Translations of Abstracts

BIOMETRIC METHODOLOGY

Des progrès substantiels ont été accomplis dans l'identification de variantes génétiques simples prédisposant à des maladies complexes communes. Cependant, l'étiologie génétique des maladies humaines reste largement inconnue. Les maladies humaines complexes sont en général sous l'influence d'effets joints d'un grand nombre de variantes génétiques et non pas d'une seule. L'analyse jointe de multiples variantes génétiques en envisageant les déséquilibres de liaison (LD) et les interactions potentielles peut améliorer le processus de découverte, et conduire à l'identification de nouvelles variantes génétiques pouvant être associées à des maladies. Motivés par les développements de la statistique spatiale, nous proposons un nouveau modèle statistique basé sur la théorie des champs aléatoires, que nous désignerons comme les modèles de champ aléatoire génétique (GenRF), pour l'analyse jointe d'association considérant de possibles interactions gène-gène et les LD. En utilisant une approche de pseudo-vraisemblance, nous développons un test GenRF d'association jointe de plusieurs variantes génétiques, et qui présente les avantages suivants : 1. Performance améliorée en prenant en compte les interactions complexes ; 2. Réduction de dimension ; 3. Renforcement de la puissance en présence de LD ; 4. Efficience pour les calculs. Des études de simulation sont menées sous divers scénarios. Notre développement a été centré sur des traits quantitatifs, mais la robustesse du test GenRF pour d'autres traits, binaires par exemple, est abordée. Comparé à une approche type machine à noyau souvent adoptée, SKAT, aussi bien qu'à d'autres méthodes standard, GenRF présente des performances globalement comparables, et meilleures en présence d'interactions complexes. La méthode est illustrée par une application à l'étude « Dallas Heart ».

En recherche sur le cancer, des études de profilage ont été abondamment conduites pour détecter les gènes/SNP associés à un pronostic. Le cancer est complexe. L'examen des similarités et des différences des bases génétiques des multiples sous-types d'un même cancer peut conduire à une meilleure compréhension de ce qui les lie et de ce qui les distingue. Les méthodes de méta-analyse classiques analysent chaque sous-type séparément et comparent ensuite les résultats entre sous-types. Les méthodes d'analyses intégratives, au contraire, analysent les données brutes des sous-types multiples simultanément et peuvent se montrer supérieures aux méthodes par méta-analyse. Dans la présente étude, les données pronostiques de sous-types multiples d'un même cancer sont analysées. Un modèle en temps accéléré est adopté pour décrire la survie. La base génétique des sous-types multiples est décrite avec un modèle d'hétérogénéité, qui

permet l'association d'un gène/SNP au pronostic de certains sous-types mais pas de certains autres. Une méthode de pénalisation composée est développée pour identifier les gènes qui contiennent d'importants SNP associés au pronostic. La méthode proposée a une formulation intuitive et est réalisée au moyen d'un algorithme itératif. Les propriétés asymptotiques sont établies rigoureusement. Les simulations montrent que la méthode proposée a des performances satisfaisantes, meilleures qu'une méthode de méta-analyse à base pénalisée et qu'une méthode de seuillage régularisé. Une étude pronostique du lymphome non-Hodgkinien avec mesures de SNP est analysée. Les gènes associés à trois sous-types principaux, à savoir DLBCL, FL, et CLL/SLL, sont identifiés. La méthode proposée identifie des gènes qui sont différents des méthodes alternatives, et ont des implications importantes et une performance de prédiction satisfaisante.

R. A. Matsouaka, J. Li, and T. Cai **489**
Evaluating Marker-Guided Treatment Selection Strategies

Une voie potentielle pour améliorer les soins de santé est d'assujettir les stratégies de traitement individualisées aux informations sur les patients, telles que leur exposition à l'environnement ou leurs caractéristiques biologiques et génétiques. Dans les dernières années, sont apparues de nombreuses méthodes statistiques permettant la dérivation de règles individualisées de traitement (RIT). Mais avant d'adopter une RIT dans la pratique clinique, il est crucial d'évaluer sa capacité à améliorer les résultats pour les patients. Si l'on considère les méthodes existantes pour une telle évaluation, on s'aperçoit que soit elles ne tiennent compte que d'un seul marqueur, soit il s'agit de méthodes semi-paramétriques sujettes à des biais lorsque le modèle est mal spécifié. Cet article considère un cadre général avec des marqueurs multiples, cadre à l'intérieur duquel nous proposons une méthode robuste en deux étapes permettant la dérivation de RITs et leur évaluation. Nous proposons également des procédures comparant différentes RITs entre elles, procédures grâce auxquelles on peut évaluer l'intérêt d'ajouter des marqueurs supplémentaires pour améliorer le choix du traitement individualisé. Après les modèles de travail de la première étape, qui permettent d'approcher les RITs optimales, nous utilisons une étape d'étalonnage destinée à se protéger d'une mauvaise spécification du modèle. Nous évaluons ensuite l'efficacité de la RIT de façon non paramétrique, afin de contrôler la validité de l'inférence. Pour tenir compte de la variabilité, liée à l'échantillonnage, des règles obtenues et de leurs propriétés, nous proposons une procédure de rééchantillonnage afin d'avoir des intervalles de confiance valides, concernant aussi bien les propriétés des RITs que l'intérêt des marqueurs supplémentaires envisagés. Ces méthodes sont évaluées au moyen de très nombreuses simulations, et illustrées sur les données d'un essai clinique étudiant les effets de deux associations de médicaments chez des patients infectés par le VIH-1.

A. Sjölander, W. Lee, H. Källberg, and Y. Pawitan **500**
Bounds on Causal Interactions for Binary Outcomes

Un objectif usuel de la recherche en épidémiologie est d'étudier comment deux expositions interagissent dans la causalité d'une issue binaire. L'interaction causale est définie comme la présence de sujets pour lesquels l'effet causal d'une des expositions dépend du niveau de l'autre exposition. Pour des expositions binaires, il a été précédemment montré que la présence d'une interaction causale peut être testée au

moyen d'une interaction statistique additive. Cependant, il a été également montré que l'amplitude de l'interaction causale, définie comme la proportion de sujets pour lesquels existe une interaction causale, n'est généralement pas identifiable. Dans cet article, nous obtenons des bornes pour les interactions causales, applicables à des issues binaires et des expositions catégorisées avec un nombre de niveau quelconque. Ces bornes peuvent être utilisées pour obtenir l'amplitude de l'interaction causale, et sont un complément important au test statistique fréquemment utilisé. Les bornes sont obtenues aussi bien avec que sans l'hypothèse d'effets monotones d'exposition. Nous présentons une application de ces bornes à l'étude d'une interaction gène-gène dans le cas de l'arthrite rhumatoïde.

L'identification causale des effets de paires (aussi appelé contagion sociale ou induction) à partir de données observées dans les réseaux sociaux est sensible à deux sources distinctes de biais: l'homophilie latente et les facteurs de confusion non observés.
Dans cet article, nous examinons comment identifier des effets de paires sur des traits et des comportements en utilisant des gènes (ou d'autres variables structurellement isomorphes) comme variables instrumentales (VI) dans un grand jeu de données à travers des modèles intégrant homophilie et facteurs de confusion.
Nous utilisons des graphes acycliques orientés pour représenter ces modèles, nous employons plusieurs stratégies de VI et nous décrivons nos trois résultats principaux.
Tout d'abord, utiliser un seul gène fixe (ou allèle) comme VI échouera généralement à identifier les effets de paires si le gène affecte les valeurs passées du traitement.
Deuxièmement, utiliser plusieurs gènes fixes/allèles ou, de façon plus prometteuse, l'expression des gènes au cours du temps peut identifier les effets de paires si les violations des conditions d'exclusion ainsi que le traitement focal sont instrumentalisés.
Troisièmement, nous montrons que l'identification des effets de paires reste possible même sous de multiples complications souvent considérées comme rédhibitoires pour l'identification par VI des effets intra-individuels, comme la pléiotropie sur les variables observables et non observables, l'homophilie sur le phénotype passé, l'homophilie passée et en cours sur le génotype, les effets de pairs inter-phénotype, la stratification de population, l'expression de génique endogène au phénotype passé et à l'expression génique passée, et d'autres.
Nous appliquons notre approche à l'estimation des effets de pairs sur l'indice de masse corporelle (IMC) chez les amis et les conjoints dans la Framingham Heart Study. Les résultats suggèrent un effet de causalité positif de l'IMC entre amis.

Beaucoup de techniques d'analyse de données fonctionnelle demande de choisir une mesure de distance entre fonctions, la plus communément choisie étant la distance L^2. Dans cet article nous montrons que choisir une distance L^2 pondérée, avec une fonction de pondération judicieusement choisie, peut améliorer les performances de méthodes statistiques variées pour des données fonctionnelles, y compris les partitionnements à *k*-

médoïdes, la classification non-paramétrique, et les tests de permutation. En supposant une représentation de base quadratiquement pénalisée (par exemple spline) pour des données fonctionnelles, nous considérons trois fonctions de pondération non triviales : plans de densité pondérée, poids d'inverse de variance, et une nouvelle fonction de pondération qui minimise le coefficient de variation du carré de la distance résultante au moyen d'une procédure itérative. Les avantages de la pondération, en particulier avec la fonction de pondération proposée, sont démontrés à la fois par des études de simulation et en application aux données de croissance de Berkeley ainsi qu'à un ensemble de données fonctionnelles obtenues par résonnance magnétique.

D. Gervini and P. A. Carter **526**
Warped Functional Analysis of Variance

Cet article présente un modèle d'analyse de variance pour des données fonctionnelles qui incorporent explicitement une variabilité de phase au travers d'une composante de déformation temporelle, permettant ainsi une approche unifiée à l'estimation et à l'inférence en présence de variabilité temporelle et en amplitude. L'intérêt est porté sur des modèles à un unique facteur aléatoire, mais l'approche peut être facilement généralisée à des modèles ANOVA plus complexes. Le comportement des estimateurs est étudié par simulation, et une application à l'analyse de courbes de croissance de scarabées de farine est présentée. Bien que le modèle suppose les trajectoires observées provenant d'un processus latent régulier, la régularité des données observées n'est pas nécessaire ; la méthode peut être appliquée à des maillages temporels irréguliers, ce qui est fréquent dans les études longitudinales.

S. Jung and X. Qiao **536**
A Statistical Approach to Set Classification by Feature Selection with Applications to Classification of Histopathology

Des problèmes de classification d'ensembles surviennent quand des tâches de classification sont fondées sur des ensembles d'observations et non sur des observations individuelles. Dans la classification d'ensembles, une règle de classification est formée de N séries d'observations, où chaque série est labélisée avec l'information de la classe. La prédiction du label de la classe est également réalisée avec un ensemble d'observations. On rencontre des séries de données pour la classification d'ensembles, par exemple, dans le cas du diagnostic d'une maladie fondé sur plusieurs images du noyau cellulaire d'un seul tissu. Des modèles statistiques pertinents pour la classification d'ensembles sont introduits, qui justifient un cadre pour la classification d'ensembles fondé sur l'extraction de caractéristiques indépendantes du contexte. En considérant un ensemble d'observations comme une distribution empirique, nous utilisons une méthode fondée sur les données pour choisir ces caractéristiques qui contiennent l'information sur la localisation et la variation majeure. En particulier, la méthode d'analyse en composantes principales est utilisée pour extraire les caractéristiques qui expliquent la variation majeure. L'analyse multidimensionnelle est utilisée pour représenter les caractéristiques comme des points à valeurs vectorielles sur lesquels des classificateurs classiques peuvent être appliqués. Les approches par classification d'ensembles proposées atteignent des résultats de classification meilleurs que les méthodes concurrentes dans un certain nombre d'exemples de données simulées. Les avantages de

notre méthode sont démontrés dans une analyse d'images histopathologiques de noyaux cellulaires liées au cancer du foie.

S. J. Shin, Y. Wu, H. H. Zhang, and Y. Liu **546**
Probability-Enhanced Sufficient Dimension Reduction for Binary Classification

Dans l'analyse de données de grande dimension, il est particulièrement intéressant de réduire la dimensionnalité sans perte d'information. La réduction suffisante de dimension (SDR) survient dans ce contexte et de nombreuses méthodes SDR couronnées de succès ont été développées depuis l'introduction de la régression inverse par tranches (SIR ; Li, 1991). En dépit de leur progrès rapides, la plupart des méthodes existantes ciblent des problèmes de régression avec une réponse continue. Pour des problèmes de classification binaire, SIR souffre de limites d'estimation à plus d'une direction puisque seulement deux tranches sont disponibles. Dans cet article, nous développons une méthode SDR améliorée probabiliste qui est nouvelle et flexible pour des problèmes de classification binaire en utilisant une machine à vecteurs de support pondérés (WSVM). L'idée clef est de trancher les données en se basant sur des probabilités d'observations de classes conditionnelles plutôt que sur des réponses binaires. Nous montrons d'abord que le sous-espace central basé sur la probabilité conditionnelle de classe est le même que celui basé sur la réponse binaire. Ce résultat important justifie le schéma de tranchage proposé à partir d'une perspective théorique et qui assure une non-perte d'information. En pratique, la vraie probabilité conditionnelle de classe n'est généralement pas disponible et le problème de l'estimation de la probabilité peut être difficile avec des données ayant des entrées de grandes dimensions. Nous observons qu'afin d'implémenter le nouveau schéma de tranchage, nous n'avons pas besoin des valeurs exactes de probabilités et la seule information nécessaire est l'ordre relative des valeurs de probabilités. Motivé par ce fait, notre procédure de nouvelle SDR évite d'utiliser l'étape d'estimation de la probabilité et emploie directement la WSVM pour estimer directement l'ordre des valeurs de probabilités, basé sur la manière de réaliser le tranchage. La performance du schéma SDR amélioré probabiliste proposé est évalué à la fois par des données simulées et des données d'exemples réels.

G. P. Levin, S. C. Emerson, and S. S. Emerson **556**
An Evaluation of Inferential Procedures for Adaptive Clinical Trial Designs with Pre-specified Rules for Modifying the Sample Size

De nombreux articles ont introduit pour les essais cliniques adaptatifs des méthodes qui permettent des modifications de la taille d'échantillon en fonction d'estimations intermédiaires de l'effet traitement. Il y a eu une discussion extensive sur le contrôle de l'erreur de type 1 et sur des considérations d'efficacité, mais peu de recherches sur l'estimation après un test d'hypothèse adaptatif. Nous évaluons la fiabilité et la précision de différentes procédures d'inférence en présence d'un dispositif adaptatif avec des règles pré-spécifiées pour modifier le plan d'échantillonnage. Nous étendons les classements séquentiels par groupe de l'espace des résultats sur la base de l'étape d'arrêt, du rapport de vraisemblance et de la moyenne observée dans le cadre adaptatif pour calculer des estimations ponctuelles non biaisées de la médiane, des intervalles de confiance exacts, et des *p*-values distribuées uniformément sous l'hypothèse nulle. Le classement par rapport de vraisemblance s'avère produire des intervalles de confiance plus courts en moyenne et

de plus fortes probabilités pour les *p*-values de tomber en dessous des seuils importants que les approches alternatives. La moyenne ajustée pour le biais conduit à la plus petite erreur quadratique moyenne parmi les estimations ponctuelles candidates. Une approche basée sur l'erreur conditionnelle a dans la littérature l'avantage d'être la seule méthode qui s'accommode d'adaptation non planifiées. Nous comparons les performances de cette méthode et d'autres, afin de quantifier le coût d'une absence de planification dans un contexte où des adaptations pourraient être pré-spécifiées de façon réaliste. Nous trouvons que ce coût est notable pour tous les dispositifs et les effets de traitement envisagés, et substantiel pour les dispositifs proposés dans la littérature.

Y. Yang, M. E. Halloran, Y. Chen, and E. Kenah **568**
A Pathway EM-Algorithm for Estimating Vaccine Efficacy with a Non-Monotone Validation Set

Nous modélisons les délais de survenue de plusieurs types d'événements qui ne peuvent être distingués qu'après confirmation, par exemple après une analyse de laboratoire. L'événement d'intérêt ne peut se produire qu'une seule fois. Les autres types d'événements sont potentiellement récurrents. Une portion d'événements dont le type est identifié constitue un ensemble de validation. Cependant, même si les événements d'un sous-ensemble aléatoire sont identifiés, les confirmations manquantes peuvent présenter une structure non monotone, d'où une incertitude sur l'appartenance ou non d'un individu à l'ensemble des sujets à risque.
Par exemple dans une étude sur l'efficacité d'un vaccin anti-grippal, un sujet peut présenter plusieurs épisodes d'infections respiratoires provoqués par divers pathogènes, mais souvent la cause de l'infection n'est recherchée au laboratoire que pour un nombre limité d'épisodes. Nous proposons deux nouvelles méthodes pour estimer l'effet de covariables sur des délais de survie de ce type et donc pour mesurer l'efficacité du vaccin. La première est un algorithme Espérance-Maximisation (EM) qui envisage toutes les séquences d'événements compatibles avec les confirmations disponibles. A chaque estimation, l'algorithme estime les risques de base et les utilise pour pondérer les différents types d'événements. La seconde, non-itérative, est une méthode de validation par morceaux des séquences qui ne nécessite pas d'estimer les risques de base. Ces méthodes sont comparées à une approche antérieure, plus simple. Une étude par simulations suggère que l'erreur quadratique moyenne sur l'estimation de l'efficacité est plus faible quand les risques de base sont estimés, en particulier lorsque ces risques de base sont élevés. Nous utilisons l'algorithme EM pour réestimer l'efficacité d'un vaccin anti-grippal en 2003-2004 à Temple-Belton, Texas, et nous comparons nos résultats à ceux d'une analyse des mêmes données, publiée antérieurement.

B. Liu, W. Lu, and J. Zhang **579**
Accelerated Intensity Frailty Model for Recurrent Events Data

Dans cet article, nous proposons un modèle de fragilité à intensité accélérée pour des données de type évènements récurrents et nous dérivons un test pour la variance de fragilité. De plus, nous développons un algorithme EM basé sur une méthode de lissage par noyaux pour estimer les coefficients de régression et la fonction d'intensité de référence. La variance de l'estimateur résultant pour les paramètres de régression est obtenue par une méthode de différenciation numérique. Des études de simulation sont

menées pour évaluer sous diverses conditions la performance de l'estimateur proposé sur des échantillons de taille fini et démontrer le gain d'efficacité par rapport à l'estimateur des rangs de Gehan basé sur le modèle de l'AFT par processus de comptage (Lin et al. 1998, Biometrika 85, 605-618). Notre méthode est ensuite illustrée par une application sur des données de récidive de tumeur de la vessie.

S. Choi and X. Huang **588**
Maximum Likelihood Estimation of Semiparametric Mixture Component Models for Competing Risks Data

En analyse des risques compétitifs, la fonction d'incidence cumulée est utile pour caractériser les probabilités brutes de survenue de chaque type d'événement. Dans cet article, nous mettons en œuvre une analyse semi-paramétrique efficace de modèles de mélange appliqués à des fonctions d'incidence cumulée. Nous effectuons alors des régressions de la survie latente associée à chaque type d'événement, dans le cadre d'une classe de modèles semi-paramétriques qui englobe le modèle à hasards proportionnels et le modèle à risques proportionnels, lesquels permettent la prise en compte de covariables dépendant du temps. Les proportions marginales des occurrences des événements liés à chaque cause sont évaluées par un modèle logistique multinomial. Notre modèle de mélange présente l'avantage de permettre l'estimation conjointe de tous les paramètres associés aux risques compétitifs étudiés°; il satisfait aussi à la contrainte qu'à l'infini, la somme des probabilités liées à chaque événement vaut 1, quelles que soient les valeurs des covariables. Nous développons une nouvelle approche du maximum de vraisemblance, basée sur une régression semi-paramétrique facilitant l'obtention d'estimateurs efficaces et fiables dont nous démontrons la consistance et la normalité asymptotique. Des inférences statistiques peuvent être réalisées aisément en utilisant l'inverse de la matrice d'information observée. Nous établissons enfin, par simulations, les bonnes propriétés de notre méthode lorsqu'on l'applique à de petits échantillons. Nous illustrons notre approche par le traitement de données issues d'une étude sur le lymphome folliculaire.

H. Lin, L. Zhou, C. Li, and Y. Li **599**
Semiparametric Transformation Models for Semicompeting Survival Data

Des critères de jugement avec risques semi-concurrents (par exemple délai avant progression de la maladie et délai de survie) sont généralement recueillis dans les essais cliniques. Cependant, le manque d'outils statistiques disponibles gêne souvent l'analyse de telles données. Nous proposons donc un modèle de transformation semi-paramétrique innovant qui améliore les modèles existants des deux manières suivantes. Premièrement, ce modèle estime les coefficients de régression et les paramètres d'association simultanément. Deuxièmement, on peut obtenir directement une mesure de la valeur de substitution, par exemple la proportion de l'effet du traitement médié par le critère substitutif et le rapport de l'effet global du traitement sur le vrai critère sur l'effet du traitement sur le critère substitutif. Nous proposons une procédure d'estimation pour l'inférence et montrons que l'estimateur proposé est consistent et asymptotiquement normal. Des simulations complètes démontrent l'utilisation valide de notre méthode. Nous appliquons la méthode à un essai sur le myélome multiple pour étudier l'impact de plusieurs biomarqueurs sur les réponses semi-concurrentes des patients, à savoir le délai

avant progression et le délai avant décès.

Y.-J. Cheng and C.-Y. Huang **608**
Combined Estimating Equation Approaches for Semiparametric Transformation Models with Length-Biased Survival Data

Les données de survies sont sujettes à l'échantillonnage avec biais de longueur quand les durées de survie sont tronquées à gauche et que la variable aléatoire sous-jacente du délai de troncature est de distribution uniforme. Des gains substantiels en efficacité peuvent être obtenus en incorporant l'information sur la distribution du temps de troncature dans la procédure d'estimation (Wang, 1989, 1996). Sous les modèles de transformation semi-paramétriques, on s'attend à ce que la méthode du maximum de vraisemblance soit pleinement efficace ; cependant elle est difficile à implémenter car la vraisemblance complète dépend de la composante non paramétrique d'une façon compliquée. De plus, ses propriétés asymptotiques n'ont pas été établies. Dans ce papier, nous étendons l'approche par équations d'estimation avec des martingales (Chen et al., 2002 ; Kim et al., 2013) et l'approche par vraisemblance pseudo-partielle (Severini & Wong, 1992 ; Zucker, 2005) pour des modèles de transformation semi-paramétriques avec des données censurées à droite pour tenir compte des données tronquées à gauche et censurées à droite. Dans le même esprit que la méthode de vraisemblance composite (Huang & Qin, 2012), nous construisons plus avant un autre ensemble d'équations d'estimation sans biais en exploitant la structure de probabilité particulière de l'échantillonnage avec biais de longueur. Ainsi le nombre d'équations d'estimation excède le nombre de paramètres et des gains d'efficacité peuvent être obtenus en résolvant une simple combinaison de ces équations d'estimation. Les méthodes proposées sont faciles à implémenter car elles ne requièrent pas d'efforts de programmation supplémentaires. De plus, on montre qu'elles sont consistantes et de distribution asymptotique normale. L'analyse de données d'une étude sur la démence illustre ces méthodes.

T. Saegusa, C. Di, and Y. Q. Chen **619**
Hypothesis Testing for an Extended Cox Model with Time-Varying Coefficients

Le test du log-rank a été largement utilisé pour tester les effets des traitements sous le modèle de Cox pour l'analyse des temps de survenue d'événements en présence de censure, mais il peut y avoir une perte de puissance substantielle lorsque l'hypothèse de proportionnalité des risques n'est pas respectée. Dans ce papier, nous considérons une extension du modèle de Cox qui utilise des B-splines ou des splines de lissage pour modéliser un effet du traitement variable au cours du temps et proposer des statistiques du test de score pour l'effet du traitement. Les nouveaux tests proposés combinent la preuve statistique à la fois sur l'amplitude et sur la forme de la fonction de rapport des risques dépendante du temps, et sont donc omnibus et puissantes contre différents types d'alternatives. En outre, le nouveau cadre de test est applicable à tout choix de fonctions de bases splines, incluant les B-splines et les splines de lissage. Des études de simulation confirment que les tests proposés ont une bonne performance pour les échantillons finis et étaient souvent plus puissants dans de nombreux contextes que les tests conventionnels seuls. Les nouvelles méthodes ont été appliquées à l'étude HIVNET 012, un essai clinique randomisé menée par le HIV Prevention Trial Network dont l'objectif était d'évaluer l'efficacité d'une dose unique de névirapine contre la transmission mère-enfant

du VIH.

D. Concordet and R. Servien **629**
Individual Prediction Regions for Multivariate Longitudinal Data with Small Samples

Le suivi des patients est de plus en plus utilisé en médecine classique ou pour le contrôle antidopage afin d'identifier des résultats anormaux. Habituellement, ces suivis sont réalisés variable par variable en utilisant des « intervalles de référence » qui contiennent les valeurs observées sur 100(1-α)% des individus sains ou non dopés. Les observations faites sur l'évolution des variables au cours du temps sur un échantillon de *N* individus sains ou non dopés permettent à ces intervalles d'être individualisés en prenant en compte les possibles effets de covariables et d'éventuelles observations de ces variables sur l'individu lorsqu'il était sain ou non dopé. Pour chacune des variables d'intérêt ces intervalles individualisés doivent contenir 100(1-α)% des valeurs observées et compatibles avec les précédentes observations sur cet individu. Des méthodes générales sont disponibles afin de construire ces intervalles, mais elles permettent seulement de réaliser le suivi variable par variable quelque soient les possibles corrélations entre elles. Dans cet article, nous proposons une méthode générale permettant de suivre de manière simultanée plusieurs variables corrélées entre elles. Cette méthodologie se base sur un modèle linéaire multivarié à effets mixtes. Tout d'abord, nous exposons une méthode pour estimer les paramètres du modèle. Ensuite, nous obtenons une région de prédiction individualisée de couverture (1-α) dans un cadre asymptotique (*N* suffisamment grand). Néanmoins, la taille de l'échantillon *N* n'est parfois pas assez grande pour donner une région de prédiction suffisamment précise dans le cadre asymptotique. Pour cette raison nous proposons et comparons trois différentes régions de prédiction qui se comportent mieux pour des petites valeurs de *N*. Enfin, toute la méthode est illustrée par un exemple sur le suivi d'insuffisance rénale chez le chat.

J. Forkman and H.-P. Piepho **639**
Parametric Bootstrap Methods for Testing Multiplicative Terms in GGE and AMMI Models

Le modèle (CGE) à effets principaux génotypiques et à effets d’interaction génotype par environnement, et le modèle (AMMI) à effets principaux additifs et à interaction multiplicative sont les deux modèles les plus courants pour l’analyse de données génotype-environnement. Ces modèles sont fréquemment utilisés par les agronomes, les éleveurs de plantes, les généticiens et les statisticiens pour l’analyse d’essais en environnement multiple. Dans de tels essais, un ensemble de génotypes, tels que les cultivars végétaux, sont comparés au travers d’une gamme d’environnements, c’est-à-dire de lieux. Les modèles CGE et AMMI utilisent la décomposition aux valeurs singulières pour partitionner l’interaction génotype par environnement en une somme ordonnée de termes multiplicatifs. Cet article traite le problème du test de signification de ces termes multiplicatifs afin de choisir combien de termes sont à retenir dans le modèle final. Nous proposons des méthodes de bootstrap paramétrique pour ce problème. Nous envisageons les modèles à effets principaux fixes, à termes multiplicatifs fixes, et distribution normale des erreurs. Deux méthodes sont obtenues : une complète, et une à simple bootstrap paramétrique. Celles-ci sont comparées avec l’utilisation de tests F approchés avec validation croisée. Dans une étude de simulation basée sur quatre essais en

environnement multiple, les méthodes de bootstrap ont un bon comportement vis-à-vis du taux d'erreur de type I et de la puissance. La méthode à simple bootstrap paramétrique est particulièrement facile à utiliser puisqu'elle requiert simplement l'échantillonnage répétitif de valeurs de distribution normale centrée réduite. Cette méthode est recommandée pour choisir le nombre de termes multiplicatifs dans les modèles CGE et AMMI. La méthode proposée peut aussi être utilisée pour tester les composantes en analyse en composantes principales.

On combine couramment les données de diverses sources spatiales pour étudier les relations entre une réponse d'intérêt et les variables susceptibles d'y être associées. Les unités géographiques d'où proviennent ces données, coordonnées du point d'observation ou éléments de découpage administratif sont souvent décalées, c'est-à-dire que les observations sont effectuées en des points ou selon des découpages distincts. En conséquence, on est souvent amené à prédire la valeur de la covariable d'intérêt là où la réponse est observée. L'incertitude sur les valeurs prédites doit être prise en compte dans l'analyse des données alignées. Nous envisageons ici le cas où le krigeage est utilisé pour aligner des données de type point-point ou point-zone quand la variable réponse est non-gaussienne. Si un modèle linéaire généralisé est utilisé pour modéliser la relation entre covariable et réponse, l'incertitude supplémentaire introduite par le krigeage de la covariable introduit une erreur de mesure de type Berkson. Nous proposons de maximiser une quasi-vraisemblance pseudo-pénalisée pour prendre en compte cette incertitude additionnelle dans l'estimation des paramètres de régression et de la précision des estimateurs. Nous appliquons cette méthode à un exemple de type point-point, la relation entre la pollution par les particules fines ($PM_{2.5}$) et le risque de petit poids de naissance pour un nouveau-né à terme, après le plus grand incendie qu'ait connu la Floride. Pour illustrer les problèmes d'alignement de type point-zone, nous modélisons la relation entre les hospitalisations pour asthme dans les différents comtés de Floride et les mêmes mesures de pollution par particules fines. Nous évaluons finalement la méthode à l'aide de simulations. Nos résultats montrent que la méthode fournit des pourcentages de recouvrement satisfaisants alors que les approches naïves ignorant l'incertitude additionnelle tendent à sous-estimer la variabilité des estimateurs. La sous-estimation est particulièrement sensible dans les modèles de régression de Poisson.

Les données spatialement agglomérées font référence à des mesures corrélées en grandes dimensions collectées sur des unités ou sujets qui sont structurées spatialement. On rencontre souvent de telles données dans les études issues des sciences sociales ou de la santé. Nous proposons un cadre de modélisation unifié, appelé GeoCopula, pour caractériser les variations à petite et large échelle pour des données de types variés, y compris les données continues, binaires et de comptage. Pour relever les défis associés à l'estimation et l'inférence sur les paramètres du modèle, nous proposons une approche de vraisemblance composite efficace en ce sens que l'efficacité d'estimation résulte de la

construction des équations jointes d'estimation composites sur-identifiées. En conséquence, la théorie statistique pour l'estimation est développée en étendant la théorie classique de la méthode des moments généralisée. Un avantage clair de la méthode d'estimation proposée est sa faisabilité en temps de calcul. Nous menons plusieurs exercices de simulation pour vérifier les performances des méthodes d'estimation et des modèles proposés sur des données gaussiennes et binaires agglomérées spatialement. Les résultats montrent une amélioration claire sur l'efficacité d'estimation par rapport à la méthode de vraisemblance composite conventionnelle. Un jeu de données est analysé pour motiver et démontrer la méthode proposée.

Il est difficile d'estimer précisément la richesse spécifique s'il y a beaucoup d'espèces presque indétectables dans une communauté à forte diversité. En pratique, une borne inférieure précise est préférable à un estimateur ponctuel imprécis. La borne inférieure non-paramétrique traditionnelle développée par Chao (1984) pour des données d'abondance à l'échelle individuelle utilise seulement l'information sur les espèces les plus rares (le nombre de singletons et doubletons) pour estimer le nombre d'espèces indétectables dans les échantillons. En appliquant une formule modifiée de fréquence de Good-Turing, nous calculons une formule approchée pour le biais de premier ordre de cette borne inférieure traditionnelle. Le biais approximatif est estimé en utilisant des informations additionnelles (à savoir le nombre de triplets and quadruplets). Ce biais approximatif peut être corrigé, et une borne inférieure améliorée est proposée. LA borne inférieure proposée est non-paramétrique dans le sens qu'elle est valide universellement pour toute distribution d'abondance d'espèce. Une version similaire de borne inférieure améliorée peut être obtenue pour des données d'incidence. Nous testons nos bornes inférieures proposées sur des jeux de données simulés générés à partir de modèles variés d'abondance d'espèces. Les simulations montrent que les bornes proposées réduisent toujours le biais par rapport aux bornes traditionnelles et améliore la précision (mesurée par l'erreur quadratique moyenne) quand l'hétérogénéité des abondances d'espèces est relativement forte. Nous appliquons aussi les nouvelles bornes inférieures proposées à des données réelles pour illustration et comparaison à des estimateurs développés précédemment.

La phylogéographie étudie le processus historique qui est responsable de la répartition géographique des populations contemporaines dans une espèce. Les analyses sont réalisées à partir des séquences moléculaires obtenues sur un échantillon de populations actuelles. Les estimations, cependant, peuvent fluctuer en fonction des régions génomiques, parce que le mécanisme d'évolution de chaque région génomique est unique, même au sein d'un même individu. Dans cet article, nous proposons un modèle d'arbres phylogénétiques qui permet l'existence d'arbres distincts pour chaque région du génome. Dans chaque arbre, des caractéristiques évolutives uniques permettent de prendre en

compte chaque génome, ainsi que leur relation homologue; par conséquent, l'approche peut distinguer l'histoire évolutive d'un génome de celle d'un autre. En plus des temps de divergence distincts, le nouveau modèle peut estimer des tailles efficaces de population, de généalogies de gènes et d'autres paramètres spécifiques de chaque génome. Pour l'inférence bayésienne, nous avons développé une méthode de Monte Carlo par chaînes de Markov (MCMC) avec un nouvel algorithme MCMC ayant de bonnes caractéristiques de mélange sur un espace complexe. La stabilité du nouvel estimateur est démontrée par simulation et comparaison avec d'autres méthodes, ainsi que grâce au diagnostic de convergence du MCMC. L'analyse des données de gorilles africains sur deux loci homologues révèle des temps de divergence discordants entre les loci et cette différence s'explique par un flux de gènes via les mâles jusqu'à la fin de la dernière ère glaciaire.

BIOMETRIC PRACTICE

C. Kang, H. Janes, and Y. Huang **695**
Combining Biomarkers to Optimize Patient Treatment Recommendations

Les marqueurs prédictifs de l’effet d’un traitement pourraient permettre d’améliorer le pronostic des patients. Par exemple, le score de rechute *Oncotype DX* a une certaine aptitude pour prédire le bénéfice d’une chimiothérapie adjuvante associée à un traitement hormonal dans les cancers du sein ayant des récepteurs d’œstrogène positifs, permettant de proposer cette chimiothérapie aux femmes qui sont le plus susceptibles d’en bénéficier. Sachant que ce score a été originalement développé pour prédire le pronostic des patientes traitées par une hormonothérapie seule, il est intéressant de développer des combinaisons différentes des gènes inclus dans le score qui soient optimisées pour une meilleure sélection des traitements efficaces. Cependant, la plupart des méthodologies permettant de combiner plusieurs marqueurs pour prédire un pronostic ne sont valides que dans le cadre d’un traitement déterminé. Nous proposons une méthode combinant plusieurs marqueurs adaptée à la sélection thérapeutique, qui exige de modéliser l’effet thérapeutique comme une fonction de ces marqueurs. Les différents modèles de l’effet du traitement sont ensuite ajustés itérativement en augmentant le poids ou en « boostant » les sujets potentiellement mal classés en ce qui concerne le bénéfice du traitement à l’étape précédente. L’approche par « boost » est comparée aux méthodes existantes dans une étude par simulations, évaluant le changement de pronostic attendu avec le traitement sélectionné à l’aide des marqueurs. Cette approche fait mieux que les autres méthodes pour plusieurs paramètres, et a des performances comparables pour les autres. Notre étude par simulation fournit également des données sur les qualités respectives des méthodes existantes. Une application dans le cancer du sein de la méthode par boost utilisant une version réduite des marqueurs originaux est ensuite présentée, proposant des combinaisons de marqueurs qui auraient pu permettre d’améliorer les performances de la décision thérapeutique.

M. Taguri, Y. Matsuyama, and Y. Ohashi **724**
Model Selection Criterion for Causal Parameters in Structural Mean Models Based on a Quasi-Likelihood

Les modèles structuraux moyens (SMM) ont été proposés pour estimer les paramètres de causalité pour des essais cliniques en présence de non compliance non ignorable. Pour

obtenir un estimateur de causalité valide, il faut imposer plusieurs hypothèses. L'une d'elles et la spécification correcte du modèle. Partant du travail de Pan (2001, *Biometrics* **57**, 120-125) développant un critère de sélection de modèle pour des équations d'estimation généralisées, nous proposons une nouvelle approche pour la sélection de modèle dans le cadre des SMM, basée sur la quasi-vraisemblance. Nous donnons un critère formel de sélection de modèle par extension du critère d'information d'Akaiké. En utilisant une sélection d'un sous-ensemble de covariables pour le point de base, notre méthode nous permet de comprendre si l'effet du traitement varie au travers des niveaux disponibles des covariables de point de base, et/ou de quantifier l'effet du traitement pour un niveau spécifique des covariables défini pour cibler la maximisation du bénéfice de traitement à certains individus. Nous présentons des résultats de simulation montrant que notre méthode se comporte correctement lorsqu'elle est comparée à d'autres méthodes de test, aussi bien en termes de probabilité de sélectionner le modèle correct, et de performance prédictive des effets individualisés du traitement. Un large essai clinique randomisé portant sur la pravastatine a motivé ce travail.

L'interférence se produit lorsque le traitement d'une personne affecte l'état d'une autre personne. Par exemple, dans les maladies infectieuses, le fait qu'une personne soit vaccinée peut avoir un impact sur le fait qu'une autre personne s'infecte ou développe la maladie. Quantifier ces effets indirects de la vaccination pourrait avoir des implications importantes pour la santé publique ou les politiques de santé. Dans ce papier, nous utilisons des estimateurs, obtenus par la méthode pondérée sur la probabilité inverse (IPW), des effets de traitements en présence d'interférence pour analyser les données d'un essai à unité de randomisation individuelle, contrôlé, de la vaccination contre le choléra versus placebo visant 121 982 personnes à Matlab au Bangladesh. Parce que ces estimateurs IPW n'ont pas encore été utilisés, une étude de simulation a également été menée pour évaluer leur comportement empirique dans des contextes similaires à l'essai vaccinal contre le choléra. Les résultats de l'étude de simulation démontrent que les estimateurs IPW produisent des estimations non biaisées des effets directs, indirects, totaux et globaux de la vaccination en cas d'interférence à condition que l'hypothèse de confusions non mesurées, et non pas invérifiables, soit valide, et que le modèle de score de propension au niveau du groupe soit correctement spécifié. L'application des estimateurs IPW à l'essai vaccinal sur le choléra indique la présence d'une interférence. Par exemple, les estimations IPW suggèrent que en moyenne 5,29 moins de cas de choléra pour 1000 personnes-années (IC à 95% : 2,61 – 7,91) se produisent chez les personnes non vaccinées dans les quartiers avec une couverture vaccinale de 60% par rapport aux quartiers avec 32% de couverture. Notre analyse montre également comment ne pas tenir compte de l'interférence peut rendre des conclusions erronées quant à l'utilité au niveau de la santé publique de la vaccination.

Estimer l'efficacité d'une nouvelle intervention est souvent l'objectif principal pour les essais de prévention du VIH. Le modèle à risques proportionnels de Cox est principalement utilisé pour estimer l'efficacité en supposant que les participants partagent le même risque conditionnellement aux covariables et que le risque est toujours non nul. En fait, le risque est non nul seulement lorsqu'un événement exposant apparaît et les participants peuvent avoir un risque de transmettre qui varie selon des événements exposants qui varient eux-mêmes. Nous proposons donc une nouvelle estimation de l'efficacité ajustée sur l'hétérogénéité de l'ampleur de l'exposition dans la population étudiée, en utilisant un processus de Poisson latent pour le chemin d'exposition de chaque participant. De plus, notre modèle considère le scenario dans lequel une proportion de participants n'a jamais été exposée et adopte une distribution modifiée en zéro pour le processus du taux d'exposition. Nous employons une approche d'estimation bayésienne pour estimer l'efficacité ajustée sur l'exposition élicitant les a priori à partir de l'information historique. Des études de simulations sont conduites pour valider l'approche et explorer les propriétés des estimateurs. Un exemple d'application est présenté à partir d'un essai de prévention du VIH.

Pour les patients dialysés, les maladies cardiovasculaires et infectieuses sont les causes majeures d'hospitalisation et de décès. Bien que des études récentes aient trouvé que le risque d'évènements cardio-vasculaires est plus élevé après une hospitalisation pour infection, ces études n'ont pas complètement élucidé comment le risque cardiovasculaire change dans le temps pour les patients dialysés. Dans ce travail, nous caractérisons la dynamique des trajectoires de risque d'événement cardiovasculaire pour des patients sous dialyse, lorsque l'on conditionne sur l'état de survie au travers de plusieurs indices temporels : (1) le temps écoulé depuis la mise en dialyse, (2) le temps écoulé depuis l'hospitalisation initiale déclenchante pour infection, (3) l'âge du patient à la mise sous dialyse. Ceci est réalisé en utilisant une nouvelle classe de modèles généralisés multi-indices à coefficients variables (GM-IVC). Les modèles GM-IVC proposés utilisent une structure multiplicative et des coefficients fonctionnels uni-dimensionnels pour chaque indice de temps ou d'âge afin de saisir la dynamique du risque cardiovasculaire avant et après l'hospitalisation initiale pour infection, et ce pour toute la cohorte des survivants. Nous développons une procédure d'estimation en deux étapes pour les modèles GM-IVC, basée sur un maximum de vraisemblance local. Nous apportons de nouveaux regards sur la dynamique des risques d'évènements cardiovasculaire en utilisant la base de données américaine « United States Renal Data System », qui collecte les données de presque tous les patients en insuffisance rénale terminale aux Etats-Unis. Pour fini, nous évaluons la performance des procédures d'estimation proposées par des études de simulation.

Motivé par le problème de la construction de réseaux co-exprimés de gènes, nous proposons un approche statistique en trois étapes pour l'estimation en grande dimension

de la matrice de corrélation partielle. Nous obtenons un premier estimateur de la matrice de corrélation partielle par pénalisation « ridge ». Ensuite, les éléments non nuls de la matrice sont déterminés par des tests d'hypothèses avant d'estimer les coefficients de corrélation non nuls. Dans la deuxième étape, la distribution des statistiques de tests dérivées des estimateurs pénalisés des coefficients de corrélation partielle est inconnue sous l'hypothèse nulle, nous l'approchons par la distribution empirique de ces estimateurs. D'importantes simulations nous ont permis de valider la bonne performance de notre approche. L'application de notre approche à des données d'expression génique du cycle des levures montre qu'elle permet d'obtenir de meilleures prédictions des interactions protéines-protéines que l'approche graphique Lasso.