Translations of Abstracts

BIOMETRIC METHODOLOGY

Dans le but d'analyser des études dans lesquelles des échantillonnages de fragments d'ADN sont réalisés sur le terrain, nous décrivons un modèle général de capture-recapture adapté au cas où les échantillons sont obtenus un par un en temps continu. Notre modèle est basé sur l'échantillonnage poissonnien dans le cas où le moment de l'‘échantillonnage peut être ne pas être toujours observé. Nous montrons que les modèles précédemment décrits correspondent à des vraisemblances partielles d'un modèle de Poisson et que leur utilisation peut être justifiée par des arguments concernant le caractère *S*-ancillaire et ancillaire au sens bayésien de l'information non prise en compte. Nous démontrons un autre lien avec les modèles de capture-recapture en temps continu et expliquons les observations faites sur cette classe de modèles du point de vue d'un caractère ancillaire partiel. L'application de nos modèles est illustrée sur une étude portant sur le blaireau européen (*Meles meles*) dans laquelle le génotypage des fragments d'ADN était susceptible d'être erroné.

La recherche portant sur la régression non paramétrique avec contrainte de forme est très active. Cependant, peu de publications considèrent explicitement l'unimodalité alors qu'il est nécessaire de recourir à de telles méthodes dans les applications, par exemple dans l'analyse des relations dose-réponse. Dans cet article, nous proposons des méthodes de régression unimodale de splines qui utilisent les splines de Bernstein et Schoenberg et leur propriété de préservation de la forme. Afin d'obtenir des solutions unimodales régulières, nous utilisons des splines pénalisés et étendons l'approche des splines pénalisés à la pénalisation contre certaines formes paramétriques générales au lieu de la simple utilisation des pénalités de différence. Pour ajuster la sélection des paramètres sous la contrainte d'unimodalité, une approche de maximum de vraisemblance avec restriction et une approche bayésienne alternative sont développées. Nous comparons les méthodologies proposées à d'autres approches communes au moyen d'une étude de simulation et les appliquons à un exemple de données d'évaluation d'une relation dose-réponse. Tous ces résultats suggèrent que la contrainte d'unimodalité ou la combinaison de l'unimodalité et d'une pénalité peuvent améliorer nettement l'estimation de la relation concernée.

Nous développons des modèles mixtes de régression linéaire dans lesquels à la fois la réponse et un prédicteur sont des fonctions. Les paramètres des modèles sont estimés en maximisant la log-vraisemblance au moyen de l'algorithme ECME. Nous montrons que les paramètres de variance ou les matrices de covariances estimés sont définis positifs à chaque itération. Une étude de simulation montre que l'approche proposée est performante pour ce qui concerne l'erreur d'ajustement et l'erreur quadratique moyenne des estimateurs des « coefficients de régression »

H. Li, J. Staudenmayer, and R. J. Carroll **802**
Hierarchical Functional Data with Mixed Continuous and Binary Measurements

Il est très important d'étudier les interactions entre différentes régions du cerveau si l'on souhaite mieux comprendre la fonction cérébrale. Dans cet article, nous nous intéressons à l'identification de schémas fonctionnels de coactivation et de réseaux fonctionnels non orientés dans les études de neuroimagerie. Nous construisons un réseau cérébral fonctionnel en utilisant une matrice de covariance creuse dont les éléments représentent des associations entre pics d'activation au niveau régional. Nous utilisons une approche de vraisemblance pénalisée afin d'imposer la structure creuse de la matrice de covariance basée sur un modèle multivarié de Poisson étendu. Nous obtenons des estimations du maximum de vraisemblance pénalisée en utilisant l'algorithme EM d'espérance-maximisation et optimisons un paramètre de réglage associé en maximisant la log-vraisemblance prédictive. Des tests de permutation sur les schémas cérébraux de coactivation fournissent une inférence au niveau des paires de régions et au niveau du réseau. Les résultats de simulations suggèrent que l'approche proposée a des biais très faibles et permet d'obtenir une probabilité de recouvrement proche de 95 % des estimations de covariance. Dans une méta-analyse de 162 études de neuroimagerie portant sur les émotions, notre modèle identifie un réseau fonctionnel qui consiste en des régions connectées à l'intérieur des noyaux gris centraux, du système limbique et d'autres régions du cerveau liées aux émotions. Nous caractérisons ce réseau en ayant recours à l'inférence statistique sur les connexions entre paires de régions ainsi qu'à des mesures de graphes.

W. Xue, J. Kang, F. D. Bowman, T. D. Wager, and J. Guo **812**
Identifying Functional Co-Activation Patterns in Neuroimaging Studies Via Poisson Graphical Models

Il est très important d'étudier les interactions entre différentes régions du cerveau si l'on souhaite mieux comprendre la fonction cérébrale. Dans cet article, nous nous intéressons à l'identification de schémas fonctionnels de coactivation et de réseaux fonctionnels non orientés dans les études de neuroimagerie. Nous construisons un réseau cérébral fonctionnel en utilisant une matrice de covariance creuse dont les éléments représentent des associations entre pics d'activation au niveau régional. Nous utilisons une approche de vraisemblance pénalisée afin d'imposer la structure creuse de la matrice de covariance basée sur un modèle multivarié de Poisson étendu. Nous obtenons des estimations du maximum de vraisemblance pénalisée en utilisant l'algorithme EM d'espérance-maximisation et optimisons un paramètre de réglage associé en maximisant la log-vraisemblance prédictive. Des tests de permutation sur les schémas cérébraux de coactivation fournissent une inférence au niveau des paires de régions et au niveau du réseau. Les résultats de simulations suggèrent que l'approche proposée a des biais très

faibles et permet d'obtenir une probabilité de recouvrement proche de 95 % des estimations de covariance. Dans une méta-analyse de 162 études de neuroimagerie portant sur les émotions, notre modèle identifie un réseau fonctionnel qui consiste en des régions connectées à l'intérieur des noyaux gris centraux, du système limbique et d'autres régions du cerveau liées aux émotions. Nous caractérisons ce réseau en ayant recours à l'inférence statistique sur les connexions entre paires de régions ainsi qu'à des mesures de graphes.

Nous considérons le problème de l'estimation robuste d'une relation de régression entre une réponse et une covariable sur la base d'un échantillon dans lequel des mesures précises sur la covariable ne sont pas disponibles mais des substituts sujets à erreur pour la covariable non observée sont disponibles pour chaque unité échantillonnée. Les méthodes existantes font souvent des hypothèses restrictives et irréalistes sur la densité de la covariable et sur les densités de la régression et des erreurs de mesure, par exemple la normalité et, pour les deux dernières, également l'homoscédasticité et donc l'indépendance par rapport à la covariable. Dans cet article, nous décrivons une méthodologie semi-paramétrique bayésienne basée sur des mélanges de B-splines et des mélanges induits par des processus de Dirichlet qui relâche ces hypothèses restrictives. En particulier, nos modèles pour les densités citées précédemment s'adaptent à l'asymétrie, aux queues lourdes et à la multimodalité. Les modèles pour les densités de régression et les erreurs de mesure s'adaptent également à une hétéroscédasticité conditionnelle. Dans des expériences simulées, notre méthode surpasse largement les méthodes existantes. Nous appliquons notre méthode à des données de épidémiologie nutritionnelle.

L'analyse de médiation est importante pour comprendre les mécanismes par lesquels une variable produit des changements sur une autre variable. L'erreur de mesure peut limiter la capacité du médiateur à expliquer de tels changements. Cet article est centré sur le développement de méthodes de correction de l'erreur de mesure sur le médiateur pour des données de délai de survenue d'un événement. Nous considérons une définition large de l'erreur de mesure qui inclut l'erreur technique et l'erreur associée aux variations temporelles. Nous supposons que le modèle sous-jacent incluant le médiateur « vrai » (sans erreur) est un modèle de Cox à risques proportionnels. Le rapport induit des risques instantanés pour le médiateur observé n'a alors plus la forme simple indépendante de la fonction de risque instantané de base du fait du conditionnement. Nous proposons une approche de régression-calibration de la moyenne et de la variance et une approche de régression-calibration sur le temps de suivi afin d'approximer la vraisemblance partielle pour la fonction induite de risque instantané. Dans une étude de simulation, les deux méthodes apparaissent intéressantes pour évaluer les effets de médiation. Ces méthodes sont généralisées à des biomarqueurs multiples pour des schémas d'étude cas-cohorte et

d'étude cas-témoin nichée. Nous appliquons ces méthodes de correction aux essais d'hormonothérapie de la « Women's Health Initiative » afin de comprendre l'effet de médiation de plusieurs mesures d'hormones sexuelles sur la relation entre l'hormonothérapie post-ménopausique et le risque de cancer du sein.

Le recours à un schéma d'étude cas-témoins nichée est fréquent dans l'analyse des études épidémiologiques de cohorte de grande taille du fait de son rapport coût-efficacité favorable. Plusieurs méthodes ont été développées pour estimer les paramètres du modèle à risques proportionnels à partir de données cas-témoins nichées dans une cohorte. Cependant, moins d'attention a été prêtée à l'évaluation de la validité des hypothèses du modèle. Dans cet article, nous proposons une classe de statistiques de tests d'adéquation pour tester l'hypothèse des risques proportionnels avec des données cas-témoins nichées. La construction des statistiques de test repose sur des processus asymptotiquement de moyenne nulle qui sont dérivés de la méthode du maximum de pseudo-vraisemblance proposée par Samuelsen. De plus, nous développons un schéma de rééchantillonnage innovant afin d'approximer la distribution asymptotique de la statistique de test tout en tenant compte de la nature dépendante du schéma d'échantillonnage des études cas-témoins nichées. Une étude de simulation est conduite afin d'évaluer les performances de notre approche et une application à une étude sur la tumeur de Wilms est présentée afin d'illustrer son application.

Lorsque l'on estime l'effet d'une exposition ou d'un traitement sur une réponse, il est important de sélectionner les facteurs de confusion pertinents devant être inclus dans le modèle. Inclure trop de covariables augmente l'erreur quadratique sur l'estimation de l'effet de l'exposition et, à l'inverse, ne pas inclure des facteurs de confusion réels entraîne un biais dans l'estimation de cet effet. Nous proposons une approche basée sur la théorie de la décision pour la sélection des facteurs de confusion et l'estimation de l'effet de l'exposition. Nous estimons d'abord les paramètres d'un modèle de régression bayésien complet standard puis obtenons la distribution a posteriori au moyen d'une fonction de perte qui pénalise les modèles omettant des facteurs de confusion importants. Notre méthode peut facilement être implémentée avec des logiciels existants et, dans beaucoup de situations, sans l'utilisation de méthodes de Monte Carlo par chaîne de Markov, ce qui aboutit à des calculs dont le niveau de complexité est de l'ordre des estimations des moindres carrés. Nous illustrons notre méthode en estimant l'effet de l'exposition aux particules fines (PM2,5) sur le poids de naissance dans le comté de Mecklenburg en Caroline du Nord (États-Unis).

Le filtrage des variables est devenu une première étape fondamentale dans l'analyse des données de haut débit mais les procédures existantes peuvent être lourdes sur le plan

calculatoire, difficiles à justifier sur le plan théorique et même inapplicables à certains types d'analyse. À partir d'un problème de grande dimension de régression quantile pour des variables censurées concernant les déterminants génomiques de l'évolution du myélome multiple, cet article apporte trois contributions. Tout d'abord, nous proposons un cadre pour le filtrage des données reposant sur un test du score qui est d'application large tout en étant très efficace sur le plan calculatoire et assez simple à justifier. Ensuite, nous proposons une procédure de rééchantillonnage pour la sélection du nombre de variables souhaité à la suite d'un filtrage en suivant le principe de reproductibilité. Enfin, nous proposons une nouvelle méthode de filtrage faisant appel à un test du score itératif qui possède des liens étroits avec la régression sparse. Dans une étude de simulation, nous appliquons nos méthodes à quatre modèles de régression différents et montrons qu'elle peut donner de meilleurs résultats que les procédures existantes. Nous appliquons également le filtrage par le test du score à une analyse de données d'expression génique de patients atteints de myélome multiple en utilisant un modèle de régression quantile pour données censurées afin d'identifier des gènes qui confèrent un risque élevé d'évolution défavorable.

Les progrès récents des biotechnologies et de leurs applications ont conduit à la génération de nombreux jeux de données d'expression génique de dimension élevée. Les méta-analyses jouent un rôle important pour résumer et synthétiser les résultats de plusieurs études. Quand les dimensions de ces jeux de données sont élevées, il est souhaitable d'incorporer un processus de sélection de variables à la méta-analyse afin d'améliorer l'interprétation du modèle et la prédiction qui en résulte. À notre connaissance, toutes les méthodes existantes conduisent à une sélection de variables à partir des données de méta-analyse qui est de type « tout ou rien », c'est-à-dire qu'un gène est soit sélectionné dans toutes les études ou n'est sélectionné dans aucune. Cependant, du fait de l'hétérogénéité des données communément rencontrée dans les méta-analyses qui provient du choix des spécimens biologiques, de la population d'étude et de la sensibilité de la mesure, il est possible qu'un gène soit important dans certaines études sans l'être dans d'autres. Dans cet article, nous proposons une méthode innovante appelée « méta-lasso » pour la sélection de variable à partir de données de méta-analyse de dimension élevée. Par l'intermédiaire d'une décomposition hiérarchique portant sur les coefficients de régression, notre méthode permet d'utiliser des jeux de données multiples pour accroître la puissance de détection de gènes importants tout en conservant une souplesse permettant de tenir compte de l'hétérogénéité des données. Nous montrons que notre méthode possède une cohérence dans la sélection des gènes, c'est-à-dire qu'elle permet, en cas de taille élevée de tous les jeux de données, d'identifier avec une probabilité élevée tous les gènes importants et de ne pas retenir les gènes non importants. Une étude de simulation montre que notre méthode est performante. Nous appliquons notre méthode méta-lasso à une méta-analyse de cinq études dans le domaine cardiovasculaire et obtenons des résultats cliniquement pertinents.

La génomique intégrative permet une approche plus puissante des études d'association génétiques. Elle offre l'espoir d'une détection plus puissante des polymorphismes de nucléotide unique (SNP). Nous présentons un nouveau test d'association basé sur un modèle statistique qui suppose explicitement que les variations génétiques affectent le phénotype par l'intermédiaire de perturbations du niveau d'expression des gènes. Nous montrons de façon analytique que le test proposé peut être plus puissant pour détecter des SNP qui sont associés à des phénotypes par l'intermédiaire d'une régulation transcriptionnelle que les tests utilisant seulement le phénotype et les données génotypiques. Des simulations montrent que notre méthode est relativement robuste à une mauvaise spécification du modèle. Nous proposons aussi une stratégie pour appliquer notre approche à des données génomiques de grande dimension. Nous utilisons cette stratégie pour identifier une nouvelle association potentielle entre un SNP et la réponse d'une cellule de levure à la tomatidine naturelle que les analyses classiques n'ont pas pu détecter.

Y. Huang and Y. Fong **891**
Identifying Optimal Biomarker Combinations for Treatment Selection via a Robust Kernel Method

Les marqueurs de sélection de traitements prédisent la réponse d'un individu à différents traitements, ce qui permet la sélection du traitement présentant la meilleure réponse attendue. Une bonne règle de sélection de traitements basée sur des marqueurs peut avoir un réel impact de santé publique en permettant une réduction coût-efficace du fardeau de la maladie. Dans cet article, notre objectif est d'utiliser des données d'essais randomisés afin d'identifier des combinaisons linéaires et non linéaires de biomarqueurs qui soient optimales pour le choix d'un traitement minimisant le fardeau total dans la population causé par la maladie ou son traitement. Nous formulons cet objectif comme un problème général de minimisation d'une somme pondérée de perte de type 0-1 et proposons une méthode innovante de minimisation pénalisée basée sur l'algorithme de différence entre fonctions convexes (DCA). L'estimateur correspondant de combinaisons de marqueurs possède une propriété de noyau qui permet une modélisation flexible de combinaisons linéaires et non linéaires de marqueurs. Nous comparons la méthode proposée aux méthodes existantes d'optimisation de schémas thérapeutiques telles que le modèle de régression logistique et la machine vectorielle de support pondéré. Nous étudions aussi les performances respectives de différentes pondérations. Nous illustrons l'application de la méthode proposée sur un exemple réel d'un essai de vaccination contre le VIH dans lequel nous cherchons à déterminer une combinaison de récepteurs Fc de gènes permettant de recommander la vaccination afin de prévenir l'infection par le VIH.

J. Perin, J. S. Preisser, C. Phillips, and B. Qaqish **902**
Regression Analysis of Correlated Ordinal Data Using Orthogonalized Residuals

Les modèles de régression semi-paramétrique pour l'estimation conjointe de la moyenne marginale et des paramètres d'association intra-grappe par paire sont utilisés dans de nombreux contextes pour modéliser les moyennes populationnelles de critères catégoriels multivariés. Récemment, une formulation de régression logistique alternative basée sur les résidus marginaux orthogonalisés a été proposée pour des données binaires corrélées.

À la différence de la procédure originelle basée sur les résidus conditionnels, l'estimateur de la covariance est invariant à l'ordre des observations à l'intérieur des grappes. Dans cet article, la méthode des résidus orthogonalisés est étendue afin de modéliser des données ordinales corrélées avec un odds ratio global. Une étude de simulation montre que cette méthode est plus efficace et moins biaisée en ce qui concerne l'estimation des paramètres d'association intra-grappe qu'une extension aux données ordinales préalablement proposée de régression logistique alternative basée sur les résidus conditionnels. Les résidus orthogonalisés sont utilisés pour estimer les paramètres d'un modèle pour trois critères corrélés mesurés de façon répétée au cours du temps dans un essai clinique évaluant une intervention visant à améliorer la récupération de la perception de sensation altérée chez des patients à la suite d'une opération chirurgicale de la mâchoire.

Les covariables auxiliaires sont souvent employées dans les problèmes de recherche biomédicale où la variable principale d'exposition n'est mesurée que sur un sous-groupe de sujets de l'étude. Cet article s'intéresse aux modèles linéaires généralisés mixtes en présence d'information sur des covariables auxiliaires pour des données en grappes. Nous proposons une approche d'estimation semi-paramétrique originale basée sur une fonction de vraisemblance par paires qui consiste en un produit de vraisemblances bivariées pour les paires d'observations d'une même grappe et développons une procédure d'inférence basée sur des équations d'estimation en traitant à la fois la structure d'erreur et les effets aléatoires comme des paramètres de nuisance. Cette méthode est robuste vis-à-vis d'hypothèses erronées du modèle concernant la structure d'erreur et la distribution des effets aléatoires et permet de tenir compte d'une dépendance entre les covariables et les effets aléatoires. Nous montrons que les estimateurs qui en résultent sont convergents et de distribution asymptotiquement normale. Une étude de simulation approfondie évalue les performances des estimateurs proposés sur des échantillons de taille finie et montrent leur avantage par rapport à la méthode basée sur un échantillon de validation et à la méthode existante. La méthode est illustrée sur deux exemples.

Une façon habituelle de modéliser des données de comptage avec surdispersion, comme par exemple le nombre de chutes rapportées pendant des études interventionnelles, est d'utiliser la distribution binomiale négative. On sait que les méthodes d'estimation classiques sont sensibles au non-respect des hypothèses du modèle consistant dans l'exemple choisi en des chutes plus fréquentes qu'attendu dans les études d'intervention analysées avec le modèle de régression binomiale négative. Dans cet article, nous étendons à la distribution binomiale négative deux approches permettant de construire des M-estimateurs robustes des paramètres de régression dans la classe des modèles linéaires généralisés. La première approche permet d'atteindre la robustesse dans la réponse en appliquant une fonction bornée aux résidus de Pearson obtenus à partir des équations de maximisation de la vraisemblance alors que la seconde approche le permet

en bornant les composants de la déviance brute. Nous explorons différents choix des fonctions de bornage avec les deux approches. À travers une notation unifiée, nous montrons que ces deux approches peuvent être très proches l'une de l'autre si les fonctions sont choisies et calibrées de façon appropriée et nous donnons les distributions asymptotiques des estimateurs qui en résultent. De plus, nous introduisons un estimateur du maximum de vraisemblance pondéré du paramètre de surdispersion qui est robuste et qui est spécifique à la distribution binomiale négative. Des simulations réalisées dans différents contextes montrent, pour les deux approches, qu'un bornage redescendant des fonctions permet d'obtenir des estimateurs peu biaisés en cas de contamination et efficaces pour le modèle retenu. Nous présentons une application à un essai randomisé récent qui évalue l'efficacité d'un programme d'exercice physique pour réduire le nombre de chûtes chez des patients atteints de la maladie de Parkinson afin d'illustrer l'utilisation diagnostique de telles procédures robustes et le besoin afférent d'inférence fiable.

Les modèles informatiques complexes jouent un rôle majeur dans la recherche sur la qualité de l'air. Ces modèles sont utilisés pour évaluer les impacts réglementaires potentiels des stratégies de contrôle des émissions et pour estimer la qualité de l'air dans des zones sans données de surveillance. Pour ces deux objectifs, il est important de calibrer les résultats fournis par le modèle sur les données de surveillance afin de contrôler les biais du modèle et d'améliorer la prédiction de nature spatiale. Dans cet article, nous proposons une nouvelle méthode spectrale pour étudier et exploiter les relations complexes entre les résultats du modèle et les données de surveillance. Les méthodes spectrales nous permettent d'estimer la relation entre les résultats du modèle et les données de surveillance séparément à différentes échelles spatiales et d'utiliser les résultats du modèle à des fins de prédiction aux échelles appropriées. La méthode proposée utilise des calculs efficients et peut être mise en œuvre en utilisant des logiciels standards. Nous appliquons cette méthode à la comparaison des résultats du modèle de « Qualité de l'Air Communautaire Multiéchelle » (CMAQ) aux mesures d'ozone effectuées aux États-Unis en juillet 2005. Il apparaît que le modèle CMAQ reproduit bien les tendances spatiales à grande échelle mais que ses résultats sont peu corrélés aux données de surveillance à plus petite échelle.

De nombreux travaux portent sur l'ajustement de modèles d'Ising sur des données binaires multivariées dans le but de comprendre les relations de dépendance conditionnelle entre ces variables. Cependant, des variables supplémentaires sont fréquemment collectées en même temps que les données binaires et peuvent influencer les relations de dépendance. Motivés par une application sur des données sur l'instabilité du génome, collectées sur des tumeurs de plusieurs types, nous proposons un modèle d'Ising clairsemé incluant des covariables pour étudier à la fois la dépendance conditionnelle au sein des données binaires et de leur relation avec les covariables supplémentaires. Ceci résulte en des modèles d'Ising sujet-spécifiques dans lesquels les

covariables du sujet influence la force de l'association entre les gènes. Comme dans toute analyse exploratoire, l'interprétabilité des résultats est importante et nous utilisons 11 pénalités pour induire de la rareté aussi bien dans les graphes ajustés que dans le nombre de covariables sélectionnées. Deux algorithmes sont proposés pour ajuster le modèle puis comparés sur données simulées. Des résultats asymptotiques sont établis. Les résultats sur les données des tumeurs et leur signification biologique sont discutés en détail.

Dans les études d'observation avec données de biopuces, la question du biais de confusion se pose toujours. Une approche permettant de prendre en compte les facteurs de confusion mesurés est de les inclure en tant que covariables dans un modèle linéaire multivarié. Cependant, avec ce modèle, l'application de procédures de tests multiples basées sur des approches de permutation pose problème parce que l'hypothèse d'échangeabilité des réponses n'est en général pas vérifiée. Néanmoins, il est possible d'obtenir que les réponses transformées aient un caractère rotationnel moyennant des hypothèses de distribution. Dans la mesure où ils permettent d'ajuster sur des facteurs de confusion, nous pensons que les tests multiples basés sur la rotation représentent une extension importante des tests multiples basés sur les approches de permutation. Nous illustrons la méthodologie proposée sur des données d'une étude d'observation avec données de biopuces dans le cancer du sein. Un programme permettant de mettre en œuvre la procédure décrite dans cet article est disponible dans l'ensemble de programmes flip de R.

Nous étudions le modèle Mt,α pour estimer l'abondance dans les études de capture-recapture en population fermée dans lesquelles les animaux sont identifiés à partir de marques naturelles telles que des profils d'ADN ou des photographies de caractéristiques individuelles distinctives. Le modèle Mt,α étend le modèle classique Mt afin de tenir compte des erreurs d'identification en spécifiant que l'identification de chaque échantillon a une probabilité α d'être correcte et une probabilité $1 - \alpha$ d'être erronée. Les informations concernant les erreurs d'identification proviennent d'épisodes de capture supplémentaires avec uniquement une entrée qui se produisent en cas d'erreur d'identification. Nous développons une expression analytique exacte de la vraisemblance du modèle et montrons qu'elle peut être calculée de façon efficiente à l'inverse des études précédentes qui ont considéré que l'expression de la vraisemblance n'était pas calculable. La rapidité de nos calculs nous a permis d'étudier en détail les propriétés statistiques des estimateurs du maximum de vraisemblance. Nous avons trouvé que l'approche indirecte pour estimer l'erreur d'estimation demande une grande richesse de données et que l'obtention de bonnes propriétés statistiques en ce qui concerne le biais et la précision requièrent des probabilités de capture élevées ou de nombreuses occasions de capture. Quand ces conditions ne sont pas remplies, l'abondance est estimée avec une très faible

précision et un biais négatif et, à l'extrême, on obtient de meilleures propriétés avec l'approche naïve qui ignore les erreurs d'identification. Nous recommandons une utilisation prudente du modèle $M_{t,\alpha}$ ainsi que la considération d'autres approches pour tenir compte des erreurs d'identification. Nous illustrons notre étude avec des enquêtes génétiques et photographiques de la population néo-zélandaise de baleines franches australes (*Eubalaena australis*).

Dans les modèles espace-état, le critère de déviance de l'information (DIC) est fréquemment utilisé pour évaluer la capacité du modèle à prédire une observation au temps *t* étant donné l'état sous-jacent à ce même temps *t*. Au vu de l'incapacité du DIC conventionnel à permettre un choix clair entre des modèles bayésiens espace-état non linéaires multivariés concurrents de dynamique de la population de saumons cohos et des difficultés calculatoires liées à des choix alternatifs de critère, ce travail propose une variante du DIC appelée « DIC avec une longueur d'avance » ou DIC*p* pour lequel la prédiction est conditionnelle à l'état au temps précédent. Une étude de simulation montre que le DIC*p* est fiable pour choisir entre différentes équations de processus ou d'observation. À l'inverse, le DIC conventionnel peut donner des résultats très trompeurs et montre une prédilection à choisir le modèle incorrect. Cela peut s'expliquer par son incapacité à tenir compte de surestimations de l'erreur de processus provenant d'une mauvaise spécification du modèle. Le DIC*p* n'est pas basé sur une véritable vraisemblance conditionnelle mais nous montrons qu'il peut être interprété comme un pseudo DIC pour lequel le comportement de compensation lié à l'augmentation des erreurs de processus est éliminé. Il peut être facilement calculé en utilisant les contrôles de DIC avec le logiciel BUGS quand les équations de processus et d'observation sont conjuguées. Les performances améliorées du DIC*p* sont illustrées par une application à la modélisation multi-état de l'abondance du saumon coho à Lobster Creek dans l'état d'Oregon aux États-Unis.

L'analyse de la distribution spatiale des données binaires multivariées fait l'objet d'un intérêt croissant qui est motivé par un grand nombre de problèmes de recherche. Deux types de corrélation sont habituellement en cause, la corrélation entre plusieurs critères au sein d'une même localisation géographique et la corrélation spatiale entre les localisations pour un critère donné. Les modèles de régression habituellement utilisés considèrent seulement un type de corrélation et ignorent ou modélisent de façon incorrecte l'autre type. Pour s'affranchir de cette limitation, nous adoptons une approche bayésienne non paramétrique afin de modéliser conjointement des données binaires spatiales multivariées en intégrant les deux types de corrélation. Un modèle probit multivarié est employé pour relier une réponse binaire à des variables latentes gaussiennes et des processus gaussiens sont utilisés pour représenter les effets aléatoires corrélés. Nous développons un algorithme efficient de méthodes de Monte Carlo par chaîne de Markov pour les calculs des distributions a posteriori. Nous illustrons le

modèle proposé à partir de simulations et des données d'une étude portant sur la tuberculose mulirésistante.

M. Höhle and M. an der Heiden **993**
Bayesian Nowcasting during the STEC O104:H4 Outbreak in Germany, 2011

Une approche bayésienne pour la prédiction des événements survenus-mais-non-encore-déclarés est développée pour une application à la surveillance de la santé publique en temps réel. La motivation est la prédiction du nombre d'hospitalisations quotidiennes pour le syndrome hémolytique et urémique lors de la grande épidémie d'Escherichia coli producteurs de Shiga-toxines (STEC) O104:H4 de mai à juillet 2011 en Allemagne. Notre nouvelle approche bayésienne traite la nature de données de comptage du problème en utilisant un échantillonnage binomial négatif et montre que la troncature à droite de la distribution du retard de déclaration sous l'hypothèse d'homogénéité de temps peut être traitée dans un cadre de prior conjugué en utilisant la distribution de Dirichlet généralisée. Depuis, avec le recul, le véritable nombre d'hospitalisations est disponible, des règles de notation appropriées aux données de comptage sont utilisées pour évaluer et comparer la qualité prédictive des procédures lors de l'épidémie. Les résultats montrent qu'il est important de prendre en compte la nature de comptage de la série chronologique et que les changements dans la répartition de retard se sont produits en raison de mesures d'intervention. En conséquence, nous étendons l'analyse bayésienne à un modèle hiérarchique, qui combine un modèle de régression de survie en temps discret pour la distribution de retard avec une spline pénalisée pour la dynamique de la courbe de l'épidémie. Finalement, nous concluons que dans les foyers émergents à temps critique, les approches de prévision immédiate sont un outil précieux pour obtenir des informations sur les tendances en cours.

M.-H. Chen, J. G. Ibrahim, D. Zeng, K. Hu, and C. Jia **1003**
Bayesian Design of Superiority Clinical Trials for Recurrent Events Data with Applications to Bleeding and Transfusion Events in Myelodyplastic Syndrome

Dans beaucoup d'études biomédicales, le même type d'événement récurrent peut survenir plusieurs fois au cours du suivi d'un patient, comme c'est le cas par exemple pour les saignements, les infections multiples et les maladies, Dans cet article, nous proposons un schéma bayésien pour un essai clinique pivot dans lequel les patients ayant un risque faible de développer un syndrome myélodysplasique (MDS) sont traités par des traitements modificateurs de ce syndrome. Un des objectifs clés de l'étude est de montrer l'effet d'un nouveau traitement sur la réduction du nombre de transfusions de plaquettes et d'événements hémorragiques chez ces patients. Dans ce contexte, nous proposons une nouvelle approche bayésienne pour la planification d'études de supériorité qui utilise des modèles de régression incorporant la fragilité individuelle. Des données antérieures sur des événements récurrents provenant d'un essai de phase 2 achevé sont incorporés dans le schéma d'étude bayésien en se basant partiellement sur la distribution a priori de type puissance proposée par Ibrahim et coll. (2012). Un algorithme efficient d'échantillonnage de Gibbs, un algorithme prédictif de génération de données et un algorithme basé sur des simulations sont développés afin d'échantillonner à partir de la distribution a posteriori ajustée aux données, ce qui permet de générer des données d'événements récurrents et de calculer différentes quantités propres au schéma d'étude telles que le risque de première

espèce et la puissance. Une étude de simulation approfondie compare la méthode proposée aux méthodes fréquentistes existantes et étudie les caractéristiques opérationnelles du schéma d'étude proposé.

Z. Wu, C. E. Frangakis, T. A. Louis, and D. O. Scharfstein **1014**
Estimation of Treatment Effects in Matched-Pair Cluster Randomized Trials by Calibrating Covariate Imbalance between Clusters

Nous nous intéressons aux effets d'interventions dans les études expérimentales en grappes dans lesquelles les interventions sont allouées au niveau de la grappe, les grappes sont sélectionnées pour former des paires appariées sur des caractéristiques observées et l'intervention est allouée aléatoirement à l'une des deux grappes dans chaque paire. Notre objectif qui a des implications en terme de politique de santé est d'estimer le résultat moyen qui serait obtenu dans le cas où l'intervention serait allouée à toutes les grappes de toutes les paires relativement au cas où l'absence d'intervention serait allouée à toutes les grappes de toutes les paires. Pour de tels schémas d'étude, l'inférence qui ignore les covariables individuelles peut être imprécise car l'allocation au niveau de la grappe peut laisser demeurer des déséquilibres importants dans la distribution des covariables entre les deux bras expérimentaux dans chaque paire. Cependant, la plupart des méthodes disponibles qui ajustent sur des covariables conduisent à des estimations qui n'ont pas d'interprétation utile en ce qui concerne la définition de politiques de santé. Nous proposons une méthodologie qui équilibre explicitement les covariables observées parmi les grappes d'une paire et fournit une estimation d'interprétation utile pour la définition de politiques de santé. Nous illustrons notre approche par l'évaluation du programme d'intervention « Guided Care ».

H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, **1023**
D. Spiegelhalter, and B. Neuenschwander
Robust Meta-Analytic-Predictive Priors in Clinical Trials with Historical Control Information

L'information historique est toujours pertinente à considérer pour la planification d'un essai clinique. De plus, si les données historiques peuvent être incorporées dans l'analyse d'un nouvel essai, cela permet d'en réduire l'effectif. La réduction du coût et de la durée de l'essai qui en résulte facilite le recrutement et peut sembler plus éthique. Cependant, si les données historiques sont conflictuelles, une utilisation trop optimiste de ces données peut se révéler inappropriée. Nous abordons ce problème en développant une distribution bayésienne méta-analytique a priori à partir de données historiques puis en combinant cette distribution aux nouvelles données. Cette approche prospective est équivalente à une méta-analyse combinant des données historiques et des données nouvelles si les paramètres sont échangeables sur les essais considérés. La version prospective bayésienne nécessite une bonne approximation de la distribution méta-analytique a priori qui n'est pas disponible de façon analytique. Nous proposons d'utiliser des mélanges à deux ou trois composants de distribution standard a priori, ce qui permet une bonne approximation et, pour la famille exponentielle à un paramètre, conduit à des calculs très simples pour dériver la distribution a posteriori. De plus, puisqu'un des composants du mélange est habituellement pris comme vague, les distributions de mélange a priori, sont souvent à queue lourde et donc robustes. Plus de robustesse encore et une réaction plus rapide à des données a priori conflictuelles peuvent être obtenues en ajoutant un

composant faiblement informatif au mélange. L'utilisation d'informations historiques est particulièrement intéressante pour les essais adaptatifs car le rapport de randomisation peut être modifié en cas de données a priori conflictuelles. En considérant des scénarios variés, il apparaît qu'à la fois les caractéristiques opérationnelles fréquentistes et les résumés distributionnels a posteriori dans un contexte bayésien montrent que ces approches ont des propriétés intéressantes. Nous illustrons cette méthodologie par un essai de phase II de preuve de concept avec des données historiques de témoins provenant de quatre études. Des distributions méta-analytiques a priori qui sont robustes réduisent les problèmes de données historiques conflictuelles. Ces résultats devraient encourager les chercheurs à utiliser mieux et plus souvent les données historiques dans les essais cliniques.

E. Kim, Z. Zhang, Y. Wang, and D. Zeng **1033**
Power Calculation for Comparing Diagnostic Accuracies in a Multi-Reader, Multi-Test Design

Les courbes ROC sont très utilisés pour évaluer la performance diagnostique d'examens donnant une réponse continue ou ordinale. Un schéma d'étude classique pour évaluer la performance diagnostique d'examens, appelé le schéma multi-examens et multi-observateurs, repose sur l'interprétation de plusieurs examens par plusieurs observateurs. Bien que plusieurs approches existent pour analyser des données provenant d'un tel schéma d'étude, peu de méthodes se sont intéressées aux questions de taille d'échantillon et de puissance statistique. Dans cet article, nous développons une formule de calcul de puissance pour comparer les aires sous la courbe ROC (AUC) corrélées dans un schéma multi-examens et multi-observateurs. Nous présentons une approche non paramétrique pour estimer et comparer les AUC corrélées qui étend l'approche de DeLong et coll. (1988). Une formule de calcul de puissance est dérivée à partir de la distribution asymptotique des AUC non paramétriques. Une étude de simulation est conduite pour évaluer la performance de la formule de puissance proposée et un exemple est donné pour illustrer la procédure proposée.

D. Huynh, O. Laeyendecker, and R. Brookmeyer **1042**
A Serial Risk Score Approach to Disease Classifi cation that Accounts for Accuracy and Cost

La performance des examens diagnostiques est souvent mesurée par leur précision diagnostique en évaluant notamment la sensibilité et la spécificité. Cependant, il est important de considérer également les coûts de ces examens. L'association de plusieurs examens diagnostiques peut conduite à une meilleure précision diagnostique mais au prix de coûts additionnels. Dans ce travail, nous considérons des approches dans lesquelles les examens diagnostiques sont réalisés de façon consécutive et qui maintiennent la précision diagnostique tout en limitant le coût. Nous présentons une approche basée sur un score séquentiel de risque. L'idée de base est d'utiliser les examens diagnostiques de façon séquentielle en augmentant le nombre d'examens utilisés jusqu'à l'obtention du classement diagnostique d'une personne. De cette manière, il n'est pas nécessaire d'utiliser tous les examens diagnostiques sur toutes les personnes. Ces méthodes sont étudiées et comparées à la régression logistique au moyen d'une étude de simulation. Elles sont appliquées à des données de cohortes de patients séropositifs pour le virus HIV

afin d'identifier les sujets récemment infectés (dont l'infection remonte à moins d'un an) en évaluant le statut des sujets vis-à-vis de différents biomarqueurs. Il apparait que l'approche basée sur le score de risque séquentiel peut maintenir la précision diagnostique tout en réduisant le coût par rapport à l'approche consistant à évaluer le statut de tous les individus vis-à-vis de tous les biomarqueurs.

A. C. McLain and P. S. Albert **1052**
Modeling Longitudinal Data with a Random Change Point and No Time-Zero: Applications to Inference and Prediction of the Labor Curve

Dans quelques études longitudinales, le temps d'initiation du processus n'est pas clairement défini alors qu'il est important de réaliser une inférence ou de faire des prédictions concernant ce processus longitudinal. L'application qui motive cet article est de fournir un cadre pour la modélisation de courbes individuelles de dilatations du col pendant le travail lors de l'accouchement (mesures longitudinales de dilatation cervicale) alors que le début du travail n'est en général pas clairement défini dans le temps. Ce problème est bien connu en obstétrique où l'origine du temps est souvent définie par la fin du processus (le moment où la dilatation du col est complète à 10 cm) et le temps est compté à rebours à partir de ce point. Cette approche conduit à une inférence valide et efficace sauf si les sujets sont censurés avant la fin du processus ou si l'objectif est de proposer des prédictions. Fournir de façon prospective des prédictions individuelles dynamiques de la courbe de dilatation du col au cours du travail (alors que le temps compté à rebours est encore inconnu) est utile aux obstétriciens pour déterminer si le travail se passe de façon normale. Nous proposons un modèle longitudinal de dilatation du col qui utilise des effets aléatoires avec un temps d'origine inconnu et un point de changement aléatoire. Nous présentons une approche de maximum de vraisemblance pour l'estimation des paramètres qui recourt à une quadrature gaussienne adaptative pour l'intégration numérique. De plus, nous proposons une approche de Monte Carlo pour la prédiction dynamique de la trajectoire longitudinale future de dilatation cervicale à partir des mesures passées de dilatation. La méthodologie est illustrée sur des données longitudinales de dilatation cervicale provenant de l'étude « Consortium of Safe Labor ».