5e Rencontres scientifiques Sherbrooke-Montpellier

Colloque de statistique et de biostatistique

Horaire et programme

Toutes les présentations auront lieu au D4-2019

Le mercredi 10 juin 2015

13h30 à 13h40 : Ouverture du colloque

13h40 à 15h20 : Président : Éric Marchand (Université de Sherbrooke)

- **Benoîte de Saporta,** Équipe de probabilités et statistique, Institut de mathématiques Alexander-Grothendieck (IMAG), Université de Montpellier. ***Asymétrie et mémoire dans la division cellulaire.***

- **Sévérien Nkurunziza,** Département de mathématiques, Université de Sherbrooke**. Inférence statistique dans un modèle linéaire multivarié avec points de rupture.**

15h20 à 16h00 : Pause-santé et séance d'affichage (Atrium-Faculté de Sciences-Édifice D8)

Les présentations dans le cadre de la séance d'affichage sont réservées aux étudiants.

16h00 à 16h50 : Président : Gilles Ducharme (Université de Montpellier)

- **Christelle Reynes,** Laboratoire de biostatistique, informatique et physique pharmaceutique, Institut de génomique fonctionnelle, Université de Montpellier. ***Aspects statistiques de la sélection de variables pour les données -omiques : Quand les statisticiens rencontrent les biologistes... et discutent.***

Le jeudi 11 juin 2015

9h à 9h50 : Président : Sévérien Nkurunziza (Université de Sherbrooke)

- **Gilles Ducharme,** Équipe de probabilités et statistique, Institut de mathématiques Alexander-Grothendieck (IMAG), Université de Montpellier**. Procédure d'extraction de diagnostique après application d'un test d'adéquation.**

9h50 à 10h20 : Pause-santé (Salon du département, D3-1027-5)

10h20 à 12h00 : Président : Ernest Monga (Université de Sherbrooke)

- **Geneviève Lefebvre,** Département de mathématiques, Université du Québec à Montréal. ***Bayesian Adjustment for Confounding : Une approche bayésienne pour l'estimation d'effets causaux.***

- **Éric Marchand,** Département de mathématiques, Université de Sherbrooke**.** ***Estimation par densités prédictives : Résultats récents.***

12h00 à 13h40 : Lunch (Salon du département, D3-1027-5)

13h40 à 15h20 : Présidente : Benoîte de Saporta (Université de Montpellier)

- **Louis-Paul Rivest,** Département de mathématiques et de statistique, Université Laval. ***Copules et estimation dans de petits domaines.***

- **David Haziza,** Département de mathématiques et de statistique, Université de Montréal. ***Estimation robuste pour des populations asymétriques.***

15h20 à 15h50 : Pause-santé (Salon du département, D3-1027-5)

15h50 à 16h40 Président : Taoufik Bouezmarni (Université de Sherbrooke)

- **Yogendra Chaubey,** Department of Mathematics and Statistics, Concordia University. ***On nonparametric density estimation for size-biased data.***

16h40 à 16h42 : Clôture du colloque

Résumés

Asymétrie et mémoire dans la division cellulaire (Benoîte DE SAPORTA)

Les organismes unicellulaires se divisent en deux cellules filles génétiquement identiques. Cependant, certaines expériences montrent une asymétrie de la division qui pourrait s'interpréter comme un effet de mémoire. Je présenterai un modèle auto-régressif de bifurcation pour décrire la structure des données, ainsi qu'un estimateur des paramètres d'intérêt pour tester la symétrie de la division. Les résultats seront appliqués et discutés sur deux jeux de données réelles.

Ce travail est en collaboration avec Bernard Delyon (Univ. Rennes 1), Anne Gégout-Petit (Univ. Lorraine), Nathalie Krell (Univ. Rennes 1) et Laurence Marsalle (Univ. Lille 1).

Inférence statistique dans un modèle linéaire multivarié avec points de rupture. (Sévérien NKURUNZIZA)

Dans cet exposé, je présenterai un problème d'estimation dans un modèle linéaire multivarié avec plusieurs points de rupture inconnus. Plus précisément, je m'intéresserai au scénario où l'on dispose d'une information a priori incertaine sur la matrice des coefficients de régression. En particulier, le paramètre d'intérêt est la matrice des coefficients de régression, tandis que les « points de rupture » sont des paramètres fantômes. Je généraliserai, de trois manières, certaines méthodes récentes connues dans la littérature. Tout d'abord, un problème d'inférence sur un vecteur sera généralisé à celui d'une matrice. Deuxièmement, par rapport à des méthodes récentes du modèle linéaire avec points de rupture, j'allégerai certains présupposés de base et, sous ces conditions plus faibles, j'établirai la normalité asymptotique conjointe entre les estimateurs restreints et sans restrictions. Troisièmement, je construirai une classe d'estimateurs de type à rétrécissement qui comprend comme cas spéciaux les estimateurs des moindres carrés avec et sans restrictions, ainsi que les estimateurs de James-Stein. Par ailleurs, afin de surmonter certaines difficultés inhérentes à l'aspect multidimensionnel, je généraliserai des identités qui sont utiles dans le calcul des risques des estimateurs de type à rétrécissement.

Ce travail est en collaboration avec Fuqi Chen.

Aspects statistiques de la sélection de variables pour les données -omiques : quand les statisticiens rencontrent les biologistes... et discutent (Christelle REYNES)

La place de la statistique en biologie est de plus en plus prépondérante mais l'établissement d'un dialogue constructif entre biologistes et statisticiens est généralement difficile. L'Institut de génomique fonctionnelle de Montpellier pratique régulièrement cette interdisciplinarité. Les enjeux de la qualité de ces interactions seront illustrés à travers un cas concret menant d'une question du biologiste jusqu'à sa réponse pratique. Plus généralement, le problème de la sélection de variables sera abordé en confrontant les points de vue de ces deux communautés. Enfin, nous parlerons de l'utilité des algorithmes génétiques pour la sélection de variables dans le contexte de collaborations biologie/statistique.

Procédure d'extraction de diagnostique après application d'un test d'adéquation (Gilles DUCHARME)

Le problème de valider un modèle statistique est d'une importance primordiale dans de nombreuses applications de la statistique. Pour cette tâche, les statisticiens ont développé et exploré, tant sur le plan théorique qu'empiriquement, de nombreuses procédures appelées « tests d'adéquation ». Pour de nombreux problèmes, ces procédures sont maintenant bien établies et leur utilité reconnue.

Quand un tel test ne rejette pas le modèle, l'utilisateur peut, avec une certaine confiance, procéder à son utilisation. Cependant, lorsque le test rejette le modèle, ce dernier reçoit essentiellement une claque au visage.

D'où l'importance de procédures qui permettent d'aller au-delà de la réponse binaire produite par le test en aidant à déterminer les aspects du modèle que les données réfutent. Ceci permet alors de réparer le modèle ou d'en proposer un nouveau, mieux adapté. De telles procédures sont appelées des procédures d'extraction d'informations diagnostiques. Certaines existent depuis longtemps, mais ce n'est que récemment qu'elles ont évolué en un paradigme dont l'exploitation est plus riche.

Dans cette présentation, je ferai un survol des méthodes d'extraction existantes et présenterai une nouvelle approche permettant d'identifier plus finement les aspects du modèle qui nécessitent des réparations. Ceci mène à la construction d'une structure « d'arbres dans un arbre » d'hypothèses nulles. J'expliquerai comment cette structure permet aussi le contrôle des risques d'erreur. Des simulations compléteront le travail théorique et montreront que la structure peut effectivement détecter les problèmes d'adéquation du modèle. Une discussion sur les perspectives de développement de ces idées formera la conclusion.

Bayesian Adjustment for Confounding : Une approche bayésienne pour l'estimation d'effets causaux **(Geneviève LEFEBVRE)**

L'estimation de l'effet causal d'une exposition sur une réponse est généralement sensible au choix des variables potentiellement confondantes incluses dans le modèle de régression utilisé à cette fin. Dans un premier temps, je présenterai l'approche appelée « Bayesian Adjustment for Confounding » (BAC) développée par Wang et coll. (*Biometrics*, 2012) et proposée pour tenir compte de l'incertitude liée à la sélection des variables confondantes dans un modèle de réponse ajusté. Par la suite, je discuterai de mes travaux récents, qui permettent une meilleure compréhension du fonctionnement de BAC et facilitent l'examen d'approches basées sur les données pour la sélection de l'hyperparamètre omega, la pierre angulaire de cette stratégie de modélisation causale.

Estimation par densités prédictives : Résultats récents **(Éric MARCHAND)**

Pour X, Y dont les lois de probabilité dépendent d'un même paramètre θ, une densité prédictive est une densité basée sur X qui estime la densité de Y. Avec une fonction de coût donnée, telle la perte Kullback-Leibler, on peut formuler une approche bayésienne et évaluer la performance fréquentiste de densités prédictives en évaluant le coût espéré par rapport à la loi de X. Une telle problématique a mené, ces dernières années, à plusieurs avancées, notamment en ce qui concerne la performance fréquentiste de densités prédictives bayésiennes pour des modèles de loi normale multivariée ou de loi de Poisson. Un résultat-type est l'inadmissibilité de la meilleure densité prédictive équivariante en trois dimensions ou plus pour X et Y de lois multinormales, de même moyenne θ avec matrice de covariance proportionnelle à la matrice identité.

Je passerai en revue de tels résultats (par exemple Komaki 2001 ; George et coll. 2006 ; Fourdrinier et coll. 2010) pour le modèle normal, pour la perte Kullback-Leibler, avec des aspects touchant l'estimation de Stein, l'estimation sous contraintes paramétriques et la performance d'estimateurs par substitution (*plug-in).* Je présenterai en outre de nouveaux résultats (obtenus en collaboration avec T. Kubokawa et W.E. Strawderman) pour des données de loi normale, mais pour des mélanges de lois normales, avec les pertes L_2 et L_1 intégrées. Ceci fait intervenir des convolutions, des identités de distance L_1 et L_2, des pertes duales intéressantes pour l'estimation ponctuelle et l'estimation de Stein pour des pertes concaves.

Copules et estimation dans de petits domaines (Louis-Paul RIVEST)

Je ferai d'abord un bref survol du développement des modèles basés sur les copules au cours des 20 dernières années. Quelques classes de copules, incluant les copules archimédiennes et les copules elliptiques, seront définies. L'estimation dans de petites régions, avec modélisation au niveau des unités, sera ensuite abordée. Elle implique un modèle de régression linéaire standard et une distribution échangeable pour capter la dépendance résiduelle intra-région. L'estimation dans les petits domaines est un problème de prévision et des prédicteurs linéaires et non linéaires seront présentés. Ces derniers sont plus complexes et on essaiera de voir s'ils permettent des gains de précision appréciables. Les modèles pour les mettre en œuvre sont semi-paramétriques. En plus des paramètres de régression, ils font intervenir une famille de copules paramétrique pour la dépendance résiduelle intra-région et une fonction de répartition marginale pour les erreurs. L'estimation de ces paramètres et la construction de prédicteurs seront ensuite abordées. Quelques exemples seront présentés.

Estimation robuste pour des populations asymétriques (David HAZIZA)

Coauteurs : Cyril Favre-Martinoz (CREST-ENSAI/IRMAR) et Jean-François Beaumont (Statistique Canada)

L'estimation de la moyenne dans le cas d'une population asymétrique est un problème important en pratique. En effet, il est courant d'observer des variables dont la loi est asymétrique ; c'est le cas par exemple du chiffre d'affaire des entreprises ou le revenu des ménages. En pratique, l'échantillon d'observations possède des unités qui sont très influentes sur la moyenne empirique, qui est l'estimateur souvent privilégié. Rivest (1994) a étudié les propriétés de l'estimateur winzorisé une fois, obtenu en remplaçant la plus grande observation par la deuxième plus grande observation ; il a montré que cet estimateur possède de bonnes propriétés en terme d'erreur quadratique moyenne. Dans cette présentation, j'introduirai un autre estimateur construit au moyen du biais conditionnel d'une unité qui est une mesure d'influence. Je donnerai les propriétés de cet estimateur en termes de biais et d'erreur quadratique moyenne et je développerai une approximation de cette erreur quadratique moyenne suivant les différents domaines d'attraction possibles pour la loi considérée. Un estimateur de l'erreur quadratique moyenne sera également présenté. Finalement, je présenterai les résultats d'une étude par simulation comparant les performances de l'estimateur proposé à celles de l'estimateur winzorisé une fois.

On nonparametric density estimation for size-biased data **(Yogendra CHAUBEY)**

This talk will highlight some recent developments in the area of nonparametric functional estimation with emphasis on nonparametric density estimation for size-biased data. Such data entail constraints that many traditional nonparametric density estimators may not satisfy. A lemma attributed to Hille, and its generalization [see Lemma 1, Feller (1965) *An Introduction to Probability Theory and Applications*, § VII.1)] will be used to propose estimators in this context from two different perspectives. After describing the asymptotic properties of the estimators, I will present the results of a simulation study comparing various nonparametric density estimators. The optimal data-driven approach of selecting the smoothing parameter will also be outlined.