Translations of Abstracts

BIOMETRIC METHODOLOGY

Nous nous intéressons au problème posé par le test de chacune de *m* hypothèses nulles avec une procédure de permutation séquentielle dans laquelle le nombre de tirages dans la distribution de permutation de chaque statistique de test est une variable aléatoire. Chaque *p*-valeur de permutation séquentielle a une distribution sous l'hypothèse nulle qui est non uniforme à support discret. Nous montrons comment utiliser une collection de telles *p*-valeurs pour estimer le nombre m_0 d'hypothèses nulles exactes parmi les m hypothèses nulles testées et comment estimer le taux de faux positifs associés à des seuils de significativité pour les *p*-valeurs. Nous utilisons des analyses de données réelles et des études de simulation pour évaluer et illustrer la performance de l'approche que nous proposons par rapport à des stratégies standard, plus intensives en calculs. Nous montrons que notre approche séquentielle donne des résultats similaires avec un coût moindre en calculs pour divers scenarii.

Les changements de taille d'une population influencent la diversité génétique de la population et par conséquent laissent une trace de ces changements dans les génomes individuels de la population. Nous nous intéressons au problème inverse de la reconstruction de la dynamique de population passée à partir de données du génome. Nous commençons dans le cadre standard de la coalescence avec un processus stochastique qui génère des généalogies liant aléatoirement les individus échantillonnés dans la population d'intérêt. Ces généalogies servent de lien entre l'histoire démographique de la population et les séquences génomiques. On obtient que seuls les instants de coalescence de lignées généalogiques apportent de l'information sur les dynamiques de taille de population. En considérant ces dates de coalescence comme une processus ponctuel, l'estimation des trajectoires de taille de population est équivalente à l'estimation de l'intensité conditionnelle de ce processus ponctuel. Cependant, notre problème inverse est similaire à l'estimation de la fonction d'intensité d'un processus de Poisson inhomogène. Nous montrons comment les récentes avancées fondées sur des processus gaussiens pour l'inférence non paramétrique de processus de Poisson peuvent être étendues à l'estimation bayésienne no paramétrique des dynamiques de taille de populations sous hypothèse de coalescence. Nous comparons notre approche par processus gaussiens à la classique approche par champs aléatoires gaussiens pour estimer les trajectoires de population. A partir de données simulées, nous montrons que notre méthode est plus précise et plus consistante. Nous analysons ensuite deux généalogies reconstruites à partir de séquences

des virus de l'hépatite C et de la grippe A. Dans les deux cas, nous retrouvons plus de faits connus de l'histoire démographique des deux virus que l'approche par champs aléatoires gaussiens. Notre approche produit aussi des estimateurs d'incertitudes plus raisonnables.

Qian Ren and Sudipto Banerjee **19**
Hierarchical Factor Models for Large Spatially Misaligned Data: A Low-Rank Predictive Process Approach

Cet article traite le problème de la modélisation conjointe d'un grand nombre de résultats géo- référencés observés sur un grand nombre de lieux. Nous cherchons à saisir les associations entre variables ainsi que la force de l'association spatiale de chaque variable. En outre, nous traitons par la mise en commun le cas où toutes les variables n'ont pas été observées sur tous les sites, ce qui conduit à un mauvais alignement spatial. Une réduction de dimension est nécessaire sous deux aspects : (i) la longueur du vecteur des résultats, et (ii) le très grand nombre d'emplacements. Des modèles à variable latente (facteur) sont généralement utilisés pour tenir compte du premier aspect, tandis que des processus spatiaux de rang faible offrent une option de modélisation flexible et riche pour traiter le cas d'un grand nombre d'emplacements. Nous fusionnons ces deux idées en proposant une classe de modèles avec un facteur spatial hiérarchique de rang faible. Le cadre de notre travail se poursuit par une sélection stochastique des facteurs latents sans recourir à des stratégies de calcul complexes (tels que les algorithmes de saut réversibles) en utilisant des caractérisations d'identification pour modéliser le facteur spatial. Nous développons un algorithme de Monte Carlo de chaîne de Markov (MCMC) pour l'estimation ce qui traite aussi le problème de mauvais alignement spatial. Nous récupérons la distribution *a posteriori* des valeurs manquantes dans un cadre prédictif Bayésien. De même, nous discutons de diverses modélisations et de mises en œuvre additionnelles. Nous illustrons notre méthode avec des expériences simulées et un jeu de données environnementales impliquant des polluants de l'air en Californie.

M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard **31**
Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension

Nous proposons une nouvelle procédure de classification de courbes non supervisée en présence de variabilité inter-individuelle. Ce thème a déjà été largement étudié, plus particulièrement par le biais des splines, utilisés pour prendre en compte la présence d'effets aléatoires fonctionnels. Cependant, l'utilisation des splines n'est pas adaptée à l'étude de données en grande dimension, ni à la modélisation de courbes fortement irrégulières, présentant des pics par exemple. L'approche développée est basée sur une décomposition en ondelettes des effets fixes et aléatoires. Nous proposons une étape de réduction de dimension inspirée des techniques de seuillage d'ondelettes classiques et adaptée à la présence de plusieurs courbes. De plus, l'utilisation d'une structure particulière pour la modélisation des variances associées aux effets aléatoires permet d'assurer que les effets fixes et aléatoires sont dans le même espace de Besov fonctionnel. Dans le domaine des ondelettes, nous obtenons alors un modèle linéaire mixte classique sur lequel nous appliquons une procédure de classification basée sur l'algorithme EM pour une estimation des paramètres du modèle par maximum de vraisemblance. Les propriétés de notre procédure sont validées par une étude de simulation approfondie. Nous illustrons ensuite notre méthode sur des données de spectrométrie de masse et nous

proposons une approche fonctionnelle originale pour l'étude de données de microarray CGH. Notre procédure est disponible sur le CRAN dans le package « curvclust » qui est à ce jour le premier package permettant de réaliser une classification de courbes non supervisée dans le cadre des modèles mixtes fonctionnels.

J. Goldsmith, S. Greven, and C. Crainiceanu **41**
Corrected Confidence Bands for Functional Data Using Principal Components

L'analyse fonctionnelle en composantes principales (FPC) est largement utilisée pour décomposer et exprimer des observations fonctionnelles. Les estimations des courbes sont implicitement conditionnées par les fonctions de base et les autres quantités déduites des décompositions par FPC; cependant ces objets sont méconnus en pratique. Dans cet article nous proposons une méthode pour obtenir des estimations correctes des courbes tenant compte des incertitudes de la décomposition par FPC. En outre on construit des intervalles de confiance, soit pour un point soit conjoints, qui tiennent compte des incertitudes dues au modèle et dues à la décomposition par FPC. On utilise des représentations en modèles mixtes standard des développements fonctionnels pour construire des estimations des courbes et les variances conditionnellement à une décomposition spécifique. Des itérations sur la distribution des décompositions donnent les espérances et les variances en combinant les estimations conditionnelles aux modèles. Une procédure bootstrap est développée pour évaluer l'incertitude attachée aux résultats de la décomposition en composantes principales. Des études de simulation montrent que notre méthode se compare favorablement aux méthodes concurrentes, aussi bien pour des observations denses que pour des observations rares. Nous appliquons notre méthode à des observations rares de décomptes de CD4 et des observations denses sur des profils de faisceaux de neurones de la matière blanche. Le code pour les analyses et les simulations est librement accessible et notre méthode est disponible dans le paquetage ***refund*** de R sur le site du CRAN.

L. Altstein and G. Li **52**
Latent Subgroup Analysis of a Randomized Clinical Trial through a Semiparametric Accelerated Failure Time Mixture Model

Cet article présente un modèle semi paramétrique de mélange avec accélération du temps pour l’estimation de l’effet d’un traitement biologique dans un sous groupe latent, dans le cadre d’un essai randomisé dont le critère de jugement est un temps jusqu’à événement (critère censuré). La latence est induite par le fait que l’appartenance à un sous groupe est observable dans un bras de l’essai mais pas dans l’autre. Cette méthode est ainsi utile dans les essais randomisés où les patients du groupe contrôle n’ont pas accès au traitement actif, alors que, comme c’est le cas dans certains essais en cancérologie, une biopsie permettant d’identifier le sous groupe latent n’est réalisée que pour les sujets randomisés dans le groupe expérimental. Nous proposons une méthode de calcul permettant d’estimer les paramètres du modèle par itérations entre un pas basé sur l’espérance et un pas basé sur une optimisation pondérée de Buckley-James. L’estimation de la variance est faite par bootstrap, et la performance de notre méthode est évaluée par des simulations ; Nous illustrons notre méthode par l’analyse d’un essai multicentrique de curage ganglionnaire sélectif dans les mélanomes.

Rena Jie Sun and John D. Kalbfleisch **62**
A Risk-Adjusted O–E CUSUM with Monitoring Bands for Monitoring Medical Outcomes

Nous proposons une méthode CUSUM ajustée sur les OE (observé-attendu) utilisant des bandes de surveillance comme critère de décision. L'objectif est de superviser des données de survie de centres médicaux en utilisant des graphiques simples. Les bandes de surveillance proposées peuvent être utilisées à la place d'une approche plus traditionnelle, mais plus compliquée, avec masque en V ou utilisant simultanément deux CUSUMs unilatéraux. Le graphique qui en résulte est conçu pour superviser simultanément les temps de survenue d'échecs concernant des évènements du type `pire que prévu' ou `mieux que prévu'. Les pentes de la CUSUM OE fournissent des estimations directes du risque relatif par rapport à un taux standard ou attendu de taux d'échecs. Les régions de rejet sont obtenues en contrôlant le taux de fausse alarmes (erreur de type I) sur une période de durée donnée. Des études de simulation sont menées pour illustrer les performances de la méthode proposée. Une étude de cas est réalisée sur 58 centres de transplantation du foie. L'utilisation de méthodes CUSUM afin d'améliorer la qualité est mises en avant.

Justin B. Post and Howard D. Bondell **70**
Factor Selection and Structural Identifi cation in the Interaction ANOVA Model

Avec une variable de réponse continue et des variables explicatives qualitatives, l'objectif de l'analyse consiste souvent en deux points : trouver quels sont les facteurs importants et déterminer quels niveaux de ces facteurs sont significativement différents des autres. Bien souvent ces taches sont réalisées séparément à l'aide d'une analyse de la variance (ANOVA) suivi d'un test de comparaison type test de Tukey. Lorsque des interactions sont incluses dans le modèle, la confusion des niveaux d'un facteur devient un problème plus difficile. Quand on teste les différences entre deux niveaux d'un facteur, cela ne concerne pas uniquement les effets principaux mais aussi l'égalité de chaque interaction impliquant ces niveaux. Cette structure entre les effets principaux et les interactions dans un modèle est similaire à la notion d'hérédité utilisée dans les modèles de régression.
Cet article présente une nouvelle approche pour mettre en œuvre les deux points simultanément à l'aide d'un modèle d'interaction qui respecte la contrainte type hérédité du modèle. Un cas défavorable est construit avec une confusion des niveaux de facteurs et des facteurs fixés à zéro. Il est montré que cette procédure possède la « oracle property », ce qui implique qu'asymptotiquement elle fonctionne aussi bien que si la structure était connue à l'avance. Nous parlons aussi de l'application pour estimer l'interaction en absence de répétition. Des études de simulation montrent que la procédure surpasse des procédures de tests d'hypothèse ainsi que d'autres méthodes similaires qui ne tiennent pas compte de contrainte structurelle. La méthode est ensuite illustrée sur un exemple réel de données.

David M. Zucker, Malka Gorfi ne, Yi Li, Mahlet G. Tadesse, and Donna Spiegelman **80**
A Regularization Corrected Score Method for Nonlinear Regression Models with Covariate Error

De nombreuses analyses de régression impliquent des variables explicatives mesurées avec erreur, et le défaut de prise en compte de cette erreur est bien connu pour conduire à des estimations biaisées, tant ponctuelles que par intervalle, pour les coefficients de régression. Nous présentons ici une nouvelle méthode générale d'ajustement à une erreur sur une covariable. Notre méthode consiste en une version approchée de l'approche du score corrigé de Stefanski-Nakamura, en utilisant la méthode de régularisation pour obtenir une solution approchée de l'équation intégrale adaptée. Nous développons cette théorie pour l'ensemble des modèles à vraisemblance classique ; cet ensemble recouvre par exemple la régression linéaire, la régression nonlinéaire, la régression logistique, et la régression de Poisson. Cette méthode est très générale en termes des types de modèles à erreur de mesure auxquels elle s'applique, et est une méthode fonctionnelle dans le sens où elle n'implique pas d'hypothèses sur la distribution de la vraie covariable. Nous discutons les propriétés théoriques de la méthode et nous présentons des résultats de simulation obtenus avec la régression logistique (univariée et multivariée). Pour l'illustration, nous appliquons la méthode à des données de l'étude « Harvard Nurses' Health » relative à l'association entre l'activité physique et la mortalité du cancer du sein dans la période suivant le diagnostic d'un cancer du sein.

Dans l'analyse des courbes ROC, l'ajustement sur les covariables est recommandé quand celles-ci influencent l'amplitude ou la précision de la réponse au test de dépistage étudié. En outre, pour de nombreux dépistages à grande échelle, le vrai statut des sujets peut être incomplètement observé car sa vérification est coûteuse ou invasive. L'analyse des cas complets peut alors être soumise à un biais de sélection, également appelé biais de vérification. Pour aborder la question de l'analyse d'une courbe ROC ajustée sur une covariable en présence d'un biais de vérification, nous proposons plusieurs estimateurs des aires sous les courbes ROC spécifiques à la covariable et ajustée sur la covariable (AUC_x et AAUC). L' AUC_x est modélisée directement sous la forme d'une régression binaire et les équations d'estimation sont fondées sur les statistiques U. L'AAUC est estimée comme une moyenne pondérée des AUC_x , où les poids dépendent de la distribution de la covariable chez les sujets atteints. Pour traiter le problème du biais de vérification, nous utilisons des techniques de repondération et d'imputation. Nos estimateurs reposent sur la supposition initiale d'un état de santé vrai manquant au hasard (MAR) puis les estimateurs sont modifiés de manière à être étendus au cas de données manquantes non au hasard (MNAR). Nous calculons les distributions asymptotiques de ces estimateurs. La performance à distance finie est évaluée à l'aide de simulations. Nous appliquons notre méthode à des données sur la maladie d'Alzheimer.

Soit les modèles SEIR les plus récents utilisent des périodes latentes et infectieuses distribuées exponentiellement, soit ils autorisent une seule distribution pour la période latente et infectieuse soit ils font des hypothèses fortes sur la quantité d'information

disponible pour les distributions des temps, en particulier le temps passé dans le compartiment exposé. Beaucoup de maladies infectieuses exigent des hypothèses plus réalistes pour les périodes latentes et infectieuses. Dans cet article, nous fournissons un modèle alternatif qui autorise des distributions générales à utiliser à la fois dans les compartiments exposé et infectieux, tout en évitant le besoin de données complètes sur les temps latents. La formulation alternative est un modèle SEIR à chemin spécifique (PS SEIR) qui suit des chemins individuels à travers les compartiments exposé et infectieux, éliminant ainsi le besoin de supposer des distributions exponentielles pour les temps latent et infectieux. Nous montrons comment le modèle PS SEIR est l'analogue stochastique d'une classe générale de modèles SEIR déterministes. Puis nous démontrons l'amélioration de ce modèle PS SEIR par rapport à des modèles marginaux plus communs à travers des résultats de simulations et nous faisons une nouvelle analyse de l'épidémie d'oreillons en Iowa à partir de 2006.

L'estimation des futures prévalences de maladies chroniques est essentielle pour les politiques de santé publique. Nous proposons une méthode, utilisant un modèle illness-death, pour estimer la prévalence d'une maladie chronique et faire des projections de celle-ci à partir de l'incidence de la maladie estimée grâce à des données de cohorte et des taux de décès dans la population générale. Contrairement aux méthodes précédemment publiées, nous ne faisons pas l'hypothèse que la mortalité des sujets non déments est égale à la mortalité générale de la population et nous comparons deux hypothèses quant à la tendance séculaire de la mortalité de sujets malades. De plus, nous avons développé un modèle permettant de faire des prédictions de prévalence en fonction de différents scénarios sur la prise en charge d'un facteur de risque en réduisant soit la prévalence, soit l'excès de risque associé à ce facteur de risque. Les méthodes sont appliquées pour évaluer la prévalence de la démence en France entre 2010 et 2030.

Dans une étude d'observations, un sujet traité peut être associé à des covariables observées sur un ou plusieurs témoins non traités. La motivation commune pour l'utilisation de plusieurs témoins plutôt que sur un seul est de vouloir augmenter la puissance d'un test de non effet sous l'hypothèse douteuse que l'appariement pour les covariables observées suffit pour supprimer le biais dans le cas où l'affectation du traitement n'est pas aléatoire. Est-ce que le choix entre un ou plusieurs témoins appariés affecte la sensibilité des conclusions aux violations de cette hypothèse douteuse? Avec des réponses continues, il est connu que réduire l'hétérogénéité des différences des paires appariées réduit la sensibilité aux biais non mesurés, mais augmenter la taille de l'échantillon a un très important effet sur la sensibilité aux biais. Faut-il utiliser plusieurs témoins plutôt qu'un analogue pour une réduction de l'hétérogénéité ou pour une augmenter la taille de l'échantillon? La question est examinée pour les m-statistiques de Huber, y compris le test *t*, l'examen comporte trois volets: un exemple, des calculs asymptotiques en utilisant la sensibilité du plan, et une simulation. L'utilisation de

témoins multiples avec réponses continues donne une réduction non négligeable de la sensibilité à des biais non mesurés. Un exemple étudie le plomb et le cadmium dans le sang des fumeurs à partir des données de 2008 de la National Health and Nutrition Examination Survey. Un sous-produit de la discussion est un nouveau résultat donnant la sensibilité du plan pour la distribution de permutation des *m*-statistiques.

Shawn E. Simpson **128**
A Positive Event Dependence Model for Self-Controlled Case Series with Applications in Postmarketing Surveillance

Un objectif essentiel dans la surveillance de la sécurité des médicaments après leur mie sur le marché est de pouvoir affirmer la relation entre des temps d'exposition différents à des produits et des effets adverses récurrents associés aux objectifs de santé. La méthode des séries de cas auto-contrôlés (SCCS) est une approche de l'analyse dans ce contexte. Elle se base sur un modèle de régression de Poisson conditionnel, qui suppose que les évènements survenus à différentes dates sont conditionnellement indépendants, connaissant le processus covarié. Cette hypothèse est problématique lorsque l'occurrence d'un événement peut altérer le risque d'un futur événement. Dans un contexte clinique par exemple, les patients qui ont un premier infarctus du myocarde (MI) peuvent être à plus haut risque pour un second. Dans ce travail nous proposons la méthode des séries de cas auto-contrôlés avec dépendance positive (PD-SCCS) : une généralisation de SCCS qui permet que l'occurrence d'un événement puisse augmenter le risque d'un futur événement, mais qui maintient les avantages du modèle original en contrôlant les valeurs de base fixées des covariables et en reposant seulement sur les données des cas. Comme dans le modèle SCCS, les niveaux de base individuels des paramètres sont écartés de la vraisemblance du PD-SCCS. Les sources de données utilisées en surveillance après mise sur le marché peuvent contenir jusqu'à dix millions de sujets, et ainsi dans cette situation il est particulièrement avantageux que PD-SCCS évite d'avoir à faire une coûteuse estimation des paramètres individuels. Nous développons des expressions pour l'inférence sur de grands échantillons et l'optimisation pour PD-SCCS, et nous comparons les résultats de notre modèle généralisé avec l'approche SCCS plus restrictive.

Irène Gijbels and Marek Omelka **137**
Testing for Homogeneity of Multivariate Dispersions Using Dissimilarity Measures

Tester l'homogénéité de dispersions peut être d'un intérêt scientifique propre et également une importante étape pour la vérification des hypothèses d'une analyse. Mais nombreuses données biologiques et écologiques sont asymétriques et chargées en zéros, et de plus le nombre de variables est souvent plus élevé que la taille de l'échantillon. Ceci conduit de nombreuses analyses à ne pas s'appuyer sur des hypothèses paramétriques, mais à utiliser une mesure particulière de dissimilarité pour calculer une matrice de différences par paires, cette matrice étant alors là a base d'inférences statistiques ultérieures. Anderson (2006) a proposé un test d'homogénéité de dispersions multivariées pour un contexte d'ANOVA à un facteur, basé sur une distance, pour lequel une matrice de dissimilarités par paires peut être pris en entrée. L'idée clef, comme dans le test de Levene, est de remplacer chaque observation par sa distance à un centre de groupe estimé. Dans cet article, nous suggérons une approche alternative, basée sur les moyennes

de distances intra-groupe et qui ne requiert pas de calculs au centre du groupe pour l'obtention de la statistique de test. Nous montrons que cette approche peut avoir des avantages autant théoriques que pratiques. Nous décrivons une procédure de permutation qui donne une erreur de type I proche de la valeur souhaitée même pour de petits échantillons.

De nombreuses études de monitorage d'espèces visent à estimer la proportion d'une certaine aire étudiée occupée par une population cible. L'aire d'étude est divisée en sites spatialement disjoints où l'on enregistre la présence ou l'absence de la population, et ceci de façon répétée pour de multiples saisons. Cependant lorsque les sites sont détectés occupés avec une probabilité p <1, le défaut de détection ne signifie pas absence d'occupation réelle. MacKenzie et al (2003, Ecology 84, 2200-2207) ont développé un modèle multi-saison pour estimer l'occupation saisonnière du site (ψ_t) en prenant en compte un p inconnu. Leur modèle a de bonnes performances lorsque les observations sont recueillies selon un dispositif robuste, où plusieurs échantillonnages sont pratiqués à chaque saison ; l'échantillonnage répété aide à l'estimation de p. Cependant leur modèle a de moins bonnes performances lorsque la robustesse du dispositif n'est pas acquise. Dans cet article, nous proposons un modèle alternatif de vraisemblance qui conduit à de meilleures estimations saisonnières de p et de ψ_t lorsque le dispositif n'est pas robuste. Nous construisons la vraisemblance marginale des donnés observées, en conditionnant sur, et en additionnant, le nombre latent de sites occupés à chaque saison. Une étude de simulation montre que dans les cas où le dispositif n'a pas de robustesse, le modèle proposé estime p avec un biais moindre que celui du modèle de MacKenzie et al. et par là améliore les estimations de ψ_t. Nous appliquons les deux modèles à un ensemble de données formé par les observations répétées de présence-absence de merles d'Amérique (Turdus migratorius) avec des périodes de recensement annuelles. Les deux modèles sont comparés à un troisième estimateur utilisable en présence de comptages répétés (de la même étude), le modèle proposé conduisant à des estimations de ψ_t plus proches des estimations obtenues avec le modèle comptage ponctuel.

BIOMETRIC PRACTICE

Cet article est motivé par l'intérêt de la comparaison des inférences faites selon une approche bayésienne ou une approche fréquentiste. Nous nous intéressons aux études relatives aux tests bayésiens de supériorité unilatérale et de non infériorité. Ces tests sont posés en termes de probabilité a posteriori que l'hypothèse nulle soit vraie pour la distribution binomiale, et en termes de limites unilatérales de crédibilité. Nous nous restreignons aux conjugués a priori bêta à paramètres entiers. Sous cette hypothèse, les probabilités a posteriori des hypothèses testées peuvent être transformées en fréquences probabilistes d'essais de Bernoulli avec des nombres ajustés d'évènements et de tailles de population. La méthode ressemble alors à une formulation standard fréquentiste. Par un

choix approprié de paramètres a priori, les probabilités a posteriori concernant l'hypothèse nulle peuvent être rendues plus petites ou plus grandes que les *p*-valeurs des tests fréquentistes.

Même si l'approche cas-contrôle en population non-apparentée constitue, en raison de la facilité de la collecte de données et des analyses statistiques, le plan d'expérience type pour la cartographie des associations génétiques avec des traits complexes, cette méthode pâtit du problème inhérent à la stratification de la population. Certes, il existe des développements méthodologiques ajustant ces études en fonction de la structure génétique interne de la population étudiée, mais il n'y a guère, jusqu'ici, de méthode statistique permettant l'estimation efficace du nombre de sous-populations (K), un paramètre génétique important de l'évolution de la population. Dans cet article, nous proposons une approche bayésienne semi-paramétrique estimant cette structure en supposant que K est aléatoire. A l'aide d'un très grand nombre de simulations, nous montrons que notre méthode n'est pas seulement plus rapide en temps de calcul que la méthode *Structure* – une méthode bayésienne existante –, mais qu'elle s'avère également plus précise dans l'estimation du nombre de sous-populations, ce qui permet d'obtenir une meilleure puissance pour détecter les associations pertinentes dans ce contexte des études cas-contrôle.

Nous nous intéressons à l'inférence pour les expériences de « phage display » pour l'humain avec trois étapes. Les données sont des comptes de tripeptides par tissu et par étape. Le but principal de l'expérience est d'identifier les ligands qui se lient à un tissu donné avec une haute affinité. Nous formalisons la question de cette recherche comme une inférence relative à la monotonicité de comptes moyens dans les différentes étapes. Le but de l'inférence est alors d'identifier une liste de paires peptide-tissu présentant un accroissement significatif au long des étapes. Nous utilisons un mélange de modèles de Poisson dans un processus semi-paramétrique de Dirichlet. La distribution a posteriori sous ce modèle permet l'inférence souhaitée sur la monotonicité des comptes moyens. Cependant, le résumé inférentiel souhaité sous forme de liste de paires peptide-tissu à augmentation significative implique un problème massif de multiplicité. Nous envisageons deux approches alternatives pour cette question de multiplicité. En premier, nous proposons une approche basée sur le contrôle du taux espéré de faux positifs a posteriori. Nous remarquons que la solution qui en dérive ignore la taille relative de l'augmentation. Ceci motive une seconde approche basée sur une fonction d'utilité qui inclut des poids explicites pour la taille de l'augmentation.

A Wavelet-Based Bayesian Approach to Regression Models with Long Memory Errors, and Its Application to fMRI Data

Cet article traite des modèles de régression linéaires avec des erreurs de mémoires longues. Ces modèles on fait leurs preuves en application dans de nombreux domaines comme l'imagerie médicale, le traitement du signal, et l'économétrie. Les ondelettes ont structurellement un lien fort avec les données à longue mémoire. Ici nous utilisons des transformées en ondelettes discrètes comme filtre blanc afin de simplifier les matrices de variances covariances importantes de données. Nous adoptons ensuite une approche bayésienne pour estimer les paramètres du modèle. Notre procédure inférentielle utilise les coefficients de variance exacts des ondelettes et conduit à une estimation satisfaisante des paramètres du modèle. Nous examinons les performances sur des données simulées et nous présentons une application sur un jeu de données fMRI. Dans l'application nous produisons des cartes de probabilité à postériori par identification avec des voxels qu'il est possible d'activer pour une confiance donnée.

Patrick J. Heagerty and Bryan A. Comstock **197**
Exploration of Lagged Associations using Longitudinal Data

Quelques approches statistiques pour l'analyse de données longitudinales exigent des modèles correctement spécifiés pour l'association entre un critère de jugement courant et l'histoire complète des critères de jugement passés et des expositions dépendantes du temps. Il est empiriquement difficile mais motivant de déterminer quels aspects spécifiques de l'histoire du critère de jugement et/ou d'une exposition sont prédicteurs du critère de jugement courant car le nombre potentiel de variables représentant l'histoire peut être relativement grand. Le but de ce manuscrit est d'exposer des méthodes statistiques qui peuvent caractériser des effets retardés et de fournir une approche structurée pour l'analyse de données avec en objectif le développement de modèles appropriés. L'une des principales contributions de ce papier est d'insister sur la possibilité qu'en pratique, des modèles de transition puissent fréquemment nécessiter plus que de simples modèles linéaires et additifs pour les prédicteurs représentant l'histoire des processus du critère de jugement et des variables explicatives. Nous illustrons les concepts à l'aide d'un exemple dans le traitement de l'anémie chez les patients en dialyse et montrons comment des modèles linéaires peuvent être spécifiés avec des dépendances flexibles sur les histoires de l'exposition et/ou du critère de jugement.

Jeremy M. G. Taylor, Yongseok Park, Donna P. Ankerst, Cecile Proust-Lima, Scott Williams, Larry Kestin, Kyoungwha Bae, Tom Pickles, and Howard Sandler **206**
Real-Time Individual Predictions of Prostate Cancer Recurrence Using Joint Models

Les patients qui ont été traités par radiothérapie pour un cancer de la prostate sont suivis à intervalles réguliers à l'aide d'un test biologique, le dosage de l'Antigène Spécifique de Prostate (PSA). Si la valeur du PSA commence à s'élever, c'est le signe que le cancer de la prostate a des risques élevés de rechute, et les patients peuvent souhaiter débuter de nouveaux traitements. Une estimation fiable de la probabilité de rechute du cancer dans

les prochaines années peut aider ces patients à prendre cette décision. Dans cet article, nous décrivons la méthodologie utilisée pour établir la probabilité de rechute pour un nouveau patient, telle que nous l'avons implantée sur le web. Nous utilisons un modèle longitudinal joint à un modèle de survie. Le modèle est développé sur un jeu de données d'apprentissage composé de 2 386 patients et testé sur un échantillon de 846 patients. Des méthodes d'estimation Bayesiennes sont utilisées, avec un algorithme de Monte Carlo-Chaine de Markov (MCMC) développé pour l'estimation des paramètres à partir du jeu de données d'apprentissage, un deuxième algorithme MCMC rapide étant développé pour prédire le risque de rechute en utilisant les mesures longitudinales du PSA d'un nouveau patient.

Les études d'intervention par essais longitudinaux sur le comportement pour réduire le risque de transmission du VIH collectent des données multivariées à plusieurs niveaux, longitudinalement pour chaque sujet, avec d'importantes structures de corrélation entre temps, niveaux et variables. Pour déterminer les interventions qui sont utiles, il est indispensable de pouvoir définir avec précision les effets de ces essais. Aussi bien le nombre de partenaires que le nombre de relations sexuelles avec chaque partenaire est enregistré à chaque date de point. Les relations sexuelles avec chaque partenaire sont distinguées entre relations protégées et relations non protégées avec les différents risques correspondants de transmission VIH/MST. Ces essais ont généralement des critères d'éligibilité limitant l'inclusion aux sujets présentant un certain niveau de risque de comportement sexuel directement lié au point d'intérêt. La combinaison de ces éléments rend difficile la quantification des comportements sexuels et les effets d'intervention. Nous proposons un modèle de comptage multivarié et multi-niveau qui modélise simultanément le nombre de partenaires, mes actes avec chaque partenaire, et prend en compte les contraintes d'éligibilité à l'inclusion. Nos méthodes sont utiles dans l'évaluation des essais d'intervention et procurent un modèle plus précis et complet de comportement sexuel. Nous illustrons les contributions de notre modèle en examinant le comportement de réduction du risque dépendant du statut sérologique du partenaire (comportement séroadaptatif). Nous quantifions plusieurs formes de comportement séroadaptatif, que nous distinguons du comportement non adaptatif.

Cet article rapporte une analyse dont le but est de quantifier l'effet du fingolimod, traitement oral pour la sclérose en plaques (SP) de forme récurrente-rémittente, sur la progression de la maladie. L'approche standard utilise des méthodes d'analyse de survie, ce qui peut être problématique dans les études SP qui présentent un nombre limité de mesures de l'invalidité dans le temps et incluent comme particularité importante à la fois des rechutes et des rémissions. Une approche préférable utilise un modèle de transition de Markov, développé à l'origine dans le cadre de données longitudinales, dont les

propriétés probabilistes spécifiques servent à estimer les courbes de survie pour le délai avant progression de la maladie. Cette approche de transition modélise le processus d'invalidité dans son ensemble et utilise toutes les données de transition disponibles pour l'inférence, tandis que les méthodes de survie se concentrent sur un événement d'intérêt unique et utilise seulement le délai jusqu'à cet événement. Cet article compare l'approche par modèle de transition aux méthodes d'analyse de survie et discute les différences dans l'interprétation des paramètres estimés. Les deux modèles sont appliqués à des données obtenues de deux essais cliniques de phase 3. L'article montre qu'ils concluent à des effets positifs pour le nouveau traitement comparé au placebo et fournissent des estimations similaires de la probabilité de progression de la maladie en fonction du temps. Le modèle de transition permet le calcul de matrices de transition spécifiques par profil de covariables qui décrivent l'effet à court terme du traitement et des autres covariables sur le processus d'invalidité.

On s'intéresse de plus en plus à l'évaluation de l'efficacité de vaccins contre la colonisation par bactéries pathogènes. Les porteurs sains dont le nasopharynx est colonisé ne présentent aucun symptôme, mais sont responsables de transmissions inter-personnelles. La colonisation diffère des autres réponses cliniques car elle présente une fréquence élevée, des récurrences, et n'est observée qu'à l'état prévalent. L'estimation des taux d'acquisition et de clairance d'une colonisation nécessite des prélèvements répétés au cours du temps sur les mêmes sujets, une entreprise coûteuse et invasive. Les contraintes de faisabilité qui pèsent sur les essais vaccinaux utilisant la colonisation comme critère d'efficacité ont amené les chercheurs à conduire des études transversales et à utiliser des méthodes insuffisamment justifiées. Nous présentons deux exemples d'études estimant l'efficacité vaccinale à partir de données transversales sur la colonisation du nasopharynx par Streptococcus pneumoniae (pneumocoque). Ces études offrent un cadre pour la définition et l'estimation de l'efficacité globale ou souche-spécifique en termes de susceptibilité à l'acquisition d'une colonisation en présence d'un grand nombre de souches en interaction mutuelle et de dynamiques de colonisation récurrentes. Nous proposons des estimateurs fondés sur une seule observation de chaque sujet, évaluons leur robustesse et réanalysons les deux essais vaccinaux. D'un point de vue méthodologique, ces estimateurs sont apparentés aux études cas-témoins nichées dans une cohorte et prennent en compte la durée de la période à risque pour définir les témoins.

Les résultats faussement positifs de tests sont l'une des nuisances les plus communes des tests de dépistage, et ils peuvent conduire à entreprendre des procédures de tests diagnostics plus invasives et plus coûteuses. Il est donc important d'estimer le risque cumulé de faux positifs dans les tests de dépistage après des sessions répétées de dépistage afin d'évaluer les régimes de dépistage possibles. Les estimateurs disponibles

pour le risque cumulé de faux positifs sont limités par des hypothèses fortes sur les mécanismes de censure et par des hypothèses paramétriques portant sur la variation du risque entre sessions de dépistage. Pour traiter ces limitations, nous proposons un modèle semi-paramétrique censuré avec biais pour le risque cumulé de faux positifs, qui accepte la censure dépendante sans spécification d'une forme fonctionnelle fixée pour la variation du risque entre les sessions. Des simulations montrent que le modèle censuré avec biais a des performances similaires à celles des modèles existants sous censure indépendante et peut largement éliminer les biais sous censure dépendante. Nous utilisons les modèles existants et le nouveau modèle pour estimer le risque cumulé de faux positifs et la variation du risque en fonction de l'âge de départ et des antécédents familiaux de cancer du sein après 10 ans de mammographies annuelles dans la base de données du Consortium de Surveillance du Cancer du Sein. Ignorer une potentielle censure dépendante dans ce contexte conduit à sous-estimer le risque cumulé de faux positifs. Des modèles qui fournissent des estimateurs justes sous censure dépendante sont cruciaux pour une information appropriée de l'évaluation des tests de dépistage.

Les analyses de tolérance en vue d'estimer l'effet d'un traitement sur la survenue d'un événement indésirable pose un problème statistique difficile, même dans les essais randomisés. En effet, ces événements sont généralement rares car observés dans des études d'efficacité sous-calibrées en terme de puissance pour l'analyse des critères de tolérance. Une méta-analyse des données groupées de plusieurs études peut augmenter la puissance, mais la problématique des données manquantes, ou bien une randomisation en échec, peut introduire un biais. Cet article montre comment un estimation du maximum de vraisemblance ciblée (TMLE) peut être appliquée à une méta-analyse afin de réduire ces biais dans l'estimation des effets de causalité. En outre, il compare les performances du nouvel estimateur avec d'autres de la littérature. Une étude de simulation dans laquelle le mécanisme des données manquantes est au hasard ou tout à fait au hasard souligne les différences entre estimateurs en terme de gains potentiels de biais et d'efficacité. La différence de risque, le risque relatif, et l'odds-ratio de l'effet du traitement sur la mortalité à 30 jours sont estimés à partir des données de huit essais randomisés. Quand un événement est rare, il y a généralement peu d'opportunités pour améliorer l'efficacité des estimateurs. En outre, les associations entre covariables et critères peuvent être difficiles à détecter. TMLE tente d'exploiter ces informations disponibles pour soit atteindre ou dépasser les performances d'un estimateur moins sophistiqué.

Les méthodes basées sur le score de propension (ou méthodes de la probabilité prédite) forment un ensemble d'outils intéressants pour les recherches d'efficacités comparatives et plus généralement pour l'estimation d'effets de causalité. Ces méthodes se composent de deux étapes distinctes : 1) une étape de score de propension où un modèle est ajusté pour prédire la propension à recevoir le traitement (le score de propension), et 2) une

étape de résultat où les réponses sont comparées entre unités traitées et unités non-traitées ayant des valeurs similaires du score de propension estimé. Les techniques traditionnelles effectuent l'estimation de façon distincte dans ces deux étapes ; les estimations de la première étape sont vues comme fixées et connues pour être utilisées dans la seconde. Les méthodes bayésiennes ont un attrait naturel dans ces dispositifs puisque des vraisemblances séparées des deux étapes peuvent être combinées en une seule vraisemblance jointe, l'estimation des deux étapes étant menée en simultané. Une caractéristique clé de l'estimation jointe dans ce contexte est le « feedback » entre l'étape de résultat et celle du score de propension, de sorte que les éléments d'un modèle pour le résultat contribuent à l'information pour la distribution a posteriori des éléments du modèle de score de propension. Nous établissons une évaluation rigoureuse de l'estimation bayésienne du score de propension pour montrer que le feedback du modèle peut conduire à de faibles estimateurs en l'absence de stratégies améliorant l'ajustement des scores de propension par un ajustement à des covariables individuelles. Nous illustrons ce phénomène avec une étude par simulation et avec une investigation d'efficacités comparatives entre la pose de dilatateurs intracarotidiens et l'endarterectomie sur 123,286 bénéficiaires de Medicare hospitalisés pour accident vasculaire cérébral en 2006 et 2007.

La modélisation de la distribution spatiale des espèces est un problème fondamental en écologie. Plusieurs méthodes de modélisation ont été développées, l'une d'entre elles MAXENT , une approche de modélisation basée sur le maximum d'entropie, étant très diffusée. Dans cet article, nous montrons que MAXENT est équivalente à un modèle de régression de Poisson et donc associée à un processus ponctuel de Poisson, dont elle diffère seulement par la constante (intercept), qui est échelle-dépendant dans le cas de MAXENT. Nous illustrons plusieurs améliorations de MAXENT qui se déduisent de ces relations. En particulier, une approche par un modèle de processus ponctuel facilite les méthodes de choix de la résolution spatiale appropriée, d'évaluation de l'adéquation du modèle, et du choix du paramètre de pénalisation LASSO, tous non accessibles par MAXENT. Le résultat d'équivalence représente un pas significatif dans le regroupement des propositions de modélisation des distributions d'espèces.

READER REACTION

Dans le contexte des essais thérapeutiques randomisés, Rosenblum et Van Der Laan (2009, Biometrics 63, 937-945) ont considéré l'hypothèse nulle comme aucun effet du traitement sur la moyenne de la variable d'intérêt à l'intérieure des strates des caractéristiques d'inclusion. Ils ont montré que les tests d'hypothèse basés sur les modèles de régression linéaires ou sur les modèles de régression linéaires généralisés garantissaient d'avoir asymptotiquement une erreur de type I correcte quelque soit la

distribution des données générées, tout en faisant l'hypothèse que l'attribution du traitement était indépendant des covariables.
Nous considérons ici une autre variable d'intérêt importante dans les essais thérapeutiques randomisés : le délai entre la randomisation et la date du premier échec, et notre hypothèse nulle est alors qu'il n'y a aucun effet du traitement sur la fonction de survie conditionnellement à une série de caractéristiques d'inclusion. Par une application directe des arguments de Rosenblum et Van der Laan (2009), nous montrons que les tests d'hypothèses basés sur les modèles de risques multiplicatifs avec une fonction de lien exponentielle i.e. modèles de hasards proportionnels et sur des modèles de risques multiplicatifs avec une fonction de lien linéaire où le risque de base peut être paramétré ou laissé non spécifié, sont asymptotiquement valides malgré de mauvaises spécifications du modèle, à condition d'avoir une distribution des censures indépendante de l'attribution du traitement pour une profil donné de covariables.