Colloque de statistique

9e Rencontre de travail et d'activité scientifique Sherbrooke-Montpellier

Colloque de statistique

Rencontre de travail et activité scientifique

- Élodie Brunel-Piccinini, professeure, IMAG (Institut Montpelliérain Alexander Grothendieck)
- Éric Marchand, professeur, Faculté des sciences, mathématiques, Université de Sherbrooke
- Gwladys Toulemonde, professeure, IMAG (Institut Montpelliérain Alexander Grothendieck)

Résumé

Les méthodes et le raisonnement statistiques jouent un rôle considérable pour l'avancement des connaissances. Que ce soit dans les enquêtes par sondages ou la mesure d'indicateurs socio-économiques, les essais cliniques pour comparer différents traitements biomédicaux ou l'étude de la survie d'une population animale en écologie, la statistique est omniprésente dans les sciences. La statistique connait une révolution dans ses techniques et son approche, stimulée par le traitement de jeux de données gigantesques d'une complexité sans cesse croissante, mais aussi par des moyens informatiques puissants. La science statistique s'attaque maintenant à des problèmes complexes, par exemple l'analyse des images du cerveau ou des données provenant du génome.

Elle développe de nouvelles méthodes, tel le forage de données (data mining), pour traiter des jeux de données de très grande taille.

La tenue de ce Colloque permettra d'enrichir nos liens et nos intérêts communs en recherche et en formation. En parallèle, ces journées seront aussi une Rencontre scientifique du Laboratoire de Statistique du CRM (voir http://www.crm.umontreal.ca/labo/stat/).

Le laboratoire inclut les meneurs de file de l'école statistique québécoise, qui travaillent sur des sujets tels que l'apprentissage statistique et les réseaux neuronaux, les méthodes d'enquête, l'analyse statistique d'images, les structures de dépendance, l'analyse bayésienne, l'analyse de séries chronologiques et de données financières et les méthodes de rééchantillonnage.

Horaire et Programme
Toutes les présentations auront lieu au D4-2024

Mercredi 19 juin 2024

10 h 30 :	Ouverture du Colloque
10 h 30 à 11 h 15 :	Élodie Brunel-Piccinini (IMAG, Montpellier)
11 h 15 à 12 h :	Erica Moodie (Université McGill)
12 h à 13 h 30 :	Lunch
13 h 30 à 14 h 15	Juliana Schulz (HÉC-Montréal)
14 h 15 à 15 h :	Ali Gannoun (IMAG, Montpellier)
15 h à 15 h 30 :	Pause
15 h 30 à 16 h 15 :	Orlane Rossini (IMAG, Montpellier)
16 h 15 à 17 h :	Sévérien Nkurunziza, (Université Windsor)
17 h 3 0 à 19 h 30 :	Coquetel au QG de l'entrepreneuriat (80 Wellington sud)

Jeudi 20 juin 2024

9 h à 9 h 50 :	Samuel Valiquette (Sherbrooke)
9 h 50 à 10 h 20 :	Pause-santé
10 h 20 à 11 h 05:	Gwladys Toulemonde (IMAG, Montpellier, INRIA)
11 h 05 à 11 h 55 :	Thierry Duchesne (Université Laval)
12 h à 13 h 40 :	Lunch
13 h 45 à 14h 30 :	Klaus Herrmann (Sherbrooke)
14 h 30 à 15 h 15 :	Mamadou Yauck (UQÀM)
15 h 20 à 15 h 50 :	Pause
15 h 50 à 16 h 35 :	Florian Maire (Université de Montréal)
17 h 30 :	Départ en autobus vers Hatley (sur invitation/réservation)
18 h :	Soirée et repas champêtre : Station du Chêne rouge, Hatley (sur invitation/réservation)

Vendredi 21 juin 2024

9 h à 11h :	Rencontre de travail et d'échanges

Programme

1. Élodie Brunel-Piccinini

Titre : Estimation non paramétrique dans un modèle de régression additif avec variables réponse et explicatives fonctionnelles

Résumé : Nous considérons le modèle de régression additif fonctionnel où la réponse est un processus unidimensionnel et les K variables explicatives sont des processus observés sur un intervalle compact. Le processus d'erreur est centré indépendant des variables explicatives et sa variance est bornée. Nous souhaitons estimer les coefficients du modèle qui sont des fonctions déterministes b_j pour $j=1, \dots K$ inconnues. Nous proposons de construire des estimateurs non paramétriques par une méthode des moindres carrés de ces K fonctions sous des conditions très générales sur les processus d'explicatives incluant, par exemple, des processus continus ou des processus de comptage fonctionnelles. Nous bornons un risque de type moindres carrés à partir duquel des vitesses de convergence sont déduites. L'optimalité des vitesses est établie. Une procédure adaptative est ensuite conçue pour mener à une sélection de modèle anisotrope pertinente, simultanément pour toutes les fonctions. Des illustrations numériques et un exemple de données réelles montrent l'intérêt pratique de la stratégie théorique.

2. Erica Moodie

Titre : Médecine de précision : estimation via une modélisation flexible des réponses censurées

Résumé : Pour atteindre l'objectif de fournir des soins optimaux à chaque patient, les médecins doivent personnaliser les traitements. La prise de décisions à plusieurs étapes au cours de la progression d'une maladie peut être formalisée sous la forme d'une stratégie de traitement adaptatif. Pour pouvoir recommander un traitement optimal, il est nécessaire d'estimer les effets causals. Dans cet exposé, je discuterai une extension de l'approche d'estimation populaire du Q-learning, adaptée aux réponses censurées, à l'aide d'arbres de régression additifs bayésiens (« Bayesian additive regression trees (BART) ») pour chaque étape dans une séquence de traitement. Les développements sont motivés et appliqués à une analyse visant à estimer les stratégies optimales de traitement immunosuppresseur pour maximiser la durée de survie sans maladie dans une cohorte de patients ayant subi une allogreffe de cellules hématopoïétiques pour traiter la leucémie myéloïde. (Travail conjoint avec Xiao Li, Brent R Logan et S M Ferdous Hossain.)

3. Juliana Schulz

Titre : Un modèle de Poisson multivarié avec dépendance flexible

Résumé : Les données de dénombrement multidimensionnelles apparaissent fréquemment dans de nombreux domaines d'étude, notamment en gestion des risques, assurance, sciences environnementales, et bien d'autres encore. Lors de l'analyse de données multivariées, il est impératif que le modèle sous-jacent reflète de manière adéquate le comportement marginal ainsi que la dépendance entre les composants. Dans ce travail, nous présentons un modèle pour les données de dénombrement multivariées permettant pour une structure de dépendance flexible en se basant sur les sommes de vecteurs aléatoires de loi de Poisson. En particulier, le modèle permet différents degrés de dépendance en incorporant des vecteurs de chocs comonotones dans la construction. Le cadre général du modèle sera présenté et diverses techniques d'estimation seront discutées. Plusieurs études de simulation seront également présentées, ainsi qu'une application avec des données réelles impliquant des événements de précipitations extrêmes.

4. Ali Gannoun

Titre : Estimation semi-paramétrique de la régression modale

Résumé : Pour certaines lois de probabilité, il existe une relation linéaire surprenante entre le mode, la médiane et la moyenne. Nous extrapolons cette relation au cas de distributions conditionnelles et nous proposons un modèle semiparamétrique pour estimer le mode conditionnel supposé unique. Ainsi la régression modale sera obtenue à partir de l'estimation non paramétrique de la régression moyenne et de la régression médiane reliées par un modèle paramétrique dont on déterminera les paramètres par la méthode des moindres carrés. Pour l'estimation non paramétrique de la moyenne et de la médiane, on utilise la méthode du noyau de convolution ou celle des polynômes locaux. La consistance et le comportement asymptotique de l'estimateur du mode conditionnel seront étudiés dans ce travail. Des exemples seront présentés pour étayer les résultats théoriques.

5. Orlane Rossini

Titre : Apprentissage par renforcement profond pour les processus de Markov déterministe par morceaux contrôlés dans le suivi du traitement du cancer

Résumé : Le cancer nécessite un suivi à long terme et se caractérise par des phases de rémission et de rechute, au cours desquelles un biomarqueur est monitoré et sert de base à une politique de traitement.
Sa dynamique est modélisée par un processus de Markov déterministe par morceaux (PDMP) caché et contrôlé. Le PDMP évolue en temps et en espace continu, le processus est observé à travers un bruit et le modèle est partiellement connu, ce qui rend le problème du contrôle particulièrement difficile.

Nous proposons une nouvelle méthode de contrôle pour ce PDMP, c'est-à-dire pour maximiser la vie du patient ou de la patiente tout en minimisant le coût du traitement et les effets secondaires.

Nous considérons des dates discrètes uniquement pour les décisions, transformant ainsi le PDMP contrôlé en un processus de décision de Markov partiellement observable (POMDP), sur lequel nous implémentons un algorithme d'apprentissage par renforcement profond. L'algorithme deep Q-network (DQN) permet de résoudre le problème de contrôle. Une des limitations de DQN est de ne pas prendre en compte l'historique complet des observations, ce qui est pourtant une caractéristique clé des POMDP. Contrairement au DQN, l'algorithme R2D2 prend en compte l'historique des observations nécessaire au contrôle optimal d'un POMDP.
Nous comparons les deux méthodes de résolution par simulation.

Ces analyses visent à éclairer les avantages et les limites de chaque approche dans le contexte du contrôle de PDMP pour une gestion optimale des maladies chroniques.

6. Sévérien Nkurunziza

Titre: Sur certaines distributions elliptiques tensorielles et leurs applications dans l'analyse de l'imagerie

Résumé : Dans cet exposé, nous considérons un problème d'inférence concernant le paramètre tensoriel d'une distribution elliptique. Plus particulièrement, on considère le scénario où le paramètre en question est susceptible de satisfaire certaines restrictions. Nous présentons certaines propriétés récentes des distributions multivariées qui sont utiles dans les méthodes d'estimation à rétrécissement. Plus précisément, nous présentons des identités remarquables ainsi que des inégalités utiles pour établir le risque et l'optimalité de certains estimateurs tensoriels. De plus, nous montrons l'utilité des résultats établis en régression tensorielle. À ce sujet, les méthodes développées devraient ouvrir de nouvelles perspectives dans l'analyse des données de l'imagerie célébrable.

7. Samuel Valiquette

Titre : Modèle multivarié discret Tree Pólya Splitting

Résumé : Dans ce travail, nous développons une nouvelle classe de distributions multivariées adaptées à des données de comptage, dénommée Tree Pólya Splitting. Cette classe résulte de la combinaison d'une distribution univariée et de distributions multivariées singulières le long d'un arbre de partition connu. Comme nous allons le montrer, ces distributions sont flexibles, permettant notamment la modélisation de dépendances complexes (positives, négatives ou nulles) au niveau des variables observées. Plus précisément, nous présentons les propriétés théoriques des distributions Tree Pólya Splitting en nous focalisant principalement sur les lois marginales, les moments factoriels et les structures de dépendance (covariance et corrélations). Une application sur un ensemble de données de trichoptères est présentée pour, d'une part, illustrer les propriétés théoriques développées dans ce travail sur un cas concret, et d'autre part, montrer l'intérêt de ce type de modèles, notamment en les comparant à d'autres modèles discrets.

8. Gwladys Toulemonde

Titre : Regroupement spatial de processus temporels multivariés basé sur la dépendance extrémale entre sites: application à des données de vent et de précipitations en Europe

Résumé : Les événements climatiques désastreux tels que les inondations, les incendies de forêt et les vagues de chaleur sont souvent dus au comportement extrême simultané de plusieurs processus en interaction. Étant donné que, dans ces événements composés, plusieurs facteurs spatio-temporels sont conjointement extrêmes et que, de par leur nature même, ils ont une dimension élevée, il est nécessaire, pour bien les comprendre, de développer des mesures de la dépendance qui soient appropriées pour les vecteurs aléatoires de valeurs extrêmes. Nous proposons une mesure qui devient alors un ingrédient clé pour proposer un algorithme de regroupement spatial de ces processus temporels. Nous illustrons cette méthode en proposant une tache de régionalisation sur la base de données maillées issues de modèles climatiques sur l'Europe. Cette approche utilise les cumuls journaliers de précipitations et les données journalières de vitesse maximale du vent provenant de l'ensemble de données de la réanalyse ERA5 de 1979 à 2022. (Travail commun avec Alexis Boulin, Elena di Bernardino et Thomas Laloé tous trois de l'Université de Côte d'Azur.)

9. Thierry Duchesne

Titre : Analyse de données longitudinales non standard – travaux passés et à venir

Résumé : Dans cet exposé je vais tout d'abord présenter certains de mes travaux antérieurs sur l'analyse de données longitudinales non standard. Ces travaux ont généralement été inspirés par des applications en écologie et en assurance et incluent l'analyse du déplacement d'animaux, la modélisation de la durée de vie de feux de forêt et la prédiction de la fin de l'association entre un client et une compagnie d'assurances. Je vais ensuite parler de l'état actuel de mes recherches sur ces problématiques et des directions dans lesquelles je compte poursuivre mes investigations. Les besoins passés et futurs de collaborations avec des personnes ayant des expertises en statistique computationnelle, statistique directionnelle, séries temporelles, analyse de texte, statistique spatiale, inférence causale ou modélisation conjointe de données longitudinales et de survie seront mis en évidence tout au long de la présentation.

10. Klaus Herrmann

Titre : Sur une classe de distorsions qui transforment les lois max-stables en lois max-stables

Résumé : Les limites faibles non dégénérées des maxima pour une séquence de variables aléatoires iid appartiennent nécessairement à la classe des distributions de valeurs extrêmes généralisées (GEV) en vertu du théorème de Fisher-Tippett-Gnedenko. Lorsque l'on considère des séquences dépendantes, les distributions asymptotiques résultantes peuvent souvent être exprimées comme des distorsions de la limite iid associée. Les distorsions de puissance qui apparaissent dans le cas de séries temporelles faiblement dépendantes en sont un exemple frappant. Le fait que les distorsions de puissance des distributions GEV restent dans la classe GEV soulève la question de savoir quelle classe générale de distorsions transforme les distributions GEV en distributions GEV. Dans cet exposé, nous répondons à cette question dans le cas univarié en établissant et en résolvant une équation fonctionnelle connexe. Nous discutons de propriétés de la solution et établissons un lien entre notre résultat et le comportement limite faible des maximums pour différents modèles de dépendance. Enfin, nous étendons notre discussion au cas multivarié et discutons d' implications.

11. Mamadou Yauck

Titre : Tests statistiques pour détecter l'homophilie et le recrutement préférentiel dans les enquêtes par traçage de liens sociaux

Résumé : Considérons un réseau d'individus partageant des liens sociaux. L'homophilie ou la tendance des individus ayant des caractéristiques similaires à former des liens sociaux – ou à devenir voisins – est une source courante de dépendance et peut être une source d'inférence invalide si elle est ignorée. Dans cette présentation, nous examinons un plan d'échantillonnage en chaîne sur le réseau social, c'est-à-dire un processus d'échantillonnage dans lequel les liens sociaux sont explorés d'un voisin à l'autre. Le recrutement préférentiel désigne la tendance des individus à recruter des voisins présentant des caractéristiques similaires. Nous abordons la question de la distinction entre l'homophilie et le recrutement préférentiel à partir d'un échantillon de référencement en chaîne. Nous présentons de nouveaux tests statistiques pour détecter l'homophilie et le recrutement préférentiel, puis analysons leurs performances.

12. Florian Maire

Titre : La chaîne de Markov occultée : théorie et application à l'inférence bayésienne

Résumé : On commencera par rappeler l'importance critique des estimateurs Monte Carlo par chaîne de Markov pour l'inférence bayésienne. L'efficacité de tels estimateurs est intimement reliée au niveau d'autocorrélation de la chaîne sous-jacente. Pour améliorer l'efficacité, on propose une altération de la chaîne initiale. Dans notre construction, chaque état peut se retrouver occulté par une variable aléatoire, indépendante du passé du processus, et dont la loi marginale garantit que la loi stationnaire du processus résultant est la loi aposteriori d'intérêt. Intuitivement, l'indépendance injectée devrait réduire l'autocorrélation et mener à des estimateurs plus efficaces. Toutefois, l'analyse formelle est complexe car l'occultation fait perdre la markovianité au processus résultant. Nous prouvons tout de même qu'il hérite d'une loi des grands nombres et un théorème de la limite centrale de la chaine initiale. D'autres résultats concernant la variance asymptotique de la chaîne occultée seront donnés. Enfin, nous présenterons une manière élégante de construire le processus en pratique qui garantit que l'estimateur initial et celui issu du processus occulté ont la même complexité computationnelle. (travail joint avec Max Hird, UCL).