a perfect classification, if the infinite-dimensionality is well exploited. Nevertheless, as functional data are inherently infinite-dimensional, consideration of dimension reduction plays a crucial role in its realization. Standard dimension reduction techniques based on functional principal component analysis or partial least squares methods are readily available for this purpose. On the other hand, there is an increasing need to incorporate interpretability within the formulation of our statistical analysis, which

tends to favour a simple and sparse solution. Such consideration is well developed for finite dimensional data with lasso type (ℓ_1) penalty but its infinite dimensional counterpart (L^1) is rarely studied for functional data. In this article, we reformulate the functional linear discriminant analysis as a regularization problem with appropriate penalty. An added advantage of using penalty formulation is the possibility of embedding some structural constraints in functional coefficient such as sparsity or smoothness as we desire. In particular, we propose a regularized functional linear discriminant analysis with L^1 functional sparsity penalty. We demonstrate that our formulation has a well defined solution and has a desirable functional sparsity property in the sense of domain selection. In addition, our solution is shown to converge to an optimal classifier. Numerical studies are included to assess finite sample performance and compare with existing methods.

Une approche par mollification au problème de déconvolution de densités

Anne VANHEMS (Toulouse Business School, France)
Léopold SIMAR (ISBA, Université Catholique de Louvain La Neuve)
Pierre MARECHAL (IMT, Université Paul Sabatier, Toulouse)
Dans cet article, nous utilisons la méthode de mollification pour régulariser le problème de la déconvolution. Cette nouvelle méthode de régularisation offre un cadre unifié et général pour comparer les avantages de différents types de régularisation comme les noyaux de déconvolution, la méthode de Tikhonov ou la méthode de spectral cutoff. En particulier, l'approche par mollification permet d'assouplir certaines hypothèses restrictives requises pour les noyaux de déconvolution, et a de meilleures propriétés stabilisantes comparées au spectral cutoff ou à Tikhonov. Nous prouvons la convergence asymptotique de notre estimateur et fournissons des simulations pour comparer les propriétés à échantillon fini de notre estimateur avec les méthodes classiques.

A mollifier approach to the nonparametric instrumental regression problem

Walter cedric SIMO TAO LEE (Institut de Mathématiques de Toulouse, Toulouse)
Pierre MARECHAL (Institut de Mathématiques de Toulouse, Toulouse)
Anne VANHEMS (Toulouse Business School)
Dans cet article, nous utilisons la méthode de mollification pour régulariser le problème de régression instrumentale non paramétrique. Cette méthode de régularisation a été utilisée dans des domaines de recherche tels que l'imagerie médicale, la tomographie, l'astrophysique, mais jamais en statistique ou en économétrie. Contrairement aux méthodes classiques de régularisation telles que Tikhonov ou Spectral Cut-off, un objet cible est clairement défini en fonction de l'objet initial inconnu : il s'agit maintenant de résoudre un nouveau problème dont la solution est une version lissée de l'objet inconnu, le lissage étant obtenu par convolution. Nous prouvons la convergence asymptotique de notre estimateur mollifié dans le contexte de la régression instrumentale non paramétrique et présentons des simulations pour étudier les propriétés de notre estimateur par rapport aux méthodes de régularisation classiques.

Statistique des processus 2 (Amphi 13)

Modération : Anna BONNET

Estimation paramétrique de séries de Hawkes localement stationnaires

Felix CHEYSSON (AgroParisTech / Institut Pasteur)
Gabriel LANG (AgroParisTech)
Laurence WATIER (Institut Pasteur)

Les processus de Hawkes sont une famille de processus stochastiques pour lesquels l'occurrence d'un événement modifie temporairement la probabilité d'occurrence des événements futurs. Dans le cas où le comptage des événements est observé en temps discret, nous proposons une approche spectrale de l'estimation du processus de Hawkes, faisant appel au spectre de Bartlett et à l'approximation de la vraisemblance proposée par Whittle. Pour permettre l'analyse de données dont la structure de probabilité varie dans le temps, nous étendons ensuite l'approche au cadre de stationnarité locale introduit par Dahlhaus. Ces approches sont accompagnées par des jeux de simulations illustrant la qualité de l'estimation, en particulier celle de la fonction d'excitation du processus de Hawkes dans le cas où les fenêtres d'observations sont grandes.

Statistical testing of the covariance matrix rank in multidimensional neuronal models

Anna MELNYKOVA (Universite de Cergy-Pontoise, Universite Grenoble Alpes)
Patricia REYNAUD-BOURET (Universite de Nice Sophia-Antipolis)
Samson ADELINE (Universite Grenoble Alpes)
Le but de ce travail est de développer une procédure de test qui détermine le rang du bruit dans un processus stochastique multidimensionnel à partir d'observations discrètes de ce processus sur un intervalle de temps fixe $[0, T]$ échantillonné avec un pas de temps Δ. Nous utilisons l'approche de perturbation aléatoire, utilisée pour l'estimation du rang de matrices non aléatoires, dans le cas d'un processus de diffusion stochastique. Nous menons une étude de simulation sur des modèles stochastiques multidimensionnels de l'activité neuronale : le modèle FitzHugh-Nagumo et une approximation stochastique du processus de Hawkes. Notre objectif principal est de contrôler le taux de perturbation, qui garantit des statistiques non dégénérées utilisées dans le test, et d'étudier son influence sur la précision du test pour une taille de pas fixe Δ.

Répartition des points critiques d'un processus ou champ gaussien isotrope

Céline DELMAS (INRA - MIAT)
Jean-Marc AZAIS (Institut de Mathématiques de Toulouse)
On étudie la répartition des points critiques d'une fonction aléatoire de $\mathbb{R}^N$ dans $\mathbb{R}$ suffisamment régulière (fonction de vraisemblance par exemple). Dans une première partie nous donnons l'espérance ainsi que la proportion exacte de chaque type de points critiques d'un champ gaussien isotrope $\mathcal{X} = \{X(t) : t \in \mathbb{R}^N\}$ sur un sous-ensemble S de $\mathbb{R}^N$. Les résultats sont obtenus sous forme de fonctions de pfaffiens de matrices antisymétriques $N \times N$ quand N est pair et $(N+1) \times (N+1)$ quand N est impair. Les calculs sont effectués dans les cas $N = 2$, $N = 3$ et $N = 4$. Dans une deuxième partie nous explorons le phénomène d'attraction, répulsion ou neutralité entre les différents points critiques de $\mathcal{X}$. Nous prouvons, pour tout champ gaussien isotrope, la répulsion entre les points critiques pour $N = 1$, la neutralité pour $N = 2$ et l'attraction pour $N > 2$. Plus généralement nous étudions la fonction de corrélation entre les points critiques d'index k et les points critiques d'index $k + i$.

Convergence du processus de Oja et ACP en ligne

Jean-Marie MONNEZ (Université de Lorraine, IECL, Inria)
Le processus de Oja est couramment utilisé pour estimer séquentiellement un vecteur propre associé à la plus grande valeur propre de l'espérance mathématique d'une matrice aléatoire symétrique en en utilisant un échantillon i.i.d., puis des vecteurs propres associés aux valeurs propres suivantes en ordre décroissant. Nous proposons deux extensions des hypothèses de convergence presque sûre de ce processus. Dans l'ACP en ligne d'un flux de données, ces extensions permettent de traiter des cas où la métrique utilisée est inconnue et est estimée en ligne, et également d'établir la convergence d'un processus où, au lieu d'utiliser plusieurs observations à chaque étape, on utilise toutes les observations jusqu'à l'étape courante, donc toute l'information contenue dans les données précédentes, sans avoir à les stocker.

Statistique et Santé (Amphi 14)

Modération : Nicolas SAVY

Distance de Fréchet et Dynamic Time Warping pour la classification non supervisée de séries chronologiques d'observance dans le syndrome d'apnées du sommeil

Guillaume BOTTAZ-BOSSON (HP2, LJK - Univ. Grenoble Alpes)
Sébastien BAILLY (Univ. Grenoble Alpes, Inserm, CHU Grenoble Alpes, HP2, 38000 Grenoble)
Agnès HAMON (LJK - Univ. Grenoble Alpes)
Adeline SAMSON (LJK - Univ. Grenoble Alpes)
Nous nous intéressons à la classification non supervisée des séries chronologiques. Les algorithmes de classification reposent sur des fonctions de dissimilarité. Nous présentons la dissimilarité discrète sommée de Fréchet (sdF) qui est une variante de la distance de Fréchet. Une étude de simulation permet de comparer les performances de la sdF avec le dynamic time warping (DTW), la distance euclidienne et la distance de Manhattan. Ces mesures sont comparées en utilisant des algorithmes hiérarchiques et à nuées dynamiques. Les meilleures performances sont atteintes en classification hiérarchique avec les dissimilarités DTW et sdF. Enfin, la classification hiérarchique est appliquée avec ces deux mesures sur des données réelles d'observance thérapeutique à la ventilation en Pression Positive Continue (PPC).

Modèle de Poisson mixte à classes latentes avec sur-représentation de zéros : application à l'identification de trajectoires hétérogènes d'intensité d'exposition vie entière

Emilie LEVEQUE (Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Biostatistics Team, UMR 1219)
Karen LEFFONDRE (Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Biostatistics Team, UMR 1219)
Cécile PROUST-LIMA (Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Biostatistics Team, UMR 1219)
De nombreuses études épidémiologiques tentent d'identifier des profils de trajectoires d'exposition avec la perspective d'estimer leur association avec le risque de survenue d'un évènement de santé, comme le risque de développer un cancer. Le modèle linéaire mixte à classes latentes permet de tenir compte de l'hétérogénéité dans les trajectoires de mesures répétées en identifiant des sous groupes d'individus ayant une évolution temporelle moyenne spécifique. Bien qu'il constitue une piste intéressante, ce modèle reste assez peu utilisé dans le cadre des expositions prolongées environnementales. Une des raisons est que les expositions environnementales ont généralement des distributions très particulières avec une large proportion d'intensités faibles ou nulles que les modèles linéaires mixtes à classes latentes classiques ne peuvent pas appréhender. Notre objectif était donc de proposer et implémenter un modèle de Poisson mixte à classes latentes avec sur-représentation de zéros (ZIP-LCMM) pour gérer ce genre de données. L'inférence est réalisée par maximum de vraisemblance. L'intégrale sur les effets aléatoires présente dans la vraisemblance est approchée par une méthode de quadrature gaussienne pseudo-adaptative en deux étapes. Ce modèle est appliqué à une étude sur la consommation de cigarettes vie entière et le cancer du poumon. Ces données proviennent de l'étude cas-témoins française, ICARE, basée en population générale ; seuls les hommes fumeurs ont été considérés ici (1938 cas / 1837 témoins). Le modèle a permis d'identifier des profils d'évolution de consommation de cigarettes bien distincts, tous associés à des risques de cancer du poumon différents. Avec le ZIP-LCMM, nous avons donc pu identifier des trajectoires d'exposition vie entière en prenant en compte une large proportion de zéros dans la distribution des mesures répétées. Ce développement méthodologique donne de nouvelles perspectives pour l'identification de trajectoires d'expositions environnementales vie entière.

Un cadre d'inférence bayésien pour l'étude de l'évolution de traits quantitatifs viraux

Paul BASTIDE (Department of Microbiology and Immunology, Rega Institute, KU Leuven)
Guy BAELE (Department of Microbiology and Immunology, Rega Institute, KU Leuven)
Marc SUCHARD (Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles)
Philippe LEMEY (Department of Microbiology and Immunology, Rega Institute, KU Leuven)
Au cours d'une épidémie, certains pathogènes viraux subissent une évolution rapide, si bien que les pressions évolutives liées à leur propagation se reflètent dans leur génome. Retrouver les traces moléculaires de ces processus de transmission est l'un des objectifs traditionnels du champ de la phylodynamique. L'évolution de traits quantitatifs viraux, comme la position géographique ou la virulence, a été relativement moins étudiée. Les méthodes comparatives phylogénétiques ont pour but d'étudier la distribution de traits quantitatifs au sein d'un ensemble d'organismes non indépendants, liés par une histoire évolutive partagée. Conditionnellement à cette histoire, les traits observés peuvent être vus comme le résultat d'un processus stochastique courant sur les branches d'un arbre phylogénétique. On décrit ici un cadre d'inférence bayésienne pour l'étude de ces modèles d'évolution. Celui-ci repose sur une méthode de type MCMC, rendue possible grâce, d'une part, à une procédure d'échantillonnage non biaisée de l'espace contraint des paramètres du modèle, et, d'autre part, à une méthode de calcul efficace de la vraisemblance, qui exploite la structure arborescente du problème. La méthode proposée peut s'appliquer à une large gamme de processus stochastiques gaussiens, rendant possible une modélisation fine des divers processus biologiques mis en oeuvre. Elle est implémentée au sein du logiciel d'inférence phylogénétique BEAST, et permet, à titre d'exemple, de jeter un regard nouveau sur le problème de l'héritabilité de la virulence du VIH.

Validation d'un processus d'identification d'événements cancéreux survenus dans une cohorte de patients diabétiques de type 2 suivis au CHU de Poitiers

Elise GAND (CIC 1402 INSERM)
Evelyne LIUU (CIC 1402 INSERM)
Gautier DEFOSSEZ (Registre des cancers Poitou-Charentes)
Marc PACCALIN (CIC 1402 INSERM)
Les établissements de santé sont des sources riches et multiples de données de santé individuelles. Dans un contexte règlementaire de valorisation des données médicales disponibles, L'axe ACDC du Centre d'investigation Clinique 1402 souhaite valider un processus d'identification de survenue de cancers dans la cohorte SURDIAGENE de patients diabétiques de type 2. Le processus mis en place comporte trois phases. La première étape consiste à recueillir différentes sources d'informations disponibles pour chacun des patients au sein du système informatique de l'hôpital : les courriers d'hospitalisation, les comptes rendus d'anatomopathologie et les réunions de concertation pluridisciplinaires. Les venues comportant des actes reliés à un cancer sont également identifiées à partir des données du Programme de Médicalisation des Systèmes d'Information (PMSI). La deuxième étape est entièrement informatisée : un algorithme de décision, utilisant notamment une lecture automatisée des documents, identifie les patients potentiellement porteurs de cancer. Enfin suit une phase d'adjudication par des médecins spécialisés en oncologie afin de valider chaque cas de cancer, et ses caractéristiques notamment leur localisation. Les cas identifiés sont confrontés à ceux enregistrés par le Registre Général des Cancers Poitou-Charentes, membre du réseau FRANCIM (réseau français des registres du cancer), qui recense de façon standardisée et exhaustive les cas incidents de cancers depuis 2008, dans l'ex-région Poitou-Charentes, sous l'égide de l'Institut de veille sanitaire (InVS) et de l'institut national du Cancer (INCa). Notre processus montre une très bonne sensibilité (91%) et spécificité (98%). Nous pensons donc développer cet outil pour la recherche d'autres évènements de santé et généraliser la démarche à d'autres cohortes de patients. Mots-clés. Cohorte, Cancer, Données Censurées, Identification Automatique

11h20-11h40 : Pause café

11h40-12h40

Prix du Dr Norbert Marx - Simon BUSSY : C-mix : un modèle de survie en grande dimension, et son application sur des données génétiques (Amphi 11, retransmis en Amphi 12)

LPSM, UMR 8001, Sorbonne University, Paris, France

Nous introduisons un modèle de mélange supervisé dans un cadre d'apprentissage en analyse de survie, le C-mix, afin de détecter des sous-groupes de patients de différents pronostiques d'une part, et de les ordonner suivant leur risque d'autre part. Notre méthode s'applique dans un contexte de grande dimension avec l'utilisation d'une pénalité Elastic-Net permettant au modèle de détecter automatiquement les covariables d'intérêt. L'inférence est faite à l'aide d'un algorithme Quasi-Newton Expectation Maximization (QNEM) pour lequel des propriétés de convergence sont démontrées. Les performances statistiques de la méthode sont ensuite examinées au travers d'une étude en simulation, puis illustrées sur trois jeux de données publiques en cancérologie. Notre approche obtient les meilleurs résultats et surpasse l'état de l'art dans un tel contexte, à la fois en terme de C-index, d'AUC(t) et de prédiction de survie.

Modération : Aurélie FISHER

12h40-14h : Repas

14h-18h : Programme social

19h30-1h : Soirée de Gala

Jeudi 6 juin

9h20-10h20

Freddy Bouchet : Climate extremes and rare trajectories in astronomy computed using rare event algorithms and large deviation theory (Amphi 11)

CNRS et ENS de Lyon

Complex dynamical systems, for instance climate models or the chaotic evolution of the solar system, are extremely difficult to study using numerical simulations because of sampling issues. I will discuss two applications where mathematical approaches developed in the field of statistical mechanics gave us huge improvement for the sampling of rare events. First, extreme heat waves, as an example of rare events with huge impacts. Using a rare event algorithm, based on large deviation theory, we were able to gain two orders of magnitude for the estimation of very rare events with a climate model (GCM). We could then study phenomena that can not be studied otherwise, for instance extreme teleconnection patterns. The second application is the study of rare trajectories that change the structure of a planetary system. Their understanding also involves large deviation and instanton theory.

Modération : Benjamin GUEDJ

Andreas Groll : A hybrid random forest approach for modeling and prediction of international soccer matches (Amphi 14)

Faculty of Statistics, TU Dortmund University

Many approaches that analyze and predict the results of international matches in soccer are based on statistical models incorporating several potentially influential covariates with respect to a national team's success, such as the bookmakers' ratings or the FIFA ranking. Based on all matches from the four previous FIFA World Cups 2002-2014, we compare the most common regression models that are based on the teams' covariate information with regard to their predictive performances. Furthermore, an alternative modeling class is investigated, so-called random forests(Breimann, 2001). Within the framework of Generalized Linear Models (GLMs), the most frequently used type of regression models in the literature is the Poisson model. It can easily be combined with different regularization methods such as penalization (see, e.g., Groll and Abedieh, 2013; Groll et al., 2015) or boosting (Groll et al., 2018). Moreover, we analyze different predictor structures, including team-specific ability parameters and extensions to smooth, non-linear effects for metric covariates, which also can be tackled by suitable boosting techniques (compare, e.g., Bühlmann and Hothorn, 2007). Finally, random forests can be seen as mixture between machine learning and statistical modeling and are known for their high predictive power. For these different modeling techniques, the predictive performance with regard to several goodness-of-fit measures is compared. Based on the estimates of the best performing method all match outcomes of the FIFA World Cup 2018 in Russia are repeatedly simulated (1,000,000 times), resulting in winning probabilities for all participating national teams.

Modération : Christophe LEY

10h20-10h40 : Pause café

10h40-12h

AMIES (Amphi 11)

Modération : Gilles STOLTZ

Les semaines d'étude maths-entreprises d'AMIES

Myriam MAUMY-BERTRAND (Université de Strasbourg)
Gilles STOLTZ (CNRS / Université Paris Sud)
Cet exposé introduit la session spéciale AMIES : il présentera d'abord brièvement l'AMIES (agence pour les mathématiques en interaction avec l'entreprise et la société), son rôle, ses activités, ses dispositifs de soutien, et se focalisera ensuite sur un dispositif en particulier, à savoir les semaines d'étude maths-entreprises (SEME). Elles ont pour vocation de rassembler pendant une semaine des industriels et des groupes de jeunes chercheurs (doctorants, post-doctorants) de tous les domaines des mathématiques. Les deux autres exposés de cette session spéciale présenteront des résultats concrets issus de telles semaines.

Système de recommandation : algorithmes et application à la plateforme KeeSeeK

Emmanuelle CLAEYS (IRMA)
Myriam MAUMY-BERTRAND (IRMA)
Frédéric BERTRAND (IRMA)
Agniel VIDAL (Laboratoire Paul Painlevé)
Alexandre DELYON (IRMA)
Titin AGUSTIN NENGSIH (IRMA)
Cet article résume le déroulement du projet proposé par la société KeeSeeK dans le cadre de la Semaine d'Etude Maths-Entreprises (SEME) de Strasbourg organisée en novembre 2018. Ce projet porte sur la recommandation d'offres d'emploi dans un moteur de recherche en ligne. Nous nous plaçons dans le cas où le moteur de recherche doit proposer des offres d'emploi susceptibles de maximiser les clics des candidats. Ces offres ont été préalablement sélectionnées et l'utilisateur (possédant le moteur de recherche) souhaite améliorer sa sélection. Pour s'intéresser à ce problème, nous avons proposé une version revisitée d'un algorithme d'apprentissage par renforcement (Thompson Sampling) lorsque le nombre d'offres d'emploi disponibles est important (100)

Détection de pathologies à partir de données de monitoring chez les vaches

Frédéric LOGE (CMAP, Polytechnique)
Yanis AMIROU (DMA, Ecole Normale Supérieure)
Florian BOURGEY (CMAP, Polytechnique)
Sean HELLINGMAN (Wilfrid Laurier University)
Malo HUARD (LMO, Université Paris Sud)
Solène THEPAUT (LMO, Université Paris Sud)
Le groupe Seenergi est un acteur majeur du monde agricole. Une de ses filiales, Médria Solutions, a équipé des troupeaux de vaches de capteurs. Nous nous intéressons à la prédictabilité de boiteries, la seconde pathologie en termes de prévalence, à partir de données comportementales des vaches. Cette étude, menée dans le cadre de la SEME Orsay 2019, a impliqué un travail important sur les données, de compréhension comme de pré-traitement et nous a amené a testé plusieurs méthodes de Machine Learning combinées à plusieurs méthodes de feature engineering.

Estimation de densité (Amphi 12)

Modération : Vincent RIVOIRARD

Adaptation sur des espaces de Besov en estimation de la densité sous contrainte de confidentialité différentielle locale.

Amandine DUBOIS (CREST-ENSAI)
Cristina BUTUCEA (CREST, ENSAE-ParisTech)
Martin KROLL (CREST, ENSAE-ParisTech)
Adrien SAUMARD (CREST-ENSAI)

Nous nous intéressons à l'estimation non-paramétrique d'une densité de probabilité sous la contrainte supplémentaire que seules des données privatisées sont disponibles. A cette fin, nous adoptons une récente généralisation de la théorie minimax classique au cadre de la confidentialité différentielle locale et nous donnons des bornes inférieures sur la vitesse de convergence sur les espaces de Besov $B^s_{p,q}$ pour le risque L_r. Les vitesses de convergence dans le cas privé sont détériorées par rapport à celles obtenues dans le cadre classique mais révèlent un changement de régime analogue. Afin de répondre à l'exigence de confidentialité, nous suggérons d'ajouter aux coefficients d'ondelettes empirique un bruit de Laplace correctement calibré. Un estimateur non linéaire adaptatif par ondelettes avec un seuillage correctement choisi atteint la vitesse donnée par la borne inférieure à un facteur logarithmique près.

Stabilité d'une procédure d'inférence de réseaux en grande dimension

Emilie DEVIJVER (UGA - LIG - CNRS)
Mélina GALLOPIN (Université Paris Sud - I2BC)
Rémi MOLINIER (UGA - Institut Fourier)

L'inférence de réseaux permet d'évaluer et de représenter les dépendences entre des variables continues. Les modèles graphique gaussiens ont été développés pour résoudre ce problème en grande dimension sous certaines hypothèses. Ce papier porte sur la stabilité d'une procédure d'inférence appelée shock et introduite dans Devijver et Gallopin (2018), qui infère un réseau modulaire via une matrice de covariance diagonale par blocs. Cette structure a beaucoup d'avantages, dont la réduction de dimension, l'interprétabilité et la stabilité. Ce dernier point est explicité dans ce papier, d'un point de vue théorique (via des arguments topologiques) et d'un point de vue numérique.

Estimation of a distribution function defined by Lagrange polynomials and Tchebytchev points

Salima HELALI (Université de Sfax Tunisie)
Yousri SLAOUI (Université de Poitiers)

Nous proposons un estimateur d'une fonction de répartition a l'aide des polynômes d'interpolation de Lagrange et la répartition de Tchebychev. Nous étudions les propriétés de cet estimateur et nous le comparons avec celui de l'estimateur de la fonction de répartition de Vitale. Nous montrons que notre estimateur domine celui de Vitale en terme de risque. Ensuite, nous confirmons ces résultats théoriques par des simulations.

Statistique et sport 2 (Amphi 13)

Modération : Christian DERQUENNE

Clustering of national elite sport policy systems and association with international sporting performance

Anne RENAUD (Swiss Federal Institute of Sport Magglingen)
Veerle DE BOSSCHER (Vrije Universiteit Brussel)

Hippolyt KEMPF (Swiss Federal Institute of Sport Magglingen)
National governments search for efficient ways to improve their results at international sporting events such as the Olympic Games. The SPLISS 2.0 project (Sport Policy factors leading to International Sporting Success) observed a great diversity between the elite sport policy systems of sixteen nations. The current study built on data from the SPLISS 2.0 project to assess the existence of groups of similar sport policy systems. Sports performances as well as the size and wealth of nations were then compared between the groups. Cluster analysis (hierarchical and k-means) and Kruskal-Wallis and Wilcoxon tests were applied to answer the research questions. Four groups of sport policy systems were identified (Leading, Challenging, Emerging and Specific). Population size, GDP per capita and sport performance in summer sports were significantly different between clusters ; performance in winter sports wasn't.

Clustering et régression de données fonctionnelles par processus Gaussiens : Application à la prédiction de performance en natation

Arthur LEROY (Université Paris Descartes)
Servane GEY (Université Paris Descartes)
Pierre LATOUCHE (Université Paris Descartes)
Une grande part des données récoltées en science du sport vient de phénomènes dépendant du temps. Récemment, plusieurs structures sportives, comme les clubs ou les fédérations, ont collecté des données longitudinales dans l'espoir qu'elles puissent aider à la détection des jeunes à haut potentiel. Cependant, plusieurs études ont mis en avant le fait que la plupart des meilleurs jeunes ne restent pas au même niveau de performance une fois adulte. C'est pourquoi le problème de la détection pourrait bénéficier de méthodes d'analyse de données objectives et notamment du domaine de l'apprentissage statistique. Lors de cette étude, l'objectif réside dans la prédiction de performances futures d'un athlète à partir de ses performances passées et de l'information apportées par apprentissage sur les autres athlètes. La progression des sportifs étant intrinsèquement continue et les temps d'observations étant très irréguliers, les données seront considérées comme fonctionnelles et lissées à l'aide de fonctions de bases B-splines. Ces observations fonctionnelles sont supposées être des réalisations de processus Gaussiens, et le problème de prédiction est également traité par régression par processus Gaussien. Plus précisément, un modèle mixte est utilisé avec un processus moyen commun à tous les individus sommé à un processus d'effets aléatoires individuels. Cette approche permet d'utiliser l'information de tous les individus pour la modélisation et règle ainsi le problème du faible nombre d'observations irrégulières. Préalablement, une étape de clustering est appliquée sur les données fonctionnelles, permettant par la suite une prédiction dépendante du groupe d'appartenance d'un individu. La procédure est estimée par une approche Bayésienne, qui permet de prendre en compte l'incertitude de modélisation et de prédiction naturellement, ainsi que le calcul d'intervalles de crédibilité. Une étude sur des simulations sera présentée ainsi que l'application sur un jeu de données réelles provenant de la Fédération Française de Natation. L'intérêt de ce travail est double, offrant une meilleure compréhension du phénomène de progression dans le sport, et fournissant un outil d'aide à la décision pour la détection de jeunes talents.

Revue de la littérature sur l'usage et mesusage des méthodes statistiques pour l'analyse de données compositionnelles d'entraînement en sciences du sport

Pauline DESNAVAILLES (Univ. Bordeaux, ISPED, F-33000 Bordeaux, France)
Carolyn INGRAM (Univ. Bordeaux, ISPED, F-33000 Bordeaux, France)
Thomas PRINCE (Univ. Bordeaux, ISPED, F-33000 Bordeaux, France)
Marta AVALOS-FERNANDEZ (Université de Bordeaux - Equipe SISTM, INSERM U1219 et INRIA BSO)
Les données de composition ou données compositionnelles (CoDa pour Compositional Data) consistent en des parts : leur somme vaut 100% pour chaque sujet, elles véhiculent une information relative. Puisqu'une composante peut être déterminée à partir de la somme du reste de la composition, les composantes sont mathématiquement et statistiquement dépendantes. Cette structure complique l'analyse et ne permet pas d'effectuer des inférences valides à partir d'analyses statistiques standard. Aitchison, 1982 et Egozcue et collègues, 2003, entre autres, ont fourni un cadre pour analyser des CoDa en projetant les données de l'espace simplex contraint à l'espace euclidien en utilisant des transformées non linéaires telles que la log-cote. Cependant, même après transformation, la nature compositionnelle reste inhérente

aux données. Le traitement statistique des CoDa est ainsi loin d'être un exercice aisé. Dans certains domaines d'application, les CoDa sont couramment utilisées alors que leur nature est ignorée. De nombreux tutoriels ont ainsi été proposés afin de guider l'analyse tout en tenant compte des spécificités du champ d'application. Qu'en est-il en sciences du sport ? L'objectif de ce travail est de faire un état des lieux sur le traitement statistique des CoDa dans un cadre précis, celui de l'analyse de la distribution de l'entraînement en termes d'intensités.

Epidémiologie II et prix SFDS-ENSAI (Amphi 14)

Modération : Jean-Michel POGGI

Phénotype de patients ayant un syndrome d'apnée du sommeil

Morgane BORDELAIS
Baptiste DELAHAIS
Thomas SAUVAGET
À partir des données issues de l'Observatoire Sommeil de la Fédération de Pneumologie concernant environ 45 000 patients atteint de Syndrome d'Apnée Obstructive du Sommeil (SAOS), nous construisons un modèle utilisant 24 variables basées sur des symptômes et des comorbidités. Une analyse en classes latentes révèle 8 classes qui se trouvent être nettement distinctes, homogènes, et intéressantes d'un point de vue clinique (avec notamment une classe composée uniquement de femmes, et trois autres en grande majorité d'hommes). Pour apprécier la robustesse de cette classification, un apprentissage supervisé par régression polytomique est mis en œuvre : le très faible taux de mal classés obtenu sur échantillons de validation est un résultat prometteur quant à l'intérêt de notre classification.

Modélisation spatio-temporelle de la chalarose (maladie fongique du frêne) en France

Coralie FRITSCH (Inria)
Anne GEGOUT-PETIT (IECL & Inria)
Benoit MARCAIS (INRA)
Marie GROSDIDIER (INRA)
La chalarose est une maladie du flétrissement du frêne apparue en Pologne en 1992 et observée pour la première fois en France en 2008. La maladie est causée par un champignon pathogène qui se développe sur les rachis infectés tombés aux pieds des arbres durant l'automne. Les champignons libèrent des spores qui se dispersent durant l'été et infectent ainsi les arbres voisins. Depuis 2008, environ 500 sites de forêt sont visités chaque année et nous disposons de la proportion d'arbres infectés parmi ceux visités pour chaque visite. Basé sur nos connaissance à propos du cycle de la maladie et afin d'estimer la propagation de la maladie dans les prochaines années, nous avons développé un modèle mécaniste spatio-temporel décrivant la propagation de la maladie. Ce modèle est basé sur un modèle paramétrique latent représentant l'infection des rachis et tenant compte des effets de l'humidité et de la température ainsi que d'un modèle de réaction-diffusion décrivant la propagation des spores. Pour l'inférence des paramètres, uniquement basée sur la proportion d'arbres infectés, nous avons utilisé une approche Bayésienne et des simulations MCMC.

Imputation multiple de comptages d'oiseaux d'eau à l'aide de covariables prédictives

Geneviève ROBIN (École Polytechnique, INRIA Saclay)
Mohamed DAKKI (Institut Scientifique, Université Mohammed V de Rabat, Maroc)
Hichem AZAFZAF (Association "Les Amis des Oiseaux" (AAO / BirdLife Tunisie), Ariana, Tunisie)

Khaled ETAYEB (Zoology Dept., Tripoli University, Tripoli, Libya)
Samir SAYOUD (Centre Cynégétique de Réghaia (C.C.R.), Alger, Algerie)
Nadjiba BENDJEDDA (Direction Générale des Forêts (DGF), Algérie)
Wed A.I. ABDOU (Egyptian Ministry of State For Environmental Affairs, Cairo, Egypt)
Pierre DEFOS DU RAU (Office National de la Chasse et de la Faune Sauvage, Unité Avifaune Migratrice, Arles, France)
Marie SUET (Institut de Recherche de la Tour du Valat, Arles, France)
Jean-Yves MONDAIN-MONVAL (Office National de la Chasse et de la Faune Sauvage, Unité Avifaune Migratrice, Arles, France)
Clémence DESCHAMPS (Institut de Recherche de la Tour du Valat, Arles, France)
Elie GAGET (Institut de Recherche de la Tour du Valat, Arles, France)
Laura DAMI (Institut de Recherche de la Tour du Valat, Arles, France)
En écologie, et en particulier en suivi de population d'espèces sauvages, nous sommes amen és à analyser des tableaux de comptages contenant une grande proportion de données manquantes. Par ailleurs, des informations complémentaires, telles que des covariables décrivant les lignes (sites écologiques) et les colonnes (années, espèces, etc.) du tableau sont souvent disponibles. Nous proposons une nouvelle méthode d'imputation multiple pour les données de comptage avec données manquantes aléatoirement (Missing At Random, MAR) qui incorpore des covariables complémentaires dans le processus d'inférence. Nous montrons empiriquement que la méthode surpasse les techniques classiquement utilisées par les écologues, en particulier lorsque le pourcentage de données manquantes est élevé. Nous appliquons la méthode à l'analyse de données d'abondance d'oiseaux d'eau d'Afrique du Nord.

Un cadre HMM spatial pour modéliser la dynamique d'espèces avec stade de dormance

Nathalie PEYRARD (MIAT, INRA)
Sebastian LE COZ (MIAT-INRA)
Pierre-Olivier CHEPTOU (CEFE, CNRS)
De nombreuses espèces ont un stade dormant dans leur cycle de vie, par exemple les graines pour les plantes. L'état de la population dormante influence la dynamique de l'espèce, cependant elle est souvent difficilement détectable. Une façon d'inclure la dormance dans un modèle dynamique est alors de considérer l'état de la population dormante comme un état caché et de se placer dans le cadre des modèles de Markov cachés. De tels modèles ont déjà été proposés mais avec plusieurs limites : les populations dormantes et non dormantes sont modélisées comme des variables binaires (présence-absence), la durée de la dormance est limitée à un seul pas de temps, la colonisation par les patchs voisins n'est pas prise en compte. Nous proposons un modèle de Markov caché qui lève ces limites et permet ainsi de mieux décrire à la fois la dynamique locale et régionale d'une espèce avec dormance : le modèle de Markov caché multidimensionnel avec rétroaction des données. Pour un modèle Markovien multidimensionnel, la complexité de l'estimation des paramètres du modèle par l'algorithme EM est a priori exponentielle en fonction du nombre de patchs, ainsi que la complexité de restauration de l'état caché et de la prédiction. Cependant nous démontrons que pour le modèle proposé ces requêtes sont réalisables pour une complexité en temps linéaire en fonction du nombre de patchs. Des tests sur des données simulées montrent que les estimateurs obtenus sont de bonne qualité, ainsi que que les performances en terme de restauration et de prédiction. Ce nouveau cadre fournit un outil simple et efficace pour l'analyse et la comparaison des dynamiques de plantes, comme par exemple les différentes stratégies de survie des espèces adventices dans les parcelles cultures.

Valeurs extrêmes multivariées (Amphi 15)

Modération : Anne SABOURIN

Modéliser des risques joints extrêmes et intermédiaires

Simon CHATELAIN (Université Lyon 1 et McGill University)
Anne-Laure FOUGERES (Université Lyon 1)
Johanna NESLEHOVA (McGill University)
Les copules Archimax fournissent un modèle souple permettant de prendre en compte tout type de dépendance asymptotique entre extrêmes, en décrivant simultanément les risques joints à des niveaux intermédiaires, sous-asymptotiques. Une copule Archimax est caractérisée par deux paramètres fonctionnels, la fonction de dépendance caudale l et le générateur Archimédien ψ, qui déforme la structure de dépendance des valeurs extrêmes. L'objectif de cette présentation est de faire état des récents développements obtenus en inférence semi-paramétrique pour les copules Archimax, à savoir l'obtention d'un estimateur non-paramétrique de l et d'un estimateur basé sur les moments de ψ, supposant que cette dernière fonction appartient à une famille paramétrique. Plusieurs aspects importants seront discutés, incluant les conditions d'identifiabilité, le comportement asymptotique des estimateurs, ainsi que leurs performances pour de petites tailles d'échantillon. Une analyse de données de maxima mensuels de pluie en trois stations de Bretagne sera finalement présentée, exhibant une excellente adéquation du modèle Archimax avec générateur de Clayton, y compris dans les queues inférieure ou supérieure.

Classification binaire sur les régions extrêmes

Hamid JALALZAI (Télécom Paristech)
Stéphan CLEMENCON (Télécom Paristech)
Anne SABOURIN (Télécom Paristech)
Parmi les diverses applications relatives à la détection d'anomalies les observations extrêmes jouent un rôle essentiel car les anomalies correspondent souvent à de grandes observations. La question clé est alors de faire la distinction entre les grandes observations issue la classe normale et celles provenant de la classe des anomalies. C'est un problème de classification binaire dans les régions extrêmes. Cependant, les observations extrêmes contribuent de manière négligeable à l'erreur empirique, au vue de leur rareté. Par conséquent, les minimiseurs du risque empirique ne bénéficient pas de garanties appropriées dans les régions extrêmes. Nous proposons un cadre général pour la classification des valeurs extrêmes. Plus précisément, dans le cadre d'hypothèses de distributions à queues lourdes non paramétriques, nous introduisons une version asymptotique du risque d'erreur mesurant la performance prédictive dans les régions extrêmes. Nous montrons que les minimiseurs d'une version empirique de ce risque ont une bonne capacité de généralisation, au moyen d'inégalités de concentration dans les régions à faible probabilité. Des expériences numériques illustrent la pertinence de l'approche développée.

Étude du support de la mesure spectrale

Nicolas MEYER (LPSM - Sorbonne Université)
Olivier WINTENBERGER (LPSM - Sorbonne Université)
Identifier les directions où des événements extrêmes se produisent est un problème majeure de la théorie des valeurs extrêmes multivariées. La mesure exposant de vecteurs à variation régulière permet de déterminer quelles coordonnées contribuent aux extrêmes. Cette mesure est définie par la convergence vague, ce qui ne fournit pas d'estimateur naturel de son support. Un problème analogue apparaît pour la mesure spectrale, qui est l'équivalent de la mesure exposant sur la sphère unité. Dans cette présentation, nous proposons une nouvelle méthode basée sur la projection ℓ^2 sur le simplexe. Cette projection permet de détecter sur quelles coordonnées se concentre la masse de la mesure spectrale. À partir de ces résultats et de la théorie de l'apprentissage, un test est proposé pour décider si une direction peut rassembler des événements extrêmes en grande dimension. Nous mettons ainsi en évidence les éventuelles coordonnées sur lesquelles la mesure spectrale ne met pas de masse et réduisons la dimension de l'étude des extrêmes.

12h-13h

Alexandra Carpentier : Adaptive inference and its relations to sequential decision making (Amphi 11)

Otto von Guericke Universitaet Magdeburg

Adaptive inference - namely adaptive estimation and adaptive confidence statements - is particularly important in high of infinite dimensional models in statistics. Indeed whenever the dimension becomes high or infinite, it is important to adapt to the underlying structure of the problem. While adaptive estimation is often possible, it is often the case that adaptive and honest confidence sets do not exist. This is known as the adaptive inference paradox. And this has consequences in sequential decision making. In this talk, I will present some classical results of adaptive inference and discuss how they impact sequential decision making.
This talk is based on joint works with Andrea Locatelli, Matthias Loeffler, Olga Klopp and Richard Nickl.

Modération : Gilles STOLTZ

Stephen Senn : In praise of small data (Amphi 14)

Consultant Statistician, Edinburgh

We have heard a lot in recent years about the power of big data. However, the revolution in statistical analysis that started one hundred years ago, with RA Fisher's appointment at Rothamsted, was all about analysing small data sets. One of the lesson learned then, which we sometimes forget, is that analysis depends on purpose and design. The latter is almost always absent from 'big data' and the former is sometimes even forgotten in looking at designed experiments. By considering various different sorts of clinical trials that might be applied to looking at treatment for asthma and by considering various sorts of questions that might be addressed in analysing clinical trials, I consider what lessons small data might have for big.

Modération : Nicolas SAVY

13h-14h20 : Repas - Déjeuners scientifiques des jeunes statisticiens

14h20-15h40

Statistique mathématique et estimation non-paramétrique (Amphi 11)

Modération : Antoine CHAMBAZ

Réduction du biais dans l'estimation non paramétrique des densités heavy tailed par la méthode du noyau

Smail ADJABI (Unité de Recherche LaMOS, Université de Béjaia, Algérie)
Nabil ZOUGAB (Unité de Recherche LaMOS, Université de Béjaia, Algérie)
On montre que les techniques multiplicatives de correction du biais (MBC) peuvent être appliquées à l'estimateur à noyau général Birnbaum-Saunders (GBS). Les propriétés des estimateurs a noyau MBS-GBS (biais, variance et erreur quadratique moyenne intégrée) sont obtenues. On utilise la méthode de validation croisée pour estimer le paramètre de lissage. Les performances des estimateurs MBC-GBS sont comparées selon les critères du biais quadratique intégré (ISB) et de l'erreur quadratique intégrée (ISE) aux performances de l'estimateur GBS standard, par une étude de simulation sur des densités heavy tailed (queue lourde) connues suivie d'une application sur des données réelles non négatives heavy tailed.

Aggregated kernel based tests in a regression model

Thi thien trang BUI (Institut de Mathématiques de Toulouse)
Béatrice LAURENT-BONNEAU (INSA, Département de Génie Mathématique)
Jean-Michel LOUBES (Institut de Mathématiques de Toulouse)
Dans un modéle de régression $Y_i = f(X_i) + \sigma\epsilon_i,\ i = 1, \cdots, n$, nous abordons la question du test de la nullité de la fonction f. Nous proposons tout d'abord une nouvelle procédure de test unique basée sur un noyau symétrique général et une estimation de la variance des observations. Les valeurs critiques correspondantes sont construites pour obtenir des tests de niveau non asymptotiques α. Nous introduisons ensuite une procédure d'agrégation afin d'optimism le choix des parametres du noyau. Les tests multiples vérifient les propriétés non asymptotiques et adaptatives au sens minimax de plusieurs classes d'alternatives classiques.

Local minimax rates for closeness testing of discrete distributions

Joseph LAM-WEIL (Magdeburg University)
Alexandra CARPENTIER (Magdeburg University)
Bharath K. SRIPERUMBUDUR (Pennsylvania State University)
Nous considérons le problème de comparaison de distributions entre deux échantillons dans un modèle de Poisson vectoriel. Ce modèle est connu pour être asymptotiquement équivalent à celui des distributions multinomiales. Le but est de distinguer si deux échantillons de données ont été tirés d'une même distribution inconnue ou si leurs distributions respectives sont séparées en norme L_1. Nous cherchons en particulier à adapter la vitesse de test à la forme des distributions inconnues. Ainsi, nous travaillons dans un cadre minimax local. A notre connaissance, nous fournissons la première vitesse minimax locale de test pour la distance de separation à des facteurs logarithmiques près, ainsi qu'un test qui l'atteint. En regard de la vitesse obtenue, le problème de test à deux échantillons est subtantiellement plus difficile que celui de test d'adéquation d'un seul échantillon à une loi connue dans de nombreux cas.

Fouille de données spatiales et de réseaux (Amphi 12)

Modération : Benjamin GUEDJ

Arbres CART pour données spatiales

Jean-Michel POGGI (LMO, Univ. Paris-Sud, Orsay)
Avner BAR-HEN (CNAM, Paris)
Servane GEY (MAP5, Univ. Paris Descartes)
En liant les partitions induites par les arbres de classification CART et les processus ponctuels marqués, nous proposons une variante spatiale de la méthode CART, SpatCART, qui se concentre sur le cas de deux populations. Alors que l'arbre CART habituel ne tient compte que de la distribution marginale de la variable réponse en chaque noeud, nous proposons de tenir compte de la position spatiale des observations. Nous introduisons un indice de dissimilarité basé sur la fonction intertype K de Ripley qui quantifie l'interaction entre deux populations. Cet indice utilisé pour l'étape de construction de l'arbre maximal de la stratégie CART, conduit à une fonction d'hétérogénéité cohérente avec l'algorithme CART original. La procédure proposée est mise en oeuvre, illustrée par des exemples classiques et comparée aux concurrents directs. SpatCART est enfin appliquée à un exemple de l'analyse de deux espèces d'arbres dans une forêt tropicale.

Clustering in weighted networks using binomial stochastic blockmodels

Abir EL HAJ (Université de Poitiers)
Yousri SLAOUI (Université de Poitiers)
Pierre-Yves LOUIS (Université de Poitiers)
Zaher KHRAIBANI (Université Libanaise)
Le modèle à blocs stochastiques est un modèle de graphe aléatoire qui vise à partitionner les sommets d'un réseau en groupes appelés blocs, ou plus généralement clusters. Dans la plupart des réseaux du monde réel, les liens entre les noeuds sont affectés par des poids qui représentent la force des relations entre ces noeuds. Il est évidemment très intéressant de pouvoir modéliser et regrouper ces réseaux pondérés en utilisant la structure du réseau et la capacité de leurs liens. Cet article présente le modèle à blocs stochastiques binomial, qui est un modèle probabiliste pour les réseaux ayant les poids sur les arêtes distribués selon une loi binomiale. Un algorithme variationnel d'espérance maximisation est proposé ici pour effecture l'inférence. Enfin, nous démontrons l'efficacité de la méthode proposée en considérant un réseau de co-citation dans un contexte de text mining.

Clustering and visualizing large cattle-trade networks using relational self-organizing maps

Madalina OLTEANU (MaIAGE, INRA)
Kevin PAME (MaIAGE, INRA)
Gael BEAUNEE (Bioepar, INRA)
Caroline BIDOT (ISPA, INRA)
Elisabeta VERGU (MaIAGE, INRA)
Farm contact networks or animal trade networks are being one of the most extensively studied mathematical objects these last few years. Indeed, research stemming from various fields – epidemiology, physical statistics, applied mathematics – aims at understanding the behavior of these complex, dynamic networks. One of the goals of these studies is to mindfully explore these massive data, extract meaningful information and structure, and eventually isolate specific patterns and clusters which may contribute to assessing phenomena such as preferential behaviors, epidemics spreading, ... The present contribution

addresses the insights that relational clustering trained with various distances computed from temporally reachable paths may bring in the exploratory study of dynamical networks. We illustrate our findings on a representative sub-network of the cattle-trade French network – Brittany –, monitored with a daily frequency between 2005 and 2009.

Données de composition (Amphi 13) (début à 14h10)

Modération : Gilbert SAPORTA

Classification de données de composition transformées et applications

Antoine GODICHON-BAGGIONI (Laboratoire de Probabilités Statistique et Modélisation)
Cathy MAUGIS-RABUSSEAU (Institut de Mathématiques de Toulouse)
Andrea RAU (INRA)
Bien qu'il y ait de nombreux algorithmes de classification dans la littérature, la question du choix de la meilleur stratégie à adopter lorsque l'on doit traiter des données de composition (i.e de données à valeurs dans le simplex) reste très largement sous explorée. Ce travail est motivé par l'analyse de deux applications basées sur la catégorisation de données de compositions ; (1) identifier des groupes de gènes 'co-exprimés' à partir de données RNA-seq ; et (2) trouver des tendances dans l'utilisation de stations Velib' à Paris. Pour chacune de ces applications, on utilise des transformations appropriées des données, telles que la transformation Centered Log Ratio, et une nouvelle transformation appelée Log Centered Log Ration conjointement avec l'algorithme des K-means.

Moyennes et fonctions de covariance pour les données compositionnelles : une approche axiomatique

Denis ALLARD (BioSP, INRA)
Thierry MARCHANT (U Ghent)
Ce travail s'intéresse à la caractérisation de la tendance centrale de données compositionnelles. Par une approche axiomatique, nous établissons de nouveaux résultats sur les propriétés théoriques de la moyenne et de la fonction de covariance pour ce type de données. Nous montrons tout d'abord que la moyenne arithmétique pondérée est la seule caractéristique de tendance centrale vérifiant les axiomes de réflexivité et de stabilité marginale. Par ailleurs, les poids intervenant dans la combinaison linéaire doivent être identiques pour toutes les composantes du vecteur compositionnel. Ce résultat a des conséquences profondes sur la structure du modèle de covariance spatiale multivariée de ces données. Dans un cadre géostatistique, nous montrons alors que le modèle de covariance proportionnelle (i.e. le produit d'une matrice de covariance et d'une fonction de corrélation) est le seul modèle de covariance pour lequel les poids du krigeage de la moyenne sont identiques pour toutes les composantes. La combinaison de ces deux résultats est que, dans le cadre des statistiques spatiales, le modèle de covariance proportionnelle est le seul modèle de covariance spatiale multivariée compatible avec les axiomes de réflexivité et de stabilité marginale.

CODA methods and the multivariate Student distribution : an application to political economy

Thi huong an NGUYEN (Toulouse School of Economics)
Thibault LAURENT (Toulouse School of Economics)
Le vecteur des proportions de vote par parti sur une subdivision donnée d'un territoire est un vecteur de

données dites de composition (mathématiquement, un vecteur appartenant à un simplexe). Les économistes politiques s'intéressent à l'impact des caractéristiques socio- économiques des unités géographiques sur le résultat des élections. Parce que les données de parts de votes présentent souvent davantage de valeurs extrêmes que des observations issues d'une loi normale, nous avons décidé d'utiliser une distribution d'erreur de Student dans le modèle de régression. Nous décrivons comment adapter le modèle de régression CODA à la distribution d'erreur multivariée de Student. Pour un vecteur d'erreur gaussien, l'hypothèse de coordonnées indépendantes équivaut celle de coordonnées non corrélées. Cependant, cette équivalence n'est plus vraie lorsqu'on envisage une distribution multivariée de Student. Dans cet article, nous nous concentrons sur la construction d'un modèle de régression CODA ayant des vecteurs d'erreurs de loi de Student multivariées indépendantes. Les modèles sont ajustés aux données électorales françaises des élections départementales de 2015.

Mesures d'impact des variables explicatives dans les modèles de régression pour variables compositionnelles

Christine THOMAS-AGNAN (Toulouse School of Economics)
Joanna MORAIS (Quantmetry)
Des vecteurs de composition peuvent intervenir dans un modèle de régression soit en tant que variables explicatives (voir Hron et al. 2012) soit en tant que variables à expliquer (voir Egozcue et al., 2012) et un modèle peut présenter les deux cas (voir Chen et al. 2016 and Morais et al. 2018b). Cependant il n'est pas facile de mesurer l'impact marginal des variables dans ces modèles car un changement sur une des composantes d'une composition affecte toutes les autres composantes. Morais et al. (2018a) ont montré que le bon outil pour mesurer ces impacts est la notion de dérivée simpliciale, c'est à dire dérivée d'une fonction dont soit l'ensemble de départ, soit l'ensemble d'arrivée, soit les deux est un simplexe (Pawlowsky-Glahn, V. and A. Buccianti, 2011). Dans le cas où les deux sont dans le simplexe, cela conduit à des interprétations en terme d'élasticités qui sont fréquemment utilisées en économétrie et dans les applications en marketing. Nous montrons ici comment ces dérivées simpliciales peuvent tre utilisées dans d'autres cadres : cas où seule la variable dépendante est une composition (basé sur le chapitre 12 de Egozcue et al. dans Pawlowsky-Glahn et al. 2011), celui où seule la variable indépendante est une composition (basé sur le chapitre 13 de Barcelo-Vidal et al. dans Pawlowsky-Glahn et al. 2011). Nous envisageons aussi les cas où un total associé a la composition intervient dans la régression en variable explicative. Les interprétations correspondantes sont en terme de semi-élasticités. Enfin, nous montrons comment construire des intervalles de confiance pour ces élasticités ou semi-élasticités qui permettent une meilleure interprétation de ces modèles. Nous illustrons sur un jeu de données réelles.

Médecine personnalisée (Amphi 14)

Modération : Jean-Marie MONNEZ

Actualisation en ligne d'un score d'ensemble

Benoît LALLOUE (Institut Elie Cartan de Lorraine ; Project-Team BIGS ; CIC-P 1433)
Jean-Marie MONNEZ (Institut Elie Cartan de Lorraine ; Project-Team BIGS ; CIC-P 1433)
Éliane ALBUISSON (Institut Elie Cartan de Lorraine ; BIOBASE CHRU de Nancy ; Faculté de Médecine de Nancy)
En construisant une collection de prédicteurs (en faisant varier les échantillons utilisés, les variables retenues, les règles d'apprentissage, ...) dont les prédictions sont ensuite agrégées, les méthodes d'ensemble permettent d'obtenir de meilleurs résultats que les prédicteurs individuels. Dans un contexte en ligne où des données arrivent de façon continue, on souhaite actualiser les paramètres d'un score construit à

l'aide d'une méthode d'ensemble. On considère le cas où il est impossible de conserver toutes les données obtenues précédemment et de recalculer les paramètres sur l'ensemble des données à chaque nouvelle observation. Nous proposons une méthode d'actualisation en ligne d'un score d'ensemble à l'aide de bootstrap Poisson et d'algorithmes stochastiques.

Régression non-linéaire pour l'analyse d'un médicament anticancéreux par diffusion Raman exaltée de surface

Tom ROHMER (INRIA)
Laetitia LE (CMAP - APHP)
Antoine DOWEK (APHP - Université Paris Sud)
Eric CAUDRON (APHP - Université Paris Sud)
Marc LAVIELLE (INRIA)
L'objectif de ces travaux est d'évaluer la faisabilité d'une technique en plein évolution, la spectroscopie Raman exaltée de surface (SERS) pour l'analyse de la concentration de médicaments anticancéreux. Cette technique utilisant des nanoparticules d'argent est appliquée à l'analyse quantitative du 5-fluorouracile, l'une des molécules les plus utilisée en cancérologie. Au vu des fortes variabilités spectrales entre les diverses répétitions de l'expérience et de l'interaction non-linéaire observée entre la concentration et l'intensité du signal, des méthodes de régressions non-linéaires permettant de prendre en compte ces variabilité ont été mis en places sur une base d'apprentissage puis comparés sur une base test.

Statistique et données complexes (Amphi 15)

Modération : Robert FAIVRE

La diffusion de Langevin comme modèle de mouvement en écologie pour le mouvement animal et la sélection d'habitat.

Pierre GLOAGUEN (AgroParisTech)
Théo MICHELOT (University of Sheffield)
Marie-Pierre ETIENNE (AgroCampus-Ouest)
En écologie, la distribution d'utilisation décrit la probabilité relative d'utilisation d'une unité spatiale par un animal. Il est naturel de penser que cette distribution est la conséquence à long terme des décisions de mouvement à court terme de l'animal : c'est l'accumulation de petits déplacements qui, avec le temps, donne lieu à des modèles globaux d'utilisation de l'espace. Cependant, la plupart des modèles de distribution d'utilisation utilisés en écologie ignorent le mouvement sous-jacent, en supposant l'indépendance des lieux observés, ou se basent sur des règles simplistes de mouvement brownien. Nous introduisons un nouveau modèle de mouvement animal en temps continu, basé sur la diffusion de Langevin. Ce processus stochastique a une distribution stationnaire explicite, conceptuellement analogue à l'idée de distribution d'utilisation, et fournit donc un cadre intuitif pour intégrer le mouvement et l'utilisation de l'espace. Nous modélisons la distribution stationnaire (utilisation) avec une fonction de sélection de ressources pour lier le mouvement à des covariables spatiales. Le choix de la fonction de sélection de ressource classique conduit à un schéma d'approximation naturel qui n'a besoin que des outils classiques du modèle linéaire.

The bivariate J-function to analyse positive association between galaxies and galaxy filaments

Kruuse MAARJA (Tartu Observatory, Tartu University)
Elmo TEMPEL (Tartu Observatory, Tartu University)

Rain KIPPER (Tartu Observatory, Tartu University)
Radu STOICA (Université de Lorraine, CNRS, IECL)
Le réseau de filaments formé par la position des galaxies est une des structures les plus fascinantes dans notre Univers. Dans cet article, nous étudions des possibles associations entre le réseau de filaments déjà détecté et des nouvelles observations. Ces observations très récentes sont obtenues à partir des mesures du décalage vers le rouge photométrique. L'utilisation de la fonction J-bivariée tend à indiquer une association positive entre les filaments existants et les nouvelles observations.

Classification de variables : une approche dynamique en grande dimension

Christian DERQUENNE (EDF R&D)
La recherche de structures dans les données représente une aide essentielle pour comprendre les phénomènes à analyser. Les méthodes de classification de variables permettent de répondre à cette problématique, mais elles peuvent être pénalisées par un trop grand nombre de variables. Nous proposons une nouvelle approche de type 'Diviser pour Régner' fondée sur le principe MapReduce pour pallier ce problème. La table de données est divisée en plusieurs sous-tableaux traités en parallèle, puis réconciliés à l'aide de l'Analyse des Correspondances Multiples. Cette approche est appliquée sur des données simulées et fournit de très bons résultats.

Bayesian Inference For A Manifold Gaussian Process Classifier : Application To Metallic Boxes

Anis FRADI (CNRS-LIMOS (UMR 6158), UCA, France)
Chafik SAMIR (CNRS-LIMOS (UMR 6158), UCA, France)
Anne-Françoise YAO (CNRS-LMBP (UMR 6620), France)
One of the challenging regression problems consists of learning relevant and meaningful relationships between high dimensional representations across a relatively few observed individuals. Since this problem could have drastic effects particularly on the classification performance, we propose a Bayesian alternative in the case of Gaussian Process classifier (GPc) depending on a covariance function. It is commonly known that methods based on GPcs are effective mainly in the case of low and medium dimensional data. On a mathematical basis, given N training inputs : $\mathbf{X} = \{x_i\}_{i=1}^{N}$ and a covariance function $c(.,.)$, the most prominent weakness of the standard GPc inference is that it suffers from time complexity because of the calculation of the inversion and determinant of the $N \times N$ covariance matrix : $\mathbf{C} = c(\mathbf{X}, \mathbf{X})$. The manifold Gaussian process classifier (MGPc) formulation has the advantage of learning a GPc in a feature space under some constraints such as : the taking into account of nonlinearity, the increase of the separability, the dimensionality reduction, etc. We illustrate the efficiency and the accuracy of our framework for classifying images of manufacturing defects.

Non-paramétrique (Amphi 16)

Modération : Thomas LALOE

Hermite density deconvolution

Ousmane B SACKO (MAP5 UMR 8145, Université Paris Descartes)
Considérons le modèle à bruit additif : $Z = X + \epsilon$, où X et ϵ sont indépendantes. Nous construisons un nouvel estimateur de la densité de X à partir d'observations de Z, fondé sur une méthode de projection en base d'Hermite sur R. Nous étudions le risque quadratique intégré de notre estimateur. Nous prouvons qu'il est consistant et atteint les vitesses classiques dans ce contexte. Nous proposons également une

sélection de modèle pour choisir la bonne dimension dans l'espace de projection. L'estimateur résultant réalise automatiquement un compromis biais-variance.

Estimation non-paramétrique d'une régression sphérique dans le cadre de l'analyse de données fonctionnelles

Papa alioune meissa MBAYE (Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne)
Chafik SAMIR (Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, Université Clermont Auvergne)
Anne-Françoise YAO (Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne)
L'analyse de données à valeurs dans une sous variété riemannienne M, de dimension finie k suscite beaucoup d'intérêt dans plusieurs domaines de la science et plus particulièrement pour analyser des données médicales. Ces données peuvent être des courbes, des contours ou des volumes. Dans ce travail, nous nous intéressons à la régression à prédicteurs fonctionnels. Nous mettrons l'accent sur le cas où les prédicteurs appartiennent à un espace non linéaire, plus précisément la sphère S^k . Nous commencerons par analyser ces observations fonctionnelles en résolvant le problème de recalage. Ensuite, nous présenterons l'intérêt de notre approche à travers différentes représentations fonctionnelles. Enfin, nous illustrerons notre méthode par des applications à des données simulées ainsi qu'à des données réelles.

Un test de détection de rupture dans un modèle de régression

Zaher MOHDEB (Ecole Nationale Polytechnique de Constantine et Laboratoire de Mathématiques et Sciences de la Décision, Université frères Mentouri)
On considère un modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixée, notre but est de construire le test de l'hypothèse d'un modèle de régression linéaire contre les alternatives de ruptures de modèle et ce sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative. On établit la normalité asymptotique de la statistique de test sous l'hypothèse nulle ainsi que sous l'hypothèse alternative de ruptures de modèle.

RKHSMetaMod : An R package to estimate the Hoeffding decomposition of an unknown function by solving RKHS Ridge Group Sparse optimization problem

Halaleh KAMARI (LaMME, Université d'Evry Val d'Essonne. MaIAGE, INRA, Jouy-en-Josas)
RKHSMetaMod est un package R qui estime un méta-modèle d'une fonction inconnue, m, dans le cadre d'un modèle de régression Gaussien. La procédure repose sur la minimisation d'un critère des moindres carrés pénalisé par une double pénalité dite Ridge Group Sparse , pour des fonctions appartenant à un espace de Hilbert a Noyau Reproduisant (RKHS). Le méta-modèle estimé est un modèle additif dont les termes estiment les termes de la décomposition de Hoeffding de la fonction m. Ce package fournit une interface conviviale entre l'environnement informatique statistique R et les bibliothèques C++ Eigen et GSL. Le temps d'exécution est optimisé via l'utilisation des packages RcppEigen et RcppGSL.

15h40-16h : Pause café

16h-17h40

Modèles à variables latentes (Amphi 11)

Modération : Vincent BRAULT

Mixture of Hidden Markov Models for Pattern-Recognition of Accelerometer data

Jonathan Y. BERNARD (Inserm)
Marie DU ROY DE CHAUMARAY (CREST-ENSAI)
Matthieu MARBAC (CREST-ENSAI)
Fabien NAVARRO (CREST-ENSAI)
L'analyse de données d'accélérométrie consiste à extraire des informations sur les temps passés à différents niveaux d'activité. Ces informations sont généralement utilisées ensuite dans un modèle prédictif. Nous proposons une modélisation de ce type de données utilisant un mélange de chaînes de Markov cachées, afin de pouvoir automatiquement détecter le nombre de niveaux d'activités ainsi que leurs caractéristiques. Pour tenir compte de la spécificité des données d'accéléromètre, les données sont modélisées par une distribution de Zero-inflated Gamma dont les paramètres sont spécifiques à l'état caché. La modélisation par un mélange permet de tenir compte de l'hétérogénéité de la population. Les propriétés de cette modélisation (identifiabilité, gestion de valeurs manquantes, probabilité de détecter la vraie partition) sont discutées.

Classification de campagnes de publicité mobile : Modèle de mélange pour données longitudinales et non gaussiennes

Faustine BOUSQUET (IMAG)
Sophie LEBRE (IMAG)
Christian LAVERGNE (IMAG)
De nombreux enjeux statistiques ont émergé du contexte de la publicité, il s'agit notamment d'afficher une publicité au bon endroit et au bon moment. Dans ce but, plusieurs métriques de performance de diffusion d'une campagne sont mesurées au cours du temps, comme le nombre ou le taux de clics sur une publicité. Un enjeu essentiel est de prédire le taux de clics. Mais, il est important de comprendre au préalable la structure des données volumineuses et hétérogènes dont nous disposons. Pour cela, nous proposons ici une méthode de classification des profils de campagnes publicitaires, i.e. de données longitudinales non gaussiennes, par un mélange de modèles linéaires généralisés (GLM).

Modèles de classification non supervisée avec données manquantes non au hasard

Fabien LAPORTE (CMAP Polytechnique, CNRS)
Christophe BIERNACKI (Inria Lille, Université Lille, CNRS)
Gilles CELEUX (Inria Saclay)
Julie JOSSE (CMAP Polytechnique, Inria XPOP)
La difficulté de prise en compte des données manquantes est souvent contournée en supposant que leur occurrence est due au hasard. Dans cette communication, nous envisageons que l'absence de certaines données n'est pas due au hasard dans le contexte de la classification non supervisée et nous proposons des modèles logistiques pour traduire le fait que cette occurrence peut être associée à la classification cherchée. Nous privilégions différents modèles que nous estimons par le maximum de vraisemblance et nous analysons leurs caractéristiques au travers de leur application sur des données hospitalières.

Données manquantes dans un modèle à blocs latents pour la recommandation

Gabriel FRISCH (Heudiasyc)
Jean-Benoist LEGER (Heudiasyc)
Yves GRANDVALET (Heudiasyc)
Nous présentons un modèle statistique basé sur le LBM pour réaliser une recommandation sociale. Le modèle utilise des variables latentes pour modéliser un processus de manquement de données de type NMAR.

Inférence Variationnelle du Modèle à Blocs Stochastiques (SBM) avec covariables en présence de données manquantes

Timothée TABOUY (AgroParisTech)
Julien CHIQUET (INRA)
Pierre BARBILLON (AgroParisTech)
Le modèle à blocs stochastiques ou Stochastic Block Model (SBM) (Nowicki and Snijders, 2001) est un modèle de graphe aléatoire généralisant le modèle d'Erdős-Reyni (Erdős and Renyi, 1959) à l'aide d'une structure latente sur les nœuds. L'utilisation de variables latentes dans le SBM permet de modéliser un large spectre de topologies de réseau, en particulier les graphes en affiliation, en étoile ou bipartite. L'inférence de ces modèles repose sur des modifications de l'algorithme EM (Expectation Maximization), comme par exemple l'approche EM variationnelle (Daudin et al., 2008) ou Bayésienne variationnelle (Latouche et al., 2012). Dans ces approches, le réseau est toujours considéré comme parfaitement observé, alors que de nombreux cas d'application (en particulier en sociologie) suggèrent que son observation est partielle et guidée par une stratégie d'échantillonnage dépendant du réseau lui-même, par exemple centrée sur les nœuds. Dans un prédécent travail (Tabouy et al., 2019) nous avions constaté qu'un échantillonnage partiel du réseau peut induire un biais d'estimation dans le modèle SBM. Notre objectif était alors la modélisation de la stratégie d'échantillonnage utilisée et son intégration dans les procédure d'inférence. S'appuyant sur la théorie des données manquantes développée par Rubin (1976), nous avons adapté les définitions de Missing At Random (MAR) et Not MAR (NMAR) aux cas de modèles à variables latentes. Nous proposons dans cette présentation de montrer que la prise en compte de covariables dans la modélisation peut induire un changement de nature (MAR ou NMAR) des données manquantes. Ce constat a pour but de faire le lien entre les différentes natures des données manquantes suivant la modélisation et de la prendre en compte dans l'inférence. En effet nous montrerons que différents modèles sans et avec covariables, couplés avec une loi d'échantillonnage donnant lieu à des donnés manquantes de natures différentes aboutissent au même modèle.

MALIA (Amphi 12)

Modération : Christine KERIBIN

Approche bayesienne et sampling pour les reseaux de neurones

Stéphane CHRETIEN (National Physical Laboratory)
Les approches Bayesiennes commencent a emerger pour l'etude des reseaux de neurones, tant du point de vue theorique que du point de vue algorithmique. Leur principal benefice est de permettre la quantification de l'incertitude associee aux decisions issues de l'utilisation de tels reseaux. Nous ferons un tour d'horizon de ces approches et montrerons comment l'approche Langevin sampler peut permettre de simuler efficacement selon la loi posterieure, afin de fournir les outils d'une utilisation statistiquement valide des reseaux de neurones.

Apprentissage PAC-bayésien et réseaux de neurones

Pascal GERMAIN (Inria Lille - Nord Europe)

When PAC-Bayesian Majority Votes meets Domain Adaptation

Emilie MORVANT (Laboratoire Hubert Curien, Université de Saint-Etienne)
De nos jours, de nombreuses données sont disponibles et de nombreuses applications requièrent l'utilisation de méthodes d'apprentissage automatique supervisées, capables de tirer avantage de différentes sources d'information. Une solution naturelle consiste à 'combiner' ces sources. Nous nous concentrons ici sur une combinaison particulière : le vote à la majorité pondérée 'PAC-Bayésien'. C'est une méthode ensembliste, fondée théoriquement, où plusieurs modèles se voient attribuer un poids spécifique. Cette approche est motivée par l'idée qu'une combinaison peut potentiellement compenser les erreurs des modèles individuels et ainsi obtenir une meilleure robustesse et performance sur de nouvelles donnée tirées selon une certaine distribution de probabilité. En apprentissage statistique, la capacité d'un modèle à généraliser sur une telle distribution de données est mesurée par des bornes de généralisation. Dans cet exposé, après avoir rappelé le cadre théorique classique des bornes en généralisation PAC-Bayesiennes, nous l'étendons à la tâche d'adaptation de domaine où l'objectif est d'adapter un modèle à partir d'une distribution de données source à une distribution cible différente, mais liée.

Large scale recommendation : a view from the trenches

Anne-Marie TOUSCH (Criteo AI Lab)
At Criteo, we are recommending billions of products to billions of users each day in the context of online advertising. There are many challenges in building such a large-scale recommender system. I'll describe some of the algorithms that have been successfully brought to production, and a few promising new lines of research, highlighting the fundamental role of mathematics.

SFB (Amphi 13)

Modération : Mounia HOCINE

Estimation de liens épidémiologiques par apprentissage statistique sur données génomiques

Samuel SOUBEYRAND (INRA - BioSP)
Maryam ALAMIL (INRA - BioSP)
Qui a infecté qui au cours d'une épidémie causée par une maladie infectieuse ? Ou plus généralement, quels individus hôtes sont liés épidémiologiquement ? Sous l'angle de la statistique, répondre à ces questions revient à estimer les liens dans un réseau dont les noeuds sont les individus hôtes, que ceux-ci soient des humains, des animaux, des plantes, des foyers, des troupeaux ou encore des champs. Diverses approches permettent de répondre à ces questions, dont des approches exploitant des données génomiques caractérisant, au niveau individuel, le pathogène causant la maladie (des individus portant des variants identiques ou proches du pathogène étant vraisemblablement liés épidémiologiquement). Nous avons récemment développé une telle approche, permettant d'estimer les liens épidémiologiques au sein d'une population d'hôtes, fondée sur un modèle semi-paramétrique dans l'espace des génomes du pathogène et sur une technique d'apprentissage exploitant des données de contact pouvant être collectées par exemple dans le cadre d'une enquête épidémiologique. Au cours de la présentation, nous détaillerons la construction du modèle semi-paramétrique et l'implémentation de la méthode d'estimation, puis nous

illustrerons l'application de notre approche à des données simulées et des données réelles portant sur Ebola, la grippe porcine et un potyvirus du salsifis sauvage.

Réseaux bayésiens et analyse de survie pour la modélisation de maladies génétiques dépendant de l'âge

Grégory NUEL (LPSM (CNRS 8001), Sorbonne Université)
Il existe de nombreuses maladies (cancers, obésité, diabète, Alzheimer, maladies rares, etc.) dans lesquels des facteurs génétiques jouent un rôle majeur (mutations BRCA1/2 pour le cancer du sein, mutation du système MMR pour cancer colorectal, APOE4 pour Alzheimer, etc.). Dès lors que cette maladie apparaît de manière répétée dans une famille, ou sous une forme sévère (à un âge précoce par exemple), il est naturel de chercher à savoir si cette histoire familiale (FH=Family History en anglais) est ou non évocatrice de la présence d'un facteur génétique au sein de la famille. Dans l'affirmative, la prise en charge des patients est souvent modifiée, ainsi que le suivi de leurs apparentés risquant eux aussi d'être exposés au(x) facteur(s) génétique(s). Nous nous proposons dans cet exposé de présenter les modèles probabilistes utilisés dans ce contexte. Il s'agit en fait de combiner un réseau bayésien pour la modélisation de la partie génétique du modèle, et une modélisation de durée de survie (typiquement : l'âge au diagnostic de la maladie) pour la partie maladie. Pour effectuer de l'inférence dans ce modèle, on a recours à l'algorithme somme-produit (sum-product en anglais, mais cet algorithme porte également de nombreux noms : Elston-Stewart, forward/backward, message-passing, belief propagation, filtre de Kalman, etc.) et à ses variantes. Le principe de l'algorithme sera présenté de manière didactique et plusieurs extensions seront présentées (calculs de dérivées, de moment generating functions, et de probability generating functions). L'intérêt de la méthode sera illustré sur plusieurs exemples concrets en collaboration avec des cliniciens : cancer du sein et de l'ovaire avec le modèle de Claus (institut Curie), syndrome de Lynch (La Pitié Salpêtrière et l'hôpital Saint-Antoine), et la neuropathie amyloïde héréditaire (hôpital Henri Mondor).

Biostatistiques 1 (Amphi 14)

Modération : Stéphane ROBIN

Influence du choix de l'arbre dans les études d'abondance différentielle

Antoine BICHAT (LaMME, UEVE, Enterome)
Mahendra MARIADASSOU (MaiAGE, INRA)
Jonathan PLASSAIS (Enterome)
Christophe AMBROISE (LaMME, UEVE, CNRS)
En métagénomique, il est courant de réaliser des études dites d'abondance différentielle pour identifier les bactéries dont l'abondance est associée à une variable donnée, comme par exemple l'état d'un patient. De telle étude s'intéressent à des centaines, voire quelques milliers de bactéries différentes, et font autant de tests. Il est donc nécessaire de corriger pour la multiplicité des tests. Après correction par des méthodes classiques (Bonferroni, Benjamini-Hochberg), quasiment aucun taxon n'est détecté, du fait de la faible puissance statistique des études (faible nombre d'échantillons, faible taille d'effet). Pour contourner ce problème, certains ont proposé d'exploiter la structure hiérarchique fournie par la taxonomie, qui constitue un a priori biologique sur la structure des données, pour augmenter la puissance des tests. Bien que la structure taxonomique semble naturelle, et justifiée biologiquement, elle présente néanmoins des désavantages et il peut être justifié de chercher d'autres structures hiérarchiques. Le présent travail met en évidence certains de ces désavantages et étudie l'impact du remplacement de l'arbre taxonomique par un arbre créé à partir des corrélations entre taxons.

Imputation multiple et prise en compte de l'incertitude pour les données de protéomique quantitative

Marie CHION (IRMA et LSMBO/DSA/IPHC, Université de Strasbourg)
Frédéric BERTRAND (IRMA, Université de Strasbourg)
Christine CARAPITO (LSMBO/DSA/IPHC, CNRS-Université de Strasbourg)
L'analyse protéomique consiste à étudier l'ensemble des protéines contenues dans un système biologique donné, à un instant donné et dans des conditions données. Les techniques les plus performantes pour déterminer l'abondance des protéines passent par la mesure des intensités peptidiques. Mais ces données peptidiques comportent des valeurs manquantes. Bien que les techniques statistiques usuelles en protéomique permettent l'imputation de celles-ci, l'incertitude liée à l'imputation n'est pas prise en compte. Nous proposons ici d'intégrer les techniques d'estimation tempérée de la variance aux méthodes d'imputation multiple, en les rendant utilisables tant au niveau peptidique qu'au niveau protéique.

Comparaison de trajectoires qualitatives avec des chaînes semi-Markoviennes : une application en analyse sensorielle

Cindy FRASCOLLA (Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Université de Bourgogne, 21000 Dijon, France)
Guillaume LECUELLE (Centre des Sciences du Goût et de l'Alimentation, UMR AgroSup Dijon-CNRS-INRA-Université de Bourgogne, 21000 Dijon, France.)
Hervé CARDOT (Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Université de Bourgogne, 21000 Dijon, France)
Michel VISALLI (Centre des Sciences du Goût et de l'Alimentation, UMR AgroSup Dijon-CNRS-INRA-Université de Bourgogne, 21000 Dijon, France.)
Pascal SCHLICH (Centre des Sciences du Goût et de l'Alimentation, UMR AgroSup Dijon-CNRS-INRA-Université de Bourgogne, 21000 Dijon, France.)
L'analyse sensorielle est très utilisée dans l'industrie agroalimentaire pour développer des nouveaux produits. La Dominance Temporelle des Sensations (DTS) est une méthode d'analyse sensorielle, utilisant une liste de descripteurs, qui permet d'indiquer comment les sensations ressenties lors de la dégustation d'un produit changent au cours du temps. En 2018, Lecuelle et al. ont introduit des chaînes semi-Markoviennes aux données DTS. Pour comparer deux produits lors d'études DTS, on introduit dans ce travail un test statistique basé sur le test du rapport de vraisemblance entre deux modèles semi-Markoviens. Pour construire la zone de rejet, trois approches sont évaluées. Une première basée sur le bootstrap paramétrique, une seconde basée sur les tests de permutation et une troisième qui repose sur la loi asymptotique du rapport de vraisemblance. Ces approches sont comparées à partir de simulations et de tests réalisés sur des jeux de données réels de dégustations de chocolats et de fromages.

Inférence de réseaux pour des données gaussiennes inflatées en zéro par double troncature

Clémence KARMANN (INRIA - Université de Lorraine)
Anne GEGOUT-PETIT (Université de Lorraine)
Aurélie GUEUDIN (Université de Lorraine)
On s'intéresse à inférer la structure de dépendances conditionnelles dans le cas de données gaussiennes inflatées en zéros par double troncature (à droite et à gauche). On dispose d'un p-vecteur gaussien X observé au travers du vecteur tronqué $Y := X1_{a \leq X \leq b}$. L'objectif est de retrouver la matrice de précision de X à partir d'observations de Y. Pour ce faire, on propose une procédure d'estimation qui consiste à estimer d'abord chacun des termes de la matrice de covariance par maximum de vraisemblance, puis la matrice de précision à l'aide du graphical Lasso. On donne un résultat théorique concernant la convergence de la matrice de précision estimée par cette méthode.

Qualité, fiabilité (Amphi 15)

Modération : Emmanuel REMY

Régression polynomiale par morceaux sous contrainte de régularité pour la propagation de fissures

Florine GRECIET (Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France et Safran Aircraft Engines)
Romain AZAIS (Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRA, Inria, F-69342, Lyon, France)
Anne GEGOUT-PETIT (Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France)
Dans ce papier, nous nous intéressons à des vitesses de propagation de fissures, phénomène physique continu et laissant observer plusieurs régimes. Dans ce but, nous proposons un modèle de régression polynomiale par morceaux à plusieurs régimes sous des hypothèses de continuité et/ou de dérivabilité ainsi qu'une méthode d'inférence permettant d'estimer les instants de transition et les lois de chaque régime. La plus efficace de nos méthodes d'inférence est basée sur un algorithme de programmation dynamique. Pour introduire la méthodologie, nous présentons d'abord le problème lorsque le nombre de régimes est 2 puis nous généralisons à un nombre quelconque de régimes.

Fiabilité de systèmes réparables en présence de covariables

Frédéric LOGE (Air Liquide R&D, Les Loges-en-Josas)
Moulay-Driss EL ALAOUI FARIS (Air Liquide R&D, Les Loges-en-Josas)
Valentin PATILEA (CREST (Ensai) & IRMAR (UEB))
Dans ce travail, nous modélisons la fiabilité de systèmes réparables, en prenant en compte les attributs des systèmes comme le modèle ou le type de technologie. Pour ce faire, nous nous appuyons essentiellement sur la littérature des modèles d'âge virtuels et des tests de tendance. Cette modélisation permet d'établir des plans de maintenance réalistes.

Optimal predictive maintenance policy for multi-component systems

Tiffany CHERCHI (Thales LAS, France)
Benoîte DE SAPORTA (IMAG, Univ Montpellier, CNRS, Montpellier, France)
François DUFOUR (INRIA CQFD, IMB, Univ Bordeaux, Bordeaux INP, CNRS, France.)
Camille BAYSSE (Thales LAS, France)
Nous présentons un problème d'optimisation pour la maintenance d'un système multi-composants sujet à des détériorations ou défaillances aléatoires de ses composants, entraînant l'évolution de l'état général du système. Le système peut être requis pour effectuer des missions de fréquences et durées déterministes. Notre objectif à long terme est de mettre en place une politique de maintenance optimisée afin de garantir le bon déroulement des missions tout en minimisant les coûts de maintenance. L'idée principale de ce travail est de proposer un modèle mathématique pour l'évolution du système en utilisant le formalisme d'un Processus Markovien Décisionnel (MDP). Par simulations de Monte Carlo, nous comparons les performances de plusieurs politiques de référence.

Détermination d'outils d'aide à la décision pour le traitement d'événements indésirables dans le cadre d'une compagnie aérienne

Aymen AMARA (UFC)

Hazard rate function estimation using generalized Birnbaum-Saunders kernel

Sylia CHEKKAL (Université de Bejaia, Algerie)
Karima LAGHA (Université de Bejaia, Algerie)
Nabil ZOUGAB (Université de Bejaia, Algerie)
Dans ce présent papier, nous nous intéressant à l'estimation non paramétrique du taux de de défaillance avec la méthode du noyau. Puisque le taux de de défaillance est défini sur un support positif, nous utilisons un noyau asymétrique afin d'éliminer le problème d'effet de bord qui engendre un biais de plus en plus élevé en se rapprochant du bord. A cet effet nous proposons d'utiliser le noyau GBS associé (Birnbaum-saunders généralisé). Nous déterminons les propriétés asymptotiques de l'estimateur proposé ainsi que le paramètre de lissage optimal. La performance de l'estimateur est étudiée par simulation des données suivants des lois de fiabilité telles que : lognormale, BS, Gamma..

Etude de cas industriels (Amphi 16)

Modération : Sébastien MARQUE

Prévision de la consommation d'électricité à l'échelle des ménages.

Fatima FAHS (IRMA)
Myriam MAUMY-BERTRAND (IRMA)
Céline CALDINI-QUEIROS (IRMA)
Frédéric BERTRAND (IRMA)
La maîtrise de la demande en energie (MDE) a provoquée depuis ces trente dernières années des travaux de type DSR (Demand Side Response : réduire la demande de consommation électrique) qui regroupent les techniques permettant de diminuer la consommation d'énergie électrique d'un ménage, d'un bâtiment, d'un territoire, d'un pays, et de réaliser ainsi des économies en ressources fossiles et de diminuer la pollution par le CO2. Les travaux de l'EPRI (Electric Power Research Institute) ont montré qu'informer en temps réel le client de sa consommation permet de la réduire, donc confronter le consommateur à un modèle de sa demande peut mener à bien cet enjeu. Notre objectif principal est de développer et d'exploiter des modèles statistiques et de machine learning permettant de donner une analyse prévisionnelle complète de la consommation électrique le jour J à l'échelle des ménages à partir des historiques de consommation et des données météorologiques afin d'alerter le consommateur en cas d'anomalie de consommation (sur-consommation, fuite d'électricité, chutes anormales de puissance...).

Prévision de production éolienne par forêts aléatoires, agrégation et alerte de rampes

Mamadou DIONE (CREST-ENSAE & ENGIE Green France)
La rupture des contrats d'obligation d'achat avec la loi de transition énergétique définie par l'État français implique la vente d'électricité éolienne sur le marché. Pour cette vente, il est nécessaire d'annoncer la quantité d'électricité à produire. Nous avons donc besoin de prévisions de la production. Nous proposons dans cet article un modèle de forêt aléatoire puis une agrégation des prévisions de plusieurs parcs éoliens pour réduire les incertitudes et des alertes de rampes pour prévenir les périodes de gosses erreurs.

Analyse des données Marketing : pré-tests publicitaire par la confrontation des slogans en utilisant l'analyse des facteurs principales

Mohamed AYAD (Laboratoire d'analyse Economique et de Modélisation)
L'information étant cruciale en marketing et son volume toujours croissant, la compétitivité des firmes

aujourd'hui dépend de la capacité de son système d'information à collecter, traiter, et analyser les données. Tout décideur a besoin de prendre la décision ou de ne pas la prendre à un moment donné et cela est largement lié à la pertinence des données collectées mais aussi à la façon par laquelle sont traitées, analysées et interprétées. La prise de décision marketing est généralement appuyée sur des données extrêmement nombreuses et diverses : études du marché, tests marketing, étude de motivation, analyse d'image et autres données pour ne pas être exhaustif. Notre étude de cas va mettre l'accent sur un prétest publicitaire des slogans sur des produits déjà émises sur le marché, il s'agit ici d'une réorientation d'une compagne publicitaire pour promouvoir certains produits. A travers le recours aux données qualitatives collectées lors d'une enquête faite : on a proposé un ensemble de spécifications (mots ou groupe de mots) à un ensemble de 40 personnes pour afin qu'ils puissent affecter ces spécifications à un ensemble de produits déjà commercialisé par la même société. Notre objectif est de faire sortir les perceptions réelles des consommateurs envers les 6 produits pour les intégrées à la nouvelle compagne publicitaire, pour la détermination des mots ou groupes de mots réels poussant le plus un consommateur d'acheter un ou plusieurs produits. Notre étude focalisera sur une analyse des correspondances factorielles (AFC) par la confrontation des évocations mieux interprétés et choisis par le responsable du département commercial de la société par moyen de créativité. Notre étude de cas est réalisée dans le cadre d'un stage effectué à l'une des agences commerciale Central Danone au nord du Maroc.

Anonymisation et confidentialité différentielle appliquées à des données spatio-temporelles : cas d'usage portant sur la billettique

Vincent THOUVENOT (Thales SIX GTS France)
Thibaut DUBOIS (Thales SIX GTS France)
Stephane LORIN (Thales SIX GTS France)
Smartphone, carte d'abonnement, compteur intelligent énergétique, etc., les sources de données personnelles sont nombreuses. Si ces données peuvent apporter des fortes valeurs ajoutées, que ce soit aux citoyens, aux collectivités ou aux entreprises, celles-ci doivent être protégées. La réglementation autour de la donnée personnelle évolue et se renforce (voire la RGPD). L'anonymisation des données peut être utilisée pour les protéger. Nous présentons ici deux méthodes d'anonymisation de données billettiques. La première méthode est fondée sur un algorithme de généralisation, alors que la seconde cherche à respecter la confidentialité différentielle. Mots-clés : Anonymisation, Billettique, Confidentialité différentielle, Données spatio-temporelles, Etude de cas, Géneralisation

Développement de modèles prédictifs dans le cadre de l'industrie manufacturière à gros volumes

Eva JABBAR (Institut de Mathématiques, Université de Toulouse, Continental Powertrain France SAS)

Philippe BESSE (IMT - Institut de Mathéematiques, Université de Toulouse, CNRS UMR5219, 118 Route de Narbonne, 31400 Toulouse)
Jean-Michel LOUBES (IMT - Institut de Mathéematiques, Université de Toulouse, CNRS UMR5219, 118 Route de Narbonne, 31400 Toulouse)
Merle CHRISTOPHE (Continental Powertrain France SAS, 1 avenue Paul Ourliac, 31036 Toulouse)
La qualité de la production des cartes électroniques atteint un excellent niveau de performance, avec un taux de rejet extrêmement faible. De ce fait, il devient de plus en plus complexe de déterminer les paramètres/facteurs sur lesquels il faut agir pour améliorer encore davantage la production. Nous avons donc eu recours aux solutions dites data driven à travers l'application des techniques de machine learning comme un moyen d'investigation pour améliorer la qualité et réduire les coûts. Une des approches classiques de détection de défaut est la mise en place de station d'inspection/contrôle en aval dans le processus de production supporté par une vérification humaine afin de valider les produits en vrais défauts et remettre en production ceux avec un faux défaut. Cette approche est caractérisée par un taux de faux défauts très élevé sur les produits écartés, autour de 95%, ce qui implique une pénibilité de tâche des opérateurs et une perte de temps importante. Une autre approche utilisée dans la production est la mise en place d'un programme d'inspection basé sur des analyses univariées à travers la définition de limites pour chaque paramètre d'un composant donné et pour chaque référence de produit. Imposer le codage d'alarme complexe limité à chaque paramètre peut engendrer l'envoi d'un grand nombre de fausses

alarmes et en même temps ne pas verrouiller les situations à risque combiné. C'est dans cette perspective que nous nous intéressons au développement de modèle prédictifs en utilisant les données historiques disponibles dans le cadre de l'industrie manufacturière à gros volumes. La chaîne de production est en effet un milieu très contrôlé, suivi par un grand nombre de capteurs divers mélangeant plusieurs types d'informations (capteurs fonctionnels, qualitatifs, booléens, etc.), générant une volumétrie de données très importante. L'étude de la première problématique à savoir la présence d'un taux élevé de faux défauts, requiert le développement d'un modèle de classification qui soit un outil d'aide à la décision pour les opérateurs leur permettant d'identifier rapidement les pièces présentant un faux défaut et ainsi se concentrer sur le repérage des pièces réellement défectueuses. Le système de production étant régi par des paramètres physiques, nous avons choisi de comparer les performances des techniques de machine learning à base d'arbres : CART, Random Forest, Adaboost, XGBoost afin de pouvoir par la suite donner une interprétabilité physique du modèle choisi. Le résultat de cette étude positionne XGBoost comme meilleur algorithme avec un score de précision de 99.4% et un rappel de 98.6%. Cette étude est présentée par Jabbar et al (2018). Une étape d'interprétabilité est ensuite étudiée à travers l'analyse des valeurs 'SHapley Additive exPlanations' présenté par Lundberg, S.M. & Lee, S.-I. (2017). Pour la problématique concernant l'amélioration des détections d'anomalies actuellement basée sur des définitions univariées des alarmes complexes et dépendante d'une référence de produit et des composants associés, nous proposons l'application de techniques de détection d'anomalies conditionnelles. En effet, la définition d'une anomalie dans ce cadre est très complexe car elle dépend de nombreuses conditions et modes de production. La méthodologie adaptée est d'identifier d'abord les sous-populations homogènes caractérisant ainsi une même condition. Puis appliquer sur chaque sous-population des méthodes de détection d'anomalie. Nous avons analysé les performances des AutoEncoders entièrement connectés (Aggarwal ; (2015)) qui utilisent l'erreur de reconstruction comme score d'anomalie, et également l'application de méthode de manifold learning pour la réduction de dimension puis One-Class Support Vector Machines (Ma et Perkins ;2003). Les résultats obtenus ont permis de détecter des anomalies dont certaines ont réellement été identifiées comme des défauts plus tard dans le processus.

18h30-19h30 : Rencontre entre jeunes statisticiens et conférenciers invités

Vendredi 7 juin

8h40-9h40

John Bacon-Shone : Compositional data analysis : choosing transformations that yield meaningful models (Amphi 11)

University of Hong Kong

The challenge with compositional data is constraints on the data, which must lie between 0 and 1, plus a linear sum constraint. These constraints ensure that standard statistical analysis assuming a Euclidean metric is inappropriate. The crucial insight of John Aitchison's seminal paper was a transformation mapping compositions onto unconstrained space. However, this transformation is not unique and only maps the interior of the simplex (excluding zeros). We examine both empirical and model-based arguments for selecting transformations in order to generate meaningful models. Specific examples include integer data, data with additional constraints, ordered components and hierarchical models.

Modération : Christine THOMAS-AGNAN

Martial Foucault : Le Grand débat national : géographie politique des réunions locales (Amphi 14)

Sciences Po CEVIPOF

A partir des données d'organisation des réunions d'initiatives locales, il est proposé de situer géographiquement et politiquement le contexte dans lequel certaines communes ont choisi d'organiser ou non une ou plusieurs réunions locales. Le travail met en évidence certaines lignes de fracture territoriale qui ne recouvrent pas celles des mobilisations des Gilets jaunes".

Modération : Aurélien Garivier

9h40-11h

Tremblements de terre (Amphi 11)

Modération : Christophe LEY

Earthquakes economic costs through rank-size laws

Valerio FICCADENTI (University of Macerata)
Roy CERQUETI (University of Macerata)
The presentation is devoted to explore the magnitude features of the earthquakes occurring in Italy between January 24th, 2016 and January 24th, 2017 in order to elaborate a proposal of cost indicator. The well known tragic seismic events with epicentres at Accumuli, Visso, Ussita, Castelsantangelo sul Nera, Norcia and Montereale have occurred during the aforementioned span of time. On this dataset, we develop two different rank-size analysis by using the standard Zipf-Mandelbrot Law (see Mandelbrot 1953, 1961) and the Universal Law proposed by Ausloos and Cerqueti (2016). The idea of designing a measure to evaluate the economic impact of earthquakes is based on the obvious evidence of a cause-effect relationship between the magnitude of earthquakes and the economic cost deriving from them. We draw attention to the role of the infrastructures resistance into the relationship between the damages and the seismic events sequences, so we conjecture different forms of cost indicators.

Some results and some challenges in statistical seismology

Isabel SERRA
Patricia PAREDES
Anna ESPINAL
Paul ROCHET
Alvaro CORRAL
The Gutenberg-Richter (GR) law is of fundamental importance in statistical seismology. It simply states that, for a given region, the magnitudes of earthquakes follow an exponential probability distribution. As the (scalar) seismic moment is an exponential function of magnitude, when the GR law is expressed in terms of the former variable, it translates into a power-law distribution. The distribution of the seismic moment is of capital interest to evaluate earthquake hazard, in particular regarding the most extreme events ; therefore we are going to analyze the tail of the seismic moment law. It is well known that there exist a range of the seismic moment such that the scale-free assumption holds, but several questions are involved. Is the scale-free range bounded ? Can we estimate it ? Moreover, which is the support of the GR law ? We are going to show some recent advances for answering each of these questions.

Statistical seismology : empirical laws and physical constraints

Yavor KAMER (D-MTEC, Scheuchzerstrasse 7, 8092 Zürich, Switzerland)
Shyam NANDAN (SED, Sonneggstrasse 5, 8006 Zürich, Switzerland)
Guy OUILLON (Lithophyse, 4 rue de l'Ancien Sénat, 06300 Nice, France ;)
Didier SORNETTE (D-MTEC, Scheuchzerstrasse 7, 8092 Zürich, Switzerland)
Nous présentons une revue des concepts et définitions propres à la sismologie statistique, les données qu'elle traite, ainsi que les principales lois statistiques décrivant le phénomène sismique. L'état de l'art dans le domaine de la simulation est également présenté, avec quelques raffinements déduits de modèles physiques ou de raisonnements sur la symétrie. Nous terminons par une brève liste des paramètres dont la détermination précise pourrait découler d'une collaboration plus intense entre statisticiens et sismologues.

Towards a general earthquake prediction platform : RICHTERX

Shyam NANDAN (ETH Zürich, SED, Sonneggstrasse 5, 8006 Zürich, Switzerland)
Yavor KAMER (ETH Zürich, D-MTEC, Scheuchzerstrasse 7, 8092 Zürich, Switzerland)
Stefan HIEMER (ETH Zürich, SED, Sonneggstrasse 5, 8006 Zürich, Switzerland)
Guy OUILLON (, 4 rue de l'Ancien Sénat, 06300 Nice, France)
Didier SORNETTE (ETH Zürich, D-MTEC, Scheuchzerstrasse 7, 8092 Zürich, Switzerland)
Avec l'ambition de repousser les limites des modèles de prévision sismique, nous (1) présentons une démarche objective d'inversion de la variation spatiale des paramètres du modèle Epidemic Type Aftershock Sequence (ETAS), qui ont jusqu'à présent été arbitrairement considérés comme homogènes ; (2) comparons la performance des prévisions du modèle ETAS avec paramètres spatialement variables à d'autres modèles de l'état de l'art ; (3) remettons en cause le protocole de test du Collaboratory for Study of Earthquake Predictability (CSEP), et introduisons une nouvelle plateforme de prédiction, RichterX.

Fouille de données : méthodologie (Amphi 12)

Modération : Vincent VANDEWALLE

L'analyse des liaisons : une nouvelle méthode de fouille de données multidimensionnelles

Jean-Luc DURAND (Laboratoire d'Éthologie Expérimentale et Comparée, Université Paris 13)
L'une des propriétés de l'Analyse des Correspondances (AC) est que la variance des nuages de points (le carré moyen de contingence ϕ^2) fournit une mesure de la liaison entre deux variables qualitatives. Ainsi, dans une AC, une proportion de variance peut être interprétée comme une proportion de la force de cette liaison. Mais une telle interprétation de la variance n'est pas possible en Analyse en Composantes Principales (ACP) ou en Analyse des Correspondances Multiples (ACM). L'objectif de cette présentation est de proposer une méthode d'analyse de données multivariées qui permette d'interpréter la variance du nuage des individus comme une mesure globale des liaisons entre variables. On définit l'indice de liaisons d'une variable comme une mesure, comprise entre 0 et 1, de la force des liaisons entre cette variable et les autres variables, basée sur les carrés des corrélations (r^2), les carrés des rapports de corrélation (η^2) ou les carrés moyens de contingence (ϕ^2) entre cette variable et les autres. Ensuite on utilise les indices de liaisons et, pour chaque variable qualitative, les résultats de l'AC d'une partie du tableau de Burt, pour construire un tableau numérique Individus × Variables (appelé tableau de liaisons) donnant les coordonnées brutes des points d'un nuage euclidien représentant les individus. La variance de ce nuage est une somme pondérée des indices de liaison des variables. Enfin, les coordonnées principales des individus sont obtenues en réalisant l'ACP non normée du tableau de liaisons. Comparés à ceux de l'ACP ou de l'ACM, les résultats de l'analyse des liaisons montrent : 1) une robustesse vis-à-vis de la présence de variables non pertinentes : l'ajout de variables aléatoires ne change pratiquement pas les résultats ; 2) pour des données quantitatives, une interprétation similaire des premiers axes ; 3) pour des données qualitatives, une interprétation similaire du premier axe, une meilleure prise en compte de l'information contenue dans les tableaux binaires croisant deux variables différentes, et des proportions de variance nettement plus importantes pour les premiers axes.

Classification ascendante hiérarchique, contrainte d'ordre : conditions d'applicabilité, interprétabilité des dendrogrammes

Nathanaël RANDRIAMIHAMISON (INRA, UR 0875 MIAT)
Pierre NEUVIAL (Institut de Mathématiques de Toulouse, Univ. Paul Sabatier, UMR 5219)
Nathalie VIALANEIX (INRA, UR 0875 MIAT)

La classification ascendante hiérarchique (CAH) avec lien de Ward, ainsi que sa version sous contrainte d'ordre, sont couramment utilisées sur des données de différents types : distances, dissimilarités, noyaux, similarités. Nous précisons dans quel cas l'utilisation de ces méthodes est justifiée théoriquement. Nous étudions les conditions garantissant la cohérence entre les résultats de la CAH et leur représentation graphique classique sous forme de dendrogramme. Cette étude révèle une distinction importante entre cette propriété de cohérence et l'absence de croisement dans le dendrogramme.

Tri-clustering pour données de comptage dynamiques

Margot SELOSSE (Université de Lyon, Lyon 2, ERIC EA 3083)
Antoine GOURRU (Université de Lyon, Lyon 2, ERIC EA 3083)
Julien JACQUES (Université de Lyon, Lyon 2, ERIC EA 3083)
Julien VELCIN (Université de Lyon, Lyon 2, ERIC EA 3083)
Les données de comptage sont très utilisées dans le monde actuel pour modéliser les occurrences d'un évènement (apparence d'un mot dans un texte, passage d'une voiture à un carrefour, contact entre utilisateurs d'un réseau social, etc.). Ce travail s'intéresse aux données de comptage dynamiques, lorsque les occurrences sont dénombrées sur plusieurs périodes de temps différentes. Dans ce cas, les données peuvent être stockées dans un cube de données ou un tenseur. L'approche proposée développe un algorithme de tri-clustering, qui va simultanément créer des clusters en ligne et en colonne mais également des clusters temporels. La distribution de Poisson est utilisée pour modéliser les données, et un algorithme EM variationnel est décrit pour inférer les paramètres du modèle.

Consensual Aggregation of Clusters Based on Bregman Divergences to Improve Predictive Models

Sothea HAS (LPSM, Université Paris-Diderot)
Dans cet exposé, nous introduisons une nouvelle approche pour construire des modèles prédictifs dans les problèmes d'apprentissage supervisé en prêtant attention à la structure de regroupement des données d'entrée. Nous nous intéressons aux situations où les données d'entrée sont composées de plusieurs grappes et qu'il existe différents modèles sous-jacents sur ces grappes. Ainsi, au lieu de construire un seul modèle prédictif sur l'ensemble des données, nous proposons dans un premier temps d'utiliser l'algorithme K-means avec différentes options de divergences de Bregman qui sont les membres d'une large classe de mesures de dissimilarité, pour approcher la structure des données d'entrée. Pour chaque divergence, nous construisons un prédicteur local sur chaque cluster observé, ce qui conduira à un modèle global qui est la collection de ces prédicteurs locaux. Dans un deuxième temps, nous proposons de combiner intelligemment tous ces modèles globaux de manière à préserver la qualité de la combinaison, voire à l'améliorer, par rapport au meilleur modèle de la combinaison. Les résultats numériques réalisés sur plusieurs types de données simulées et une donnée réelle de Air Compressor montrent qu'il est très intéressant de prendre en compte la structure de clustering des données d'entrée, ainsi que d'utiliser des méthodes d'estimation combinées pour améliorer les performances de modèles prédictifs.

Biostatistiques et grande dimension (Amphi 13)

Modération : Mélisande ALBERT

A methodology to select and rank covariates in high-dimensional data under dependence

Aurélie GUEUDIN (IECL, INRIA)
Anne GEGOUT-PETIT (IECL, INRIA)

Nous proposons une méthode de sélection et de tri de covariables associées à une variable d'intérêt, dans le cadre de données dépendantes et de grande dimension, mais avec peu d'observations. Une primière étape consiste à décorréler les covariances : après avoir effectué un clustering des covariables, nous décorrélons les covariables de chaque cluster via l'analyse en facteurs latents. La seconde étape sélectionne et trie les covariables en utilisant une agrégation de méthodes et tests statistiques. Après quelques simulations, nous appliquons notre méthode sur des données transcriptomiques (p = 6810 covariables) de n = 37 patients atteints d'un cancer du poumon, et ayant reçu un traitement. Notre méthode permet de sélectionner les covariables liées à la réussite ou non du traitement. Nous obtenons différents profils de patients suivant leur temps de survie. ls de patients suivant leur temps de survie.

ASICS : identifier et quantifier des métabolites à partir d'un spectre RMN ^{1}H

Gaëlle LEFORT (Inra)
Laurence LIAUBET (Inra)
Hélène QUESNEL (Inra)
Cécile CANLET (Inra)
Nathalie VIALANEIX (Inra)
Rémi SERVIEN (Inra)

La résonance magnétique nucléaire du proton (^{1}H-RMN) est une technologie haut-débit permettant d'obtenir des profils métaboliques, sous forme de spectres, à un coût relativement faible. C'est un outil prometteur pour détecter des biomarqueurs facilement mesurables. Cependant, les métabolites présents dans un mélange complexe ne sont pas identifiables et quantifiables directement, ce qui limite l'interprétabilité de ces approches. Pour faciliter l'utilisation de ces données, nous avons développé une méthode d'analyse automatique, encapsulée dans un nouveau package R/Bioconductor, ASICS, qui permet l'identification et la quantification globale et automatique des métabolites dans un spectre RMN. Le package permet d'enchaîner facilement toutes les étapes de l'analyse (pré-traitements, quantification, outils de diagnostic pour juger de la qualité des quantifications, analyses statistiques post-quantification). La méthode de quantification, préexistante (Tardivel et al., 2017), a été testée sur un jeu de données réel (ANR PORCINET). Le but était de juger des performances de la méthode comparativement à celles existantes et de l'améliorer grâce à un paramétrage plus fin.

Réseaux de neurones pour l'analyse de survie en grande dimension

Mathilde SAUTREUIL (Laboratoire MICS, CentraleSupélec)
Sarah LEMLER (Laboratoire MICS, CentraleSupélec)
Paul-Henry COURNEDE (Laboratoire MICS, CentraleSupélec)

L'analyse de survie est l'étude du temps écoulé jusqu'à la survenue d'un événement d'intérêt qui peut correspondre par exemple au décès ou à la rémission dans une étude médicale. Dans ce contexte, l'objectif de notre travail est de prédire la durée de survie de patients à partir de données génomiques et cliniques. L'approche par réseaux de neurones pour l'analyse de survie n'est pas récente, mais seulement des données d'entrées de faible dimension ont été considérées par le passé. Cependant, depuis l'arrivée du séquençage à haut-débit le nombre de covariables potentiellement intéressantes pour les modèles de prévision est devenu très important. Le cadre statistique a donc changé. Nous présenterons et testerons donc quelques approches récentes de l'analyse de survie par de réseaux de neurones, adaptées à l'analyse de survie en grande dimension.

Décorrélation adaptative pour la prédiction en grande dimension

Florian HEBERT (Agrocampus Ouest, IRMAR)
Mathieu EMILY (Agrocampus Ouest, IRMAR)
David CAUSEUR (Agrocampus Ouest, IRMAR)

Dans les procédures de tests en grande dimension, la prise en compte ou non de la dépendance donne lieu à de nombreux développements méthodologiques et discussions, notamment sur l'impact de la décorrélation des statistiques de tests. Pourtant, dans une optique d'estimation d'un modèle pour la prédiction, la question de la décorrélation de grands profils de variables prédictrices n'est pas abordée dans les mêmes termes, bien que de nombreuses études comparatives aient rapporté la supériorité de méthodes

de prédiction dites naïves, au sens où elles ignorent la dépendance. Sous l'hypothèse classique en analyse linéaire discriminante d'un mélange de lois gaussiennes, nous montrons que pour une structure de dépendance des prédicteurs donnée, les performances de classification ignorant ou non cette dépendance peuvent être très variables et opposées selon la forme du signal d'association entre les prédicteurs et la classe. Afin de minimiser le risque maximal d'erreur de classification, nous proposons donc une prise en compte adaptative de la dépendance et montrons sur des simulations que les performances de la règle de classification proposée sont généralement au moins aussi bonnes que la meilleure des règles parmi celles ignorant la dépendance ou au contraire basées sur une décorrélation des prédicteurs.

11h-11h20 : Pause café

11h20-12h40

Etude de cas scientifiques (Amphi 11)

Modération : Frédéric BERTRAND

Modeling Heaping in Marine Surveys : A Bayesian Approach.

Eric PARENT (UMR MIA 518 INRA/AgroParisTech)
Matthieu AUTHIER (Université de la Rochelle)
Sophie DONNET (UMR MIA 518 INRA/AgroParisTech)
Isabelle ALBERT (INRA)
Daouda BA (UMR MIA 518 INRA/AgroParisTech)
Marine ecosystems have been considerably disrupted, maybe due to climate change, but also certainly because of the unstoppable advancement of fishing efficiency. Studying variations in species abundance through scientific surveys helps to better estimate and anticipate the stock variations on which appropriate management policies are defined. The MEGASCOPE campaigns for instance are regularly conducted on board a scientific vessel in the North-East Atlantic. Using binoculars, observers try to identify the different species of marine mammals. But observation is uneasy and perturbed in many manners : the unpredictable trajectory of mammals, the blurring effect of depth, the distance to the boat, the angle of observation, etc. . In addition, observers are used to rounding the number of observed animals. These counting difficulties make stock inference rather tricky and, consequently, abundance assessment unreliable. In this paper, we develop a Bayesian approach to distangle the uncertainty due to the counting process and the uncertainty stemming from environmental stochasticity.

Reconstructing the evolutionary history of the desert locust by means of ABC random forest

Louis RAYNAL (IMAG, Univ Montpellier, CNRS, Montpellier, France)
Marie-Pierre CHAPUIS (CIRAD, CBGP, Montpellier, France)
Jean-Michel MARIN (IMAG, Univ Montpellier, CNRS, Montpellier, France)

Arnaud ESTOUP (CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France)
Le criquet pèlerin Schistocerca gregaria est une espèce d'insecte répartie dans deux régions d'Afrique, une au nord et une au sud. Dans cette présentation nous nous intéressons aux processus évolutifs qui ont façonné la distribution géographique et les différences génétiques de ces deux sous-espèces. Pour ce faire, nous utilisons des données génétiques obtenues par marqueurs microsatellites sur les populations actuelles. Nous utilisons les méthodes récentes de calcul bayésien approché (Approximate Bayesian Computation (ABC)) par forêt aléatoire (ABC-RF, Pudlo et al., 2016 ; Estoup et al., 2018 ; Raynal et al., 2019) pour comparer des scénarios évolutifs et estimer des paramètres tels que le temps de divergence entre ces sous-espèces sud et nord. D'un point de vue méthodologique, nous discutons de mesures d'erreurs pour l'inférence de paramètres et montrons comment les calculer par forêts aléatoires.

Classification et modélisation de particules de cendres volcaniques

Sophie MIALLARET (LMBP, UCA)
Julia EYCHENNE (LMV, UCA)
Jean-Luc LEPENNEC (LMV, UCA)
Anne-Françoise YAO (LMBP, UCA)
La typologie, la distribution de taille et la forme des particules donnent de nombreuses informations sur les conditions de fragmentation, de transport et de sédimentation des cendres, et renseignent sur le style et la dynamique des éruptions accompagnées d'émissions de cendres. Les travaux de Julia Eychenne, Jean-Luc Lepennec et Sébastien Leibrant ont permis de mesurer la taille et la forme de particules de cendres provenant du Tungurahua, un strato-volcan andésitique situé en Équateur. L'objectif de cette étude est de faire une analyse statistique sur la morphologie des cendres et d'en déduire des profils en fonction de la localisation géographique.

Analyse géométrique de données barométriques

Frédérik CASSOR (CEVIPOF Sciences Po)
Brigitte LE ROUX (MAP5, Université Paris Descartes)
Les données d'une enquête barométrique comportent plusieurs groupes d'individus représentatifs d'une même population. Chaque groupe est interrogé à des périodes différentes. Chaque groupe répond à la même batterie de questions. Nous comparons, en utilisant les méthodes d'analyse géométrique des données, deux vagues d'enquête du « Baromètre de la confiance politique » initié par le CEVIPOF, qui ont eu lieu en décembre 2017 et en décembre 2018. L'étude est menée à partir de 29 questions bipolaires qui relèvent de 4 composantes de confiance (politique, institutionnelle, économique et interpersonnelle). Nous avons construit un espace de la confiance en effectuant une analyse des correspondances après avoir dédoublé les questions et les avoir pondérées selon leur thème. Les individus interrogés en 2017 (vague 9) on été pris comme ensemble de référence, ceux interrogés en 2018 (vague 10) ont été mis en éléments supplémentaires. Pour étudier l'évolution des réponses aux questions, nous avons appliqué les formules de transition aux individus supplémentaires, ce qui permet de visualiser les écarts entre les modalités des questions des deux vagues. Par ailleurs, nous avons effectué une classification des individus interrogés en 2017 (méthode de Ward). Nous avons identifié 5 classes d'individus au regard de la confiance ('hyperconfiants', 'confiants modérés', 'défiants altruistes', 'défiants autoritaires', 'hyperdéfiants'). Nous avons ensuite affecté les individus interrogés en 2018 aux classes en prenant un critère d'affectation basé sur la distance de Mahalanobis associée à chaque classe afin de tenir compte de la forme des classes ce qui nous a permis d'étudier les transferts entre classes.

Fouille de données et analyse en composantes principales (Amphi 12)

Modération : Chloé FRIGUET

Mixture of Multinomial PCA

Nicolas JOUVIN (Université Paris 1 Panthéon-Sorbonne)
Pierre LATOUCHE (Université Paris 5 Paris Descartes)
Charles BOUVEYRON (Université Côte d'Azur - INRIA)
Nous proposons le modèle MMPCA (pour mixture of multinomial PCA) pour la classification de données de comptages. Basé sur le modèle Latent Dirichlet allocation, il permet d'allier la flexibilité des topic models, tout en identifiant une partition explicite des données. L'inférence des paramètres et le clustering sont réalisés via un algorithme C-VEM (classification variational expectation-maximisation) qui optimise une vraisemblance complétée, couplé avec une stratégie de type branch & bound sur la borne variationnelle. Un critère ICL est proposé pour la sélection de modèles, et la performance de la méthodologie proposée est évaluée sur données simulées.

Imputation multiple pour données mixtes par analyse factorielle

Vincent AUDIGIER (Laboratoire Cedric MSDMA, CNAM)
François HUSSON (Laboratoire de mathématiques appliquées, Agrocampus Ouest)
Julie JOSSE (Centre de Mathématiques Appliquées, Ecole Polytechnique, France / XPOP, INRIA)
Matthieu RESCHE-RIGON (Service de Biostatistique et Information Medicale, Hopital Saint-Louis, AP-HP / Universite Paris Diderot - Paris 7, Sorbonne Paris Cite, UMR-S 1153,/ INSERM, UMR 1153, Equipe ECSTRA, Hopital Saint-Louis)
La prise en compte de données toujours plus nombreuses complexifie sans cesse leur analyse. Cette complexité se traduit notamment par des variables de types différents, la présence de données manquantes, et un grand nombre de variables et/ou d'observations. L'application de méthodes statistiques dans ce contexte est généralement délicate. L'objet de cette présentation est de proposer une nouvelle méthode d'imputation multiple basée sur l'analyse factorielle des données mixtes (AFDM). L'AFDM est une méthode d'analyse factorielle adaptée pour des jeux de données comportant des variables quantitatives et qualitatives, dont le nombre peut excéder, ou non, le nombre d'observations. En vertu de ses propriétés, le développement d'une méthode d'imputation multiple basée sur l'AFDM permet l'inférence sur des variables quantitatives et qualitatives incomplètes, en grande et petite dimension. La méthode d'imputation multiple proposée utilise une approche bootstrap pour refléter l'incertitude sur les composantes principales et vecteurs propres de l'AFDM, utilisés ici pour prédire (imputer) les données. Chaque réplication bootstrap fournit alors une prédiction pour l'ensemble des données incomplètes du jeu de données. Ces prédictions sont ensuite bruitées pour refléter la distribution des données. On obtient ainsi autant de tableaux imputés que de réplications bootstrap. Après avoir rappelé les principes de l'imputation multiple, nous présenterons notre méthodologie. La méthode proposée sera évaluée par simulation et comparée aux méthodes de références : imputation séquentielle par modèle linéaire généralisé, imputation par modèle de mélanges et par "general location model". La méthode proposée permet d'obtenir des estimations ponctuelles sans biais de différents paramètres d'intérêt ainsi que des intervalles de confiance au taux de recouvrement attendu. De plus, elle peut s'appliquer sur des jeux de données de nature variée et de dimensions variées, permettant notamment de traiter les cas où le nombre d'observations est plus petit que le nombre de variables.

Modélisation statistique d'un procédé de centrifugation

Zhanhao LIU (IECL)
Marion PERRODIN (Saint-Gobain Recherche Paris)
Thoams CHAMBRION (IECL)
Radu STOICA (IECL)
Cet article présente une analyse statistique des données issues d'un procédé de centrifugation utilisé à Saint-Gobain. Les différentes corrélations entre les variables enregistrées ont été analysées via une ACP, et sur cette base plusieurs modèles statistiques ont été proposés. L'objectif final est de proposer un processus de contrôle de procédé de centrifugation à travers cette analyse statistique. Ce travail est actuellement

en cours mais les résultats obtenus indiquent déjà quelles étapes du procédé industriel pourraient jouer de manière prépondérante sur la qualité du produit final.

Une analyse des correspondances multiples topologique

Rafik ABDESSELAM (Laboratoire COACTIS-ISH Université de Lyon, Lumière Lyon 2)
L'objectif de ce papier est de proposer une méthode topologique d'analyse des données qui consiste à explorer, analyser et représenter les associations entre plusieurs variables qualitatives dans un contexte d'analyse des correspondances multiples. Les mesures de similarité jouent un rôle important dans de nombreux domaines de l'analyse des données. Les résultats de toute opération de structuration, de classification ou de classement d'objets dépendent fortement de la mesure de proximité choisie. Basées sur la notion de graphes de voisinage, certaines de ces mesures de proximité sont plus ou moins équivalentes. La notion d'équivalence topologique entre deux mesures est définie et statistiquement testée selon leur degré de description des associations entre les modalités de ces variables qualitatives. Un exemple sur données réelles illustre cette méthode.

Statistique computationnelle (Amphi 13)

Modération : Jean-Michel MARIN

Test de permutation pour comparer des groupes indépendants : cas d'un nuage euclidien

Brigitte LE ROUX (MAP5, université Paris Descartes)
Solène BIENAISE (data science)
Jean-Luc DURAND (LEEC, Université Paris 13)
Dans ce papier, nous présentons un test de permutation applicable à des nuages euclidiens qui sont, par exemple, construits par une méthode d'analyse géométrique des données (AGD). Les tests de permutation appartiennent à l'ensemble des méthodes de ré-échantillonnage. Ils ne sont pas basés sur un modèle aléatoire, mais sur des procédures de permutation formulées dans un cadre combinatoire. Les tests statistiques classiques s'appuient, la plupart du temps, sur des hypothèses invérifiables et sont donc souvent inapplicables aux données d'observation, en particulier en AGD. Les tests de permutation ne dépendent que des données observées et il n'est fait aucune hypothèse sur la distribution des données, c'est pourquoi, l'approche combinatoire est le plus en harmonie avec l'analyse inductive des données. Les méthodes que nous proposons s'appliquent à des nuages euclidiens multidimensionnels et traitent de la comparaison des points moyens de plusieurs groupes d'observations (tests d'homogénéité pour des groupes indépendants). Nous présentons une application des méthodes à une enquête portant sur les 'députés et la mondialisation'.

Co-clustering de courbes fonctionnelles multivariées : Une aide pour l'étude de profils consommateurs

Amandine SCHMUTZ (ERIC, LBMC, Lim France)
Julien JACQUES (ERIC)
Charles BOUVEYRON (Université Côte d'Azur, INRIA Sophia-Antipolis, Laboratoire J.A Dieudonné, UMR CNRS 7351 & Equipe Epione)
Laurence CHEZE (LBMC)
Pauline MARTIN (Lim France, CWD-Vetlab)
La croissance exponentielle des objets connectés présents maintenant dans tous les aspects de la vie quotidienne entraine une collecte de données à haute fréquence pour un même individu. Ces objets

facilitent aussi la collecte de plusieurs variables simultanément pour un même individu, entrainant des besoins croissants de méthodes pour résumer et interpréter ces données fonctionnelles multivariées. Ce travail propose une nouvelle méthode de co-clustering fonctionnelle de façon à faciliter la mise en évidence de groupes d'individus et de variables se ressemblant au sein de bases de données multivariées. Cette méthode s'appuie sur un modèle à blocs latents fonctionnels et l'inférence du modèle est faite à l'aide d'un algorithme SEM-Gibbs. L'efficacité de ce modèle sera testée sur un exemple de suivi de consommation électrique et de température au sein de maisons intelligentes connectées.

Estimation avec des données incomplètes informatives dans un cas de faible rang

Aude SPORTISSE (LPSM, CMAP)
Claire BOYER (LPSM, DMA)
Julie JOSSE (CMAP, INRIA)
Pour traiter les données manquantes, la complétion de matrices basée sur des modèles de faible rang est récemment devenue très populaire puisqu'elle s'appuie sur des garanties algorithmiques, méthodologiques et théoriques. Cependant, les méthodes existantes ne s'appliquent pas au cas de valeurs manquantes informatives de type MNAR (Missing Not At Random). Le cas MNAR nécessite de spécifier la distribution du mécanisme de données manquantes et donc d'avoir un fort a priori sur les causes du manque. Par conséquent, cette approche est difficile à mettre en oeuvre en pratique. Nous reformulerons le problème de données manquantes dans une matrice de faible rang à l'aide de modèles graphiques à variables latentes. Ce nouveau cadre établi nous permet d'obtenir des garanties théoriques d'estimation d'espérance d'une variable avec données manquantes informatives de type MNAR auto-masqué, sans avoir à modéliser le mécanisme de données manquantes. Nous présentons aussi une méthode pour l'estimation de la variance et des covariances liées d'une telle variable. Nous proposons également une méthode d'imputation et d'estimation de la matrice paramètre de faible rang. Nous comparerons notre proposition avec les méthodes classiques, comme les algorithmes itératifs de décomposition en valeurs singulières qui ignore le mécanisme de données manquantes, ou la méthode qui le modélise, le plus souvent grâce à un modèle paramétrique de régression logistique.

ABC pour le choix de modèle de formation stellaire des galaxies

Grégoire AUFORT (Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France)
Pierre PUDLO (Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France)
Laure CIESLA (Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France)
Véronique BUAT (Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France)
Nous nous intéressons au problème du choix de modèle bayésien dans le cas où un nombre important de jeux de données, ou d'objets doivent être traités. Nous proposons une extension de l'algorithme ABC-RandomForest pour le choix de modèle, basée sur du boosting d'arbres (minimisation de l'entropie croisée) sur le catalogue de simulations ABC. Cet algorithme d'apprentissage nous permet de contourner l'emploi de statistiques résumées dans notre algorithme ABC. Nous présentons une application à l'astrophysique. À partir de données photométriques, nous montrons la pertinence de la complexification d'un modèle d'histoire de formation stellaires pour une proportion non négligeable de jeux de données parmi des dizaines de milliers de galaxies.

Parcimonie et grande dimension (Amphi 14)

Modération : Efoevi KOUDOU

Exploitation conjointe de plusieurs bases de données dans la détection de signaux en pharmacovigilance via l'utilisation du lasso pondéré

Émeline COURTOIS (Inserm UMR 1181)
Ismaïl AHMED (Inserm UMR 1181)
Pascale TUBERT-BITTER (Inserm UMR 1181)
La pharmacovigilance a pour objectif de détecter le plus précocement possible les effets indésirables des médicaments commercialisés. Ce travail de détection s'apparente à une problématique de sélection de variables en grande dimension. Classiquement il s'effectue sur les bases de notifications spontanées, mais récemment un intérêt croissant s'est porté sur l'exploitation des bases médico-administratives. Nous proposons dans ce travail d'intégrer l'information issue d'une stratégie de détection réalisée à partir d'un référentiel de témoins fournis par les bases médico-administratives, dans les modèles d'analyses de la base des notifications spontanées. Cette intégration de l'information est effectuée via l'utilisation de pénalités différentiées pour chaque covariable médicament dans un lasso pondéré afin de guider la sélection de variables opérée. Ces pénalités différentiées sont obtenues à partir de deux types d'informations : des odds ratio et des p-valeurs corrigées pour les tests multiples. Les performances des méthodes basées sur le lasso pondéré sont comparées à celle d'un lasso classique et sont évaluées empiriquement grâce à un ensemble de signaux de référence concernant l'évènement indésirable 'lésion hépatique aiguë'.

An l1-version of the spectral clustering to promote sparse eigenvectors basis

Camille CHAMPION (IMT / I2MC)
Les graphes jouent un rôle central dans la modélisation des systèmes complexes. Leur analyse est une problématique importante qui couvre une grande variété de domaines et d'applications. Dans ce contexte, nous proposons une variante d'une des méthodes les plus connues d'analyse de graphe, le spectral clustering. Cette nouvelle méthode, appelée l1-spectral clustering, ne requiert pas l'utilisation du k-means pour regrouper les nœuds du graphe, mais estime directement les indicateurs des communautés en déterminant une base propre spécifique à partir d'une pénalité l1.

Joint-Lasso applied to sparse group Partial Least Square and application to pleiotropy

Camilo BROC (UPPA)
Thérèse TRUONG (INSERM)
Borja CALVO (UPV/EHU)
Benoit LIQUET (UPPA)
L'élaboration de données de grande dimension peut être menée en rassemblant des données provenant de différents jeu de données indépendants. Mais que se passe-t-il si l'on s'intéresse à l'effet global d'un prédicteur dans le cas où le type de variable ou la direction de l'effet dépend du jeu de données ? Parmi les modèles parcimonieux, le "Joint Lasso" permet de construire un modèle spécifique à chaque jeu de données tout en liant les modèles par une pénalité Lasso. Cela permet de traiter les différents jeux de données de manière indépendante tout en ayant une sélection globale de prédicteurs. La Régression des moindres carrés partiels (PLS) est une méthode populaire dans l'étude des données Omics. Une de ses extensions parcimonieuses est la "sparse group Partial Least Square". Une application de l'idée du "Joint Lasso" à la sgPLS est proposée, ce qui permet d'ouvrir de nouvelles perspectives en pleiotropie, où une variable Omics peut avoir un effet sur plusieurs variables, et ce, même si, la nature du phénotype ou la direction des effets varie d'un jeu de données à l'autre.

Lasso concomitant avec répétitions (CLaR) : au-delà de la moyenne pour des réalisations multiples en présence d'un bruit hétéroscédastique

Quentin BERTRAND (INRIA, Université Paris-Saclay)
Mathurin MASSIAS (INRIA, Université Paris-Saclay)
Alexandre GRAMFORT (INRIA, Université Paris-Saclay)
Joseph SALMON (IMAG, Univ Montpellier, CNRS)
Les régularisations par normes induisant de la parcimonie sont fréquemment utilisées pour la régression en grande dimension. Une limite des estimateurs ainsi obtenus (le Lasso étant l'exemple canonique) est que le paramètre de régularisation dépend du niveau de bruit, qui varie entre les jeux de données et

les expériences.Des estimateurs comme le concomitant Lasso Owen [2007], Sun and Zhang [2012] ou le square-root Lasso Belloniet al. [2011] résolvent cette dépendance en estimant conjointement le niveau de bruit et les coefficients de régression. Cependant, dans de nombreuses applications expérimentales, les données sont obtenues en faisant la moyenne de plusieurs mesures. Cela aide à réduire la variance du bruit, mais diminue considérablement la taille des échantillons, empêchant ainsi une modélisation ranée du bruit. Dans ce travail, nous proposons un estimateur capable de gérer des structures de bruit complexes (hétéroscédastique) en utilisant l'ensemble des mesures (non moyennées) et une estimation concomitante de structure du bruit.Le problème d'optimisation qui en résulte reste convexe, ce qui permet d'utiliser des algorithmes efficaces pour le résoudre. Grâce à la théorie du lissage Nesterov [2005], Beck and Teboulle [2012], il est donc possible de recourir à des techniques de descente de coordonnées (qui sont état-de-l'art pour ces approches dans un contexte d'apprentissage statistique en grande dimension) pouvant tirer parti de la parcimonie attendue des solutions. Les avantages pratiques sont illustrés sur des données simulées sur des applications de neuroimagerie.

12h40-13h : Clôture des journées

13h-14h20 : Repas

Index