Translations of Abstracts

BIOMETRIC METHODOLOGY

Beaucoup d'études de survie sont compliquées par la présence de risques compétitifs et par l'appartenance d'individus à un cluster, tels que des patients dans le même centre pour une étude multicentrique. Plusieurs méthodes ont été proposées pour modéliser la fonction d'incidence cumulée avec des observations indépendantes. Cependant, quand des patients appartiennent à un même cluster, il est nécessaire de prendre en compte la présence d'un effet cluster soit à travers la modélisation de la fragilité du risque ou sa sous-distribution, ou par l'ajustement de la corrélation intra-groupe dans un modèle marginal. Nous proposons une méthode pour modéliser directement la fonction d'incidence cumulée marginale par une validation croisée des pseudo-observations de l'incidence cumulée à différents délais. Elles sont ensuite utilisées dans une équation généralisée (GEE) pour modéliser la courbe d'incidence cumulée marginale, et ainsi obtenir des estimations cohérentes des paramètres du modèle. Un estimateur à variance sandwich est dérivé pour ajuster la corrélation intra-groupe. La méthode, qui est une généralisation de plusieurs modèles existants, est facile à implémenter en utilisant des logiciels standards une fois que les pseudo-valeurs sont obtenues. Des études de simulation montrent que la méthode est performante pour ajuster l'erreur standard pour la corrélation intra-groupe. La méthode est appliquée sur un jeu de données portant sur les résultats après transplantation de moelle osseuse.

Dans cet article nous proposons un modèle de Cox avec fragilité positive stable partagée, pour des données de durée en grappe, dans lesquelles la distribution de la fragilité varie en fonction de covariables liées à la grappe. Le modèle proposé tient compte de la corrélation intra-grappe dépendant des covariables, et permet l'inférence conditionnelle aussi bien que marginale. Nous obtenons l'inférence marginale directement à partir d'un modèle marginal, puis nous utilisons une vraisemblance stratifiée pseudo-partielle de type modèle de Cox, afin d'estimer le coefficient de régression associé au paramètre de fragilité. Les estimateurs proposés sont consistants et de distribution asymptotique normale, et nous donnons un estimateur consistant de la matrice de covariance. Nous montrons par simulations que la procédure d'estimation proposée est applicable avec un nombre réaliste de grappes. Finalement, nous appliquons cette méthode à des données de transplantation rénale provenant du « registre scientifique des receveurs de greffes ».

H. Zhang, D. E. Schaubel, and J. D. Kalbfleisch **18**
Proportional Hazards Regression for the Analysis of Clustered Survival Data from Case–Cohort Studies

Lors d'analyses sur de grandes cohortes, il n'est pas rare de trouver des études cas-témoin nichées. La plupart des méthodes d'analyse des études cas-témoin n'existent qu'en univarié. D'autre part, des données de survie avec cluster se rencontrent fréquemment dans les données de santé. Par exemple, des sujets traités dans un même centre ne sont pas nécessairement indépendants. Dans ce papier, nous étudions des méthodes fondées sur les équations d'estimation pour analyser des données censurées, cas-témoins avec cluster. Nous partons d'un modèle marginal, avec une fonction de risque de base et des coefficients de régressions communs aux clusters. Nous proposons un estimateur pour les paramètres du modèle et un estimateur de la fonction de risque base cumulé et nous démontrons que ces estimateurs sont consistants et asymptotiquement normaux. Nous en tirons un estimateur consistant de la matrice de variance covariance. Les estimateurs des paramètres de régression sont obtenus à partir d'un programme classique intégrant le modèle de Cox avec la possibilité d'ajouter des termes offset. Les estimateurs proposés sont étudiés par simulation et montrent, empiriquement, une plus grande efficacité que d'autres méthodes existantes. Nous appliquons ensuite notre méthode à une étude de mortalité chez des patients Canadiens dialysés.

D. E. Schaubel and G. Wei **29**
Double Inverse-Weighted Estimation of Cumulative Treatment Effects Under Nonproportional Hazards and Dependent Censoring

Dans les études médicales prenant en compte le temps jusqu'à événement, la non proportionnalité des risques et la non indépendance de la censure sont deux problèmes fréquemment rencontrés dans l'estimation de l'effet du traitement.
Une méthode classique pour prendre en compte les effets dépendant du temps du traitement est d'utiliser un modèle paramétrique prenant en compte cette dépendance dans le temps.
Les limites de cette approche comprennent la difficulté de vérifier l'adéquation de la forme de la fonction spécifiée, ainsi que le fait qu'en présence d'un effet du traitement qui varie avec le temps, les cliniciens sont en général intéressés par les effets cumulés du traitement plus que par ses effets instantanés. De plus, dans beaucoup d'applications, le temps de censure n'est pas indépendant du temps jusqu'à l'événement d'intérêt.
Nous proposons donc des méthodes permettant d'estimer l'effet cumulé du traitement en présence de risques non proportionnels et de censure non indépendante.
Trois mesures sont proposées, incluant le ratio des risques cumulés, le risque relatif, et la différence de moyenne d'espérance de vie.
Pour chacune de ces mesures, nous proposons un estimateur à double pondération inverse, construit d'abord en utilisant les probabilités inverses de pondération du traitement (IPTW) pour équilibrer les distributions des covariables spécifiques au traitement, puis en utilisant la probabilité inverse de pondération de la censure (IPCW) pour compenser la dépendance de la censure. Les estimateurs proposés se montrent cohérents et asymptotiquement normaux.
Nous étudions leurs propriétés sur des échantillons par des simulations. Les méthodes

présentées sont appliquées à la comparaison des attentes de transplantation rénale en fonction de la race.

R. B. Geskus **39**
Cause-Specific Cumulative Incidence Estimation and the Fine and Gray Model Under Both Left Truncation and Right Censoring

L'estimateur standard de la fonction d'incidence cumulée spécifique à la cause dans le cadre des risques compétitifs avec troncature à gauche et/ou censure à droite peut s'écrire deux façons équivalentes. Une sous la forme d'une fonction de répartition cumulative empirique pondérée et l'autre sous la forme d'un estimateur produit-limite de type Kaplan-Meier. Cette équivalence suggère un point de vue alternatif pour l'analyse des temps de survenue d'un événement avec prise en compte de la troncature à gauche et de la censure à droite : les individus qui sont encore à risque ou ont connu un événement concurrent plus tôt reçoivent des poids supplémentaires provenant des mécanismes de censure et de troncature. Par conséquent, l'inférence sur l'échelle cumulative peut être effectuée en utilisant des versions pondérées des procédures classiques. Ceci est valable aussi bien pour l'estimation de la fonction d'incidence cumulée spécifique à la cause que pour l'estimation des paramètres de régression du modèle à subdistributions de risque proportionnelles de Fine et Gray. Nous montrons que, avec une filtration appropriée, la propriété de martingale est valide et permet d'obtenir des résultats asymptotiques pour les subdistributions de risque proportionnelles de la même façon que pour le modèle à risques proportionnels classique de Cox. L'estimation de la fonction d'incidence cumulée spécifique à la cause et des paramètres de régression sur la subdistribution de risque peut se faire en utilisant un logiciel standard d'analyse de données de survie, si ce logiciel permet d'inclure des poids dépendant du temps. Nous proposons une implémentation de cette technique sous le logiciel R. Le modèle à subdistributions de risque proportionnelles est utilisé pour étudier l'effet de la période calendaire, considérée comme une covariable déterministe externe variant avec le temps et qui peut aussi être vue comme un cas particulier de troncature à gauche, sur la mortalité cumulée associée ou non au VIH/Sida.

X. Liao, D. M. Zucker, Y. Li, and D. Spiegelman **50**
Survival Analysis with Error-Prone Time-Varying Covariates: A Risk Set Calibration Approach

Les épidémiologistes s'intéressant aux problèmes liés au travail, à l'environnement ou à l'alimentation souhaitent très souvent estimer l'effet prospectif de variables d'exposition dépendant du temps, comme par exemple l'exposition cumulée ou la moyenne mise à jour d'expositions cumulées, sur des indicateurs liés à une maladie chronique, tels que l'incidence et la mortalité du cancer. Dans les études visant à valider l'exposition, il est évident que plusieurs des variables d'intérêt vont être mesurées avec des erreurs plus ou moins importantes. L'approche de calibration par régression ordinaire est approximativement valide et efficace pour la correction de l'erreur de mesure dans l'estimation du risque relatif à partir du modèle de Cox avec variables d'exposition ne dépendant pas du temps lorsque la maladie est rare, mais elle n'est pas adaptable au cas où les expositions varient avec le temps. En re-calibrant le modèle pour l'erreur de mesure dans chaque ensemble à risque, une méthode de régression par calibration de l'ensemble à risque est proposée. Un algorithme pour une estimation ponctuelle avec

correction de biais du risque relatif, utilisant l'approche RRC, est présenté, ainsi que l'estimation par la méthode « sandwich » de sa variance. L'accent est mis sur des méthodes applicables à l'étude principale/à la conception d'étude externe de validation, assez courant dans des applications importantes. Des études de simulation sous différentes hypothèses sur la modélisation de l'erreur de mesure ont été conduites, et montrent la validité et l'efficacité de la méthode pour des échantillons finis. La méthode a été appliquée à une étude sur le régime alimentaire et le cancer à partir de la cohorte de professionnels de santé de Harvard (HPFS).

J. Guedj, R. Thi´ebaut, and D. Commenges **59**
Joint Modeling of the Clinical Progression and of the Biomarkers' Dynamics Using a Mechanistic Model

Les modèles conjoints sont utilisés pour explorer rigoureusement le lien entre la dynamique de biomarqueurs et des événements cliniques. Dans le contexte de l'infection par le VIH, où les dynamiques multivariées de l'ARN VIH et des CD4 sont complexes, une approche mécanistique basée sur un système d'équations différentielles prend en compte naturellement la corrélation entre les marqueurs. En utilisant des données issues d'un essai clinique randomisé (DELTA) comparant une bithérapie d'antirétroviraux à une monothérapie, une approche par maximisation de vraisemblance complète est proposée pour explorer le lien entre l'évolution des biomarqueurs et du temps jusqu'à la survenue d'un événement clinique. Le rôle de chaque marqueur comme prédicteur indépendant de la progression clinique a été évalué. Nous montrons que la dynamique conjointe de la charge virale et des CD4 capture l'effet du traitement antirétroviral ; la dynamique des CD4 seule capture la plupart mais pas l'ensemble de l'effet du traitement.

G. Y. Yi, W. Liu, and L.Wu **67**
Simultaneous Inference and Bias Analysis for Longitudinal Data with Covariate Measurement Error and Missing Responses

Des données longitudinales sont fréquentes dans les études cliniques et elles sont généralement analysées par des modèles linéaires mixtes généralisés. De tels modèles nous permettent de prendre en compte différentes sources d'hétérogénéité dont celles inter et intra sujets. Les procédures d'inférence se compliquent grandement en cas d'observations manquantes ou d'erreurs de mesure. Dans la littérature, une très grande attention a été portée à la prise en compte soit d'observations manquantes, soit d'erreurs de mesure par des modèles à effets aléatoires. Cependant, relativement peu de travaux ont été consacrés à la prise en compte simultanée des deux situations. Le besoin existe de combler ce manque car les données longitudinales présentent souvent ces deux caractéristiques. Dans ce papier nos objectifs sont d'étudier l'impact simultané d'observations manquantes et d'erreurs de mesure de covariables sur les procédures d'inférence et de développer une méthode à la fois faisable sur le plan du calcul et en accord avec la théorie. Des simulations ont été réalisées pour évaluer la performance de cette méthode qui a été par ailleurs appliquée à un cas réel.

R. M. Hillary **76**
A New Method for Estimating Growth Transition Matrices

La grande majorité des modèles de population utilisent l'âge ou l'état et non la longueur mais il existe de nombreux cas où l'âge des animaux ne peut pas être déterminer raisonnablement ou précisément. Pour ces cas, les modèles basés sur la longueur représentent une alternative logique mais peu de travaux ont été réalisés pour développer et comparer différentes méthodes pour estimer les matrices de transition de croissance qui sont utilisés dans de tels modèles. Ce papier montre comment une approche bayésienne consistante pour l'estimation des paramètres de croissance et une nouvelle méthode pour construire les matrices de transitions de longueurs prend en compte les variations de croissance d'une manière claire et consistante et évite de potentiels choix subjectifs en utilisant des méthodes mieux établies. L'inclusion de l'incertitude de croissance dans les modèles d'évaluation de populations et les impacts potentiels sur les décisions sont aussi discutés..

L'analyse de classes latentes (LCA) et la régression de classes latentes (LCR) sont largement utilisées pour modéliser des réponses catégorielles multivariées dans les sciences sociales et les études biomédicales. Les analyses standard font l'hypothèse que les données de différents répondeurs sont mutuellement indépendantes, excluant l'application de ces méthodes aux données familiales et autres plans dans lesquels les participants sont groupés. Dans ce papier, nous considérons des modèles à classe latente multi-niveaux, dans lesquels les probabilités de mélange de sous-population sont traitées comme des effets aléatoires qui varient entre les grappes en accord avec une distribution de Dirichlet commune. Nous appliquons l'algorithme espérance-maximisation (EM) pour estimer le modèle par maximum de vraisemblance (ML). Cette approche marche bien, mais est numériquement intensive quand le nombre de classes ou la taille des grappes sont grands. Nous proposons une approche de maximisation de la vraisemblance par paires (MPL) via un algorithme EM modifié pour cette situation. Nous montrons aussi qu'une analyse de classes latentes simple, combinée avec des écart-types robustes, fournit une autre procédure consistante, robuste, mais moins inférentiellement efficace. Les études de simulation suggèrent que les trois méthodes fonctionnent bien dans les échantillons de taille finie, et que les estimations de MPL offrent souvent une précision comparable à celles de ML. Nous appliquons nos méthodes aux analyses de signes de comorbidité dans l'étude des désordres obsessionnels compulsifs. La structure des effets aléatoires de nos modèles a une interprétation plus directe que celles des méthodes concurrentes, et ainsi enrichit utilement les outils disponibles pour les analyses de classes et de données multi-niveaux.

Comme la plupart des ensembles de données géoréférencés sont de nature multivariée, et concernent des variables de types différents, les méthodes de cartographie spatiale doivent pouvoir prendre en compte de telles données. Les principales difficultés sont la prédiction des variables non gaussiennes et la modélisation des dépendances entre processus. Cet article présente une nouvelle approche bayésienne hiérarchique permettant

la modélisation simultanée de champs spatiaux dépendant gaussiens, ordinaux et de comptage. Cette approche est basée sur les modèles linéaires généralisés spatiaux. Nous utilisons une approche moyenne mobile pour modéliser la dépendance spatiale entre les processus. La méthode est d'abord validée par une étude de simulation. Nous y montrons que le modèle multivarié a de meilleures capacités prédictives que le modèle univarié. Puis nous appliquons le modèle hiérarchique spatial multivarié à un ensemble de données réelles recueillies en Guyane française pour la prévision des types de couches arables.

T. Tango, K. Takahashi, and K. Kohriyama **106**
A Space–Time Scan Statistic for Detecting Emerging Outbreaks

Comme une méthode analytique majeure pour la détection des épidémies, la statistique de balayage spatio-temporel de Kulldorff (2001) a été implémentée dans de nombreux systèmes de surveillance syndromique. Cependant, comme elle est basée sur des fenêtres circulaires dans l'espace, il est difficile de détecter correctement des clusters non circulaires. Takahashi *et al.* (2008) ont proposé une statistique de balayage spatio-temporel flexible qui permet de détecter des aires non circulaires. Il nous semble, cependant, que la détection de la plupart des clusters définis par ces statistiques de balayage spatio-temporel n'est pas la même que la détection d'épidémies localisées de maladies émergentes car elle compare le nombre de cas observés au nombre de cas attendus conditionnel.
Dans ce papier, nous proposons une nouvelle statistique de balayage spatio-temporel qui compare le nombre de cas observés aux nombre de cas attendus *non-conditionnel*, en prenant en compte une sur-dispersion temporelle et en implémentant un modèle épidémique pour capturer des épidémies de maladies émergentes localisées au moment opportun et correctement. Les modèles proposés sont illustrés à partir de données hebdomadaires de surveillance du nombre d'absents dans des écoles primaires à Kita-Kyushu shi, Japan, 2006.

D. I. Warton **116**
Regularized Sandwich Estimators for Analysis of High-Dimensional Data Using Generalized Estimating Equations

Nous proposons une méthodologie pour modifier l'estimation des équations généralisées (GEE) pour le test d'hypothèses sur des données définies dans de grandes dimensions, avec un intérêt tout particulier pour les données multidimensionnelles d'abondance en écologie, application importante dans des milliers d'études environnementales. De telles données sont typiquement des comptages caractérisés par une dimensionnalité élevée (dans le sens où la taille de la classe dépasse leur nombre, $n>K$) et par une sur-dispersion par rapport à la distribution poissonienne. Les méthodes habituelles GEE ne peuvent pas être utilisées dans cette situation essentiellement parce que les estimateurs à matrice de covariance robuste (ou « *estimateur sandwich* ») deviennent numériquement instables dès que n croît. Nous proposons d'utiliser à la place un estimateur sandwich qui implique une matrice de corrélation **R** commune et rétrécit l'estimation d'échantillonnage de **R** vers une matrice de corrélation de travail qui améliore sa stabilité numérique. Nous montrons théoriquement et par simulation que ceci améliore de façon substantielle la puissance de la statistique de Wald quand la taille de la classe n'est pas trop petite. Nous appliquons cette approche pour étudier l'effet d'addition de nutriments sur des communautés de

nématodes, et ce faisant nous discutons des aspects importants de son utilisation, comme celle de statistiques ayant de bonnes propriétés quand les estimations des paramètres se rapprochent de la frontière ($\mu_i=0$), et nous faisons du ré-échantillonnage pour fournir une inférence valable robuste pour une dimensionnalité élevée et pour de possibles mauvaises spécifications du modèle.

M. Jin and Y. Fang **124**
Variable Selection in Canonical Discriminant Analysis for Family Studies

Dans les études familiales, l'analyse discriminante canonique peut être utilisée pour identifier des combinaisons linéaires de phénotypes qui montrent des rapports élevés entre la variabilité inter-famille et la variabilité intra-famille. Toutefois, en présence d'un grand nombre de phénotypes, l'analyse discriminante canonique peut être source de sur ajustement. Pour estimer les rapports associés avec les coefficients obtenus avec l'analyse discriminante canonique, deux méthodes sont proposées ici : l'une est fondée sur la correction du biais et l'autre sur la validation croisée. Parce que la validation croisée est très gourmande en temps-calcul, une approximation de la validation croisée est également proposée. De plus, ces méthodes peuvent être utilisées pour la sélection de variable dans l'analyse discriminante canonique. Les méthodes proposées sont illustrées au travers d'études de simulation et appliquées à deux exemples réels.

Y.-Y. Ho, G. Parmigiani, T. A. Louis, and L. M. Cope **133**
Modeling Liquid Association

En 2002, Ker-Chau Li a proposé une mesure dite "d'association liquide" pour caractériser des interactions d'ordre 3 entre des mesures d'expression de gènes et a développé le calcul d'un estimateur efficace qui permet d'explorer ce genre d'interaction dans des données d'expression issues de puces à ARN. Cette étude, et d'autres publiées depuis, ont validé de manière biologique les résultats que peut fournir cette méthode et ont démontré qu'elle était un outil très utile pour analyser des données génomiques.
Dans le prolongement de cette méthode, nous avons cherché une famille paramétrique de distributions multivariées permettant de modéliser avec une grande flexibilité l'espace des possibles pour les corrélations d'ordre 3 qui peuvent être définies à partir d'une "association liquide". Un tel modèle permettrait de formaliser la méthode d'association liquide dans un contexte théorique d'inférence statistique.
Dans ce papier, nous décrivons une famille de distributions trivariées ayant pour distribution marginales des lois Gaussiennes et dont la loi trivariée Gaussienne est un cas particulier. La particularité la plus intéressante de cette distribution est que sa paramétrisation sépare la structure des dépendances d'ordre 3 en un nombre de composantes distinctes et interprétables.
Une de ces composantes s'apparentant beaucoup à l'association liquide est qualifiée de d'association liquide modifiée.
Nous proposons deux méthodes pour estimer cette quantité et des tests statistiques pour vérifier l'existence de ce type de dépendance. Une étude de simulations est réalisée pour évaluer les propriétés inférentielles de ces méthodes et nous illustrons leur applications à partir de données expérimentales publiques.

F. Yu, M.-H. Chen, L. Kuo, P. Huang, and W. Yang **142**

Bayesian Hierarchical Modeling and Selection of Differentially Expressed Genes for the EST Data

Le séquençage de marqueur de séquence exprimée ou Expressed Sequence Tag (EST) est une lecture d'un seul séquençage d'ADNc clonés provenant d'un certain tissu. La fréquence d'étiquettes uniques à partir de différentes banques d'ADNC non biaisés est utilisée pour inférer le niveau d'expression relatif de chaque étiquette. Dans ce papier, nous proposons un modèle hiérarchique multinomial avec un distribution a priori de Dirichlet non linéaire pour les données d'EST avec des banques multiples et des types de tissus multiples. Une nouvelle distribution a priori hiérarchique est développée et les propriétés de cette distribution a priori proposée sont examinées. Un algorithme de Monte-Carlo par chaînes de Markov efficace est développé pour exécuter le calcul de la distribution a posteriori. Nous proposons également un nouveau critère de sélection pour détecter les gènes différentiellement exprimés entre deux types de tissus. Nous démontrons à l'aide de plusieurs simulations que notre nouvelle méthode avec ce nouveau critère de sélection des gènes a des taux faibles de faux positifs et faux négatifs. Un jeu de données d'EST réel est utilisé pour motiver et illustrer la méthode proposée.

Le "ChIP-seq" est une méthode qui consiste à appliquer une méthode de séquençage à haut débit au produit obtenu par une expérience d'immunoprécipitation de la chromatine. Elle permet ainsi d'obtenir un profil à l'échelle du génome des interactions *in vivo* entre les facteurs de transcription et leurs séquences d'ADN cibles avec une meilleure sensibilité, spécificité et résolution spatiale que la technique ChIP-chip (basée sur l'utilisation d'une puce microarray). Elle pose en revanche de nouveaux problèmes statistiques liés à la complexité des systèmes biologiques étudiés et à la variabilité et aux biais des données de séquence obtenues. Nous proposons une méthode appelée PICS ("Probabilistic Inference for Chip-seq) pour identifier des régions de fixation de facteurs de transcription à partir de l'alignement des séquences obtenues. PICS identifies la localisation des sites de liaison en modélisant les concentrations locales des séquences et utilises l'information a priori sur la longueur des fragments d'ADN dans un modèle hiérarchique bayésien de mélanges de distribution de Student pour discriminer entre des événements de liaison adjacents. L'utilisation de cartographies de l'ensemble des séquences du génome pré-établies et d'une distribution de *t* tronquée permets de modéliser les séquences manquantes dues à la présence locales de régions répétées. L'incertitude des paramètres du modèle permet de définir des régions de confiance pour la location des sites de fixation et de proposer des filtres sur ces paramètres. Enfin, PICS calcule un score d'enrichissement par rapport à une échantillon contrôle et peut également obtenir une estimation du taux de fausses découvertes (FDR) à partir d'un échantillon contrôle. En utilisant des données publiques sur les facteurs de transcription GABP et FOXA11 obtenues à partir de lignées cellulaires humaines, nous montrons que la méthode PICS fournit des prédictions de site de fixation plus efficaces que les autres méthodes disponibles que sont MACS, QuESt, CisGenome et USeq. Enfin, une étude de simulation confirme les bonnes performances de PICS par rapport à ces méthodes alternatives et sa robustesse par rapport à une mauvaise spécification du modèle

statistique.

P. Wang, D. L. Chao, and L. Hsu **164**
Learning Oncogenic Pathways from Binary Genomic Instability Data

L'instabilité génomique, qui est la propension à des aberrations chromosomiques, joue un rôle critique dans le développement de plusieurs maladies. Des expériences de génotypage à haut débit ont été réalisées pour étudier l'instabilité génomique dans les maladies. Les résultats de telles expériences peuvent se résumer à des vecteurs binaires de grande dimension, dans lesquels chaque variable binaire représente le statut d'un locus marqueur vis-à-vis de l'aberration. Il est essentiel de comprendre comment les aberrations peuvent interagir entre elles, car cela aide à mieux comprendre le mode de développement de la maladie. Dans cet article, nous proposons une méthode nouvelle, ***LogitNet***, pour inférer de telles interactions entre les événements aberrants. La méthode se base sur une régression logistique pénalisée, et étendue à la prise en compte des corrélations spatiales des données d'instabilité génomique. Nous présentons des simulations à grande échelle et montrons que la méthode proposée est très performante dans les situations examinées. Enfin, nous illustrons la méthode sur des données d'instabilité génomique dans des échantillons de cancer du sein.

L. Finos and A. Farcomeni **174**
k-FWER Control without p-value Adjustment, with Application to Detection of Genetic Determinants of Multiple Sclerosis in Italian Twins

Nous présentons une nouvelle approche pour le contrôle de l’erreur d’ensemble de famille (*k*-FWER) qui n’implique aucune correction, mais seulement le test des hypothèses dans un certain ordre (pouvant être déterminé par les données) jusqu’à ce qu’un nombre adapté de *p*-valeurs supérieures au niveau non corrigé soient observées. Les *p*-valeurs peuvent provenir d’un quelconque modèle linéaire dans une représentation paramétrique ou non-paramétrique. Cette approche est non seulement très simple et peu consommatrice en calculs, mais de plus le recours possible à un ordre données-dépendant donne une meilleure puissance lorsque la taille d’échantillon est faible (ainsi que lorsque k et/ou le nombre de tests est élevé). Nous illustrons la méthode par une étude originale sur la recherche des gènes dans la sclérose multiple, dans laquelle on disposait d’un petit nombre de couples de jumeaux, distincts selon la maladie. Les méthodes sont implémentées dans un programme écrit en R (someKfwer), disponible librement sur le serveur CRAN.

J. Qin and K.-Y. Liang **182**
Hypothesis Testing in a Mixture Case–Control Model

Nous envisageons le problème posé par le test de proportion de mélange fait au travers de deux échantillons de données, l’un venant du groupe 1 et l’autre venant d’un mélange des groupes, un et deux, avec une proportion inconnue, □, du groupe deux. Plusieurs applications statistiques, telles que l’analyse de micropuces, les études d’épidémiologie infectieuse, les études cas-témoin avec des témoins contaminés, les essais cliniques prenant en compte la non-réponse, les études génétiques portant sur la mutation, ou encore les questions portant sur les pêches, peuvent se formuler dans ce cadre. Sous

l'hypothèse que le logarithme du ratio des densités de probabilité dans les deux groupes est linéaire par rapport aux observations, nous proposons une statistique généralisée de test du score pour tester la proportion dans le mélange. Avec quelques conditions de régularité, on montre que cette statistique converge vers un khi-deux pondéré sous l'hypothèse nulle = 0, le poids ne dépendant que de la fraction d'échantillonnage des groupes. Nous utilisons une méthode de permutation pour obtenir une approximation plus fiable pour échantillon fini. Nous présentons des résultats de simulation et des applications sur données.

H. Zhou, R. Song, Y. Wu, and J. Qin **194**
Statistical Inference for a Two-Stage Outcome-Dependent Sampling Design with a Continuous Outcome

Le schéma cas-témoin à deux étapes a été largement utilisé dans les études épidémiologiques en raison de son bon rapport coût-efficacité et de l'amélioration de l'efficacité. L'évolution des études biomédicales modernes a nécessité des schémas coût-efficaces avec une réponse continue et des variables d'exposition. Dans ce papier, nous proposons un nouveau schéma à deux étapes avec échantillonnage dépendant d'une réponse continue dans lequel à la fois les données de la première et de la deuxième étape proviennent de schémas d'échantillonnage dépendant de la réponse. Nous développons une estimation semi-paramétrique empirique de la vraisemblance pour l'inférence sur les paramètres de la régression. Des simulations ont été réalisées pour explorer le comportement de cette estimation pour de petits échantillons. Nous démontrons que, pour une puissance donnée, le schéma proposé nécessite un effectif nettement plus petit que les autres schémas. La méthode proposée est illustrée à partir d'une étude santé-environnement conduite par l'Institut National de la Santé.

E. Kulinskaya, M. B. Dollinger, and K. Bjørkestøl **203**
Testing for Homogeneity in Meta-Analysis I. The One-Parameter Case: Standardized Mean Difference

La méta-analyse cherche à combiner les résultats de plusieurs expériences de manière à améliorer la précision des décisions. Il est habituel d'employer un test d'homogénéité pour déterminer si les résultats de plusieurs expériences sont assez semblables pour justifier leur combinaison en un résultat commun. La statistique *Q* de Cochran est souvent utilisée pour effectuer ce test d'homogénéité. On suppose souvent que *Q* suit une distribution du *chi2* sous l'hypothèse nulle d'homogénéité, mais il est connu depuis bien longtemps que la distribution asymptotique de *Q* n'est pas précise pour des tailles modérées d'échantillon. Ici, nous présentons un développement pour la moyenne de *Q* sous l'hypothèse nulle qui est la valable quand l'effet et le poids de chaque étude dépend d'un seul paramètre, sans que l'on ait besoin ni de la Normalité ni de l'indépendance des estimateurs de l'effet et du poids. Ce développement représente une correction d'ordre *O(1/n)* du moment classique du *chi2* dans le cas d'un seul paramètre. Nous appliquons ce résultat au test d'homogénéité pour des méta-analyses où les effets sont mesurés par la différence moyenne standardisée (la statistique *d* de Cohen). Dans cette situation, nous recommandons d'approcher la distribution nulle de *Q* par une distribution du *chi2* avec des degrés de liberté fractionnaires estimés à partir des données en utilisant notre développement de la moyenne de *Q*. L'homogénéité qui en résulte est substantiellement

plus précise que celle du test habituel employé. Nous fournissons un programme disponible sur le site de *Biometrics* (http://www.biometrics.tibs.org) pour les calculs nécessaires.

C. Fan, D. Zhang, and C.-H. Zhang **213**
On Sample Size of the Kruskal–Wallis Test with Application to a Mouse Peritoneal Cavity Study

Généralisation non-paramétrique du modèle d'analyse de variance (ANOVA) à un facteur, le test de Kruskall-Wallis est valide pour tester la différence entre plusieurs échantillons, les populations sous-jacentes ayant des distributions non gaussiennes ou inconnues. Bien que le test de Kruskall-Wallis ait été largement utilisé, la puissance de ce test et la méthodologie de tailles d'échantillons nécessaires ont été beaucoup moins étudiées. Cet article présente de nouvelles méthodes de calcul de la puissance et des tailles d'échantillons nécessaires adaptées au test de Kruskall-Wallis, basées sur l'étude pilote, dans un contexte de modèle complètement non-paramétrique comme dans un contexte de modèle semi-paramétrique de position. Il n'est fait aucune hypothèse sur la forme des distributions des populations sous-jacentes. Des résultats de simulation montrent que, en termes de calculs de tailles d'échantillons, les méthodes proposées sont plus fiables et donc préférables aux méthodes plus traditionnelles. Une étude portant sur la cavité péritonéale de la souris est utilisée pour démontrer l'application de ces méthodes.

C. C. Drovandi and A. N. Pettitt **225**
Estimation of Parameters for Macroparasite Population Evolution Using Approximate Bayesian Computation

Nous estimons les paramètres d'un modèle de processus stochastique pour une population de macro-parasites à l'intérieur d'un hôte à l'aide du calcul bayésien approché (ABC). L'immunité de l'hôte est un variable latente du modèle et seuls les macro-parasites matures lors du sacrifice de l'hôte sont comptabilisés. Malgré un nombre de données très limité, les taux du processus sont estimés avec une précision raisonnable. Notre modèle comprend un processus de Markov à trois variables pour lequel la vraisemblance construite à partir des données réelles est incalculable. Les méthodes ABC sont particulièrement utiles lorsque la vraisemblance ne peut être calculée ni analytiquement ni informatiquement. L'algorithme ABC que nous proposons est basé sur une méthode de Monte Carlo séquentielle, est par construction adaptatif et surpasse certains inconvénients des approches précédentes de l'ABC. L'algorithme est d'abord validé par un test sur des données simulées à partir d'un modèle autologistique avant d'être utilisé pour estimer les paramètres du modèle de processus de Markov sur les données expérimentales. Le modèle estimé explique la variation extra-binomiale comme la présence éphémère chez l'hôte d'une variable binaire d'immunité.

R. S. McCrea and B. J. T. Morgan **234**
Multistate Mark–Recapture Model Selection Using Score Tests

L'importance des modèles de marquage-capture est reconnue, mais il manque une procédure simple de choix du meilleur modèle. Cet article propose et évalue une

procédure par étapes pour choisir des modèles de marquage-capture appropriés sur la base de statistiques de test. Seuls les modèles acceptés par les données ont besoin d'être ajustés, si bien que les structures de modèle trop compliquées, avec trop de paramètres, sont écartées. Dans les situations habituelles, seul un petit nombre de modèles sont retenus, et la procédure permet de repérer les modèles à paramètres redondants et proches de la redondance.
La valeur de la technique est démontrée par simulation, et la méthode est illustrée sur des données relatives à l'oie du Canada dans trois régions. Dans ce cas, la méthode conduit à un nouveau modèle beaucoup plus simple que le meilleur modèle considéré jusqu'à présent pour de telles données.

La régression quantile, qui modélise les quantiles conditionnels d'une variable réponse étant données des covariables, suppose habituellement un modèle linéaire. Cependant ce type de linéarité est souvent non réaliste. Une situation dans laquelle la régression quantile linéaire n'est pas appropriée est rencontrée quand la variable réponse est linéaire par morceaux mais toujours continue en les covariables. Pour analyser de telles données, nous proposons un modèle de régression quantile en ligne brisée. Nous déduisons les estimations des paramètres, démontrons qu'ils sont asymptotiquement valides étant donnée l'existence d'un point de rupture et discutons plusieurs méthodes pour tester l'existence d'un point de rupture dans une régression quantile en ligne brisée ainsi qu'une comparaison de puissance par simulation.
Un exemple de vitesses maximales de courses des animaux terrestres est donné pour illustrer une application de la régression quantile en ligne brisée pour laquelle ce modèle est justifié théoriquement et ses paramètres sont directement d'intérêt biologique.

Les études de la dynamique du VIH, basée sur des systèmes d'équations différentielles, ont amélioré de façon significative les connaissances sur l'infection par le VIH. Alors que les premières études utilisaient de simples modèles dynamiques à court terme, des travaux plus récents ont considéré des modèles complexes à plus long terme, combinés avec une analyse globale de l'ensemble des patients grâce à des modèles non linéaires mixtes, et améliorant ainsi la qualité de l'analyse de la dynamique du VIH. Vu la complexité du problème, des difficultés d'ordre statistique subsistent néanmoins. Nous avons proposé d'utiliser l'algorithme SAEM (approximation stochastique de EM), un algorithme puissant pour l'estimation par maximum de vraisemblance, pour analyser simultanément la décroissance des charges virales et l'augmentation des CD4 chez les patients traités, à partir d'un modèle de dynamique du VIH à long terme. Nous avons appliqué cette méthodologie à l'analyse de l'essai clinique COPHAR2-ANRS 111. De très bons résultats ont été obtenus avec un modèle défini par cinq équations

différentielles et intégrant des cellules CD4 latentes. Seul un des paramètres a été fixé alors que les dix autres (dont huit avec une variabilité inter-sujets) ont été estimés avec succès. Nous avons montré que l'efficacité du nelfinavir était plus réduite que celle de l'indinavir et du lopinavir.

Des modèles mixtes non linéaires ont récemment été proposés pour modéliser les données longitudinales complexes. Des covariables sont généralement introduites dans le modèle pour expliquer une part de la variation entre individus. Cependant, on suppose souvent que l'erreur résiduelle aléatoire et les effets aléatoires sont distribués normalement, ce qui peut parfois produire des résultats erronés lorsque leur distribution est en fait asymétrique. De plus, certaines covariables telles que les comptages de lymphocytes T-CD4 peuvent parfois faire l'objet d'erreurs de mesure significatives. Dans cet article nous traitons ces deux problèmes à la fois, en modélisant conjointement la distribution de la variable réponse et des covariables, par une approche bayésienne des modèles mixtes non linéaires avec erreurs de mesure sur les covariables et une distribution normale asymétrique. Nous illustrons ces méthodes avec un exemple réel, et nous comparons divers modèles possibles avec des distributions différentes. Nous montrons que les modèles faisant l'hypothèse d'une distribution normale asymétrique peuvent fournir des résultats crédibles en cas d'asymétrie dans les données. Les résultats de nos travaux peuvent être importants pour les études sur le SIDA/VIH, en fournissant une indication quantitative pour mieux comprendre les réponses virologiques aux traitements antirétroviraux.

Les modèles statistiques incluant des effets aléatoires sont couramment utilisés pour analyser des données longitudinales et des données corrélées, souvent avec l'hypothèse d'une distribution gaussienne pour les effets aléatoires. A travers des calculs théoriques et numériques et des simulations nous étudions l'impact d'une mauvaise spécification de cette distribution, aussi bien sur la façon dont les valeurs prédites correspondent à la véritable distribution sous-jacente, que sur la précision de la prédiction des effets aléatoires. Nous montrons que, bien que les valeurs prédites varient avec la distribution supposée, la précision de la prédiction, mesurée par l'erreur quadratique moyenne, est peu affectée pour des violations faibles à modérées des hypothèses. Par conséquent, les approches standard, aisément disponibles dans les logiciels statistiques, sont souvent suffisantes. Les résultats sont illustrés à partir de données de l'étude « Cœur et substitution oestrogènes/progestatifs » utilisant des modèles pour prédire les valeurs de pression artérielle.

En 2001, l'Office Américain d'Administration du Personnel a demandé tous les plans de santé participant au Programme de Bénéfices pour la Santé des Employés Fédéraux en vue d'offrir des bénéfices en termes de santé mentale et d'abus de drogues à parité avec les autres bénéfices médicaux. L'évaluation initiale a montré que, en moyenne, la parité ne créait ni de gros accroissements de dépenses ni d'augmentation de l'usage des services sur une période d'observation de quatre ans. Néanmoins certains groupes de recrutés pourraient avoir bénéficié de la parité plus que d'autres. Afin de traiter cette question nous proposons un modèle bayésien à classes latentes avec 2 parties pour caractériser l'effet de la parité sur l'usage des services et les coûts de la santé mentale. A l'intérieur de chaque classe, nous ajustons un modèle à effets aléatoires à 2 parties pour modéliser séparément la probabilité de l'usage de services de santé mentale ou d'abus de drogues et les trajectoires de coût moyen parmi ceux qui ont utilisé ces services. Les coefficients de la régression et les covariances des effets aléatoires varient entre les classes, permettant ainsi des structures de corrélation variables d'une classe à l'autre entre les deux composantes du modèle. Notre analyse a identifié trois classes de sujets : un groupe peu dépensier, plutôt masculin, faisant peu appel aux services, et dont les dépenses décroissent au cours du temps ; un groupe modérément dépensier, principalement féminin, qui a eu un accroissement à la fois de l'usage et des dépenses moyennes après l'introduction de la parité ; et un groupe très dépensier tendant à avoir un usage chronique du service et des schémas de dépense constants. En examinant les régions 95% de plus hautes densités jointes de changement attendu dans l'usage et la dépense de chaque classe, nous confirmons que la parité a eu un impact seulement dans la classe des dépensiers modérés.

J. A. Dupuis, F. Bled, and J. Joachim **290**
Estimating the Occupancy Rate of Spatially Rare or Hard to Detect Species: A Conditional 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 grandeur étant définie comme la proportion de quadrats occupés. Nous nous intéressons tout particulièrement aux espèce spatialement rares, ou difficilement détectables, qui sont typiquement détectées dans un très petit nombre de quadrats, et pour lesquelles estimer le taux d'occupation (avec une précision raisonnable) 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 que celui-ci rend identifiable les paramètres d'occurrence et de détectabilité, indépendamment du nombre de visites effectuées dans les quadrats échantillonnés. Comparée avec une approche inconditionnelle, elle s'avère être complémentaire. Deux analyses bayésiennes des données sont réalisées: l'une est non informative, et l'autre tire avantage de ce que de l'information a priori sur la détectabilité est disponible. Il ressort de cette étude que la prise en compte d'un tel a priori améliore de façon significative la précision de l'estimation quand l'espèce cible a été détectée dans peu de quadrats et qu'on sait par ailleurs qu'elle est facilement détectable.

J. Chen, J. Xie, and H. Li **299**

A Penalized Likelihood Approach for Bivariate Conditional Normal Models for Dynamic Co-expression Analysis

L'observation de gènes coexprimés a été largement utilisée dans l'analyse des puces d'expression. Cependant, les motifs de coexpression entre deux gènes peuvent être médiés par des états cellulaires comme le reflètent l'expression d'autres gènes, les polymorphismes mononucléotidiques et l'activité des protéines kinases. Dans ce manuscrit, nous proposons un modèle normal bivarié conditionnel pour identifier les variables qui peuvent médier ces motifs de coexpression entre deux gènes. Sur la base de ce modèle, nous proposons un test du rapport de vraisemblance et une procédure de vraisemblance pénalisée pour identifier les médiateurs qui influencent les motifs de coexpresions géniques. Nous proposons un algorithme informatique efficace fondé sur la méthode des moindres carrés itératifs repondérés et une descente de type « cyclic coordinate » . Nous montrons que lorsque le paramètre de réglage dans la vraisemblance pénalisée est convenablement sélectionné, une telle procédure a la propriété « de l'oracle » dans la capacité à sélectionner les variables. Nous présentons des résultats de simulations pour comparer notre méthode aux méthodes existantes et nous montrons que l'approche fondée sur le rapport de vraisemblance à des performances au moins égales que la méthode d'association liquide tandis que la procédure de vraisemblance pénalisée peut être assez performante dans la sélection et l'identification des médiateurs. Nous appliquons notre méthode à des données d'expression chez la levure afin d'identifier les kinases et/ou les polymorphismes mononucléotidiques qui influence les motifs de coexpression entre les gènes.

S. Haneuse and J. Chen **309**
A Multiphase Design Strategy for Dealing with Participation Bias

Une étude récemment financée sur l'impact de l'utilisation d'une contraception orale sur le risque de fracture osseuse utilise un plan de recrutement randomisé de Weinberg et Wacholder (1990). Le risque d'une potentielle complication dans l'étude d'une fracture osseuse provient des taux de réponse différentiels entre cas et témoins ; les taux de participation dans des études précédentes ont été d'environ 70%. Tandis que des données de plans de recrutement randomisés peuvent être analysées dans le cadre d'une étude à deux étapes, ne pas prendre en compte la participation différentielle potentielle peut mener à des estimations biaisées de l'association. Pour surmonter cet aspect, nous nous fondons sur la structure à deux phases et proposons une extension en introduisant une étape additionnelle de collectes des données qui vise spécifiquement la participation différentielle potentielle. Quatre estimateurs qui corrigent à la fois l'échantillonnage et le biais de participation sont proposés ; deux d'entre eux sont généraux et deux autres pour le cas particulier où les covariables sous-jacentes au mécanisme de participation sont discrètes. Puisque l'étude sur le risque de fracture est en cours, nous montrons les méthodes en utilisant des données de mortalité infantile de la Caroline du Nord.

READER REACTION

S. G. Baker **319**
Estimation and Inference for the Causal Effect of Receiving Treatment on a Multinomial Outcome: An Alternative Approach

Cheng (*Biometrics*, 2009) a récemment proposé un modèle de l'effet traitement pour les essais cliniques où la randomisation est parfaitement respectée dans un bras tandis que dans le second, certains patients ne reçoivent pas le traitement alloué mais l'autre (cette configuration est appelée « all-or-none compliance in one randomization group »). Son modèle est estimé par maximisation de la vraisemblance via un algorithme de programmation convexe. Nous présentons et discutons ici un modèle alternatif pour le cas plus général où ce sont des patients des deux groupes qui reçoivent l'autre traitement que celui qui leur avait été alloué (cas appelé « all-or-none compliance in two randomization groups ») : ce modèle est estimé via un ajustement exact (car le modèle est saturé), ou via un algorithme EM en cas de données de comptage. Cette approche nous semble plus simple à mettre en œuvre, facilitant ainsi la reproductibilité des calculs.