
Translations of Abstracts

BIOMETRIC METHODOLOGY**Min Zhang and Douglas E. Schaebel**

999

Double-Robust Semiparametric Estimator for Differences in Restricted Mean Lifetimes in Observational Studies

Le temps moyen de survie restreint présente souvent un intérêt direct dans les études épidémiologiques avec des temps de survie censurés. Les différences portant sur cette quantité peuvent servir de base pour la comparaison de plusieurs groupes. Par exemple dans le cas des transplantations, les chirurgiens, les néphrologues et naturellement les patients sont intéressés par la comparaison des temps de survie post-transplantation de différents types de reins afin d'aider à la décision clinique. Comme le facteur d'intérêt n'est pas randomisé, il faut ajuster au moyen de covariables pour prendre en compte les déséquilibres des facteurs confondants. La méthode proposée comprend la construction de modèles de travail pour les mécanismes de temps jusqu'à l'évènement et d'incomplétion (à savoir l'affectation au groupe et la censure). Nous montrons que l'estimateur proposé possède la propriété de double robustesse ; c'est-à-dire quand soit le processus de temps jusqu'à l'évènement, soit le processus d'incomplétion est correctement modélisé, l'estimateur est consistant et asymptotiquement normal. Des études de simulation sont effectuées pour évaluer sa performance en taille finie et la méthode est appliquée aux données nationales de transplantation rénale.

Baquin Zhang, Anastasios A. Tsiatis, Eric B. Laber and Marie Davidian

1010

A Robust Method for Estimating Optimal Treatment Regimes

Un « régime thérapeutique » correspond à une règle qui assigne, parmi un ensemble de traitements possibles, un traitement donné à un patient donné, en fonction de ses caractéristiques observées, en « personnalisant » ainsi le traitement. L'objectif est donc d'identifier le régime thérapeutique optimal, qui, s'il était suivi par l'ensemble de la population des patients, conduirait en moyenne au meilleur résultat. Pour des données d'un essai clinique ou d'une étude observationnelle, et pour une décision thérapeutique unique, le régime optimal peut être trouvé en utilisant un modèle de régression sur le résultat attendu, conditionnellement aux traitements et aux covariables, où, pour un profil donné de covariables, le traitement optimal est celui qui est associé à l'espérance de résultat attendu la plus élevée. Cependant, l'assignation d'un traitement via une telle règle est discutable si le modèle de régression est mal spécifié. Sachant que, même s'il est mal spécifié, ce modèle de régression définit une classe de régimes, nous proposons plutôt de trouver le régime optimal à l'intérieur d'une telle classe en sélectionnant le régime qui optimise un estimateur du résultat moyen sur l'ensemble de la population. Pour prendre en considération les facteurs de confusion dans une étude observationnelle et pour améliorer la précision, nous utilisons un estimateur pondéré sur la probabilité inverse

doublément robuste. Des simulations et une application aux données d'un essai thérapeutique dans le cancer du sein démontrent les performances de cette méthode.

Stijn Vansteelandt and Tyler J. VanderWeele

1019

Natural Direct and Indirect Effects on the Exposed: Effect Decomposition under Weaker Assumptions

Nous définissons les effets naturels direct et indirect sur les exposés. Nous montrons qu'ils autorisent la décomposition des effets sous des conditions d'identification plus souples que les effets naturels direct et indirect sur la population. Quand l'exposition n'affecte aucune des variables de confusion de l'association médiateur-réponse, les conditions d'identification sont essentiellement les mêmes que pour les effets directs contrôlés. Dans les autres cas, l'identification reste possible avec des informations complémentaires sur une fonction de biais de sélection non identifiable qui exprime la mesure dans laquelle l'effet du médiateur dépend de l'exposition à un niveau donné du facteur de confusion, mais qui reste nulle dans une grande famille de mécanismes réalistes de génération des données.

Nous soutenons que les effets naturels direct et indirect sur les exposés sont pertinents dans diverses applications. Nous montrons en outre qu'ils sont identiques aux effets naturels direct et indirect sur la population quand l'exposition est randomisée. Dans ce cadre, nos résultats permettent également d'apprécier les effets naturels direct et indirect quand l'exposition confond la relation médiateur-réponse, ce que la méthodologie disponible ne permettait pas.

**Michael J. Daniels, Jason A. Roy, Chanmin Kim, Joseph W. Hogan,
and Michael G. Perri**

1028

Bayesian Inference for the Causal Effect of Mediation

Nous proposons une approche bayésienne non-paramétrique pour estimer les effets directs et indirects par médiation, dans un dispositif avec une variable médiatrice continue et une réponse binaire. Plusieurs hypothèses d'indépendance conditionnelle sont introduites (avec les paramètres de sensibilité correspondant) pour rendre ces effets identifiables à partir des données observées. Nous suggérons des stratégies d'élicitation des paramètres de sensibilité, et nous réalisons des simulations pour évaluer les violations des hypothèses. Cette approche est utilisée pour évaluer l'effet médiateur dans un récent essai clinique portant sur le contrôle du poids.

Michael E. Sobel and Bengt Muthén

1037

Compliance Mixture Modelling with a Zero-Effect Complier Class and Missing Data

Les études randomisées sont la référence pour l'évaluation des thérapeutiques proposées. On s'intéresse en intention de traiter (ITT) à la mesure de l'effet du traitement affecté, mais pas l'effet du traitement si le sujet prend des traitements qui n'ont pas été affectés. Le besoin d'estimer l'efficacité du traitement dans ce cas a impulsé une littérature abondante sur la compliance dans les 15 dernières années. Dans les articles traitant de ce sujet, il est classiquement supposé qu'il existe différents types de sujets, par exemple ceux qui suivront le traitement affecté (compliers), et ceux qui vont toujours prendre un traitement particulier sans tenir compte du traitement qui leur aura été affecté. Les

quantités les plus importantes à estimer sont la proportion de compliants et l'effet moyen du traitement pour un compliant (CACE). Pour estimer CACE, les chercheurs ont utilisé diverses méthodes, par exemple les variables instrumentales et les modèles paramétriques de mélange, traitant les compliants comme une seule classe. Cependant il est souvent peu raisonnable de penser que tous les compliants seront affectés. Cet article par conséquent, traite les compliants comme un mélange de deux catégories, ceux appartenant à une classe sans effet, et les autres appartenant à une classe à effet. Ensuite, dans la plupart des études, quelques sujets sont exclus ou simplement ne donnent pas lieu au report de la variable mesurée, et l'impossibilité de prendre en compte ces données manquantes peut conduire à des estimations biaisées des effets de traitement. Des travaux récents sur la compliance dans les études randomisées se sont intéressés à ce point en supposant que les données manquantes sont de type « missing at random (MAR) » ou ignorables sur le plan latent. Nous étendons ce travail au cas où les compliants sont un mélange, et nous examinons également des modèles alternatifs d'hypothèses de données manquantes non ignorables.

Chung-Wei Shen and Yi-Hau Chen

1046

Model Selection for Generalized Estimating Equations Accommodating Dropout Missingness

Les équations d'estimation généralisées (GEE) ont été un outil populaire pour les analyses de régression marginales avec des données longitudinales, et son extension, l'approche GEE pondérée, peut de plus s'accommoder de données manquantes au hasard (MAR). Cependant, les méthodes de sélection de modèles pour les GEE, n'ont pas été systématiquement développées pour prendre en compte les données manquantes. Nous proposons le critère d'information longitudinal manquant (MLIC) pour la sélection du modèle de moyenne, et le critère d'information longitudinal manquant pour la corrélation (MLICC) pour la sélection de la structure de corrélation dans les GEE quand les critères de jugement sont sujet à des sorties d'étude ou un mécanisme de données manquantes monotone et sont MAR. Nos résultats de simulation révèlent que les MLIC et MLICC sont efficaces respectivement pour la sélection de variable dans le modèle de moyenne et dans la sélection de la structure de corrélation. Nous démontrons aussi les défauts importants de traiter naïvement des données incomplètes comme si elles étaient complètes et d'appliquer les méthodes de sélection existantes pour les GEE. L'utilité de la méthode proposée est ensuite illustrée par deux applications réelles impliquant des données longitudinales manquantes.

Michael J. Daniels, Arkendu S. Chatterjee, and Chenguang

1055

Wang

Bayesian Model Selection for Incomplete Data Using the Posterior Predictive Distribution

Nous proposons et évaluons un critère de sélection de modèle basé sur la perte prédictive *a posteriori*, pour des données longitudinales incomplètes. Nous identifions tout d'abord une propriété importante des critères de sélection de modèles pour données incomplètes. Ensuite, nous démontrons qu'une généralisation directe du critère de Gelfand et Ghosh (1998) aux données incomplètes poserait deux problèmes. Premièrement, elle introduirait un terme supplémentaire (outre les termes de qualité d'ajustement et de pénalité) rendant

le critère résultant non fiable. Deuxièmement, le critère résultant n'aurait pas la propriété importante évoquée ci-dessus. Nous proposons donc un critère alternatif et nous en explorons les propriétés par simulations ainsi que sur un jeu de données réel, en le comparant avec le critère DIC. Le critère DIC reste en général meilleur que celui basé sur la prédiction *a posteriori*, mais ce dernier semble bien fonctionner dans l'ensemble. De plus son calcul est très simple, contrairement au calcul du DIC dans certaines classes de modèles pour données incomplètes.

Silvia Montagna, Surya T. Tokdar, Brian Neelon, and David B. Dunson

1064

Bayesian Latent Factor Regression for Functional and Longitudinal Data

Dans les études impliquant des données fonctionnelles, on s'intéresse en général à modéliser l'impact des prédicteurs sur la distributions des courbes, ce qui permet une certaine flexibilité des effets, non seulement sur la courbe moyenne, mais aussi sur la distribution autour de la moyenne. Nous caractérisons la courbe de chaque sujet par une combinaison linéaire de bases de fonctions appartenant à un espace de grande dimension. Sur les coefficients de la base, nous plaçons un modèle de régression à facteurs latents, peu denses. La sélection des prédicteurs s'effectue alors en choisissant un *a priori* contractant, qui fait que la plupart des poids sont proches de zéro. Le nombre de facteurs latents est traité par un algorithme de Gibbs comme une quantité inconnue. Notre algorithme est adaptatif par blocs; il est en outre hautement efficace. Les prédicteurs se situant au niveau des variables latentes, les différents prédicteurs peuvent alors impacter les différents facteurs latents. Ce modèle induit un cadre pour la régression fonctionnelle dans lequel la distribution des courbes permet, par le biais des prédicteurs, de modifier la flexibilité. Nous évaluons notre méthode à partir de simulations, puis nous l'appliquons sur des données de trajectoires de pression sanguine mesurée lors de la grossesse.

Liya Fu and You-Gan Wang

1074

Efficient Estimation for Rank-Based Regression with Clustered Data

L'inférence basée sur les rangs est largement utilisée en raison de sa robustesse. Cet article propose des fonctions d'estimation optimales, fondées sur les rangs, pour l'analyse de données classées avec des effets de classe aléatoires. Les études extensives de simulation qui ont été menées pour évaluer la performance de la méthode proposée, montrent qu'elle est robuste aux données éloignées (outliers) et très efficace par rapport à l'existence de fortes corrélations de classes. La performance de la méthode est satisfaisante même lorsque la structure de corrélation est mal spécifiée, ou lorsqu'on est en présence de variances hétéroscédastiques. Enfin, un ensemble de données réelles est analysé pour illustrer.

Martha Skup, Hongtu Zhu, and Heping Zhang

1083

Multiscale Adaptive Marginal Analysis of Longitudinal Neuroimaging Data with Time-Varying Covariates

Les données de neuroimagerie collectées de façon répétées (dites longitudinales) ont attiré une attention croissante dans la communauté de la neuroimagerie, en raison de leur potentiel à répondre à des questions concernant le développement du cerveau, le

vieillesse ou la neuro-dégénérescence. Ces jeux de données, larges et complexes, sont caractérisés par une structure de dépendance spatiale compliquée dans chaque image réponse, un grand nombre d'images réponses par sujet et des covariables susceptibles de varier au cours du temps. Afin d'estimer des modèles de régression marginaux pour des données d'imagerie aussi complexes (données spatio-temporelles avec covariables dépendantes du temps) nous proposons une méthode des moments généralisés multi-échelle adaptée (MA-GMM pour multiscale adaptive generalized method of moments).

Notre méthode repose sur une catégorisation des covariables, chaque type correspondant à des conditions de moments valides. De plus, contrairement à de nombreuses techniques actuelles d'analyse de neuroimagerie, nous ne supposons pas l'indépendance des voxels (les voxels étant les éléments qui composent l'image réponse de chaque individu à chaque instant) : notre méthode lisse de façon adaptative les données et plus précisément calcule les estimateurs des paramètres en construisant de façon itérative des sphères autour de chaque voxel et en combinant avec pondération les observations au sein des sphères aux questions.

Le développement de la méthode MA-GMM vient s'ajouter au petit nombre de méthodes disponibles pour analyser les données d'imagerie longitudinales. Les performances de la méthode MA-GMM sont illustrées sur données simulées et sur un jeu de données réelles d'imagerie longitudinales fourni par l'ADNI (Alzheimer's Disease Neuroimaging Initiative).

Na Cai, Wenbin Lu, and Hao Helen Zhang

1093

Time-Varying Latent Effect Model for Longitudinal Data with Informative Observation Times

Dans l'analyse de données longitudinales, il n'est pas rare que les dates d'observation de mesures répétées soient sujet-spécifiques et corrélées avec les résultats longitudinaux sous-jacents. La prise en compte de cette dépendance entre les dates d'observation et les mesures objectives est nécessaire pour garantir la validité de l'inférence statistique. Dans cet article nous proposons un modèle joint flexible pour l'analyse de données longitudinales dans le cas de dates d'observations informatives. En particulier, la nouvelle procédure considère le modèle à effets aléatoires partagés et suppose un coefficient dépendant du temps pour la variable latente, permettant ainsi une approche flexible de la modélisation de résultats longitudinaux tout en ajustant à leur association aux dates d'observation. Des équations d'estimation sont développées pour l'estimation des paramètres. Nous montrons que les estimateurs qui en résultent sont consistants et asymptotiquement gaussiens, avec une matrice de variance-covariance de forme fermée et qui peut être estimée de façon consistante par la méthode usuelle de « plug-and-play ». Un avantage supplémentaire de cette procédure est qu'elle fournit un cadre unificateur pour tester si l'effet d'une variable latente est nul, constant ou dépendant du temps. Des études de simulation montrent que l'approche proposée est adaptée à un usage pratique. Une application à des données du cancer de la vessie est également fournie pour illustrer la méthodologie.

Y. Fong, J. Wakefield, S. De Rosa, and N. Frahm

1103

A Robust Bayesian Random Effects Model for Nonlinear Calibration Problems

Dans le cadre d'un essai biologique ou d'un dosage immunologique, l'étalonnage signifie l'ajustement d'une courbe, généralement non linéaire, à travers des observations recueillies sur un ensemble d'échantillons contenant des concentrations connues d'une substance cible, et ensuite l'utilisation de la courbe ajustée et des observations collectées sur des échantillons d'intérêt pour prédire les concentrations de la substance cible dans ces échantillons. Les récents progrès technologiques ont grandement amélioré notre capacité à quantifier des quantités infimes de substances à partir d'un minuscule volume de l'échantillon biologique. Cela a à son tour conduit à un besoin d'améliorer les méthodes statistiques pour l'étalonnage. Dans cet article, nous nous concentrons sur le développement de méthodes d'étalonnage robustes aux valeurs aberrantes dépendantes. Nous introduisons un nouveau modèle de mélange normal avec des termes d'erreur dépendants pour modéliser le bruit expérimental. En outre, nous proposons une reparamétrisation des cinq paramètres du modèle de régression logistique non linéaire qui nous permet de mieux intégrer l'information a priori. Nous examinons la performance de nos méthodes avec des études de simulation et montrons qu'elles conduisent à une augmentation substantielle de la performance mesurée en termes d'erreur quadratique moyenne d'estimation et de mesure de la précision de la prédiction moyenne. Un exemple de données réelles issues du HIV Vaccine Trials Network Laboratory est utilisé pour illustrer les méthodes.

Yuanjia Wang and Huaihou Chen

1113

On Testing an Unspecified Function Through a Linear Mixed Effects Model with Multiple Variance Components

Dans ce travail, nous évaluons un test pour fonctions non-paramétriques construit sur une statistique de Fisher dérivée de la reformulation d'une fonction spline en un modèle linéaire à effet mixte. De part cette réécriture, ce test d'une fonction non-spécifiée devient le test d'un vecteur d'effets fixes et de certains composants de la variance d'un modèle linéaire à effets mixtes où des composants de la variance sont des paramètres de nuisance sous l'hypothèse nulle. Cette approche peut être utilisée pour tester une fonction non-paramétrique ou un coefficient d'effet dans des analyses en sous-groupes, pour comparer 2 splines, pour tester la significativité d'une fonction non-spécifiée dans un modèle additif avec plusieurs composants ou encore pour tester un effet ligne ou colonne dans une analyse de variance à 2 facteurs. Par une décomposition spectrale de la somme des carrés des résidus, nous proposons un algorithme rapide déterminant la distribution du test sous l'hypothèse nulle, ce qui en améliore de façon significative l'efficacité computationnelle par rapport à la méthode bootstrap. Par surcroît, cette étape de décomposition spectrale trouve une connexion entre le test du rapport de vraisemblance (LRT) pour un modèle à composante de la variance multiple et un modèle à composante unique. Ces méthodes sont évaluées par une étude de simulations où il est montré que le test de Fisher généralisé est plus puissant que le LRT selon l'hypothèse de test et le vrai modèle simulé sous l'hypothèse alternative. Ces méthodes sont utilisées pour calculer la statistique et la P-value seuil dans une analyse « génome entier » où la méthode bootstrap est beaucoup trop coûteuse en calculs (jusqu'à 10^8 simulations nécessaires) et l'approximation asymptotique invalide et conservative.

Sangbum Choi and Xuelin Huang

1126

A General Class of Semiparametric Transformation Frailty Models for Nonproportional Hazards Survival Data

Nous proposons une estimation semi-paramétrique efficace d'une large classe de transformation de modèles de régression pour des données à risques non-proportionnels. Les modèles transformés classiques doivent être considérés sous le paradigme du modèle à fragilité et la méthode proposée fournit une approche unifiée qui est valable pour les modèles à fragilité discrète et continue. Les modèles proposés sont suffisamment flexibles pour modéliser les données de survie avec suivi à long terme lorsque l'effet du traitement diminue avec le temps, un cas pour lequel l'hypothèse des risques ou des cotes proportionnels est violée, ou une situation dans laquelle une proportion importante de patients reste guéris après le traitement. L'estimation du paramètre de lien dans la distribution de la fragilité, considéré comme inconnu et possiblement dépendent de covariables dépendent du temps, est automatiquement incluse dans les méthodes proposées. La matrice d'information observée est calculée pour évaluer les variances des estimateurs de tous les paramètres. Notre approche basée sur la vraisemblance fournit un moyen naturel de construire des statistiques simples pour tester l'hypothèse des risques ou des cotes proportionnels pour des données de survie usuelles, ou de tester les effets à court ou à long terme pour des données de survie avec fraction guérie. Des études de simulation montrent que les procédures d'inférence proposées sont performantes dans des contextes réalistes. Deux applications à des études médicales sont présentées.

Jianchang Lin, Debajyoti Sinha, Stuart Lipsitz, and Adriano Polpo

1136

Semiparametric Bayesian Survival Analysis using Models with Log-Linear Median

Nous proposons un nouveau modèle de survie semi-paramétrique basé sur une fonction de régression log-linéaire de la médiane. En alternative aux modèles semi-paramétriques existants, notre vaste classe de modèles présente en pratique beaucoup d'avantages, intégrant l'interprétation simple des paramètres de régression via la médiane et la possibilité de s'intéresser à l'hétérosédasticité. Nous démontrons que notre technique de modélisation facilite l'établissement d'hypothèses a priori et les calculs dans le cadre d'une analyse Bayésienne de données de survie, que ce soit dans un cadre semi-paramétrique ou paramétrique. Nous illustrons les avantages de notre modélisation, ainsi que l'interprétation des paramètres du modèle en ré-analysant les données d'une étude sur le cancer du poumon à petites cellules. Les résultats d'une étude de simulation permettent des recommandations pour l'utilisation de nos modèles.

Karen Lostritto, Robert L. Strawderman, and Annette M. Molinaro

1146

A Partitioning Deletion/Substitution/Addition Algorithm for Creating Survival Risk Groups

L'estimation précise du risque d'un événement donné pour un patient est essentielle pour faire des choix thérapeutiques éclairés. Une approche est de stratifier les patients en deux ou plusieurs groupes de risques distincts par rapport à un résultat spécifique en utilisant à la fois des variables cliniques et démographiques. Les résultats peuvent être de nature catégorielle ou continue; des exemples importants dans les études en oncologie

pourraient inclure le niveau de toxicité ou le délai jusqu'à la récurrence. Des méthodes de partitionnement récursif sont idéales pour la construction de tels groupes à risque. Deux de ces méthodes sont les arbres de classification (Classification and Regression Trees, CART) et une plus récente connue sous le nom de *algorithme de partitionnement Suppression/Substitution/Addition* (*partDSA*), ces deux méthodes utilisent les fonctions de perte (par exemple, l'erreur quadratique pour un résultat continu) comme base pour la construction, la sélection et l'évaluation des prédicteurs mais diffèrent dans la manière dont les arbres de régression sont construits. Récemment, nous avons montré que *partDSA* est souvent plus performant que CART dans les situations appelées « *full data* » (par exemple, non censurée). Toutefois, lorsqu'elles sont confrontées à des données censurées, les fonctions de perte utilisées par les deux procédures doivent être modifiées. Il y a eu plusieurs tentatives pour adapter CART pour des données censurées à droite. Cet article décrit deux extensions pour *partDSA* qui permettent l'utilisation de fonctions de perte de données observées (i. e. possiblement censurées). Ces fonctions de perte de données observées, construites en utilisant la probabilité inverse de poids censurés, sont des estimations cohérentes de leurs homologues non censurés à condition que le modèle de censure correspondant soit correctement spécifié. La performance relative de ces nouvelles méthodes est évaluée par des études de simulation et illustrée par une analyse des données d'un essai clinique sur des patients atteints d'un cancer du cerveau. L'implémentation de *partDSA* pour données non censurées et censurées à droite est accessible dans le package R, **partDSA**.

Joshua Warren, Montserrat Fuentes, Amy Herring, and Peter Langlois

1157

Spatial-Temporal Modeling of the Association between Air Pollution Exposure and Preterm Birth: Identifying Critical Windows of Exposure

L'exposition à des niveaux élevés de pollution atmosphérique pendant la grossesse est associée à une plus grande probabilité de naissance prématurée (NPM), une cause majeure de morbidité infantile et de mortalité. De nouvelles méthodologies statistiques sont nécessaires pour déterminer spécifiquement à quelle époque un polluant particulier impacte la NPM, pour déterminer le rôle des différents polluants, et pour caractériser la variabilité spatiale de ces résultats. Nous développons un nouveau modèle spatial Bayésien pour la NPM qui identifie des fenêtres de susceptibilité tout au long de la grossesse conjointement pour plusieurs polluants (PM_{2.5}, ozone) en permettant à ces fenêtres de varier de façon continue à travers le temps et l'espace. Nous géo-codons les données de naissance de l'état civil du Texas (2002-2004) et les liions aux données classiques de surveillance de la pollution ainsi qu'à un produit de l'EPA nouvellement développé, basé sur la sortie calibrée d'un modèle de pollution atmosphérique. Nous appliquons le modèle spatial complet à une région de 13 comtés de l'est du Texas consistant en des zones très urbaines ou rurales. Nos résultats indiquent un signal significatif dans les deux premiers trimestres de la grossesse avec différents polluants menant à des fenêtres critiques différentes. L'introduction de l'aspect spatial met en évidence des fenêtres critiques restées non-identifiées lorsque l'aspect spatial était ignoré. Une procédure d'inférence est présentée pour analyser correctement ces fenêtres.

Long Qu, Dan Nettleton, and Jack C. M. Dekkers

1168

A Hierarchical Semiparametric Model for Incorporating Intergene Information for Analysis of Genomic Data

Pour l'analyse de données de puces à ADN, dans le cadre d'expérience visant à mesurer des profils d'expression de gènes, il est en pratique très répandu de modéliser la distribution de ces profils par un modèle de mélange à deux composantes afin d'identifier des gènes différentiellement exprimés.

Cependant, cela impose naïvement de formuler des hypothèses fortes d'interchangeabilité des gènes et ne fait pas activement usage de l'information a priori sur les relations entre gènes qui est actuellement mise à disposition, comme par exemple les annotations des gènes disponibles à travers le projet *Gene Ontology* (GO).

Nous proposons une stratégie générale qui génère en premier lieu un ensemble de covariables qui résumant l'information entre gènes et qui ensuite étend le modèle de mélange à deux composantes en un modèle hiérarchique semi-paramétrique prenant en compte les covariables générées à travers une régression latente non paramétrique.

Des analyses menées à la fois sur des données simulées et des données réelles de puces à ADN permettent de montrer que notre méthode surpasse le modèle naïf de mélange à deux composantes.

Long Qu, Dan Nettleton, and Jack C. M. Dekkers

1178

Improved Estimation of the Noncentrality Parameter Distribution from a Large Number of t -Statistics, with Applications to False Discovery Rate Estimation in Microarray Data Analysis

Étant donné le grand nombre de statistiques t , nous considérons le problème de l'approximation de la distribution des paramètres de non-centralité (PNC) par une densité continue. Ce problème est étroitement lié au contrôle du *False Discovery Rate* (FDR) dans des applications considérant une très grande quantité de test d'hypothèses, par exemple, l'analyse de l'expression des gènes par biopuces. Notre méthodologie est similaire, mais améliore, l'approche proposée par Ruppert, Nettleton, et Hwang (2007, *Biometrics*, 63, 483-495). Nous fournissons des estimateurs paramétrique, non paramétrique, et semi-paramétrique pour la distribution des PNCs, ainsi que des estimations du FDR et FDR local. Dans la situation paramétrique, nous supposons que les paramètres de non-centralité suivent une distribution qui conduit à une distribution marginale disponible analytiquement pour les statistiques de test. Dans la situation non-paramétrique, nous utilisons des combinaisons convexes de fonctions de densité de base pour estimer la densité des PCNs. Une procédure de programmation quadratique séquentielle est développée pour maximiser la vraisemblance pénalisée. Le paramètre de lissage est sélectionné avec l'approximation du critère d'information. Un estimateur semi-paramétrique est également développé pour combiner les deux ajustements paramétrique et non paramétrique. Des simulations montrent que, dans plusieurs situations différentes, nos estimations de la densité sont plus proches de la vérité sous-jacente et nos estimations de FDR sont améliorées par rapport à d'autres méthodes. Des données simulées et les analyses de deux jeux de données issus de biopuces sont utilisées pour évaluer la performance dans des situations réalistes.

Stefano Favaro, Antonio Lijoi, and Igor Prünster

1188

A New Estimator of the Discovery Probability

Les questions d'échantillonnage d'espèces ont une longue histoire dans les études biologiques et écologiques, et nombre de thématiques, parmi lesquelles l'évaluation de la richesse des espèces, le schéma des expériences d'échantillonnage, l'estimation de la variété des espèces rares, sont à approfondir. De tels problèmes inférentiels ont aussi récemment été posés dans des applications en génomique, présentant cependant quelques particularités qui les rendent moins attractifs : ainsi, ils concernent alors de grandes populations (les bibliothèques génomiques) contenant un très grand nombre d'espèces distinctes (les gènes) et seule une petite partie de la bibliothèque a été échantillonnée (séquencée). Ces aspects motivent une approche non-paramétrique bayésienne que nous avons choisie, car elle permet d'atteindre le degré de flexibilité nécessaire dans ce cadre. En se basant sur un échantillon observé de taille n , nous nous concentrons sur la prédiction d'un aspect essentiel du résultat d'un échantillon additionnel de taille m , appelé la probabilité de découverte. En particulier, conditionnellement à un échantillon de base de taille n , nous obtenons un nouvel estimateur de la probabilité de détection, à la $(n+m-1)^{\text{e}}$ observation, des espèces qui avaient été observées à une fréquence quelconque donnée dans l'échantillon élargi de taille $n+m$. Un tel estimateur admet une expression qui peut être exactement évaluée. Le résultat obtenu nous permet de quantifier à la fois le taux de détection des espèces rares, et pour les espèces abondantes le taux de couverture d'échantillonnage lorsque m augmente. Nous présentons des applications naturelles avec l'estimation de la probabilité de découverte de gènes rares dans des bibliothèques génomiques, et nos résultats sont illustrés par deux ensembles de données de marqueurs de séquences exprimées.

Hokeun Sun and Hongzhe Li

1197

Robust Gaussian Graphical Modeling Via l_1 Penalization

Les modèles graphiques Gaussiens sont très souvent utilisés comme une méthode efficace pour l'étude de la structure d'indépendance conditionnelle entre gènes et la construction de réseaux génétiques. Cependant, les données d'expression de gène sont caractérisées par des distributions présentant de plus longues traînes ou plus de valeurs exceptionnelles que la distribution Gaussienne standard. La présence de telles valeurs exceptionnelles peut mener à des erreurs lors de l'inférence de la structure de dépendance entre gènes. Nous proposons une procédure d'estimation avec une pénalité l_1 visant à estimer des modèles graphiques Gaussiens épars qui est rendue robuste aux valeurs aberrantes. La fonction de vraisemblance est pondérée en fonction de la déviation des observations, la déviation d'une observation étant mesurée à partir de sa propre vraisemblance. Nous développons un algorithme efficace de calcul basé sur une méthode de descente de gradient par coordonnée afin d'obtenir le minimum de la vraisemblance robuste pénalisée négative, dans laquelle les coefficients non nuls de la matrice de concentration représentent des arêtes dans le graphe des liens entre gènes. Une fois la structure du graphe obtenue, nous ré-estimons la matrice de concentration définie positive à l'aide d'un algorithme itératif d'adéquation proportionnelle. Des simulations nous permettent de montrer que les performances de la méthode robuste proposée sont bien meilleures que celles de la méthode *graphical Lasso* pour des modèles graphiques Gaussiens en termes à la fois de sélection de la structure du graphe et d'estimation en présence de valeurs aberrantes. Nous appliquons également cette méthode d'estimation robuste à l'analyse de données d'expression de gènes de la levure et montrons que le

graphe résultant est mieux interprétable biologiquement que celui obtenu avec la méthode *graphical Lasso*.

Mei-Cheng Wang and Shanshan Li

1207

Bivariate Marker Measurements and ROC Analysis

Cet article s'intéresse à l'analyse des courbes ROC (receiver operating characteristic) pour des mesures bivariées de marqueurs. L'intérêt de cette recherche est d'étendre les outils et les règles connus pour les marqueurs univariés au cas bivarié pour l'évaluation de la capacité prédictive des marqueurs en utilisant une règle de classification arborescente. En utilisant un classifieur et/ou, nous proposons une fonction ROC ainsi qu'une fonction ROC pondérée (WROC) avec leurs contreparties conjuguées pour l'examen des performances de marqueurs bivariés. Les fonctions proposées évaluent la performance des classifieurs et/ou parmi toutes les combinaisons possibles des valeurs de marqueurs, et sont ainsi des mesures idéales pour comprendre la capacité prédictive des biomarqueurs dans la population cible. Nous étudions et discutons les particularités spécifiques des fonctions ROC et WROC comme d'autres statistiques associées, en comparaison avec les propriétés usuelles des marqueurs univariés. Nous développons des méthodes non paramétriques pour l'estimation des fonctions associées aux ROC, de l'aire sous la courbe, et de la probabilité de concordance. Avec une attention particulière sur les performances moyennes des marqueurs, les procédures proposées et les résultats inférentiels sont utiles à l'évaluation de la prédictabilité des marqueurs basée sur des mesures uni- ou bivariées pour différents choix de marqueurs, ainsi que pour l'évaluation de différentes combinaisons de classifieurs et/ou. Les résultats inférentiels développés dans cet article sont aussi étendus aux marqueurs multivariés avec une suite de classifieurs et/ou arbitrairement combinés.

Dandan Liu, Tianxi Cai, and Yingye Zheng

1219

Evaluating the Predictive Value of Biomarkers with Stratified Case-Cohort Design

L'identification de nouveaux biomarqueurs pour l'évaluation des risques est importante aussi bien pour une prévention efficace des maladies que pour la recommandation d'un traitement optimal. La recherche de biomarqueurs repose sur la précieuse ressource encore limitée des échantillons biologiques conservés dans le cadre de grandes études de cohortes prospectives. Les études cas-cohorte constituent un outil coût-efficace dans le contexte de l'évaluation des biomarqueurs, en particulier lorsque l'événement clinique d'intérêt est rare. Les méthodes statistiques existantes se concentrent sur une inférence efficace sur les risques relatifs dans le modèle de Cox. En nous basant sur le développement théorique récent de la vraisemblance pondérée pour les modèles semi-paramétriques dans le cadre des études à deux phases (Breslow and Wellner, 2007), nous proposons des méthodes statistiques pour évaluer l'exactitude et la prédictivité du risque de prédiction d'un biomarqueur, avec temps de survenue d'un événement censuré dans le cadre d'une étude cas-cohorte stratifiée. Nous considérons des méthodes non paramétriques et une méthode semi-paramétrique. Nous établissons les propriétés asymptotiques des estimateurs proposés et évaluons leurs performances à échantillon fini sur la base d'études numériques. Nous illustrons les nouvelles procédures en utilisant les données de l'étude Framingham Offspring pour évaluer l'exactitude d'un score de risque récemment proposé incorporant l'information sur les biomarqueurs pour prédire les

maladies cardiovasculaires.

P. Nguyen, P. E. Brown, and J. Stafford

1228

Mapping Cancer Risk in Southwestern Ontario with Changing Census Boundaries

Cartographier le risque lié à une pathologie requiert souvent d'analyser des données spatiales regroupées arbitrairement par circonscriptions de recensement ou postales, pour des raisons administratives comme d'anonymisation. Dans l'étude de pathologie rare, les données sont recueillies sur de longue période afin d'accumuler un nombre conséquent de cas. Sur ces longues périodes, les circonscriptions peuvent être amenées à changer de démarcations, comme dans le cas ayant motivé ce travail qui explore la structuration spatiale du risque lié au mésothéliome dans les comtés de Lambton et du Middlesex du sud-ouest de l'Ontario au Canada. Cet article décrit un algorithme de lissage de type EM-local à noyau permettant de combiner les données de différentes cartes en prenant en compte les variations temporelles des démarcations des différentes circonscriptions. Les estimateurs et leur erreur d'estimation sont dérivés par bootstrap et une validation croisée est utilisée pour sélectionner la largeur des bandes de découpe des cartes. On montre les résultats pour l'étude sur le mésothéliome.

BIOMETRIC PRACTICE

Yang Yang, Ira M. Longini Jr., M. Elizabeth Halloran, and Valerie Obenchain

1238

A Hybrid EM and Monte Carlo EM Algorithm and Its Application to Analysis of Transmission of Infectious Diseases

Pour des épidémies de maladies infectieuses telles que la grippe, un individu peut être dans quatre états : immunité antérieure, non infecté, infecté et ayant des symptômes, infecté asymptomatique. L'état exact n'est souvent pas observé. De plus, les dates de transmission des infections asymptomatiques ne sont pas observées ce qui complique l'analyse. Sous l'hypothèse de données manquantes au hasard, les techniques d'augmentation des données peuvent être utilisées pour intégrer de telles incertitudes. Nous adaptons un algorithme EM de Monte Carlo (MCEM) basé sur un échantillonnage préférentiel pour une maladie infectieuse transmise dans des groupes à contacts proches. En supposant l'indépendance entre les groupes à contacts proches, nous proposons un algorithme EM-MCEM hybride qui applique l'algorithme MCEM et l'algorithme EM traditionnel à chaque groupe, selon l'étendue des données manquantes dans chaque groupe, et discutons de l'estimation de la variance. De plus, nous proposons une approche par bootstrap pour mesurer l'erreur totale de Monte Carlo et la prendre en compte dans l'estimation de la variance. Les méthodes proposées sont évaluées par des études de simulations. Nous utilisons l'algorithme hybride EM-MCEM pour analyser deux épidémies de grippe à la fin des années 70 afin d'estimer les effets de l'âge et du niveau d'anticorps sur la transmissibilité et la pathogénicité des virus.

Devan V. Mehrotra, Xiaoming Li, Jiajun Liu, and Kaifeng Lu

1250

Analysis of Longitudinal Clinical Trials with Missing Data Using Multiple Imputation in Conjunction with Robust Regression

Dans un essai randomisé classique, la variable d'intérêt continue (ex. la densité de l'os) est mesurée à l'entrée puis, après l'entrée, à des temps fixés. Cela conduit à des données longitudinales, souvent incomplètes en raison de perdus de vue ou d'autres raisons, que l'on analyse classiquement par des méthodes paramétriques basées sur la vraisemblance qui supposent la normalité multivariée du vecteur des réponses. Si l'hypothèse de normalité n'est pas réaliste alors, des méthodes semi-paramétriques telles que les équations d'estimations généralisées (pondérées) sont considérées. Nous proposons une approche alternative dans laquelle le problème des données manquantes est pris en compte par imputations multiples, et chaque échantillon imputé est analysé à l'aide d'une régression robuste (M-estimation ; Hubert, 1973) pour se protéger de la présence potentielle d'une non-normalité/valeurs aberrantes dans l'échantillon original ou imputé. Les résultats de l'analyse robuste de chaque échantillon imputés sont ensuite combinés pour l'ensemble des estimations et des inférences en utilisant la méthode simple de Rubin (1987), ou la méthode plus complexe mais potentiellement plus précise de Robins et Wang (2000). Nous utilisons des simulations pour montrer que notre approche est au moins aussi bonne que les méthodes classiques sous l'hypothèse de normalité, mais qu'elle est particulièrement meilleure pour des distributions à la fois symétrique elliptique et asymétrique non gaussiennes. Un exemple d'essai clinique est utilisé pour illustration.

Hongxiao Zhu, Philip J. Brown, and Jeffrey S. Morris

1260

Robust Classification of Functional and Quantitative Image Data Using Functional Mixed Models

Cet article introduit des méthodes nouvelles utilisant le cadre du modèle fonctionnel mixte (FMM) pour réaliser une classification de données fonctionnelles multidimensionnelles et complexes. Le FMM est associé à une réponse fonctionnelle à un ensemble de variables prédictives, via des effets fonctionnels fixes et aléatoires, permettant de tenir compte de divers facteurs et corrélations inter-fonctions. La classification se fait à travers un apprentissage du modèle en traitant l'identification des classes comme l'un des effets fixes, puis chaque nouvel objet est classé à l'aide des probabilités prédictives à posteriori d'appartenance aux classes. Grâce à un schéma bayésien, nous sommes en mesure d'ajuster des facteurs affectant à la fois les fonctions et l'identification des classes. Bien que la méthode puisse être appliquée à n'importe quel modèle FMM, nous fournissons des résultats détaillés pour deux approches bayésiennes spécifiques : la FMM par ondelettes gaussienne (G-WFMM) et sa version robuste (R-WFMM). Les deux méthodes modélisent dans l'espace des ondelettes, aboutissant à des représentations parcimonieuses pour les fonctions, et peuvent s'adapter naturellement aux caractéristiques locales et aux non-stationnarités complexes des fonctions. Pour des fonctions indexées par des coefficients d'ondelettes particuliers, le R-WFMM autorise des queues potentiellement plus importantes, ce qui aboutit à une sous-pondération des valeurs aberrantes et rend la méthode robuste aux fonctions aberrantes. Les modèles sont appliqués à des données de spectroscopie de masse dans le cancer du pancréas et sont comparés à d'autres méthodes de classification fonctionnelle récemment mises au point.

H. P. Piepho, E. R. Williams, and L. V. Madden

1269

The Use of Two-Way Linear Mixed Models in Multitreatment Meta-Analysis

Les méta-analyses permettent de résumer les résultats d'une série d'essais. Lorsque plus de deux traitements sont inclus dans les essais et lorsque la série de traitements testés diffère d'un essai à l'autre, la synthèse des résultats entre essais demande une attention toute particulière. Quelques méthodes ont été proposées pour cela sous différentes appellations, comme les méta-analyses en réseau ou les comparaisons de traitements mixtes. Deux types de modèles linéaires mixtes peuvent être utilisés dans les méta-analyses. Le premier exprime le résultat prédit pour les traitements comme un contraste par rapport au traitement de référence. Le deuxième utilise un prédicteur linéaire à deux facteurs classique avec des effets principaux pour le traitement et l'essai. Dans cet article, nous comparons les deux types de modèles et explorons sous quelles conditions ils donnent des résultats équivalents. Nous illustrons les avantages pratiques du modèle à deux facteurs à partir de deux jeux de données publiés. En particulier, nous montrons que l'hétérogénéité entre les essais ainsi que la divergence entre les différents types d'essais est facile à prendre en compte.

Han Chen, Alisa K. Manning, and Josée Dupuis

1278

A Method of Moments Estimator for Random Effect Multivariate Meta-Analysis

La méta-analyse est une approche puissante pour combiner des données provenant de plusieurs études afin d'en tirer des conclusions sur un ou plusieurs paramètres intéressants, tels que des coefficients de régression. La validité du modèle à effet fixe d'une méta-analyse repose sur la supposition sous-jacente que toutes les études de la méta-analyse ont le même effet de taille. En présence d'hétérogénéité, le modèle à effet fixe ignore de manière incorrecte la variance entre les études et peut donner des résultats faussement positifs. Le modèle à effets aléatoires tient compte à la fois des deux niveaux de variances, intra-études et inter-études. Il est plus conservateur que le modèle à effet fixe et devrait être utilisé quand il y a hétérogénéité. Dans cet article, nous développons une méthode non itérative de l'estimateur des moments pour la matrice de covariance inter-études du modèle multivariable à effets aléatoires d'une méta-analyse. À notre connaissance, c'est la première méthode de l'estimateur des moments sous forme matricielle. Nous montrons que notre estimateur est une extension multivariable de la méthode univariable de l'estimateur des moments de DerSimonian et Laird, et il est invariant par transformations linéaires. Dans l'étude de simulation, notre méthode donne de bons résultats si nous la comparons aux approches qui existent de la méta-analyse du modèle à effet aléatoire multivariable. Nous appliquons également notre méthode pour analyser un exemple réel de données.

Nandini Dendukuri, Ian Schiller, Lawrence Joseph, and Madhukar Pai

1285

Bayesian Meta-Analysis of the Accuracy of a Test for Tuberculous Pleuritis in the Absence of a Gold Standard Reference

Dans les études diagnostiques, l'absence de référence parfaite est une source connue de biais. Dans le cas de la pleurésie tuberculeuse, les tests standard de référence, comme, la microscopie de frottis, la culture ou la biopsie sont peu sensibles. Cependant, les méta-analyses de nouveaux tests de cette maladie supposaient toujours que le test de référence était parfait et fournissaient des estimations biaisées des performances du nouveau test. Nous décrivons une méthode pour la méta-analyse conjointe de la sensibilité et de la spécificité d'un test diagnostique adapté à la nature imparfaite du standard de référence.

Nous utilisons un modèle hiérarchique bayésien qui prend en compte les variabilités intra- et inter-études. Nous montrons comment obtenir des estimations groupées de la sensibilité et de la spécificité et comment tracer un résumé hiérarchique de la courbe ROC. Nous montrons comment étendre le modèle aux situations où divers tests de référence sont utilisés et où une dépendance conditionnelle existe entre le test index et les tests de référence. Nous évaluons par simulations les performances du modèle et nous l'illustrons avec les données d'une méta-analyse des tests par amplification d'acide nucléique (NAAT) pour le diagnostic de la pleurésie tuberculeuse. Par référence au modèle qui supposait une référence parfaite, l'estimation de la spécificité était augmentée et celle de sa sensibilité diminuée.

Bo Zhang, Zhen Chen, and Paul S. Albert

1294

Estimating Diagnostic Accuracy of Raters Without a Gold Standard by Exploiting a Group of Experts

L'estimation de la précision d'un diagnostic médical d'un ensemble d'évaluateurs ou de tests médicaux par rapport à un standard est souvent d'un intérêt essentiel. Lorsqu'on ne dispose pas d'un standard, on utilise souvent les modèles à classes latentes où le standard inconnu est traité comme une variable latente. Cependant ces modèles ont fait l'objet de critiques dans la littérature sur le plan conceptuel comme sur le plan de la robustesse. Nous proposons une approche alternative dans laquelle nous utilisons une référence standard imparfaite, avec une précision diagnostique inconnue, et nous menons une analyse de sensibilité en faisant varier cette précision dans un domaine scientifiquement raisonnable. Dans cet article, un modèle à classes latentes à effets aléatoires croisés est proposé pour l'estimation de la précision diagnostique de gynécologues et obstétriciens (OB/GYN) régionaux dans le diagnostic de l'endométriose. Pour éviter les écueils des modèles sans standard, nous exploitons les résultats d'un groupe de gynécologues obstétriciens (OB/GYN) de réputation internationale dans le diagnostic de l'endométriose. Nous construisons une référence standard ordonnée basée sur la discordance entre ces experts internationaux et nous proposons un schéma pour mener une analyse de sensibilité relative à la précision diagnostique inconnue parmi eux. Un algorithme EM Monte-Carlo est proposé pour l'estimation des paramètres et une procédure de sélection de modèle de type BIC est présentée. Nous montrons, au travers de simulations et d'analyses de données, que cette nouvelle approche est une alternative utile aux approches traditionnelles de modèles à classes latentes utilisées dans ce cadre.

Robert M. Dorazio

1303

Predicting the Geographic Distribution of a Species from Presence-Only Data Subject to Detection Errors

Plusieurs modèles ont été développés pour prédire la distribution géographique d'une espèce en combinant des mesures de covariables en des sites où l'espèce est connue pour être présente avec des mesures des mêmes covariables en des sites où le statut de présence de l'espèce (présence ou absence) est inconnu.. En l'absence d'erreurs de détection, des modèles de processus ponctuels spatiaux et des modèles de régression binaire avec données augmentées conduisent à des estimateurs efficaces de la distribution géographique d'une espèce en l'absence de connaissances a priori sur la prévalence de l'espèce. En outre, ces modèles de régression peuvent être modifiés pour

produire des estimateurs d'abondance de l'espèce asymptotiquement équivalents à ceux des modèles de processus ponctuels spatiaux. Cependant, si les sites de présence de l'espèce sont sujets à des erreurs de détection, aucune des deux classes de modèles ne fournit un estimateur efficace des effets des covariables, sauf si les covariables liées à l'abondance sont distinctes et indépendantes des covariables liées à la détectabilité. Ces résultats analytiques sont illustrés par des études de simulation de jeux de données contenant une large gamme de tailles d'échantillons de présence. Des analyses de présence de 3 espèces d'oiseaux observés dans une enquête sur les oiseaux terrestres de l'Ouest du Montana et de l'Est de l'Idaho sont comparées avec des analyses de détections et non-détections par des modèles d'occupations de sites.

Roland C. Deutsch and Walter W. Piegorsch

1313

Benchmark Dose Profiles for Joint-Action Quantal Data in Quantitative Risk Assessment

L'analyse comparative est un outil largement utilisé dans l'analyse de risque en santé publique. Dans cette application, quand on a affaire à une réponse indésirable due à un stimulus unique, on maîtrise bien l'estimation d'un niveau d'exposition minimum, appelé dose de référence (BMD – Benchmark Doses), induisant un niveau prespécifié de réponse (BMR - Benchmark Response). Dans les cas où deux agents sont étudiés en association, l'approche comparative est beaucoup moins développée. Cet article montre comment le paradigme de l'approche comparative peut être étendu de la dose unique aux études de l'effet conjoint de deux agents. On se focalise sur les réponses exprimées en proportions. En s'appuyant sur ce modèle, pour définir une quantification et une caractérisation du risque, on élabore le concept de profil de comparaison (BMP – Benchmark Profile) – un analogue bidimensionnel du BMD pour la dose unique pour lequel les deux agents atteignent le BMR donné. Les directives pour faibles doses, conjointes, qui en résultent pourraient améliorer la planification sanitaire et la réglementation dans le cas d'exposition à des faibles doses de combinaisons d'agents potentiellement dangereux.

READER REACTION

E. E. Giorgi and T. Bhattacharya

1313

A Note on Two-Sample Tests for Comparing Intra-Individual Genetic Sequence Diversity between Populations

Gilbert, Rossini et Shankarappa (2005, *Biometrics* **61**, 106-117) présentent quatre tests basés sur des U-statistiques pour comparer la diversité génétique de différents échantillons. Les tests proposés améliorent les méthodes précédemment utilisées grâce à la prise en compte des corrélations entre les données. Mais nous trouvons que ces mêmes corrélations introduisent un biais inacceptable dans les estimateurs de variance-covariance des distances génétiques entre séquences pour le cas d'échantillons de taille modeste. Nous en calculons ici des estimateurs sans biais et nous évaluons l'amélioration qui en résulte en utilisant des données simulées. Nous montrons aussi que, contrairement aux assertions de Gilbert et al., il n'est pas toujours possible d'appliquer le test t approximatif de Welch-Satterthwaite. Nous donnons par ailleurs des formules explicites pour les degrés de liberté à utiliser lorsque le test approximatif est applicable.