Translations of Abstracts

BIOMETRIC METHODOLOGY

Nous étudions la problématique de l'estimation paramétrique de la fonction d'incidence cumulative (CIF) de risques compétitifs en présence de données censurées par intervalle. Nous adaptons les modèles paramétriques existants estimant la CIF en présence de risques compétitifs dans le cas de données censurées à droite à la problématique de la censure par intervalle. Des estimateurs par maximum de vraisemblance de la CIF sont proposés sous ces hypothèses, suite à nos travaux antérieurs d'estimation non paramétrique. Un estimateur de la vraisemblance naïf et simple est également présenté, n'utilisant qu'une partie des données observées. L'estimateur naïf permet des estimations séparées modélisant chaque risque compétitif, contrairement au modèle complet de maximisation de la vraisemblance dans lequel tous les modèles sont estimés conjointement. Le modèle naïf se révèle valide dans le cas où le processus de censure par intervalle est mixte (c'est-à-dire variable d'un patient à l'autre), mais pas dans le cas d'un processus d'inspection indépendant, contrairement au modèle complet qui reste valide dans les deux types de censure par intervalle. Dans des simulations, nous montrons que l'estimateur naïf a de bonnes performances et a une efficacité comparable à celle du modèle complet dans un certain nombre de situations. Ces méthodes sont appliquées aux données d'un vaste essai thérapeutique randomisé récent de prévention de la transmission du HIV de la mère à l'enfant.

Cet article vise l'estimation d'une mesure d'association dépendant du temps pour des temps d'échecs bivariés, le rapport de risques conditionnel spécifique d'une cause (CCSHR), qui est une généralisation du rapport de risques conditionnel (CHR) pour une adaptation à des données de risques compétitifs. Nous modélisons le CCSHR comme une fonction de régression paramétrique du temps et des causes d'évènements, et laissons tous les autres aspects de la distribution jointe des temps d'échecs non spécifiés. Nous développons une procédure d'estimation à base de pseudo-vraisemblance pour l'ajustement du modèle et l'inférence, et établissons les propriétés asymptotiques des estimateurs. Nous évaluons les propriétés non asymptotiques des estimateurs proposés en comparaison des estimateurs obtenus au moyen d'équations d'estimation basées sur les moments. Les données de l'étude Cache County sur la démence sont utilisées pour illustrer la méthodologie proposée.

Samiran Sinha and Yanyuan Ma **21**
A Semiparametric Analysis of Linear Transformation Models with Covariate Measurement Errors

Nous adoptons une approche semi-paramétrique pour l'ajustement d'un modèle de transformation linéaire à des données censurées à droite lorsque les variables prédictives sont sujettes à des erreurs de mesure. Nous construisons des équations d'estimation consistantes lorsque des mesures répétées d'une variable subrogée à un prédicteur non observé sont disponibles. L'approche proposée s'applique sous des conditions minimales de distribution de la véritable covariable ou des erreurs de mesure. Nous obtenons les propriétés asymptotiques de l'estimateur, et nous illustrons ses caractéristiques pour des échantillons de taille finie par des études en simulation. Nous appliquons la méthode à l'analyse de données d'un essai clinique sur le SIDA, à l'origine de ce travail.

Kiranmoy Das and Michael J. Daniels **33**
A Semiparametric Approach to Simultaneous Covariance Estimation for Bivariate Sparse Longitudinal Data

L'estimation de la structure de covariance pour des données longitudinales éparses et irrégulières a été étudiée par de nombreux auteurs ces dernières années, mais en utilisant des spécifications paramétriques complètes. De plus, lorsque les données sont collectées à partir de plusieurs groupes au cours du temps, il est bien connu que de supposer la même matrice de covariance ou des matrices totalement différentes pour les groupes peut conduire à une perte d'efficience et/ou à des biais. Des approches non-paramétriques ont été proposées pour estimer la matrice de covariance pour des données longitudinales régulières univariées en partageant l'information entre les groupes étudiés. Pour le cas irrégulier, avec des mesures longitudinales bivariées ou multivariées, la modélisation devient plus difficile. Dans cet article, pour modéliser des données longitudinales bivariées éparses provenant de plusieurs groupes, nous proposons une structure flexible de covariance au travers d'une nouveau processus matriciel de type « stick-breaking » pour la structure de covariance résiduelle, et un processus de Dirichlet, mélange de gaussiennes pour les effets aléatoires. Des études de simulations sont réalisées pour étudier l'efficacité de cette approche par rapport aux approches plus traditionnelles. Nous analysons également un sous-ensemble de données issu de l'étude de Framingham pour examiner comment les trajectoires de pression artérielle et les structures de covariance changent entre les patients des différents groupes BMI (ou IMC, indice de masse corporelle) à leur valeur de base (haute, moyenne, élevée).

John M. Neuhaus, Alastair J. Scott, Christopher J. Wild, Yannan Jiang, Charles E. McCulloch, and Ross Boylan **44**
Likelihood-Based Analysis of Longitudinal Data from Outcome-Related Sampling Designs

Les enquêteurs rassemblent généralement des données longitudinales pour évaluer des changements de réponses au cours du temps et lier ces changements aux changements intra-sujets dans les prédicteurs. Avec des variables à expliquer rares ou chères comme des maladies rares et des mesures radiologiques coûteuses, dépendantes de la variable à expliquer, et plus généralement liées à la variable à expliquer, les plans d'échantillonnage peuvent améliorer l'efficacité de l'estimation et réduire le coût. Le suivi longitudinal de

sujets rassemblés dans un échantillon initial lié à la variable à expliquer peut alors être utilisé pour étudier les trajectoires de réponses au cours du temps et évaluer l'association de changements dans les prédicteurs intra-sujets avec le changement dans la réponse. Dans cet article nous développons deux approches basées sur la vraisemblance pour l'ajustement de modèles linéaires généralisés à effets mixtes (GLMMs) aux données longitudinales d'une large variété de plans d'échantillonnage liés à la variable à expliquer. La première est une extension de l'approche du maximum de vraisemblance semi-paramétrique développée dans Neuhaus et al. (2002, 2006) et s'applique tout à fait généralement. La deuxième approche est une adaptation des méthodes de vraisemblance conditionnelle standard et est limitée aux modèles à constante aléatoire avec une fonction de lien canonique. Les données d'une étude sur le déficit d'attention chez les enfants hyperactifs motivent le travail et illustre les résultats.

Eric B. Laber, Daniel J. Lizotte, and Bradley Ferguson **53**
Set-Valued Dynamic Treatment Regimes for Competing Outcomes

Les régimes dynamiques de traitement rendent opérationnel le processus de décision clinique comme une séquence de fonctions, une pour chaque décision clinique, où chaque fonction fait correspondre l'information actualisée du patient à un unique traitement recommandé. Les méthodes actuelles pour estimer des régimes dynamiques de traitement optimaux, par exemple l'apprentissage par renforcement de type Q, requiert la spécification d'une variable de réponse unique pour laquelle on mesure l'« adéquation » de régimes dynamiques de traitement concurrents. Cependant, ceci constitue une sur-simplification de l'objectif de la prise de décision clinique qui a pour but de mettre en balance plusieurs réponses potentiellement concurrentes, par exemple l'apaisement des symptômes et le poids des effets secondaires. Quand il y a des réponses concurrentes et que les patients ne connaissent pas ou ne peuvent communiquer leurs préférences, établir une variable de réponse composite unique qui équilibre correctement les réponses concurrentes n'est pas possible. Ce problème survient également quand les préférences du patient évoluent dans le temps. Nous proposons une méthode pour construire des régimes dynamiques de traitement qui s'adaptent aux réponses concurrentes en recommandant des ensembles de traitements à chaque point de décision. Formellement, nous construisons une séquence de fonctions à valeurs multiples qui prennent en entrée l'information actualisée du patient et fournissent en sortie un sous-ensemble recommandé de traitements possibles. Pour une histoire de patient donnée, l'ensemble de traitements recommandé contient tous les traitements produisant des vecteurs de réponse non inférieurs. La construction de ces fonctions plurivalentes nécessite de résoudre un problème d'énumération non trivial. Nous proposons un algorithme d'énumération exact en reformulant le problème comme un programme linéaire mixte en nombres entiers. Les méthodes proposées sont illustrées à l'aide de données de l'étude CATIE sur la schizophrénie.

M.J. Daniels, C. Wang, and B.H. Marcus **62**
Fully Bayesian Inference under Ignorable Missingness in the Presence of Auxiliary Covariates

Afin de rendre plus réaliste l'hypothèse Missing At Random (MAR) comme quoi les données manquantes sont « ignorables », on s'appuie souvent sur des covariables auxiliaires dont on ne souhaite cependant pas qu'elles figurent dans le modèle

d'inférence. Or, les méthodes courantes d'imputation multiple utilisant ces covariables auxiliaires ne prétendent pas permettre de retomber sur le modèle marginal d'inférence. Modèle d'imputation et modèle d'inférence, pourrait-on dire, sont « en bisbille » (‘*uncongenial’*, Meng, 1994). Pour les rendre compatibles, il faudrait éviter d'utiliser un modèle paramétrique pour la distribution marginale des covariables auxiliaires, tout en notant qu'en général, il n'y a pas non plus assez de données pour obtenir une estimation non paramétrique correcte de la distribution conjointe. Qui plus est, si le modèle d'imputation s'écrit avec une fonction de lien non linéaire (par exemple, le lien logistique pour une réponse binaire), le recours aux variables auxiliaires pour dériver le modèle marginal d'inférence rend difficilement interprétable l'effet des covariables dans le modèle d'inférence. Le présent article propose une approche entièrement bayésienne qui garantit la compatibilité des modèles dans le cas de données longitudinales incomplètes en incorporant un modèle d'inférence interprétable dans le modèle d'imputation, prenant ainsi en compte les deux complications décrites précédemment. Cette approche est évaluée à l'aide de simulations et est appliquée sur un essai clinique récent.

Inspiré par des exemples d'études d'associations en génétique, cet article s'intéresse au problème de sélection de modèle dans le cadre général des systèmes complexes de modèles linéaires et d'une approche bayésienne. Nous évoquons les problèmes de formulation de la sélection de modèle et de l'incorporation d'une information a priori contexte-dépendante au travers de différents niveaux de spécifications d'a priori. Nous obtenons des facteurs de Bayes analytiques et leurs approximations pour faciliter la sélection de modèle, et nous discutons leurs propriétés théoriques et calculatoires. Nous mettons en évidence la valeur de notre approche bayésienne basée sur un algorithme MCMC implémenté au travers de simulations et d'une application sur des données réelles de cartographie pour des eQTL tissu-spécifiques. Nos résultats sur les facteurs de Bayes aboutissent à un cadre général pour effectuer des comparaisons efficientes dans les systèmes complexes de modèles linéaires.

L'analyse par enrichissement fonctionnel est usuellement conduite sur des données génomiques de haut-débit pour permettre l'interprétation fonctionnelle d'une liste de gènes ou de protéines partageant une même propriété, par exemple la propriété d'être différentiellement exprimé(e)s (DE). La *p*-valeur issue d'un test d'hypothèse basé sur une loi hypergéométrique est couramment utilisée pour étudier le sur-enrichissement - dans cette liste de gènes différentiellement exprimés - de gènes correspondant à des termes prédéfinis, par exemple des termes d'ontologie génomique (GO). La *p*-valeur hypergéométrique est sujette à trois limitations : 1) elle est calculée indépendamment pour chaque terme, sans prendre en compte la dépendance biologique entre les différents termes ; 2) elle est contrainte par le nombre de variables contenues dans chaque terme (la taille), ce qui a pour conséquence un biais de sélection des termes moins spécifiques ; 3) l'application de la « règle du vrai chemin » (*True path rule*) provoque l'utilisation d'information redondante entre catégories chevauchantes. Nous proposons une approche

bayésienne basée sur un modèle hypergéométrique non central. La structure de dépendance du modèle GO est incorporée dans le modèle à travers un *a priori* sur les paramètres de décentrage.la fonction de vraisemblance ne contient pas d'information redondante. L'inférence de l'enrichissement d'une catégorie est basée sur les probabilités *a posteriori* qui ne sont pas contraintes par une information sur la taille. Notre méthode permet de détecter des signaux d'enrichissement modérés mais consistants et d'identifier des ensembles de termes fonctionnels proches les uns des autres et biologiquement sensés plutôt que des termes isolés. Nous décrivons également des idées de base pour des hypothèses et des implémentations de différentes méthodes afin de donner un aperçu théorique de notre approche, le tout est démontré à l'aide de résultats sur des données simulées. Nous présentons également une application sur données réelles.

Dans un cadre Bayésien nous développons une modélisation par mélange non-paramétrique pour les essais en tout ou rien. Nous adoptons comme prior un mélange non paramétrique structuré qui impose la monotonicité de la courbe dose-réponse. On porte un intérêt particulier à la quantification d'un risque clé dans la détermination d'un niveau de dose fournissant un niveau de réponse spécifié. La méthodologie proposée donne une façon adaptable de faire des inférences, notamment sur la relation dose-réponse. Les résultats sont illustrés par deux exemples tirés de la littérature.

Cet article propose une nouvelle approche par test multiple pour l'estimation de la dose minimale effective permettant des courbes de dose-réponse non monotones. L'approche présentée combine les avantages de deux méthodes communément utilisées. Nous montrons que cette nouvelle approche contrôle le taux d'erreur de la sous-estimation de la dose minimale effective réelle. Des simulations de Monte Carlo indiquent que la méthode proposée surpasse les méthodes alternatives dans de nombreux cas, et est à peine marginalement inférieure dans les autres situations.

Nous proposons un nouveau critère de sélection de variables conçu pour une utilisation avec des algorithmes de sélection en avant: le critère du score d'information (SIC). Ce critère est basé sur un score statistique qui prend en considération les réponses corrélées. Le principal avantage du SIC est qu'il est beaucoup plus rapide à calculer que les critères de sélection existants lorsque le nombre de variables prédictives ajoutées dans le modèle est important. La raison est que le SIC peut être calculé pour tous les modèles candidats sans avoir besoin de les ajuster. Un deuxième avantage est qu'il tient compte de la corrélation entre les variables dans sa quasi-vraisemblance, ce qui induit des meilleures propriétés que les autres critères de sélection. La consistance et les propriétés de

prédiction du SIC sont étudiées. En particulier, nous avons mené des études de simulation pour évaluer les performances de sélection et de prédiction, et nous comparons celles-ci, ainsi que les temps de calcul, aux critères habituels de sélection de variables. Nous appliquons le SIC sur des données recueillies sur les arthropodes en considérant la sélection de variables sur un grand nombre de termes d'interactions constitués de traits d'espèces et de covariables environnementales.

Jian Zhang, Chao Liu, and Gary Green **121**
Source Localization with MEG Data: A Beamforming Approach Based on Covariance Thresholding

La reconstruction des activités neurales utilisant une sonde non envahissante hors du cerveau est un problème inverse malade-posé puisque les mesures observées de la sonde pourraient résulter d'un nombre infini de sources neuronales possibles. La cartographie de faisceaux électroniques basée sur la covariance représente une solution populaire et simple au problème ci-dessus. Dans cet article, nous proposons une famille de filtres de faisceaux électroniques à l'aide du seuillage de la covariance. Une théorie générale est développée sur la façon dont leurs dimensions spatiales et temporelles déterminent leur représentation. Des conditions sont données pour le taux de convergence de l'estimation associée des filtres de faisceaux électroniques. Les implications de la théorie sont illustrées par des simulations et une analyse de données réelles.

F. S. Nathoo, A. Babul, A. Moiseev, N. Virji-Babul, and M. F. Beg **132**
A Variational Bayes Spatiotemporal Model for Electromagnetic Brain Mapping

Dans cet article nous présentons une nouvelle approche bayesienne variationnelle pour résoudre le problème inverse neuroélectromagnétique dans des études impliquant l'électroencéphalographie (EEG) et la magnétoencéphalographie (MEG). Ce problème d'estimation spatiotemporelle de grande dimension implique la récupération de l'activité neurale variant dans le temps en un grand nombre de positions à l'intérieur du cerveau, à partir de signaux enregistrés en un nombre relativement faible de positons externes ou près du cuir chevelu. Formulant ce problème à l'intérieur du contexte de la sélection spatiale de variable d'un modèle linéaire fonctionnel indéterminé, nous proposons une formulation spatiale de mélange dans laquelle le profil de l'activité électrique dans le cerveau est représenté par une position spécifique basée sur une spécification spatiale logistique. Les spécifications a priori adaptent le regroupement spatial dans l'activation de cerveau, alors qu'elles tiennent également compte de l'inclusion d'information auxiliaire provenant de modalités alternatives de représentation, telles que la représentation de résonance magnétique fonctionnelle (fMRI). Nous développons une nouvelle approche bayesienne variationnelle pour calculer des estimations de l'activité de la source neurale et nous incorporons un bootstrap nonparamétrique de l'intervalle d'estimation. La méthodologie proposée est comparée à plusieurs approches alternatives par des études de simulation, et appliquée à l'analyse d'une image neurale en examinant la réponse neurale de perception du visage utilisant EEG, MEG et fMRI.

BIOMETRIC PRACTICE

Mireille E. Schnitzer, Erica E. M. Moodie, Mark J. van der Laan, Robert W. Platt, and Marina B. Klein **120824P**

Modeling the Impact of Hepatitis C Viral Clearance on End-Stage Liver Disease in an HIV Co-Infected Cohort with Targeted Maximum Likelihood Estimation

Malgré un traitement actuel efficace contre le VIH, la co-infection par le virus de l'hépatite C (HCV) est associée à un risque élevé de progression vers un stade terminal de la maladie hépatique (ESLD), devenu la première cause de décès dans cette population. La question clinique est de déterminer l'effet de la clairance du HCV sur le risque de ESLD. Nous étudions, dans cette étude de cas, si la clairance du HCV influence le risque de ESLD, en utilisant les données de l'Etude de Cohorte multicentrique Canadienne sur la co-infection. L'analyse de survie est compliquée par la nature dépendante du temps des données, la présence de facteurs de confusion initiaux, des perdus de vue, et des facteurs de confusion variant au cours du temps ; tous ces paramètres peuvent dissimuler l'effet causal d'intérêt. Le manque de certaines variables non-censurées et les événements rares sont des défis supplémentaires.
Afin d'estimer correctement les probabilités de survie sans ESLD, dans le cadre d'une histoire spécifique de clairance du HCV, nous démontrons les propriétés de la méthode semi-paramétrique, efficace et doublement robuste de l'Estimation Ciblée par le Maximum de Vraisemblance (TMLE). Des modèles structuraux marginaux (MSM) peuvent être utilisés pour modéliser l'effet de la clairance virale (exprimé comme un rapport de risque) sur la survie sans ESLD ; et nous mettons en évidence une façon d'estimer avec le TMLE, les paramètres d'une régression logistique pour la fonction de risque. Nous étudions les courbes d'influence pour les paramètres de deux modèles MSM différents, et comment elles peuvent permettre d'avoir des approximations des variances des paramètres estimés. Enfin les données ont été analysées pour évaluer le rôle de l'HCV sur ESLD, avec des imputations multiples pour prendre en compte les données manquantes non monotones.

Jing Qian, Seyedmehdi Payabvash, André Kemmling, Michael H. Lev, Lee H. Schwamm, and Rebecca A. Betensky **153**
Variable Selection and Prediction Using a Nested, Matched Case-Control Study: Application to Hospital Acquired Pneumonia in Stroke Patients

Les designs cas-témoins appariés sont couramment utilisés dans les études épidémiologiques pour leur plus grande efficacité. Ces designs ont été récemment introduits dans le contexte de l'imagerie moderne et des études génomiques, qui se caractérisent par des covariables de grande dimension. Toutefois, les analyses statistiques appropriées qui ajustent l'appariement n'ont pas encore été largement adoptées. Une étude appariée cas-témoin de 430 patients victimes d'AVC ischémiques aigus a été menée au Massachusetts General Hospital (MGH) afin d'identifier les régions cérébrales spécifiques de l'infarctus aigu associées avec une pneumonie nosocomiale (HAP) chez ces patients. Il y a 138 régions du cerveau dans lesquelles un accident vasculaire a été mesuré, qui introduisent près de 10.000 interactions bidirectionnelles, et impactent l'analyse statistique. Nous étudions les approches de régression logistique conditionnelle et inconditionnelle pénalisées dans cette problématique de sélection de variables qui différencient correctement les effets principaux et les interactions, et qui tiennent compte de l'appariement. Cette étude de neuroimagerie est extraite d'une étude prospective plus large (HAP) chez 1915 patients victimes d'AVC au MGH, qui comporte des variables cliniques, mais pas de neuro-imagerie. Nous montrons comment la grande étude, avec celle imbriquée, nous donne la possibilité de dériver un score de prédiction de l'HAP

chez les futurs patients victimes d'AVC sur la base de données d'imagerie et cliniques. Nous évaluons les méthodes proposées via des études de simulation et nous les appliquons à l'étude MGH HAP.

Adam A. Szpiro, Lianne Sheppard, Sara D. Adar, and Joel D. Kaufman **164**
Estimating Acute Air Pollution Health Effects from Cohort Study Data

Les études traditionnelles des effets à court terme de la pollution de l'air sur la santé utilisent des données de séries temporelles tandis que les études de cohorte se concentrent généralement sur les effets à long terme. L'exploitation de données individuelles de cohorte pour évaluer les effets sur la santé à court terme suscite un intérêt croissant afin de comprendre les mécanismes et les échelles temporelles d'action. Nous étendons les méthodes de régression semi-paramétriques utilisées pour ajuster sur des facteurs de confusion non mesurés dans les études de séries temporelles au cadre de la cohorte.
Les méthodes des séries temporelles ne sont pas directement applicables puisque les données de cohorte sont typiquement recueillies au cours d'une période de temps pré-spécifiée et incluent des mesures d'exposition à des jours sans observations de santé. En conséquence, la théorie asymptotique pour une longue période de temps n'est pas appropriée et il est possible d'améliorer l'efficacité en exploitant les données d'exposition supplémentaires. Nous montrons que la flexibilité du modèle d'ajustement semi-paramétrique devrait correspondre à la complexité de la tendance dans la réponse de santé, en contraste avec le cadre des séries temporelles où il lui suffit de correspondre à la structure temporelle de l'exposition. Nous démontrons aussi qu'ajuster préalablement les expositions concomitantes des mesures de santé en utilisant des tendances dans la série temporelle complète des expositions conduit à une estimation non biaisée des effets sur la santé et peut améliorer l'efficacité sans ajustement supplémentaire sur des facteurs de confusion.
Une étude publiée récemment a mis en évidence une association entre exposition à court terme aux matières particulaires fines ($PM_{2.5}$) ambiantes et diamètre rétinien artériel mesuré par photographie rétinienne dans l'étude multiethnique sur l'athérosclérose (MESA). Nous réanalysons les données de cette étude afin de comparer les méthodes décrites ici et nous évaluons nos méthodes dans une étude de simulation s'appuyant sur les données MESA.

Paul S. Albert, Aiyi Liu, and Tonja Nansel **175**
Efficient Logistic Regression Designs Under an Imperfect Population Identifier

Cet article s'intéresse à la conception de plans d'échantillonnage efficaces en régression logistique, quand la population est identifiée à l'aide d'un test diagnostique binaire entaché d'erreur, une situation courante. Nous considérons le cas où le résultat du test imparfait est connu pour l'ensemble des participants alors que la réponse au test de référence n'est connue que pour un sous échantillon sélectionné. Nous évaluons le plan optimal pour l'estimateur du maximum de vraisemblance en termes de sélection de l'échantillon et de vérification. Nous montrons qu'un gain substantiel d'efficacité peut découler de la sélection d'un pourcentage limité de sujets jugés négatifs par le test imparfait (par exemple en vérifiant 90% des sujets jugés positifs. Nous montrons également que dans certaines situations, un schéma d'étude en deux phases peut représenter une bonne alternative à un schéma fixe. Dans le cadre de schémas optimaux ou presque optimaux, nous comparons à l'aide de simulations les performances des

estimateurs efficaces du maximum de vraisemblance ou semiparamétrique lorsque le modèle est correct ou mal spécifié. La méthodologie est illustrée avec les données d'un essai d'intervention comportementale en diabétologie.

Les investigateurs d'essais ont souvent comme intérêt principal l'estimation de la courbe de survie dans une population pour laquelle il existe des informations historiques acceptables dont on peut tirer de la puissance. Cependant, tirer de la puissance de données historiques qui ne sont pas échangeables avec celles de l'essai en cours peut produire des conclusions biaisées. Dans cet article nous proposons une méthode semiparamétrique complètement Bayesienne qui a pour objet d'atténuer les biais et d'augmenter l'efficacité quand on modélise conjointement des temps d'événements provenant de deux sources d'information éventuellement non échangeables. Nous illustrons le mécanisme de nos méthodes en les appliquant à une paire de fichiers de données de surveillance post-marketing d'effets adverses chez des personnes sous dialyse, qui sont porteuses soit d'un stent de métal nu, soit d'un stent imprégné de médicament, implanté au cours d'une chirurgie de revascularisation cardiaque. Nous terminons par une discussion des avantages et des limitations de cette approche de synthèse de preuves, ainsi que par des indications pour de futurs travaux dans ce domaine. Le matériau complémentaire à l'article contient des simulations pour montrer les propriétés de biais, d'erreur quadratique moyenne, et de probabilité de couverture de nos procédures dans des conditions variées.

L'adjonction d'un polynôme de Bernstein transformé aux familles paramétriques standards, comme celle de Weibull ou log-logistique, est proposée pour les modèles à temps accélérés. Cette famille de polynômes offre une façon pratique de créer une loi a priori non paramétrique pour les densités lissées, cumulant les avantages de la méthode paramétrique mais aussi de la méthode non paramétrique, ce qui assouplit les approches d'estimation standard.
Par exemple, les méthodes d'optimisation implémentées dans les logiciels SAS ou R permettent ainsi d'obtenir la loi a posteriori ainsi que la matrice de variance-covariance asymptotique. Cette nouvelle loi a priori non paramétrique est utilisée dans les modèles à temps accélérés, ce qui a également été généralisé aux variables dépendantes du temps.
L'approche proposée se comporte nettement mieux que les approches précédentes au vu des simulations ; des données concernant l'efficacité de polymères biodégradables de la carmustine sur les gliomes cérébraux en rechute sont également étudiées.

Les études épidémiologiques impliquant des bio-marqueurs sont souvent entravées par des tests de laboratoire de coûts prohibitifs. Il a été montré que l'assemblage stratégique de spécimens comme préalable à la réalisation de ces tests de laboratoire pouvait réduire le coût avec une perte d'information minimale dans un cadre de régression logistique. Lorsque l'objectif est de réaliser une régression dont le résultat est donné par un bio-marqueur continu, l'analyse de régression de spécimens assemblés pouvait ne pas être adaptée, particulièrement lorsque cette variable de résultat est dissymétrique à droite. Dans de tels cas, nous montrons qu'une légère modification du modèle de régression linéaire multiple standard pour des données assemblées peut donner des estimations valides et précises des coefficients à condition que les assemblages soient formés en combinant les bio-spécimens de sujets possédant les mêmes valeurs de covariables. Lorsque ces assemblages x-homogènes ne peuvent être formés, nous proposons un algorithme MCEM (Monte Carlo Espérance Maximisation) pour calculer les estimations au maximum de vraisemblance. Des études par simulation montrent que ces méthodes analytiques fournissent principalement des estimations non biaisées des coefficients ainsi que leurs erreurs standard lorsque les hypothèses appropriées sont remplies. De plus, nous montrons comment utiliser la totalité des données des covariables observées pour renseigner la stratégie d'assemblage, permettant ainsi un niveau élevé d'efficacité statistique pour une fraction seulement du coût total de laboratoire.

Après avoir établi l'intérêt d'un marqueur diagnostique quantitatif, les chercheurs se posent classiquement la question de la définition du seuil à utiliser pratiquement pour porter un diagnostic. Dans le contexte de l'analyse des courbes ROC, le critère le plus couramment utilisé pour déterminer ce seuil est la maximisation de l'indice de Youden. Les performances du marqueur diagnostique sont quantifiées par la sensibilité et la spécificité observées au seuil choisi. Pour obtenir les intervalles de confiance de cette sensibilité et de cette spécificité, on suppose généralement qu'il s'agit de variables binomiales indépendantes puisqu'elles sont observées sur des populations distinctes, les sujets sains et les malades. Mais le seuil défini par l'indice de Youden est estimé à partir des données et par conséquent la sensibilité et la spécificité qui lui sont associées sont corrélées. Cette corrélation doit être prise en compte dans le calcul des intervalles de confiance. C'est ce que nous faisons dans cet article en envisageant des approches paramétrique et non paramétrique. Les simulations réalisées sous différents scénarios conduisent à des pourcentages de recouvrement corrects pour les intervalles fournis par ces deux approches. Nous observons qu'une approche paramétrique utilisant une normalisation par transformation de Box et Cox donne souvent de bons résultats. Lorsque la distribution du biomarqueur est trop complexe, une procédure non paramétrique qui estime cette distribution par logspline est également disponible.

Nous étudions les différences entre les analyses en composantes indépendantes (ICA),

découlant de différentes hypothèses, de mesures de dépendance et des valeurs initiales des algorithmes. ICA est une méthode populaire avec diverses applications, y compris l'élimination des artefacts dans les données d'électrophysiologie, l'extraction de caractéristiques dans les données de biopuces , et l'identification des réseaux cérébraux dans l'imagerie par résonance magnétique fonctionnelle (fRMI). ICA peut être considéré comme une généralisation de l'analyse en composantes principales (ACP) qui tient compte des corrélations croisées d'ordre supérieur. Alors que la solution de l'ACP est unique, il existe de nombreuses méthodes ICA dont les solutions peuvent être différentes. Infomax, FastICA , et Jade sont couramment appliquées aux études d'fRMI, FastICA étant sans doute la plus populaire. Hastie et Tibshirani (2003) ont démontré que ProDenICA est plus performante que FastICA dans les simulations avec deux composantes. Nous introduisons l'application de ProDenICA à des simulations avec plus de composantes et des données d'fIRM. ProDenICA était plus adapté dans les simulations, et nous avons identifié des différences entre IC biologiquement significatives à partir de ProDenACI en comparaison avec d'autres méthodes dans l'analyse fIRM. Les méthodes ICA nécessitent une optimisation non convexe, mais les pratiques actuelles ne reconnaissent pas l'importance des valeurs initiales, ni suffisamment
face à la sensibilité aux valeurs initiales, Nous avons constaté que des optimums locaux a conduisent à des estimations considérablement différentes à la fois des simulations et groupe ICA de l'IRMf, et nous apportons la preuve que l'optimum global à partir de ProDenICA est la meilleure estimation. Nous avons introduit une modification à l'algorithme hongrois (Kuhn- Munkres) pour accorder les IC à partir d'estimations multiples, gagnant ainsi de nouvelles perspectives sur la façon dont les réseaux du cerveau varient dans leur sensibilité aux valeurs initiales et la méthode ICA.

Ander Wilson, David M. Reif, and Brian J. Reich **237**
Hierarchical Dose-Response Modeling for High-Throughput Toxicity Screening of Environmental Chemicals

Le balayage à haut débit (HTS) en chimie environnementale est utilisé pour identifier les produits à haut potentiel de réactions adverses en santé humaine et pour l'environnement, parmi des milliers de substances non encore testées. La prévision d'activités significatives sur le plan physiologique avec des données HTS nécessite l'estimation de la réponse pour un grand nombre de produits au travers d'une batterie de tests de balayage basés sur des données de dose-réponse éparses sur toutes les combinaisons produit-test. Beaucoup de méthodes standard de type dose-réponse sont inappropriées car elles traitent chaque courbe isolément et sont sous-performantes lorsque l'on a un nombre aussi petit que six à dix points par courbe. Nous proposons un modèle semi-paramétrique bayésien qui gagne en force au travers des produits et des tests. Notre méthode paramètre directement l'efficacité et la force des produits, aussi bien que la probabilité de réponse. Notre démarche est motivée par les données ToxCast de l'Agence de Protection de l'Environnement Américaine (EPA). Nous montrons, par une étude de simulation, que notre méthode hiérarchique conduit à des estimations plus précises de la probabilité de réponse, de l'efficacité, et de la puissance par rapport à des estimations sur courbes séparées. Nous utilisons notre méthode semi-paramétrique pour comparer l'efficacité de produits des données ToxCast et de produits de référence spécifiques sur les tests basés sur les récepteurs α à l'œstrogène (ERα) et sur le récepteur γ activé par les proliférateurs

de peroxisomes (PPARγ), puis estimer la probabilité que d'autres produits soient actifs à des concentrations moindres que celles des produits de référence.

Daniel Gerhard, Melanie Bremer, and Christian Ritz **247**
Estimating Marginal Properties of Quantitative Real-Time PCR Data Using Nonlinear Mixed Models

Nous proposons un cadre de modélisation unifié sur un ensemble de modèles non linéaires mixtes pour la modélisation de l'expression génique dans des expériences de PCR (Réaction en chaîne par polymérase) en temps réel. Nous nous concentrons sur l'estimation marginale ou sur l'ensemble de la population des paramètres suivants : cycle seuil et $\Delta\Delta c(t)$, tout en conservant la structure du modèle mixte conditionnel afin de refléter de manière adéquate le plan expérimental. De plus, le calcul de l'estimateur du modèle moyen permet d'incorporer l'incertitude sur la sélection du modèle. Notre méthode est appliquée à l'estimation de l'expression différentielle du gène du transporteur de phosphate OsPT6 dans le riz en comparaison avec l'expression d'un gène de référence à différents stades lors du processus de réapprovisionnement en phosphate. Nous comparons les performances de la méthode proposée à celles d'une méthode standard dans une petite étude sur des données simulées.