Translations of Abstracts

PRESIDENTIAL ADDRESS

Plus de 60 ans après la fondation de la Société, le Conseil Législatif a massivement approuvé une structure de gouvernance modifiée pour la Société Internationale de Biométrie (IBS), devant prendre effet à dater du 1er janvier 2012. Les responsabilités de la gouvernance et de la direction de la Société seront placées ensemble entre les mains d'un Bureau Exécutif avec le support d'un Conseil Représentatif élargi. Le Conseil représentatif sera formé par des membres sélectionnés des différentes régions (ou composantes géographiques). Il sera responsable pour superviser la nomination et l'élection (par la Société entière) du Bureau Exécutif, et pour assurer l'échange entre les régions et cette équipe de direction. Les membres du Conseil Représentatif présideront également les Comités Permanents. Le processus de transition vers cette nouvelle structure de gouvernance est décrit, ainsi que les objectifs essentiels de la nouvelle décennie.

BIOMETRIC METHODOLOGY

Ce papier explore l'implémentation des dispositifs expérimentaux avec parcelles divisées (« split-plot ») dans les bioessais de dilution en série utilisant des robots. Nous montrons que le plus court chemin pour qu'un robot remplisse correctement les plaques dans un dispositif avec parcelles divisées est équivalent au problème de plus courte superséquence commune en combinatoire. Nous développons un algorithme pour trouver la plus courte superséquence commune, fournissons une implémentation R, et explorons la distribution du nombre d'étapes nécessaires pour implémenter des dispositifs avec parcelles divisées pour un bioessai à travers une simulation. Nous montrons également comment construire des collections de « split-plots » qui peuvent être remplis en un nombre minimal d'étapes, ainsi nous démontrons que les dispositifs avec parcelles divisées peuvent être implémentés avec à peu près le même effort que les dispositifs en blocs aléatoires complets et bandes croisées (« strip-plot »). Enfin, nous fournissons des indications pour modéliser les données qui résultent de ce type de dispositifs expérimentaux.

Les puces à protéines provenant de lysats sont une technologie émergente pour quantifier les ratios de concentration en protéines dans des échantillons biologiques multiples. L'inférence statistique pour une procédure de quantification paramétrique a été abordée insuffisamment dans la littérature, principalement parce que la théorie asymptotique appropriée entraîne un problème avec le nombre de paramètres augmentant avec le nombre d'observations. Dans cet article, nous développons une procédure multi-étapes pour les modèles sigmoïdaux, garantissant une estimation consistante des niveaux de concentration avec une efficacité asymptotique complète. Les résultats obtenus dans le papier justifient les procédures inférentielles basées sur des approximations « grand-échantillon ». Des études de simulation et l'analyse de données réelles sont utilisées dans ce papier pour illustrer la performance des méthodes proposées dans les échantillons finis. La procédure multi-étapes est adaptée pour travailler en asymptotique et est recommandée pour son efficacité statistique dans l'estimation de la concentration en protéine et la stabilité numérique améliorée par une focalisation sur l'optimisation des fonctions objectifs de faible dimension.

Les technologies de séquençage de nouvelle génération sont sur le point de révolutionner le milieu de la recherche biomédicale. La résolution augmentée de ces données promet de garantir une meilleure compréhension des processus moléculaires qui contrôlent la morphologie et le comportement d'une cellule vivante. Cependant, le volume des données générées par ces nouvelles techniques est bien plus important, cela requiert le développement de nouvelles méthodes d'analyse statistique qui doivent être à la fois puissantes et peu gourmandes en ressources de calcul. Dans cet article, nous présentons une méthode capable d'identifier de petites molécules d'ARNs, appelés miARNs, qui régulent la transcription en dégradant ou réprimant la transcription des mARNs qu'ils ciblent. La première étape de notre algorithme consiste à identifier, grâce à un modèle linéaire dynamique, des miARNs candidats dans les données de séquençage haut-débit. Le modèle est souple et permet d'identifier des caractéristiques biologiques d'importance tout en permettant de prendre en compte le décompte des séquences lues ainsi que leur espacement et la profondeur de séquençage. De plus, les miARNs candidats sont analysés avec un algorithme d'alignement de séquences de Smith et Waterman. Cette méthode permet notamment de reconnaître les régions propices à l'existence d'« épingles à cheveux », caractéristiques des miARNs. Nous illustrons notre méthode sur des données simulées ainsi que sur un jeu de données de séquençage haut-débit (plateforme Illumina) de l'organisme *Caenorhabditis elegans*. Ces exemples montrent que notre méthode identifie de façon très sensible des gènes de miARNs connus ou inconnus.

Le séquençage de nouvelle génération (NGS) génère des millions de séquences courtes ce qui fournit des informations précieuses pour étudier les divers aspects de l'activité cellulaire et des fonctions biologiques. Une étape clé pour les applications du NGS (par

exemple, de l'ARN-Seq) est la cartographie des séquences courtes par rapport à leur position réelle sur le génome d'origine. Alors que la plupart de ces séquences courtes ont une localisation unique, une proportion significative s'aligne à plusieurs positions sur le génome avec un nombre égal ou similaire d'erreurs; ceux-ci sont appelés alignements multiples. L'ambiguïté dans la cartographie des alignements multiples peut entraîner un biais dans les analyses en aval. Actuellement, la plupart des utilisateurs ignore les alignements multiples dans leurs analyses, ce qui entraîne une perte d'informations précieuses, en particulier pour les gènes ayant des séquences semblables. Pour améliorer l'exploitation des données NGS, nous avons développé un modèle bayésien qui calcule la probabilité a posteriori d'une séquence à chaque localisation concurrente. Les probabilités sont utilisées pour les analyses en aval, telles que la quantification de l'expression des gènes. Nous montrons à travers des études de simulation et d'analyse d'ARN-Seq de données réelles que notre méthode bayésienne donne de meilleurs rendements que les méthodes de cartographie actuelles. Nous fournissons un programme en C++ dans un logiciel convivial à télécharger.

S. Matsui and H. Noma **1225**
Estimating Effect Sizes of Differentially Expressed Genes for Power and Sample-Size Assessments in Microarray Experiments

Dans les études de puces d'expression qui visent à identifier les gènes différentiellement exprimés en utilisant des tests multiples, l'évaluation de la puissance ou de la taille de l'échantillon est d'une importance particulière pour éviter que des gènes pertinents ne soient pas écartés prématurément d'investigations ultérieures. Dans cette évaluation, une estimation précise de la taille des effets des gènes différentiellement exprimés est crucial en raison de son impact substantiel sur l'estimation de la puissance et de la taille d'échantillon. Cependant, les méthodes classiques qui utilisent les gènes les plus différentiellement exprimés sont sources de surestimation en raison de la variation aléatoire. Dans cet article, nous proposons une méthode d'estimation simple fondée sur des modèles de mélanges hiérarchiques avec une distribution a priori non paramétrique afin de prendre en compte la variation aléatoire et la possible diversité des tailles d'effet au sein des gènes différentiellement exprimés séparées du bruit, i.e. les gènes non différentiellement exprimés. Sur la base d'une estimation bayésienne empirique de la taille des effets, la puissance et le taux de fausses découvertes peuvent être estimés pour les contrôler simultanément dans le criblage des gènes. Nous proposons aussi un indice de puissance, concernant la sélection des gènes les plus différentiellement exprimés avec les tailles d'effet les plus grandes, que nous appelons la puissance partielle'. Ce nouvel indice de puissance est un compromis pratique pour résoudre la difficulté à atteindre des puissances globales élevées dans la plupart des analyses de puces d'expression. Nous appliquons notre méthode à deux jeux de données réelles issus d'études cliniques sur le cancer.

A. Roverato and F. Marta L. Di Lascio **1236**
Wilks' Λ Dissimilarity Measures for Gene Clustering: An Approach Based on the Identification of Transcription Modules

Les méthodes classification sont largement utilisées dans l'analyse des données de micropuces, en raison de leur aptitude à révéler des profils d'expression coordonnées. Un

but important de la classification est de découvrir des gènes co-régulés car il a été postulé que les gènes impactés par les mêmes facteurs de transcription ont tendance à montrer des modèles d'expression similaires. Nous nous intéressons particulièrement à la classification hiérarchique ascendante et considérons le problème du choix de la mesure de dissimilarité sur la base de son pouvoir d'identification des modules fonctionnels formés par un facteur de transcription et les gènes impactés associés. Nous proposons d'abord deux critères qui constituent un cadre théorique pour affirmer l'adéquation et pour comparer différentes mesures de dissimilarité. Nous montrons que les critères proposés permettent d'améliorer l'étude du comportement des mesures de dissimilarité et conduisent à un classement des mesures de dissimilarité les plus utilisées. Puis nous introduisons deux mesures de dissimilarité basées sur la statistique de Wilks et nous montrons que, en nous basant sur les critères précédents, elles ont de meilleures performances que les autres mesures considérées. Les résultats théoriques sont éclairés par une analyse appliquée portant sur des données simulées et des données réelles.

Nous nous intéressons aux tests du rapport des vraisemblances (LRT) et à leur modification pour tester l'homogénéité dans les modèles de mélange par adjonction. Le modèle de mélange par adjonction est un modèle de mélange à deux composants, où un composant est indexé par un paramètre inconnu, tandis que le paramètre de l'autre composant est connu. Ce modèle est largement utilisé en analyse de liaison génétique avec hétérogénéité pour laquelle la distribution du noyau est binomiale. Pour de tels modèles, il est connu depuis longtemps que le test de l'homogénéité n'est pas standard, et que la statistique de test du rapport des vraisemblances ne converge pas vers une distribution usuelle du χ^2. Dans cet article, nous étudions le comportement asymptotique de la statistique du LRT pour le modèle général de mélange par adjonction et nous montrons que sa distribution limite est équivalente à la borne supérieure du carré d'un processus gaussien. Nous discutons aussi les lins et la comparaison entre le LRT et les approches alternatives telles que les modifications du LRT ou le test du score, en y incluant le LRT modifié (Fu et al., 2006). Le LRT est un test omnibus qui exprime sa puissance pour détecter une hypothèse alternative générale. Les approches alternatives contrastent et leur puissance peut être meilleure ou moins bonne selon le type d'alternative à détecter. Nos résultats sont illustrés par des études de simulation et une application à une étude sur les liaisons génétiques pour la schizophrénie.

Il est utile de disposer de tests robustes d'interaction gène-environnement qui puissent utiliser diverses structures familiales avec efficience. Cet article se concentre sur les tests de l'interaction gène-environnement en présence d'effets principaux aussi bien génétiques que environnementaux. L'objectif est de développer des tests puissants pouvant combiner des données ternaires basées sur les génotypes des parents et des

fratries discordantes lorsque les génotypes des parents sont manquants. Nous faisons d'abord une modeste amélioration d'une méthode adaptée à des fratries discordantes (discordance des phénotypes), mais cette approche ne permet pas d'utiliser des familles lorsque l'ensemble des résultats est affecté, c'est-à-dire les trios. Nous effectuons alors une légère amélioration pour une approche basée sur la transmission mendélienne qui est inefficace lorsque des fratries discordantes sont présentes, mais peut être appliquée à toute famille nucléaire. Enfin, nous proposons une approche hybride qui utilise la méthode la plus efficiente pour un type familial spécifique, puis recombine au sein des familles. Nous utilisons cette approche hybride pour l'analyse d'un ensemble de données de pathologie pulmonaire obstructive chronique afin de tester l'interaction gène-environnement concernant le gène Serpine2 avec le tabagisme. Les méthodes sont disponibles librement dans le package **R** nommé fbati.

A. Maity and X. Lin **1271**
Powerful Tests for Detecting a Gene Effect in the Presence of Possible Gene–Gene Interactions Using Garrote Kernel Machines

Nous proposons dans ce papier une procédure de test, puissante, pour détecter les effets d'un gène sur une variable quantitative en présence d'interaction gène*gène (i.e. épistasie) dans un jeu de gènes donné, par exemple une cascade ou un réseau particulier. Les tests classiques dans ce contexte requièrent très généralement un grand nombre de degrés de liberté car ils testent les effets marginaux et toutes les interactions correspondantes dans un contexte paramétrique et de fait sont peu puissants. Nous proposons ici un test puissant fondé sur les machines à noyau. Spécifiquement, notre test est basé sur une méthode à noyau de type garrote, construit comme un test du score. Le type garrote réfère à un paramètre supplémentaire non négatif qui est multiplié à la covariable quantitative d'intérêt de façon à ce que notre test du score puisse être formulé en termes de paramètre non négatif. Une caractéristique essentielle du test que nous proposons est sa flexibilité et le fait i) qu'il soit développé à la fois pour les modèles paramétriques et non paramétriques au sein d'un cadre unifié et ii) qu'il est plus puisant que les tests classiques puisqu'il prend en compte la corrélation entre les gènes et qu'il repose donc sur un plus petit nombre de degrés de liberté. Nous étudions les propriétés théoriques de ce test, nous étudions ses performances dans un échantillon de taille finie par des études de simulations et nous l'appliquons à des données d'expression issues de la cohorte des cancers de la prostate du Michigan.

G. Yi, J. Q. Shi, and T. Choi **1285**
Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data

Le modèle basé sur un processus gaussien (GP) a priori et une fonction de covariance par morceaux peut être utilisé pour ajuster des données non linéaires avec covariables multidimensionnelles. Il a été utilisé comme une approche non paramétrique souple pour l'ajustement de courbes, en classement, en classification et pour d'autres problèmes statistiques. Il a aussi été largement utilisé pour l'étude de systèmes non linéaires complexes dans beaucoup de domaines différents et en particulier en apprentissage automatique. Il existe cependant un problème non résolu lorsque le modèle est utilisé avec des jeux de données importants et de dimension élevée, par exemple dans le cas des

données viande de cet article avec une centaine de covariables hautement corrélées. Pour de telles données le modèle est défaillant en présence d'un nombre important de paramètres de variance et des erreurs de prédictions élevées, et il est confronté à des instabilités dans les calculs. Dans cet article l'approche de la vraisemblance pénalisée est appliquée au modèle basé sur des Processus gaussiens. Différentes pénalités ont été étudiées et leurs capacités à s'appliquer correctement à des modèles ayant les caractéristiques des modèles GP ont été discutées. Leurs propriétés asymptotiques ont été discutées avec des preuves pertinentes. Plusieurs applications à des jeux de données bio-mécaniques et bioinformatiques ont été rapportées.

B. Zhu, P. X.-K. Song, and J. M.G. Taylor **1295**
Stochastic Functional Data Analysis: A Diffusion Model-Based Approach

Cet article présente une nouvelle stratégie de modélisation pour l'analyse de données fonctionnelles. Nous considérons le problème de l'estimation d'une fonction lisse inconnue en partant de données fonctionnelles bruitées. La fonction inconnue est traitée comme la réalisation d'un processus stochastique incorporée dans un modèle de diffusion. La méthode d'estimation par spline de lissage est liée à un cas particulier de cette approche. Les modèles résultants offrent une grande flexibilité pour représenter les aspects dynamiques des données fonctionnelles, et permettent une interprétation aisée et riche. La vraisemblance des modèles est obtenue par l'approximation d'Euler et une augmentation de données. On développe une méthode bayésienne unifiée d'inférence par un algorithme MCMC incluant un lisseur par simulation. Les méthodes et modèles proposés sont illustrés par des données sur un antigène prostatique spécifique, avec lesquelles nous montrons aussi comment les modèles peuvent être utilisés pour de la prévision.

J. Cao, L. Wang, and J. Xu **1305**
Robust Estimation for Ordinary Differential Equation Models

Les scientifiques appliqués aiment souvent utiliser des équations différentielles ordinaires (ODE) pour modéliser des processus dynamiques complexes qui se présentent en biologie, ingénierie, médecine et dans bien d'autres domaines. Il est intéressant mais pose problème d'estimer les paramètres des ODE à partir de données bruitées, particulièrement quand certaines des données sont aberrantes. Nous proposons une méthode robuste pour résoudre ce problème. Le système dynamique est représenté avec une fonction non-paramétrique, qui est combinaison linéaire de fonctions de base. La fonction non-paramétrique est estimée avec une méthode de lissage pénalisée. Le terme de pénalisation est défini avec un modèle ODE paramétrique, lequel contrôle la rugosité de la fonction non-paramétrique et conserve la fidélité de la fonction non-paramétrique du modèle ODE. Les coefficients de base et les paramètres de l'ODE sont estimés à deux niveaux d'optimisation. Les estimateurs des coefficients sont traités comme une fonction implicite des paramètres de l'ODE, lesquels permettent d'en déduire les gradients analytiques pour l'optimisation en utilisant le théorème des fonctions implicites. Des études de simulation montrent que cette méthode robuste donne des estimations satisfaisantes pour les paramètres de l'ODE à partir de données bruitées avec valeurs aberrantes. On applique la méthode robuste en estimant un modèle proie-prédateur à partir de données écologiques.

T. Gaugler and M. G. Akritas **1314**
Testing for Interaction in Two-Way Random and Mixed Effects Models: The Fully Nonparametric Approach

Dans un papier récent, Gaugler et Akritas (2010), ont considéré le test d'absence d'effet principal dans un dispositif mixte à deux facteurs quand les suppositions classiques ne sont pas vérifiées. Nous généralisons ici la modélisation non paramétrique à des dispositifs à effets aléatoires et considérons le problème de l'absence d'effet d'interaction. Les nouveaux modèles pour ces dispositifs tiennent compte des dépendances entre les effets aléatoires, de l'hétéroscédasticité des erreurs et dans les termes d'interaction et ne nécessitent pas la normalité. A un niveau plus systématique, ces modèles se différencient des approches classiques dans le fait qu'ils ne considèrent pas le terme d'interaction aléatoire comme une source additionnelle de variabilité exogène. La procédure de test proposée s'applique à des jeux de données où le facteur aléatoire dans le cas d'un modèle mixte ou au moins un des facteurs aléatoires dans le cas d'un modèle à effets aléatoires a beaucoup de niveaux. Le nombre de répétitions peut être petit et inégal. De plus le modèle et la procédure de test sont assez généraux pour prendre en compte des données manquantes au hasard (« MAR ») à la condition que le mécanisme de censure soit le même pour chaque niveau de l'effet aléatoire. La distribution limite de la statistique de test est la loi normale. Des simulations importantes indiquent que notre procédure de tests, avec ou sans données manquantes, maintient le taux d'erreur nominal de type I dans tous les jeux de simulations. Par contre, les procédures standard (tests F de PROC GLM de SAS et les méthodes ML et REML de PROC MIXED de SAS), ainsi que le test F exact de Khuri (1998) sont extrêmement des procédures libérales en présence de données hétéroscédastiques alors que sous l'homoscédasticité et la normalité la procédure de tests proposée leur est comparable. Une analyse d'un jeu de données du projet Mussel Watch est présenté.

Y. Li, E. P. Wileyto, and D. F. Heitjan **1321**
Prediction of Individual Long-term Outcomes in Smoking Cessation Trials Using Frailty Models

Dans les essais cliniques de cessation du tabac, les sujets reçoivent couramment un traitement accompagné d'un relevé de la consommation journalière de tabac sur un période de plusieurs semaines. Bien que le résultat à la fin de la période soit un indicateur important du succès du traitement, une incertitude demeure sur la façon dont le comportement individuel à l'égard du tabac va évoluer au cours du temps. Il est donc intéressant de prédire le succès de l'arrêt du tabac à long terme basé sur des observations cliniques à court terme. Nous développons une méthode bayésienne de prédiction basée sur un modèle de fragilité à mélanges guéris-récidive que nous avons proposé précédemment qui décrit le processus de transition entre l'abstinence et le tabac. Spécifiquement, nous proposons un algorithme de prédiction à deux étapes qui d'abord utilise l'échantillonnage stratégique pour générer des fragilités spécifiques du sujet à partir des distributions a posteriori conditionnelles aux données observées puis échantillonne les trajectoires de comportement de tabac futures à partir des paramètres du modèle estimé et des fragilités échantillonnées. Nous appliquons la méthode à des données de deux essais cliniques randomisés de cessation du tabac comparant le bupropion à un placebo. Les comparaisons du statut actuel par rapport au tabac à un an

avec les prédictions de notre modèle et d'un grand nombre de méthodes empiriques suggèrent que notre méthode donne d'excellents prédictions.

P. Du, Y. Jiang, and Y. Wang **1330**
Smoothing Spline ANOVA Frailty Model for Recurrent Event Data

Pour estimer le risque de récurrence en utilisant une échelle de temps par intervalle, nous proposons une approche non-paramétrique, fondée sur une décomposition du logarithme du risque relatif par des splines et d'une fragilité dont la distribution n'est pas pré-spécifiée. Cette décomposition fonctionnelle de type ANOVA permet de modéliser les composantes d'hétérogénéité et de corrélation entre intervalles. L'estimation des fonctions de risques sur chaque intervalle successif et la variance de la fragilité repose sur une optimisation stochastique combinant la procédure de Newton-Raphson, l'algorithme d'approximation stochastique (SAA) et une méthode de Monte Carlo Chaînes de Markov (MCMC). Une procédure de sélection des composants négligeables de la décomposition est proposée. Cette modélisation est illustrée sur des données historiques de cancer de la vessie.

J. Cai and D. Zeng **1340**
Additive Mixed Effect Model for Clustered Failure Time Data

Le but de cet article est de proposer un modèle mixte avec une structure additive afin d'analyser les données de survie groupées. Ce modèle fait l'hypothèse d'une structure additive et ajoute un effet aléatoire à la composante fixe habituelle. Il est proche du modèle mixte souvent utilisé dans l'analyse de données répétées, mais adapté au contexte de la modélisation d'un risque ; il peut aussi être considéré comme le développement de modèles de fragilité gamma avec structure additive. Dans cet article, nous avons développé un système d'équation afin d'estimer les paramètres du modèle et avons proposé une façon d'approcher la distribution des effets aléatoires latents en présence de clusters de taille importante.
Nous avons ensuite établi les propriétés asymptotiques de l'estimateur proposé. Les performances de notre méthode sur des échantillons de faible taille ont été démontrées par un grand nombre d'étude par simulation. Pour finir, nous avons appliqué le modèle proposé pour analyser les données d'une étude sur le diabète et d'un essai thérapeutique sur les insuffisances cardiaques congestives.

J. Zhang, Y. Peng, and O. Zhao **1352**
A New Semiparametric Estimation Method for Accelerated Hazard Model

Il y a déjà plus d'une décennie que le modèle à hasards accéléré a été proposé. Cependant son application reste très limitée, en particulier en raison de la complexité de la méthode d'estimation semiparamétrique existante. Nous proposons une nouvelle méthode d'estimation semiparamétrique basée sur une approximation d'une fonction de vraisemblance profilée du modèle par lissage par noyau. La méthode conduit à des équations d'estimations régulières et est facile à mettre en œuvre. Les estimations obtenues par cette méthode sont consistantes et asymptotiquement gaussiennes. Notre étude numérique montre que cette nouvelle méthode est plus efficiente que celle existant jusqu'ici. La méthode proposée est appliquée à une réanalyse des données provenant

d’une étude sur le traitement de tumeurs cérébrales.

On utilise des modèles à risques additifs pour quantifier les effets de prédicteurs génétiques et environnementaux sur des données de floraison de la betterave sucrière se présentant sous forme de données de temps jusqu'à survenue d'un événement, regroupées par catégories. Les effets estimés ont une interprétation intuitive provenant de la dynamique temporelle implicite dans le processus de floraison. Cependant, les expériences agronomiques sont souvent structurées par des parcelles agricoles qui contiennent un grand nombre de plantes suivies régulièrement. Dans cet article, nous utilisons un modèle à risques additifs avec un effet des parcelles induit par des variables de fragilité latentes partagées. Cette approche nous permet de dériver une méthode d'évaluation de la qualité des prédicteurs en termes de proportion de variation des parcelles expliquée. Nous appliquons notre méthode à un important jeu de données explorant la floraison de la betterave à sucre et concluons que le prédicteur du biotype génétique, qui a un effet fort, explique aussi une proportion substantielle de la variation de la parcelle. La méthode est aussi appliquée à un jeu de données de recherche médicale concernant le nombre de jours jusqu'à séropositivité de patients atteints du SIDA.

Nous présentons une généralisation naturelle de l'estimateur de type Buckley-James pour données de survie, applicable à des données censurées à droite et biaisées sur la durée, sous un modèle d'accélération du risque. Le biais de sélection sur la durée apparaît souvent dans les études de prévalence par cohortes ou de dépistage de cancers. La censure à droite informative induite par un échantillonnage biaisé par la durée apporte une difficulté supplémentaire à la modélisation d 'effets des facteurs de risque sur la durée de vie réelle dans la population sous-jacente.
Dans cet article nous évaluons les effets de covariables sur les durées de vie dans la population, sous un modèle d'accélération du risque, conditionnellement aux données biaisées par la durée. Nous construisons une équation d'estimation du type Buckley-James, développons un algorithme itératif, et déterminons les propriétés asymptotiques des estimateurs. Nous comparons les propriétés en échantillon fini des estimateurs proposés à celles d'estimateurs obtenus par les méthodes existantes. Nous illustrons la méthode présentée par une étude de prévalence sur une cohorte de patients atteints de démence.

La réduction de dimension, la sélection de modèles et de variables sont des concepts omniprésents dans les statistiques modernes, le développement de nouvelles méthodes dérivées d’approches classiques est monnaie courante. Cette note passe en revue brièvement les méthodes existantes de régularisation par les moindres carrés pénalisés et par la vraisemblance pour les données de survie et leurs extensions à certaines classes de

fonctions d'estimation pénalisée. Nous montrons que, si l'un des objectifs est d'estimer les contraintes de régularisation des coefficients en utilisant les données de survie observées, aucune stratégie ne parvient à estimer la fonction de Buckley-James. Nous proposons une nouvelle méthode en 2 étapes pour estimer et réévaluer les contraintes de régularisation des coefficients de Dantzig pour données censurées dans le cadre des moindres carrés.
Appliquée à une étude de données de génomique dans l'adénocarcinome du poumon avec une taille d'échantillon de N=200 et de p=1036 gènes de prédisposition, notre méthode permet de sélectionner 10 gènes cohérents avec différents critères et 14 gènes supplémentaires qui méritent de plus amples investigations. Dans une étude de simulation, nous avons trouvé que la technique proposée pour la restauration des contraintes et la sélection de variables a la capacité de faire jeu égal avec les méthodes existantes qui reposent sur la vraie fonction de perte convexe.

Les procédures quantitatives visant à évaluer la plus-value d'un nouveau marqueur sur la base d'un système conventionnel de score de risque pour prédire des taux d'évènements à des moments spécifiques ont été largement étudiées. Toutefois, une simple statistique descriptive, par exemple l'aire sous la courbe ROC ou ses dérivées, peut ne pas apporter une vision claire de la relation entre les systèmes de score de risque conventionnels et nouveaux. Lorsqu'il n'y a pas d'observations censurées, deux simples nuages de points des scores individuels conventionnels et nouveaux pour les `cas' et les `témoins' fournissent une information intéressante concernant le niveau d'incrément global et spécifique par sujet du nouveau marqueur. Malheureusement, en présence de censures, il n'est pas évident de construire de tels graphiques. Dans cet article, on propose une procédure d'estimation non-paramétrique des distributions des différences entre deux scores de risque conditionnellement au score conventionnel. Les courbes de quantile de ces différences rapportées au score conventionnel spécifique aux sujets fournissent une information supplémentaire sur la plus-value globale du nouveau marqueur. Elles nous aident également à identifier un sous-groupe de sujets futurs nécessitant des nouveaux prédicteurs, en particulier lorsque aucune fonction d'utilité n'est disponible pour des décisions sur la base des rapports entre coûts, bénéfice et risque. La procédure est illustrée avec deux jeux de données. Le premier est issu d'une étude largement diffusée dans le cancer du foie menée à la clinique Mayo. Le second provient d'une étude récente dans le cancer du sein pour évaluer la plus-value apportée par un score de gène, qui est relativement cher à mesurer comparé au biomarqueurs cliniques utilisés en routine, pour prédire la survie du patient après chirurgie.

Dans beaucoup d'études avec un résultat de survie, des censures administratives apparaissent quand les suivis se terminent à une certaine date et que beaucoup de sujets sont encore en vie. Il y a aussi une complication supplémentaire dans certains essais

quand il y a refus du traitement assigné. Pour ce cas, nous étudions l'estimation de l'effet causal du traitement sur la probabilité de survie jusqu'à un certain moment pour les sujets qui accepteraient le contrat à la fois du traitement et du contrôle. Nous discutons d'abord sur la méthode standard avec variable instrumentale pour des résultats de survie et les méthodes paramétriques du maximum de vraisemblance, puis nous développons une approche qui se fonde sur un estimateur du maximum de vraisemblance empirique efficace (PNEMLE). La méthode PNEMLE ne nécessite aucune hypothèse sur les distributions des résultats fait utiliser la structure mixte sur les données pour gagner plus d'efficacité qu'avec la méthode standard de variable instrumentale. On donne les résultats théoriques de la PNEMLE et la méthode est illustrée par une analyse de données issues d'un essai de criblage sur le cancer du sein. À partir de notre analyse de mortalité limitée avec une censure administrative de 10 ans de suivi, nous trouvons qu'un bénéfice significatif de criblage est présent après 4 ans (au niveau 5%) et il persiste jusqu'à 10 ans de suivi.

T. J. VanderWeele and I. Shpitser **1406**
A New Criterion for Confounder Selection

Nous proposons un nouveau critère pour la sélection des facteurs de confusion lorsque la structure de causalité sous-jacente est inconnue. Nous supposons de plus que toutes les covariables considérées sont des variables de pré-traitement et que, pour chaque covariable, on sait (i) si la covariable est une cause du traitement, et (ii) si la covariable est une cause du résultat. Les relations causales entre les covariables sont supposées inconnues. Nous proposons de contrôler pour toute covariable qui est soit une cause du traitement, soit du résultat, soit des deux. Nous montrons que, indépendamment de la structure causale sous-jacente réelle, si il existe un sous-ensemble de covariables qui suffisent à contrôler les facteurs de confusion alors les covariables choisies par notre critère suffiront également. Nous montrons que d'autres critères couramment utilisés pour contrôler les facteurs de confusion n'ont pas cette propriété. Nous utilisons la théorie formelle des diagrammes de causalité pour prouver notre résultat, mais l'utilisation de notre critère ne demande pas de connaissance des diagrammes de causalité. Un enquêteur a simplement besoin de demander, «Est-ce que la covariable est une cause du traitement? "et "Est-ce que la covariable est une cause du résultat ?". Si la réponse à l'une des deux questions est «oui», la covariable est inclue comme contrôle de la confusion. Nous discutons d'autres résultats de la sélection de covariables préservant la non-confusion et qui peuvent être intéressants à utiliser avec notre critère.

T. J. VanderWeele, Y. Chen, and H. Ahsan **1414**
Inference for Causal Interactions for Continuous Exposures under Dichotomization

En recherche médicale ou épidémiologique, les variables d'exposition continues sont souvent dichotomisées. Le coût de cette pratique en termes d'efficacité et de biais a été souligné.Nous étudions ici les conséquences de la dichotomisation d'une variable continue sur l'étude des interactions. Nous montrons qu'il est alors possible d'effectuer certaines inférences causales impliquant la mesure d'exposition sur l'échelle continue initiale. Dans le contexte des analyses d'interaction, une dichotomisation et les résultats présentés dans cet article permettent d'éviter l'interprétation en termes d'interaction de résultats dus à une erreur de modélisation des variables d'exposition. Considérer

différents seuils de dichotomisation permet de révéler l'existence d'interactions causales à différents niveaux d'exposition. Ces résultats sont appliqués à une étude de l'interaction entre les expositions à l'arsenic et au tabac dans le développement de lésions cutanées.

Y. Zhao, D. Zeng, M. A. Socinski, and M. R. Kosorok **1422**
Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer

Les régimes typiques pour les stades avancés métastatiques (IIB/IV) du cancer du poumon à cellules non petites (NSCLC) consistent en traitement à intentions multiples. Nous proposons une approche par apprentissage adaptatif par renforcement pour découvrir les régimes de traitements individualisés optimaux à partir d'un essai clinique conçu spécialement (essai à renforcement clinique) d'un traitement expérimental de patients avec un NSCLC avancé qui n'ont pas été traités précédemment par une thérapie systémique. En plus de la complexité du problème de sélectionner les composantes optimales du traitement en première et seconde intention basé sur des facteurs pronostiques, un autre but est de déterminer le moment optimal pour débuter le traitement de seconde intention, soit immédiatement ou retardé après une thérapie d'induction pour obtenir le temps de survie total le plus long. Une méthode d'apprentissage par renforcement, appelée Q-apprentissage qui comporte l'apprentissage d'un régime optimal à partir de données d'un patient générées à partir de l'essai clinique de renforcement. Approximer la Q-fonction à partir de paramètres indexés sur le temps peut être obtenu avec une méthode de régression à support vecteur qui peut utiliser des données censurées. Dans ce cadre, une étude de simulation montre que cette procédure peut extraire les régimes optimaux pour un traitement à deux intentions directement à partir des données cliniques sans connaissance a priori du mécanisme d'effet des traitements. En complément, nous démontrons que le traitement sélectionne de façon fiable le meilleur temps initial pour la thérapie de seconde intention en prenant en compte l'hétérogénéité du groupe de malades NSCLC.

Y. Li, J. M. G. Taylor, and R. J. A. Little **1434**
A Shrinkage Approach for Estimating a Treatment Effect Using Intermediate Biomarker Data in Clinical Trials

Dans un essai clinique, un biomarqueur (S) mesuré après randomisation et fortement corrélé avec le critère principal (T) peut souvent fournir des informations sur T et, par conséquent, des informations sur l'effet d'un traitement (Z) sur T. Un biomarqueur s'avère utile s'il est mesuré avant T, et à un coût inférieur. Dans cet article, nous étudions l'utilisation de S en tant que variable auxiliaire : lorsque S est complètement observé alors que T ne l'est que partiellement, S permet de récupérer de l'information pour estimer l'effet traitement de Z sur T. Idéalement, si S satisfaisait à la définition – irréaliste, dans la pratique – de Prentice de ce qu'est un critère de substitution parfait, on obtiendrait un gain considérable de précision à utiliser les données de S pour estimer l'effet du traitement sur T. Cependant, dès que S s'éloigne un tant soit peu de ce statut de « critère de substitution parfait », il ne peut apporter d'information conséquente que dans des circonstances particulières. Nous proposons d'utiliser une approche ad hoc de régression avec estimateur à rétrécissement, approche qui amène des gains potentiels d'efficacité, par son adaptivité aux données, tout en évitant de poser des hypothèses fortes sur les qualités du critère de substitution. Des simulations montrent que cette approche permet un bon compromis entre biais et efficacité. Les erreurs quadratiques moyennes obtenues

sont meilleures que celles des méthodes concurrentes, et l'on observe une efficacité nettement accrue, en particulier dans le cas très courant où S rend compte d'une partie importante, mais partielle, de l'effet du traitement, et lorsque la taille de l'échantillon est relativement faible. Nous appliquons la méthode proposée à des données concernant un essai dans le glaucome.

Y. Huang and P. B. Gilbert **1442**
Comparing Biomarkers as Principal Surrogate Endpoints

Dans le domaine des critères de substitution, une récente définition utilise les associations causales entre les effets respectifs du traitement sur le biomarqueur (« critère principal de substitution » potentiel) et sur le critère d'évaluation clinique. Bien que l'interprétation de tels critères de substitution soit particulièrement intéressante, peu de recherches ont été menées pour les évaluer. Qui plus est, les méthodes existantes se concentrent sur des modèles de risque ne prenant en compte qu'un seul biomarqueur. De ce fait, pouvoir comparer les mérites, en tant que critères principaux de substitution, de différents biomarqueurs, ou bien envisager des modèles de risque plus généraux incluant simultanément plusieurs biomarqueurs sont autant de questions ouvertes. Pour notre part, nous proposons de caractériser la valeur de substitution d'un marqueur ou d'un critère, dans un modèle de risque, à l'aide de la distribution de la différence des risques estimée entre les groupes de traitement. En outre, nous proposons une nouvelle mesure globale (le gain normalisé total), qui peut permettre de comparer les marqueurs entre eux et de mesurer l'apport d'un nouveau marqueur. Nous développons également une méthode semi-paramétrique de vraisemblance pour estimer une valeur de substitution commune à plusieurs biomarqueurs. Cette méthode, qui peut s'adapter au cas où l'on échantillonne les biomarqueurs en deux étapes, est d'application plus large que les méthodes non paramétriques existantes – parce qu'elle peut intégrer des covariables continues pour prédire le(s) biomarqueur(s) – et s'avère également plus robuste que les méthodes paramétriques – car on n'a pas à faire d'hypothèse sur la distribution des erreurs des marqueurs. Nous illustrons la méthodologie sur des données simulées, ainsi que sur des données réelles recueillies dans le cadre d'essais de vaccins contre l'HIV.

T. Wang and L. Wu **1452**
Multiple Imputation Methods for Multivariate One-Sided Tests with Missing Data

On rencontre fréquemment en pratique des problèmes de tests unilatéraux d'hypothèse multivariées. Des tests variés ont été proposés. En pratique, les données multivariées comprennent des valeurs manquantes. Dans ce cas, les procédures standard de test, basées sur des données complètes, ne peuvent être appliquées ou bien donnent de mauvais résultats si les données manquantes ne sont pas prises en compte. Dans cet article, nous proposons plusieurs méthodes d'imputation multiple adaptées au problème de test unilatéraux d'hypothèses multivariées en cas de données manquantes. Quelques résultats théoriques sont présentés. Les méthodes proposées sont évaluées à l'aide de simulations. Un exemple sur données réelles est présenté pour illustrer les méthodes.

L. Thomas, L. Stefanski, and M. Davidian **1461**
A Moment-Adjusted Imputation Method for Measurement Error Models

Les études de caractéristiques cliniques mesurent fréquemment les covariables par une unique observation. Celle-ci peut être une version avec erreur de mesure du « vrai » phénomène du fait des sources de variabilité comme des fluctuations biologiques ou des erreurs d'appareil. En général, les analyses descriptives et les modèles de réponse qui se fondent sur ces données avec erreur de mesure ne reflèteront pas les analyses correspondantes fondées sur la « vraie » covariable. Beaucoup de méthodes statistiques sont disponibles pour ajuster sur les erreurs de mesure. Des méthodes d'imputation comme la régression calibration et la reconstruction des moments s'implémentent facilement mais ne sont pas toujours adéquates. Des méthodes sophistiquées ont été proposées pour des applications spécifiques telles que l'estimation de densité, la régression logistique et l'analyse de survie. Cependant, il est fréquemment irréalisable pour un analyste d'ajuster chaque analyse séparément, notamment dans les études préliminaires où les ressources sont limitées. Nous proposons une approche par imputation appelée imputation ajustée sur les moments (*Moment Adjusted Imputation*, MAI) qui est flexible et relativement automatique. Comme les autres méthodes d'imputation, elle peut être utilisée pour ajuster une variété d'analyses rapidement et ses performances sont bonnes dans un large ensemble de circonstances. Nous illustrons la méthode par simulation et l'appliquons à une étude sur pression artérielle systolique et issues de santé chez des patients hospitalisés pour insuffisance cardiaque aiguë.

Les erreurs de mesure sur les covariables peuvent induire des estimations biaisées dans les analyses de régression. La plupart des méthodes pour corriger ce biais supposent des erreurs de mesure indépendantes de la variable expliquée (erreurs de mesure non-différenciées). Cependant, dans les modèles de régression pour des données de comptage tronquées en zéro, pour une observation donnée le nombre de mesures de covariables sujettes à erreur peut être égal à la valeur de comptage, ce qui réalise de fait une situation d'erreurs de mesure différenciées. Pour prendre en compte cette difficulté, nous développons une approche de score modifié conditionnel afin d'obtenir une estimation consistante. La méthode proposée est une technique novatrice, avec des gains d'efficacité reposant sur l'augmentation des erreurs aléatoires, et qui montre de bonnes performances dans une étude par simulation. Une application en écologie démontre l'application de la méthode.

Il est important en écologie de pouvoir tester l'homogénéité des assemblages d'espèces. Les données d'abondance récoltées dans les études écologiques ont souvent une structure particulière, qui ne permet pas d'appliquer directement les tests statistiques classiques.
Dans cet article, nous proposons deux nouveaux tests non paramétriques permettant de comparer des assemblages d'espèces, en se basant sur le concept de profondeur des données. Ces tests peuvent être considérés comme la généralisation naturelle des tests de Kolmogorov-Smirnov et de Cramér-von Mises, dans le contexte de la comparaison d'assemblages d'espèces. Nous montrons par simulations que le test proposé est plus puissant que les approches existantes, sous différents scenarii. Enfin nous montrons

comment la méthode s'applique concrètement pour comparer des assemblages d'espèces, sur des données de communautés végétales dans une forêt tropicale de grande diversité spécifique sur l'île de Barro Colorado, Panama.

J. A. Dupuis and M. Goulard **1489**
Estimating Species Richness from Quadrat Sampling Data: A General Approach

Le problème examiné dans cet article est celui de l'estimation du taux d'occupation d'une espèce cible dans une région divisée en unités spatiales (appelées quadrats): cette quantité étant par définition égale à la proportion de quadrats occupés. Nous nous intéressons essentiellement aux espèces spatialement rares ou difficiles à détecter qui sont typiquement détectées dans un très petit nombre de quadrats, et pour lesquelles estimer le taux d'occupation (avec une précision acceptable) est problématique. Nous développons une approche conditionnelle pour estimer la quantité d'intérêt; le conditionnement portant sur la présence de l'espèce cible dans la région d'étude. Nous montrons qu'un tel conditionnement rend identifiable les paramètres du modèle, indépendamment du nombre de visites faites dans les quadrats échantillonnés. Comparée à une approche non conditionnelle, elle s'avère être complémentaire, au sens où elle permet de traiter des questions biologiques qui ne peuvent pas être abordées par l'approche alternative. Deux analyses bayésiennes sont effectuées: la première est non informative, alors que la seconde exploite le fait que de l'information a priori
sur les probabilités de détection est disponible. Il en ressort que la prise en compte d'un tel a priori peut améliorer de façon significative la précision des estimations quand l'espèce cible a été détectée dans un petit nombre de quadrats et que l'on sait par ailleurs que celle-ci est facilement détectable.

S. J. Bonner and C. J. Schwarz **1498**
Smoothing Population Size Estimates for Time-Stratified Mark–Recapture Experiments Using Bayesian P-Splines

Les dispositifs de marquage-recapture de type Petersen sont souvent utilises pour estimer le nombre de poissons ou d’autres animaux d’une population se déplaçant sur un ensemble de traces migratoires. Un premier échantillon d’individus est capturé à un endroit, marqué, et restitué ensuite à sa population. Un second échantillon est pris plus tard plus loin sur le tracé de migration, et on réalise alors l’inférence à partir des nombres d’individus marqués et non marqués observés dans ce second échantillon. Les données de tels dispositifs sont souvent stratifiées selon le temps (le jour ou la semaine) pour tenir compte de modifications possibles des probabilités de capture, ce qui rend les méthodes précédentes d’analyse inaptes à utiliser pleinement les relations temporelles dans les données stratifiées. Nous présentons une méthode bayésienne, semi-paramétrique, qui modélise explicitement le nombre moyen de poissons dans chaque strate comme fonction lisse du temps. Nous utilisons les résultats de l’analyse de données historiques de la migration du jeune saumon Atlantique (*Salmo salar*) le long du fleuve Conne, à Terre-Neuve, ainsi que celles d’une étude de simulation pour montrer que cette nouvelle méthode fournit des estimations plus précises de la taille de la population, et de meilleures estimations de l’incertitude que les méthodes habituelles.

Y. Zhao, D. Zeng, A. H. Herring, A. Ising, A. Waller, D. Richardson, and M. **1508**

R. Kosorok
Detecting Disease Outbreaks Using Local Spatiotemporal Methods

Une méthode de surveillance en temps réel est développée mettant l'accent sur la détection rapide et précise de l'émergence d'épidémies. Nous proposons un modèle avec des hypothèses relativement faibles concernant les processus latents générant les données observées, permettant une prévision robuste de la surface spatio-temporelle d'incidence. L'estimation se fait par une méthode de régression linéaire locale avec prise en compte des effets jour-de-semaine, où le lissage spatial est effectué à l'aide d'une nouvelle distance métrique qui s'ajuste à la densité de la population. La détection d'épidémies émergentes est effectuée par l'intermédiaire de l'analyse des résidus. Les résidus journaliers et les résidus basés sur un modèle autorégressif auquel on a enlevé la composante de tendance sont utilisés pour détecter des anomalies dans les données, sachant qu'un résidu journalier large ou une tendance temporelle croissante dans les résidus indique une potentielle épidémie ; le seuil de signification statistique étant déterminé en utilisant une approche par rééchantillonnage.

R. M. Fewster **1518**
Variance Estimation for Systematic Designs in Spatial Surveys

Lors d'analyses spatiales visant à estimer la densité d'objets dans une région, un échantillonnage systématique donne en général une variance plus faible qu'un échantillonnage aléatoire. Mais il est bien connu que l'estimation de la variance systématique est un problème difficile. Les méthodes existantes tendent à surestimer la variance, ce qui conduit à ce paradoxe que, bien que la variance soit véritablement réduite, elle est surestimée, et le gain de l'échantillonnage plus efficace est perdu. Les méthodes courantes d'estimation de la variance systématique dans les analyses spatiales sont d'approcher l'échantillonnage systématique par un échantillonnage aléatoire ou par un échantillonnage stratifié. Des travaux antérieurs ont montré que l'approximation par un échantillonnage aléatoire pouvait donner de très mauvais résultats, et que l'approximation par un échantillonnage stratifié, bien que constituant une amélioration notable, pouvait aussi être fortement biaisée dans certains cas. Nous développons ici un nouvel estimateur basé sur la modélisation d'un processus de rencontre dans l'espace. Ce nouvel estimateur « striplet » a un biais négligeable et une excellente précision dans un grand nombre de scénarios simulés, dont l'échantillonnage par bandes, par distances, et par quadrats, et y compris pour des populations d'objets montrant des tendances ou des agrégations très fortes. Nous avons appliqué cet estimateur à des données de distribution des hyènes tachetées (*Crocuta crocuta*) dans le parc national du Serengeti (Tanzanie). Nous avons trouvé que le coefficient de variation pour les densités estimées est de 20% en utilisant l'approximation par un échantillonnage aléatoire, 17% pour l'approximation par un échantillonnage stratifié, et de 11% avec notre nouvel estimateur. Cette importante réduction de la variance estimée est aussi vérifiée par simulation.

Y. Guo **1532**
A General Probabilistic Model for Group Independent Component Analysis and Its Estimation Methods

L'analyse en composantes indépendantes (ICA) est devenue un outil important pour

l'analyse des données en provenance d'études d'imagerie fonctionnelle à résonance magnétique (fMRI). ICA a été appliquée avec succès aux données de fMRI sur un seul sujet. L'extension de l'ICA à l'inférence de groupe reste toutefois difficile en raison de l'indisponibilité à priori d'une matrice d'incidence indicatrice des groupes et de l'incertitude sur la variabilité inter-sujets des données de fMRI. Dans cet article, on présente un modèle probabiliste ICA général (PICA) pouvant accepter différentes structures de groupes de processus spatio-temporels multi-sujets. Un avantage du modèle proposé est qu'il peut modéliser avec flexibilité plusieurs types de structures de groupes pour différents signaux sources de neurones et dans différentes conditions expérimentales pour des études dans la fMRI. Une méthode de maximisation de la vraisemblance est utilisée pour estimer ce modèle général de groupes ICA. On propose deux algorithmes EM pour obtenir les estimateurs de maximum de vraisemblance. La première méthode repose sur un algorithme EM exact qui fournit une étape 'E' exacte et une étape 'M' explicite et non-itérative. La seconde méthode est une approximation variationnelle de l'algorithme EM plus efficace que l'EM exact. Dans des études de simulation, on compare d'abord la performance du modèle général de groupes PICA proposé et de l'approche probabiliste de groupes existante. On compare ensuite les deux algorithmes EM proposés et on montre que l'approximation EM variationnelle permet d'atteindre une précision comparable à celle de l'EM exact avec significativement moins de temps de calcul. Un exemple de données de fMRI est utilisé pour illustrer les méthodes proposées.

Y. Yuan and G. Yin **1543**
Dose–Response Curve Estimation: A Semiparametric Mixture Approach

Pour l'estimation d'une courbe dose-réponse, les modèles paramétriques sont pratiques et efficients mais sujets à des défauts de spécification ; les méthodes non-paramétriques sont robustes mais moins efficients. Nous proposons, en compromis, une approche semi-paramétrique qui combine les avantages des deux modes, paramétrique et non-paramétrique, d'estimation de courbe. Avec une forme de mélange, notre estimateur prend une moyenne pondérée des estimateurs paramétrique et non-paramétrique de la courbe, avec une poids plus élevé donné à l'estimateur donnant le meilleur ajustement. Lorsque l'hypothèse de modèle paramétrique est valide, l'estimateur semi-paramétrique de courbe converge vers l'estimateur paramétrique et donne ainsi une efficience élevée ; en présence d'erreur de spécification du modèle paramétrique, l'estimateur semi-paramétrique converge vers l'estimateur non-paramétrique et la consistance est assurée. Nous considérons également un schéma de pondération adaptative pour permettre aux poids de varier selon l'ajustement local des modèles. Nous avons mené des études extensives de simulation pour étudier la performance des méthodes proposées, et nous les illustrons avec deux exemples réels.

N. A. Wages, M. R. Conaway, and J. O'Quigley **1555**
Continual Reassessment Method for Partial Ordering

La plupart des approches statistiques pour les plans expérimentaux d'essais cliniques de Phase I en oncologie sont destinées à l'étude d'un seul agent cytotoxique. Le but de ces essais est d'identifier la dose maximale tolérée, c'est à dire la dose la plus élevée qu'on puisse administrer sans dépasser une limite acceptable de toxicité. Une hypothèse de

travail importante pour ces approches concerne la supposée monotonicité de la courbe dose-toxicité, une hypothèse raisonnable lorsqu'il s'agit d'un seul agent où la proportion de patients présentant une toxicité inacceptable augmente avec la dose. Pour les agents multiples cette hypothèse ne tient plus puisque l'ordre des probabilités de toxicité n'est pas connu pour plusieurs des combinaisons des agents. En même temps certains des ordres sont connus, d'où un problème d'ordre partiel. Dans cet article nous proposons une nouvelle approche bi-dimensionnelle pour les agents multiples. Cette approche se réduit à la méthode CRM lorsque l'ordre est connu parfaitement. Elle a l'avantage d'assouplir l'exigence d'avoir des courbes dose-toxicité monotone croissantes. Une étude comparative est réalisée pour évaluer la performance de la nouvelle méthode face à un design CRM quand l'ordre est connu, ainsi que d'autres méthodes proposées pour ce type de problème.

BIOMETRIC PRACTICE

H. R. Al-Khalidi, Y. Hong, T. R. Fleming, and T. M. Therneau **1564**
Insights on the Robust Variance Estimator under Recurrent-Events Model

Lors de l'analyse de la survenue d'événements récurrents, le modèle de Andersen-Gill est couramment employé pour estimer l'effet d'un traitement sur le risque de récurrence. Il en est ainsi en pathologie cardiovasculaire pour des patients avec un défibrillateur implantable (ICD) qui présentent des arythmies récurrentes qui se terminent par des chocs ou une stimulation antitachycardique délivrée par l'appareil. Dans un essai clinique randomisé publié, un modèle à événement récurrent était utilisé pour étudier l'effet d'un traitement médicamenteux chez des sujets avec ICD et présentant des événements symptomatiques et récurrents d'arythmie. La variance de l'estimateur de l'effet traitement tend à décroître avec les survenues de récurrences, qui augmentent le nombre total d'événements observés. Cependant nous illustrons à partir des données de deux essais clinique randomisés que cette décroissance de la variance n'est pas systématique. Une décomposition analytique de la variance robuste est proposée et permet de confirmer les résultats empiriques en établissant quel terme domine en fonction de l'hétérogénéité de l'échantillon. Nous étendons ce résultat au calcul du nombre de sujets nécessaire lorsque le critère de jugement principal est un événement récurrent.

C.-N. Wang, R. Little, B. Nan, and S. D. Harlow **1573**
A Hot-Deck Multiple Imputation Procedure for Gaps in Longitudinal Recurrent Event Histories

Nous proposons une méthode d'imputation de type Hot-Deck basée sur la régression pour des intervalles de données manquantes dans les études longitudinales, dans lesquelles les sujets ont un processus d'événements récurrents et un événement terminal. Des exemples d'application sont les épisodes répétés d'asthme et le décès, ou les menstruations et la ménopause comme dans l'application ayant motivé ce travail. L'intérêt de cette recherche est le temps de survenue d'un événement marqueur, défini par un processus d'événements récurrents, ou le délai entre cet événement marqueur et l'événement final. La présence d'intervalles de temps sans aucune donnée dans les enregistrements de l'événement rendent difficile la détermination du temps de survenue de l'événement marqueur et donc du délai entre cette survenue et l'événement final. Des approches

simples comme le fait de sauter les intervalles de temps manquants ou d'éliminer les sujets ayant des intervalles de temps manquants ont des limites évidentes. Nous proposons une procédure pour imputer l'information dans l'intervalle de temps manquant en substituant l'information dans cet intervalle par celle d'un individu apparié ayant un enregistrement complet de ses données dans l'intervalle correspondant. La technique d'appariement sur la moyenne prédite est utilisée pour incorporer l'information sur les caractéristiques longitudinales du processus récurrent et du temps d'événement final. L'imputation multiple est utilisée pour diffuser l'incertitude de l'imputation. La procédure est appliquée à un grand jeu de données pour évaluer le temps et la durée de la transition vers la ménopause. Les performances de la méthode proposée sont évaluées par une étude de simulations.

Lorsque des questions scientifiques nouvelles se posent après que des données binaires longitudinales ont été recueillies, la sélection ultérieure des sujets de la cohorte auxquels on demandera des données détaillées supplémentaires est souvent nécessaire pour recueillir efficacement de nouvelles informations. Des exemples clefs de collecte de données supplémentaires incluent des données rétrospectives obtenues par questionnaire, des données nouvelles par croisement, ou l'évaluation des spécimens biologiques conservés. Dans ces cas, toutes les données requises pour les nouvelles analyses sont disponibles sauf pour le nouveau prédicteur ou la nouvelle exposition visé(e). Nous proposons une classe de plans d'échantillonnage dépendant d'une réponse longitudinale et détaillons une analyse par maximum de vraisemblance conditionnelle corrigée du schéma d'étude pour une estimation hautement efficace des coefficients des covariables variant dans le temps et indépendantes du temps quand les contraintes de ressources interdisent la mesure de l'exposition chez tous les participants. De plus, nous détaillons une phase importante de la conception de l'étude qui exploite les données disponibles de la cohorte pour examiner de façon pro-active la faisabilité de toute sous-étude proposée, ainsi que pour informer les décisions quant au schéma d'étude le plus souhaitable. Les schémas d'étude proposés et les analyses associées sont discutés dans le contexte d'une étude visant à examiner l'effet modificateur d'un polymorphisme nucléotidique simple (SNP) de la cytokine interleukine-10 sur la régression des symptômes de l'asthme chez des adolescents participant à l'étude continuant le programme de gestion de l'asthme dans l'enfance (*Childhood Asthma Management Program Continuation Study*). Dans cet exemple, nous supposons que toutes les données nécessaire pour conduire l'étude sont disponibles sauf les données de génotypage sujet-spécifique. Nous supposons aussi que ces données seraient obtenues après analyse d'échantillons de sang conservés, dont le coût limite la taille d'échantillon.

Les mesures de charge virale ARN VIH sont souvent sujettes à des limites de détection supérieure ou inférieure selon la technique de quantification. Ainsi, les réponses sont

censurées soit à gauche soit à droite. Des modèles linéaires (ou non linéaires) à effets mixtes (avec des modifications pour prendre en compte la censure) sont couramment utilisés pour analyser ce type de données et sont basés sur des hypothèses de normalité pour les termes aléatoires. Cependant, ces analyses peuvent ne pas donner d'inférence robuste quand les hypothèses de normalité sont remises en cause. Dans cet article, nous développons une approche Bayésienne pour des modèles linéaires (et non linéaires) remplaçant les hypothèses Gaussiennes pour les termes aléatoires par des distributions normales/indépendantes (NI). La NI est une classe de densités symétriques avec des queues de distributions longues intéressante qui inclus en particulier les distributions normale, *t* de Student, slash et normale contaminée. La vraisemblance marginale est calculable (en utilisant des approximations pour les modèles non-linéaires) et peut être utilisées pour développer des diagnostics d'influence par délétion de cas basés sur la divergence de Kullback-Leibler. Les procédures nouvellement développées sont illustrées avec des simulations ainsi que deux études VIH/SIDA sur des charges virales qui étaient initialement analysées avec des modèles à effets mixtes utilisant une distribution normale (censurée).

J. F. Bobb, F. Dominici, and R. D. Peng **1605**
A Bayesian Model Averaging Approach for Estimating the Relative Risk of Mortality Associated with Heat Waves in 105 U.S. Cities

L'estimation des risques que les vagues de chaleur posent à la santé des personnes constitue une étape importante dans l'évaluation de l'impact futur du changement climatique. Dans cet article, nous proposons une classe flexible des modèles pour séries chronologiques pour estimer le risque relatif de mortalité lié aux vagues de chaleur, et nous mettons en œuvre un moyennage de modèle Bayésien (MMB) pour tenir compte de la multiplicité de modèles potentiels. En appliquant ces méthodes aux données provenant de 105 villes des États-Unis pour la période 1987-2005, nous identifions les villes ayant une probabilité a posteriori élevée de risque accru de mortalité pendant les vagues de chaleur, examinons l'hétérogénéité des distributions a posteriori du risque de mortalité entre les villes, évaluons la sensibilité des résultats au choix des distributions a priori, et comparons nos résultats de MMB à une approche de sélection de modèle. Nos résultats montrent qu'aucun des modèles seul ne permet de prédire le risque de façon satisfaisante sur la majorité des villes, et pour certaines villes l'estimation du risque lié à la vague de chaleur est sensible au choix du modèle. Alors que le moyennage de modèle conduit à des distributions a posteriori avec une variance croissante comparé à l'inférence statistique conditionnelle sur un modèle obtenu par sélection de modèle, nous constatons que la moyenne a posteriori du risque de mortalité lié à la vague de chaleur est robuste pour la prise en compte de l'incertitude due au modèle parmi une large classe de modèles.

L. Ruan and M. Yuan **1617**
An Empirical Bayes' Approach to Joint Analysis of Multiple Microarray Gene Expression Studies

La prévalence des études sur l'expression des gènes et la faible reproductibilité liée à des échantillons de taille insuffisante conduisent naturellement à envisager l'analyse combinant des données de différentes expériences dans le but réel d'améliorer la précision des résultats. Dans cet article nous présentons une approche modélisante, pour

une meilleure identification de gènes à expression différenciée en rassemblant des données de différentes études. Le modèle peut s'adapter à une grande variété d'études incluant celles réalisées sur différents environnements en ajustant chaque donnée par plusieurs ensembles de paramètres, et/ou sous différentes conditions biologiques pouvant être partiellement recouvrantes. Des inférences peuvent être réalisées par une approche bayésienne empirique. En raison de l'information partagée entre les études, l'analyse jointe améliore de manière extrêmement importante les inférences faites sur les analyses individuelles et non regroupées. Des études par simulation et des exemples sur données réelles sont présentés pour mettre en évidence l'efficacité de l'approche proposée en envisageant des complications survenant fréquemment en pratique.

L'anticipation génétique, décrite par un âge de survenance précoce et des symptômes de plus en plus agressifs dans les générations successives, est un phénomène remarqué dans certaines maladies héréditaires. Son extension peut varier entre les familles et/ou entre les sous-types de mutations connues pour être associées au phénotype de la maladie. Dans ce papier, nous postulons une approche Bayésienne pour inférer l'anticipation génétique sous des modèles à effets aléatoires pour données censurées flexibles qui capturent l'effet des générations successives sur l'âge de début de maladie. L'intérêt principal réside dans les effets aléatoires. Une mauvaise spécification de la distribution des effets aléatoires peut entrainer des conclusions inférentielles incorrectes. Nous comparons l'ajustement de 4 distributions candidates d'effets aléatoires via des diagnostics Bayésiens d'ajustement de modèle. Une question statistique apparentée est l'isolement de l'effet confondu des changements de tendances séculières, dépistages et pratiques médicales qui peuvent influer sur la détection de la maladie à travers les cohortes de naissance. En utilisant un registre de données historiques de cancer, nous empruntons aux méthodes d'analyses de survie relative pour ajuster les changements sur l'incidence spécifique à l'âge à travers les cohortes de naissance. L'étude cas-témoins qui nous motive est issue d'un registre de cancer Danois de 124 familles avec des mutations dans les gènes réparant les erreurs de duplication, et connues pour causer un cancer colorectal héréditaire sans polypose, aussi appelé syndrome de Lynch. Nous trouvons la preuve d'une baisse de l'âge de début de maladie entre générations dans cette étude. Notre modèle prédit les effets de l'anticipation au niveau familial qui sont potentiellement utiles dans les cliniques de conseil génétique pour les familles à haut risque.

Nous considérons les régimes de traitement dans lesquels un agent est administré en continu à une concentration spécifiée jusqu'à l'obtention d'une certaine réponse ou jusqu'à ce qu'un temps maximum de délivrance prédéterminé soit atteint. La réponse envisagée est un événement défini pour caractériser l'efficacité thérapeutique. Une portion de la quantité maximale totale planifiée pour l'administration est donnée par un bolus initial. Pour de tels régimes, la quantité reçue par le patient dépend du temps écoulé

à la réponse. Une complication supplémentaire lorsque la réponse est évaluée périodiquement et non pas continument est que le temps à la réponse est alors une variable censurée par intervalle. Nous nous attachons au problème de la définition du schéma d'un essai clinique dans lequel de telles données de temps à la réponse ainsi qu'un indicateur binaire de toxicité sont utilisées conjointement pour optimiser la concentration et la quantité du bolus. Nous proposons un schéma bayésien adaptatif et séquentiel qui choisit le traitement optimal pour des patients successifs en maximisant l'utilité moyenne a posteriori du résultat conjoint efficacité-toxicité. La méthodologie est illustrée par un essai dans lequel l'activateur tissulaire du plasminogène est administré en intra-artériel pour le traitement rapide de l'accident ischémique aigu.

Dans les pays occidentaux où l'approvisionnement alimentaire est satisfaisant, le régime des consommateurs est agencé à partir d'un grand nombre d'aliments. L'objectif de ce travail est d'étudier comment une technique récente d'analyse en variables latentes, la factorisation positive de matrices ("Nonnegative Matrix Factorization" en anglais ou NMF), peut être appliquée aux données de consommation afin de mieux comprendre cet agencement. Les données de consommation sont positives par nature et de grande dimension. Le modèle NMF fournit une représentation de ces données au moyen de vecteurs latents dont les coefficients sont positifs ou nuls, appelés ici systèmes de consommation, et en nombre très petit. L'approche NMF favorisant la parcimonie de la représentation des données produite, les systèmes de consommations obtenus sont facilement interprétables. Au delà d'une illustration des propriétés fournie au travers de résultats simulés, la méthode NMF est appliquée à des données issues d'une enquête de consommation française. Les résultats numériques ainsi obtenus sont présentés et discutés en profondeur. Un regroupement, fondé sur la méthode des k plus proches voisins dans le sous-espace latent de consommation en découlant, est effectué afin de recouvrer des groupes de consommateurs facilement interprétables par les nutritionnistes.

En pratique pour analyser des données d'expériences de capture-recapture il est tentant d'appliquer des méthodes statistiques avancées modernes à des observations historiques de capture. Cependant à moins que l'analyse ne prenne en compte que les données collectées sur des individus ayant été capturés au moins une fois, les résultats peuvent être biaisés. Sans le développement de nouveaux programmes, des méthodes comme les Modèles Additifs Généralisés, les Modèles Linéaires Mixtes Généralisés, et la Simulation-Extrapolation ne pourraient être facilement implémentés. Contrastant avec ceci, l'approche par vraisemblance partielle permet de réaliser l'analyse d'expériences de capture-recapture à l'aide des programmes les plus communément disponibles. Nous examinons ici l'efficience de cette approche, et nous l'appliquons à plusieurs ensembles de données.

READER REACTION

Les tests exacts non conditionnels ont largement été utilisés pour tester la différence entre deux probabilités avec des données binaires appariées par paire 2x2 dans le cadre de petits échantillons. Dans ce contexte, Lloyd (2008, *Biometrics* **64**, 716-723) a proposé une *p*-valeur E+M qui montre des performances meilleures que les *p*-valeur M et *p*-valeur C. Cependant, le calcul analytique de la *p*-valeur E+M nécessite que la condition de convexité de Barnard soit respectée ; ceci peut être difficile à prouver théoriquement. Dans cet article, par une simple reformulation, nous montrons qu'une condition plus faible, la monotonie conditionnelle, est suffisante pour calculer les trois *p*-valeurs (M, C et E+M) et leur taille exacte correspondante.
De plus, cette condition de monotonie conditionnelle est applicable pour les tests de non-infériorité.