
Translations of Abstracts

SPECIAL SECTION - INTERNATIONAL BIOMETRIC SOCIETY HISTORY**L. Billard** 259*Sir Ronald A. Fisher and The International Biometric Society*

L'année 2012 a marqué le cinquantième anniversaire de la mort de Sir Ronald A. Fisher, l'un des deux pères de la science statistique et l'un des fondateurs de la Société Internationale de Biométrie (la « Société »). Pour commémorer le génie extraordinaire de Fisher, et la vision à long terme de Fisher et de Chester Bliss lorsqu'ils organisèrent et encouragèrent la création de la Société, cet article étudie des origines et la croissance de la Société, certains de ses acteurs et événements les plus marquants, et en particulier le rôle de Fisher lui-même en tant que premier Président. Nous portons aussi une attention renouvelée à Fisher, en tant qu'homme plutôt que comme scientifique de génie.

O. Mayo. 266*Fisher in Adelaide*

R.A. Fisher a passé la plupart de ses trois dernières années à Adelaide. C'était un endroit agréable pour vivre et travailler, et il était très demandé comme conférencier en Australie et au-delà. C'était cependant une période difficile pour lui, en raison des critiques nourries sur l'inférence fiduciaire à partir du début des années 50. Cet article présente certains des travaux de Fisher sur l'inférence, partant de sa vision d'Adelaide. Nous considérons également certaines des réussites issues de cette période, pour les statistiques de l'expérimentation de terrain et pour la génétique de l'évolution. On présente aussi quelques souvenirs personnels de Fisher, en tant qu'invité. Cet article est la forme écrite d'une présentation faite le 31 août 2012 à la 26^{ème} Conférence Internationale de Biométrie, à Kobe au Japon.

BIOMETRIC METHODOLOGY**S. Hattori and M. Henmi** 270*Stratified Doubly Robust Estimators for the Average Causal Effect*

Considérons l'estimateur de l'effet causal moyen à partir d'une étude observationnelle. Un estimateur doublement robuste, qui est un hybride de régression et d'un score de propension pondéré, est plus robuste que des estimateurs obtenus par l'une ou l'autre méthode dans le sens où, si au moins l'un des deux modèles tient, l'estimateur doublement robuste est cohérent. Cependant un estimateur doublement robuste peut quand même pâtir d'une spécification incorrecte du modèle car il n'est pas cohérent si au moins l'un des modèles n'est pas spécifié correctement. Dans cet article, nous proposons un estimateur de remplacement, appelé estimateur stratifié doublement robuste, obtenu en combinant une stratification du score de propension avec la régression et le score de propension pondéré. Cet estimateur offre deux modèles-candidats pour le score de propension, et il est plus robuste que les estimateurs doublement robustes existants, dans le sens où il est cohérent soit si le modèle de régression est correct, soit si l'un des deux modèles pour le score de propension est correct. Les propriétés asymptotiques de l'estimateur sont étudiées et la valeur de l'estimateur proposé sur des échantillons de taille limitée est évaluée par simulation. La méthode que nous proposons est illustrée avec l'étude « Tone », qui est une enquête conduite au Japon.

A. Mattei, F. Mealli, and B. Pacini

278

Identification of Causal Effects in the Presence of Nonignorable Missing Outcome Values

Nous considérons une nouvelle approche pour identifier des liens de causalité dans le cadre d'un traitement binaire. Nous supposons que l'output est manquant sur un sous-ensemble d'unités et que la dépendance de la non réponse vis à vis de l'output ne peut pas être exclue quand on travaille conditionnellement aux covariables observées. Nous donnons des conditions suffisantes pour lesquelles le fait de disposer d'une variable instrumentale pour la non-réponse nous permet, d'une part, d'obtenir une meilleure inférence quant à l'existence de liens de causalité dans la population globale, et d'autre part, d'identifier partiellement des liens de causalité dans des sous-groupes latents d'unités, appelés strates principales et définies par un comportement de non-réponse sur toutes les combinaisons possibles de la variable instrumentale et du traitement. Une étude par simulation est réalisée afin d'évaluer les bénéfices de la présence (versus l'absence) de la variable instrumentale. La simulation est basée sur des données médicales réelles provenant d'un essai randomisé d'un auto-examen de la poitrine; cet essai étant caractérisé par un fort taux de données manquantes. La variable instrumentale est simulée en considérant plusieurs scénarios afin de pouvoir discuter du rôle clé de la variable instrumentale dans l'identification des liens de causalité, et ce, en présence de données manquantes supposées non-ignorables. Au niveau de l'inférence, nous examinons également les gains potentiels attendus suite à l'usage d'une variable instrumentale pour la non-réponse, en adoptant une approche Bayésienne. Sur la base de résultats théoriques et empiriques, nous donnons quelques indications sur la façon de conduire une inférence causale.

B. Chen and J. Qin

289

Test the Reliability of Doubly Robust Estimation with Missing Response Data

Dans l'inférence statistique on doit s'assurer que le modèle de régression sous-jacent est spécifié correctement, car autrement l'estimation obtenue peut être biaisée. Cette vérification est importante pour détecter tout écart du modèle de régression au modèle

vrai. Les données manquantes constituent une difficulté courante dans les études sociales et médicales. Si le modèle de régression sous-jacent est correctement spécifié, les études récentes montrent qu'on utilise fréquemment des estimations doublement robustes en cas de données manquantes, à cause de la robustesse de ces méthodes vis-à-vis d'une spécification incorrecte soit du modèle de données manquantes soit du modèle de moyenne conditionnelle, *i. e.* le modèle de l'espérance du vrai modèle de régression conditionnée aux valeurs observées. Cependant, le test d'ajustement des données au modèle a été peu étudié dans le cas des estimées doublement robustes. Dans cet article, nous proposons une méthode de test pour évaluer la fiabilité de l'estimateur obtenu par l'équation d'estimation doublement robuste avec la possibilité de réponses manquantes et avec des variables auxiliaires toujours observées. Les études numériques démontrent que le test proposé sait bien contrôler les erreurs de type I. De plus, la méthode est puissante dans la détection des écarts au modèle de moyenne marginale qui s'applique. Des résultats d'une étude de démence servent à illustrer la méthode pour le diagnostic d'une spécification incorrecte du modèle, dans le cas de données manquantes et avec une variable auxiliaire toujours observée pour des données transversales.

S. Sinha, K. K. Saha, and S. Wang

299

Semiparametric Approach for Non-Monotone Missing Covariates in a Parametric Regression Model

Résumé. Il arrive souvent qu'il y ait des données manquantes pour des covariables dans les études biomédicales, et analyser de telles données en ne prenant pas en compte les sujets pour lesquels l'information est incomplète peut aboutir à des estimations inefficaces ou biaisées. Beaucoup d'attention a été accordée au traitement des cas avec une seule covariable manquante ou bien avec un schéma monotone de données manquantes quand les données manquent au hasard. Dans cet article nous proposons une méthode semi-paramétrique pour traiter les cas où les données manquantes sont distribuées de façon non monotone. La méthode proposée repose sur l'hypothèse que le mécanisme de l'absence ne dépend pas de la variable manquante elle-même mais peut dépendre des autres variables manquantes. Ce mécanisme est un peu moins général que le mécanisme complètement non-ignorable mais il est parfois plus flexible que le mécanisme d'absence au hasard où le mécanisme d'absence peut seulement dépendre des variables complètement observées. L'approche proposée est robuste vis-à-vis de l'erreur de spécification de la distribution des covariables manquantes, et le mécanisme proposé aide à annuler (ou réduire) les problèmes dus à la non-identifiabilité qui résulte du mécanisme d'absence non-ignorable. On en dérive les propriétés asymptotiques de l'estimateur proposé. Ses performances à distance finie sont évaluées par des études de simulations. Enfin, pour illustrer cette méthode nous analysons un ensemble de données de cancer endométrique et un ensemble de données de fracture de la hanche.

L. S. Chen, R. L. Prentice, and P. Wang

312

A Penalized EM Algorithm Incorporating Missing Data Mechanism for Gaussian Parameter Estimation

Le taux de données manquantes peut dépendre des valeurs ciblées dans de nombreux contextes, y compris les études protéomique basée sur la spectrométrie de masse. Ici, nous considérons l'estimation de la moyenne et de la covariance sous l'hypothèse d'une

distribution Gaussienne multivariée avec des données manquantes non-ignorables, incluant des scénarios dans lesquels la dimension (p) du vecteur réponse est supérieur ou égal au nombre (n) d'observations indépendantes. Une procédure d'estimation des paramètres est développée par maximisation d'une classe de fonctions de vraisemblance pénalisées qui conduit à une modélisation explicite des probabilités des données manquantes. La performance de la procédure d'estimation de l'algorithme EM pénalisé qui incorporant le mécanisme de données manquantes (PEMM) est évaluée à l'aide de simulations et illustrée sur des données de protéomique.

Z. Wen, K. H. Pollock, J. D. Nichols, P. M. Waser, and W. Cao 323

Using Imputation and Mixture Model Approaches to Integrate Multi-State Capture-Recapture Models with Assignment Information

Dans cet article nous généralisons d'abord le modèle de capture-recapture en super-population au cas d'états (sites ou sous-populations) multiples pour deux classes d'âge. L'approche proposée par Wen (2011 ; 2013) combine des données de capture-recapture avec de l'information d'affectation à un état, pour estimer les contributions relatives des naissances sur le site et de l'immigration à la croissance d'une population étudiée isolément. Ici, nous généralisons cette approche à un système composé de plusieurs populations étudiées (système multi-état) avec deux classes d'âge, et nous utilisons une méthode d'imputation pour tenir compte de l'incertitude inhérente à l'information d'affectation à une sous-population. Nous développons ensuite une autre approche par mélanges, au niveau individuel, afin d'incorporer l'information d'affectation obtenue au niveau individuel avec les données de capture-recapture. Nos analyses par simulations et sur données réelles montrent que la combinaison d'information d'affectation et de données de capture-recapture permet l'estimation du recrutement de nouveaux animaux dans le système séparément par origine, ainsi que l'estimation du processus de dispersion entre sous-populations au sein du système étudié. Par comparaison avec un modèle standard de capture-recapture, nos nouveaux modèles améliorent la justesse et la précision des estimations des paramètres démographiques, dont la probabilité de survie, la probabilité d'entrée spécifique à chaque origine, et en particulier la probabilité de mouvements entre populations.

C.-C. Wen and Y.-H. Chen 335

Semiparametric Analysis of Incomplete Current Status Outcome Data under Transformation Models

Ce travail, issu d'une étude de suivi d'ostéoporose, s'intéresse à l'analyse de régression avec des données incomplètement observées sur l'état courant. Par conséquent l'état courant, entendu comme la date d'évaluation et l'indicateur de survenue ou non de l'événement d'intérêt à la date d'évaluation, n'est pas observé pour l'ensemble des sujets. Par contre un état subrogé, pouvant donner lieu à erreur de classification est disponible pour tous les individus. Nous nous concentrons sur des modèles de régressions semi-paramétriques avec transformation, incluant les modèles de hasards proportionnels et les modèles de chances proportionnelles comme cas particuliers. Sous le principe de l'absence aléatoire (*missing at random*) où l'absence de l'état courant dépend seulement du subrogé observé et de covariables, nous proposons une approche basée sur la vraisemblance du sous-échantillon de validation où les données sont complètes, avec des

ajustements de la probabilité d'observer le résultat courant, ainsi que la distribution du subrogé dans le sous-échantillon de validation. Nous proposons un algorithme pour l'implémentation, efficient sur les aspects calculatoires, et nous obtenons la consistance et la normalité asymptotique pour inférer avec l'estimateur proposé. L'application aux données d'enquête sur l'ostéoporose et à des simulations montrent le bon fonctionnement de la vraisemblance de validation ; elle élimine le biais de l'analyse sur « cas complets » écartant les individus à données manquantes, et aboutit à une meilleure efficacité que l'analyse par pondération sur probabilité inverse.

J. Knapé, K. M. Daane, and P. de Valpine

346

Estimation of Stage Duration Distributions and Mortality under Repeated Cohort Censuses

De nombreux processus dans la nature peuvent être considérés comme résultant de la progression de sujets à travers des étapes séquentielles et peuvent être décrits par des modèles multi-étapes. Des exemples concernent le développement d'une maladie et le développement physiologique des plantes et des animaux. Nous développons un modèle multi-étapes pour des schémas d'études où un petit nombre de sujets est suivi et le nombre de sujets dans chaque état est évalué de façon répétée au cours du temps, mais pour lesquels les sujets ne peuvent être identifiés. L'exemple qui motive cette recherche concerne l'étude en laboratoire du développement d'arthropodes à travers les données de fréquence d'étapes. Notre modèle suppose que les mêmes individus sont recensés à chaque temps, introduisant des dépendances entre les échantillons. Ce type de données se rencontre souvent dans des études de laboratoire de petits arthropodes, mais leur analyse détaillée n'a reçu que peu d'attention. La vraisemblance du modèle est obtenue à partir d'un modèle stochastique de développement et de mortalité des individus dans la cohorte. Nous présentons un schéma MCMC ciblant la distribution a posteriori des temps de développement et de mortalité des individus. Il s'agit d'un nouveau type de MCMC qui utilise des propositions personnalisées pour explorer une distribution a posteriori avec support discontinu découlant du fait que les identités des individus sont inconnues. L'algorithme MCMC peut être utilisé pour l'inférence sur les paramètres régissant les distributions des durées de séjour dans les étapes et les taux de mortalité. Le modèle est illustré en ajustant le modèle de développement aux données de fréquence d'étapes d'une cohorte de cochenilles placées dans une vigne.

P. Xu, L. Zhu, and Y. Li

356

Ultrahigh Dimensional Time Course Feature Selection

Des défis statistiques proviennent d'études biomédicales modernes qui fournissent des données génomiques au cours du temps d'ultra-haute dimension. Dans une étude sur le cancer du rein qui a motivé cet article, les mesures pharmacocinétiques d'un suppresseur de tumeur (CCI-779) et les niveaux d'expression de 12625 gènes ont été mesurés sur 33 patients 8 et 16 semaines après le début des traitements, avec l'objectif d'identifier des transcrits prédictifs et les interactions avec le temps dans des cellules mononucléaires du sang périphérique pour la pharmacocinétique au cours du temps. La base de données qui en résulte défie l'analyse même par régression régularisée. Bien que des solutions aient été proposées pour les modèles linéaires et linéaires généralisés, il n'existe quasiment pas de solution dans le contexte d'une évolution au cours du temps. C'est

pourquoi, une nouvelle procédure de screening basée sur des GEE est proposée. Cette procédure dépend uniquement des spécifications des deux premiers moments marginaux et d'une structure de corrélation de travail. Cette nouvelle procédure est différente des méthodes existantes qui ajustent des modèles marginaux séparés ou calculent des mesures de corrélation deux-à-deux ; elle nécessite simplement de faire une seule évaluation des fonctions d'estimation et ainsi est computationnellement extrêmement efficace. Cette nouvelle méthode est robuste à la mauvaise spécification des structures de corrélation et bénéficie de propriétés théoriques, qui sont validées par des simulations de Monte Carlo. La procédure est utilisée pour analyser les données de l'étude sur le cancer du rein mentionnée plus haut et identifier des transcrits et des possibles interactions avec le temps qui seraient pertinents pour le métabolisme du CCI-779 dans le sang périphérique.

S. Liu, Y. Yuan, R. Castillo, T. Guerrero, and V. E. Johnson

366

Evaluation of Image Registration Spatial Accuracy Using a Bayesian Hierarchical Model

Pour évaluer l'utilité s'algorithmes d'enregistrement automatisé d'image déformable (DIR), il faut évaluer à la fois la précision de l'algorithme DIR lui-même, mais aussi la précision de l'enregistrement par des lecteurs humains à partir desquels la référence est obtenue. Nous proposons un modèle bayésien hiérarchique pour évaluer la précision spatiale des lecteurs humains et des méthodes DIR automatisées, basé sur des données d'enregistrement multiples produites par des lecteurs humains et par des méthodes DIR automatisées. Pour tenir totalement compte des localisations des repères dans toutes les images, nous traitons les vraies positions des repères comme des variables latentes, et nous imposons une structure hiérarchique sur l'amplitude des erreurs d'enregistrement observées entre paires d'images. Les erreurs d'enregistrement par DIR sont modélisées à l'aide de processus gaussiens et de densités a priori pour déterminer les matrices de covariance associées. Nous développons un algorithme échantillonneur de Gibbs pour ajuster efficacement nos modèles à des données de haute dimension, et nous appliquons la méthode proposée à l'analyse d'un ensemble de données image obtenues dans une étude cardio-thoracique 4D.

A. Kapelner and A. Krieger

378

Matching On-the-Fly: Sequential Allocation with Higher Power and Efficiency

Nous proposons une procédure d'allocation dynamique qui améliore la puissance et l'efficacité de la mesure de l'effet moyen d'un traitement dans le cadre des essais randomisés de taille fixe à allocation séquentielle. Les patients sont inclus itérativement et sont soit randomisés, soit appariés à un patient déjà randomisé à l'aide d'un critère d'appariement, et on leur donne alors le traitement alternatif. Nous développons des estimateurs de l'effet moyen du traitement combinant l'information des paires de sujets et des sujets non appariés, ainsi qu'un test exact. Des simulations montrent la meilleure efficacité et l'amélioration de puissance de notre méthode sur différentes méthodes d'allocation concurrentes, à la fois sur des données simulées et sur des données réelles d'un essai thérapeutique.

BIOMETRIC PRACTICE

Bien que l'objectif principal des études cliniques de phases précoces est d'identifier une dose acceptable en vue d'investigations plus approfondies, les études modernes de recherche de dose et leurs designs sont très spécifiques des contextes cliniques individuels. En outre, comme les méthodes de recherche de dose avec adaptation des critères d'évaluation utilisent souvent des algorithmes complexes, il est essentiel de disposer d'outils de diagnostic pour évaluer la plausibilité de la performance simulée d'une méthode et l'adéquation de l'algorithme. Dans cet article, nous proposons une technique simple qui fournit une limite supérieure, ou un indice de référence, de la précision des méthodes de recherche de dose pour un design donné. L'indice de référence proposé est non-paramétriquement optimal, dans le sens de O'Quigley, Paoletti, et Maccario (2002), et on démontre via des exemples qu'il constitue une borne supérieure de précision pour les méthodes de recherche de dose basées sur des modèles. Nous illustrons la mise en œuvre de la méthode dans un contexte d'essais de phase I qui considèrent des toxicités multiples et d'essais de phase I/II où le choix des doses est basé sur la toxicité et l'efficacité. Ensuite, nous appliquons l'indice de référence à plusieurs exemples cliniques pris dans la littérature. En comparant les caractéristiques opérationnelles d'une méthode de choix de la dose à celle de l'indice de référence, nous pouvons faire une première évaluation rapide pour savoir si la méthode est correctement calibrée et évaluer sa sensibilité aux relations dose-réponse.

Les régressions linéaires sont couramment utilisées pour étalonner les mesures du signal en analyse protéomique par spectrométrie de masse. Cependant avec ou sans transformation monotone (par ex. log), les données de ces expériences de protéomique fonctionnelles ne sont pas nécessairement une fonction linéaire ni même monotone de la concentration en protéine sauf sur une étendue restreinte. Une procédure de calcul d'efficacité de spline améliore la régression linéaire. Cependant les données de spectrométrie ne sont pas nécessairement homoscédastiques. Le plus souvent la variation des concentrations mesurées s'accroît de façon non proportionnelle au voisinage des limites de capacité des instruments de mesures (étendue dynamique), à savoir les bornes supérieures et inférieures de quantification. Ces difficultés d'étalonnage existent avec d'autres applications de spectrométrie de masse ainsi que pour d'autres étalonnages à grande échelle. En conséquence la méthode proposée utilise une approche fonctionnelle des données pour définir la courbe d'étalonnage et aussi les limites de quantification sous les deux hypothèses :

- i) la variance est une fonction convexe bornée de la concentration,
- ii) la courbe d'étalonnage elle-même est monotone au moins entre les limites de quantification, mais pas nécessairement à l'extérieur de ces limites.

Avec ce paradigme, la limite de détection qui existe mais qui n'est pas mesurable avec exactitude est ainsi définie. Une approche itérative s'appuyant sur des méthodes de lissage existantes pour tenir compte des deux restrictions est présentée pour obtenir une convergence globale et optimale sous des conditions peu restrictives. Cette approche peut être aussi mise en œuvre quand la propriété de convexité est remplacée par d'autres

restrictions (bornées).

Des exemples de Addona et al. illustrent et motivent l'efficacité de cette méthodologie sur des données fonctionnelles comparée à des régressions linéaires plus simples et aux techniques de spline.

J. Viladomat, R. Mazumder, A. McInturff, D. J. McCauley, and T. Hastie **409**
Assessing the Significance of Global and Local Correlations under Spatial Autocorrelation: A Nonparametric Approach

Nous proposons une méthode permettant de tester la corrélation entre deux champs aléatoires eux-mêmes autocorrélés. Dans ce cas de figure, l'hypothèse d'indépendance pour une paire d'observations du test standard n'est pas vérifiée, et a pour conséquence un rejet de l'hypothèse nulle dans des cas où il n'y a pourtant aucun effet (la précision de la distribution nulle est surestimée). Notre méthode recalcule la distribution nulle en prenant en compte l'autocorrélation. Elle est basée sur la méthode de Monte-Carlo : on applique successivement des permutations, un lissage et une normalisation d'une des variables pour annuler la corrélation avec la seconde, tout en maintenant l'autocorrélation initiale. Avec ce modèle de simulation, tout est basé sur l'indépendance de deux ou plusieurs champs aléatoires peut être construit. Cette recherche a été motivée par un projet en conservation de la biodiversité du département de biologie de l'Université de Standford (Californie, USA).

D. Lee, A. Rushworth, and S. K. Sahu **419**
A Bayesian Localized Conditional Autoregressive Model for Estimating the Health Effects of Air Pollution

L'estimation des effets à long-termes sur la santé de la pollution est défi, particulièrement lorsque l'on modélise l'incidence d'une maladie dans des petites zones géographiques dans les études écologiques. La difficulté vient de la structure d'autocorrélation spatiale sous-jacente non observable dans ces données, qui est prise en compte par des effets aléatoires modélisés par un modèle autorégressif conditionnel globalement lissé. Ces effets aléatoires lissés confondent les effets de la pollution atmosphérique, qui sont aussi globalement lissés. Pour éviter cette colinéarité un modèle autorégressif conditionnel Bayésienne localisé est développé pour les effets aléatoires. Ce modèle localisée spatialement flexible, dans le sens où il est non seulement capable de modéliser des zones de lissage spatial, mais aussi il est capable de capturer des changements de seuil dans la surface des effets aléatoires. Ce développement méthodologique nous permet d'améliorer la performance de l'estimation des effets des covariables par rapport à l'utilisation classique des modèles autorégressifs conditionnels. Ces résultats sont établis à l'aide d'une étude de simulation, et sont ensuite illustrés avec notre étude sur la pollution atmosphérique et les problèmes de santé respiratoire à Greater Glasgow, en Écosse en 2011. Le modèle montre des effets importants sur la santé des particules la pollution atmosphérique et du dioxyde d'azote, dont les effets ont été systématiquement atténués par les modèles lissés globaux actuellement disponibles.

J. Yun, T. Wang, and G. Xiao **430**
Bayesian Hidden Markov Models to Identify RNA-Protein Interaction Sites in PAR-CLIP

La réticulation et immunoprécipitation renforcée à ribonucléoside photoactivable (PAR-CLIP) est de plus en plus utilisée pour la cartographie globale des sites d'interaction ARN-protéines. Il existe deux caractéristiques essentielles des expériences PAR-CLIP : il est probable que les tags de lecture de séquences forment un pic enrichi autour de chaque site d'interaction ARN-protéine, et il est probable que la procédure de réticulation introduise une mutation spécifique à chaque tag de la lecture de séquences au site d'interaction. Quelques méthodes ad hoc ont été développées pour identifier les sites d'interaction ARN-protéines en utilisant soit le nombre de lectures de séquences soit le nombre des mutations. Cependant, des méthodes statistiques rigoureuses pour l'analyse des PAR-CLIP font encore défaut. Dans cette étude, nous proposons un modèle intégratif pour établir la distribution conjointe des comptes de lectures et de mutations observés. Pour localiser les sites d'interaction avec une résolution au niveau des paires de bases, nous développons une nouvelle approche de modélisation qui adopte des modèles de Markov cachés non-homogènes pour inclure la séquence de nucléotides à chaque localisation génomique. Des études de simulation et les données d'application montrent que notre méthode a de meilleures performances que les méthodes ad hoc, et fournit des inférences fiables pour les sites de liaison ARN-protéines à partir de données de PAR-CLIP.

N. A. Gouskova, S. R. Cole, J. J. Eron, and J. P. Fine

441

Viral Suppression in HIV Studies: Combining Times to Suppression and Rebound

Dans les essais cliniques sur le VIH-1, l'intérêt est souvent de comparer la façon dont les traitements suppriment l'ARN du VIH-1. La pratique actuelle dans l'analyse statistique de ces essais est de définir un seul événement composite ad hoc qui combine des informations à la fois sur la suppression de la charge virale et le rebond viral ultérieur, puis d'analyser les données en utilisant des techniques classiques d'analyse de survie univariable. La principale faiblesse de cette approche est que les résultats de l'analyse peuvent être facilement influencés par des détails mineurs dans la définition de l'événement composite. Nous proposons un critère alternatif simple basé sur la probabilité de suppression au cours du temps, et nous suggérons de résumer les différences de traitement en utilisant le temps moyen restreint qu'un patient passe dans l'état de suppression virale. Une analyse non-paramétrique est basée sur les méthodes pour études à multiples critères. Nous démontrons l'utilité de notre stratégie analytique en utilisant un essai thérapeutique récent dans lequel le protocole spécifiait une analyse primaire utilisant une approche pour critère composite.

S. R. Seaman, M. Pavlou, and A. J. Copas

449

Methods for Observed-Cluster Inference When Cluster Size is Informative: A Review and Clarifications

Les données en cluster surviennent fréquemment en épidémiologie. Nous supposons que chaque membre du cluster a une variable résultat Y et des covariables X . Quand il y a des données manquantes dans Y , la distribution de Y sachant X sur tous les membres du cluster ('clusters complets') peut être différente de la distribution sur les membres avec Y observée ('clusters observés'). Souvent, la première distribution est d'intérêt, mais lorsque les données sont manquantes car Y ne peut pas exister (e.g., qualité de vie chez une personne décédée), la deuxième distribution peut être plus pertinente (e.g., qualité de

vie conditionnellement au fait d'être en vie). Des modèles basés sur les équations d'estimation généralisées pondérée et doublement pondérées et des modèles à effets aléatoires partagés ont été proposés pour des inférences sur les clusters observés lorsque la taille de cluster est informative, à savoir, la distribution de Y sachant X dans les clusters observés dépend de la taille du cluster observé. Nous montrons que ces méthodes peuvent être finalement considérées comme produisant des inférences sur les clusters complets et ne peuvent donc pas produire des inférences sur les clusters observés. Cela reste valable même si les clusters observés sont intrinsèquement complets plutôt que de n'être que la partie observée de clusters complets plus grands: dans ce cas, les méthodes peuvent décrire des clusters complets imaginaires plutôt que des clusters observés. Nous montrons dans quelles conditions des modèles à effets aléatoires partagés proposés pour des inférences sur les clusters observés décrivent finalement des membres avec Y observée. Un jeu de données dans l'arthrite psoriasique est utilisé pour illustrer le danger de mauvaise interprétation des estimations des modèles à effets aléatoires partagés.

P. Fearnhead, V. Giagos, and C. Sherlock

457

Inference for Reaction Networks Using the Linear Noise Approximation

Nous nous intéressons à l'inférence sur les taux de réaction dans des réseaux observés à temps discret comme ceux rencontrés dans les modèles pour la biologie des systèmes, l'écologie des populations et les épidémies. La plupart de tels réseaux ne sont ni suffisamment lents, ni suffisamment petits pour permettre une inférence au travers d'un véritable processus de Markov à saut, état-dépendant. Typiquement, l'inférence est menée en approximant la dynamique au travers d'une équation différentielle ordinaire (ODE) ou d'une équation différentielle stochastique (SDE). Dans le premier cas, on ignore la trait stochastique du véritable modèle, ce qui peut conduire à des inférences inappropriées. Dans le second cas, plus approprié, la difficulté est son implémentation car la densité de transition du modèle SDE est en général inconnue. L'approximation par bruit linéaire (LNA) est un développement de Taylor au premier ordre de l'approximation de la SDE par une solution déterministe. On peut la voir comme un compromis entre les modèles ODE et SDE. C'est un modèle stochastique, mais des probabilités de transition en temps discret pour le LNA sont disponibles à travers la solution d'une série d'équations différentielles ordinaires. Nous décrivons comment la réinitialisation du LNA peut être efficacement utilisée pour inférer dans un cadre général des réseaux de réaction et pour l'évaluation de la précision d'une telle approche, et nous montrons comment et quand cette approche est plus efficiente que les méthodes ODE ou SDE sur le plan statistique ou sur le plan calculatoire. Nous appliquons le LNA à l'analyse des tendances sur les données de grippe recueillies par Google sur les îles du Nord et du Sud de la Nouvelle Zélande, pour lesquelles nous obtenons des prévisions à court-terme plus précises des nouveaux cas de grippe par rapport à une autre méthode récemment proposée, mais au prix d'un coût en calcul plus élevé.