
Translations of Abstracts

Presidential Address**A. Mead****1***Biometry and the IBS – Strength through Diversity*

La Société Internationale de Biométrie (IBS) rassemble des membres d'origines culturelles diverses, organisés en Régions définies sur une base géographique et en Groupes Nationaux, s'intéressant à des centres d'intérêts variés aussi bien dans les thèmes méthodologiques que dans les domaines d'applications. Nous regardons brièvement comment le développement historique de notre science, de notre société et des conférences internationales reflète cette diversité, en nous attardant particulièrement sur l'histoire des régions Britannique et Irlandaise de l'IBS. Puis, en partant de la diversité culturelle et géographique de notre Société ainsi que de la diversité scientifique de la Société et des biométriciens, nous identifions certaines de nos forces (sujets diversifiés pour les rencontres organisées partout dans le monde, application des méthodes biométriques à des domaines divers, organisation de notre Société par des membres d'origines variées), ainsi que certains enjeux du moment (la diffusion électronique du journal et d'autres informations, le spectre des domaines d'application concernant les membres de la Société, l'amélioration des relations avec les sociétés scientifiques rassemblant ceux qui motivent nos recherches). Pour finir, nous mettons en évidence la diversité des problèmes scientifiques que chacun d'entre nous rencontre dans son activité de biométricien.

Biometric Methodology**S. Lin and J. Ding****9***Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies*

Un des défis principaux rencontrés par les chercheurs étudiant les systèmes biologiques complexes est l'intégration des données issues de plateformes « omiques ». Les données à échelle « omique » incluent les variations d'ADN, les profils transcriptomiques et les données RA-omiques. La sélection d'une approche appropriée pour l'intégration des données dépend du problème considéré et est en premier dictée par l'information contenue dans les données. Dans des situations où la modélisation conjointe de multiples

jeux de données brutes est extrêmement complexe du fait de leurs grandes différences, les classements obtenus avec chaque jeu de données représentent une standardisation permettant ensuite l'intégration des résultats. L'agrégation de cibles microRNA prédites à l'aide de différents algorithmes de calcul est un problème de ce type. L'intégration de résultats provenant de plusieurs études mRNA basées sur différentes plateformes est un second exemple qui sera également discuté. En formulant le problème de l'intégration de classements comme la minimisation d'un critère objectif, nous explorons l'usage d'une méthode de type Monte-Carlo basée sur l'entropie croisée pour résoudre un tel problème combinatoire. Au lieu de considérer une loi uniforme discrète sur l'ensemble des solutions potentielles, une technique d'échantillonnage itérative est utilisée pour « resserrer peu à peu le filet », de façon à placer plus de poids sur la solution optimale et ses voisins. Des simulations approfondies ont été conduites pour évaluer la performance de cette méthode, avec des résultats satisfaisants. La méthode a été ensuite appliquée à des données microRNA et mRNA pour illustrer son utilité.

W. N. van Wieringen and M. A. van de Wiel

19

Nonparametric Testing for DNA Copy Number Induced Differential mRNA Gene Expression

Le dogme de la biologie moléculaire relie l'ADN avec l'ARNm. Les puces CHG (hybridation génomique comparative) mesurent le nombre de copie d'ADN et les micropuces mesurent la quantité d'ARNm. Les méthodes qui intègrent les données issues de ces deux plateformes peuvent permettre de découvrir des interactions biologiques significatives favorisant notre compréhension du cancer. Nous développons des tests non paramétriques pour la détection d'expressions différentielles de gènes induites par des différences dans le nombre de copie. Ces tests prennent en compte l'incertitude de lecture des aberrations génomiques. Le test est précédé par un 'algorithme de réglage' qui ne tient pas compte de certains gènes afin d'augmenter la puissance globale de la procédure de sélection par FDR. En outre, les statistiques de test sont 'écrasées' pour capturer l'information parmi des gènes voisins qui partagent la même signature de puce CGH. Pour chaque gène, nous estimons également son effet, la quantité de sa différence d'expression qui est due à des changements de nombre de copie et nous calculons le coefficient de détermination. Notre méthode est illustrée par l'analyse de données sur le cancer du sein où elle retrouve les résultats précédemment rapportés mais avec des arguments statistiques plus profonds.

J. Yang, R. Wu, and G. Casella

30

Nonparametric Functional Mapping of Quantitative Trait Loci

La cartographie fonctionnelle est un outil utile pour l'identification des QTL (quantitative trait loci) qui contrôlent les traits dynamiques. Elle incorpore les aspects mathématiques des processus biologiques dans le cadre de mélange de vraisemblances modèles-dépendantes pour la cartographie des QTL ce qui augmente la puissance de détection des QTL et la précision de l'estimation des paramètres. Cependant, dans de nombreuses situations il n'y a pas de forme fonctionnelle évidente et dans ce cas cette stratégie n'est pas optimale. Nous proposons d'utiliser une fonction d'estimation non paramétrique,

typiquement implémentée avec des B-splines pour estimer la forme fonctionnelle sous-jacente des trajectoires phénotypiques et construire un test non paramétrique pour identifier des QTL. Utilisant une représentation de la régression non paramétrique comme un modèle mixte, la statistique finale du test est un rapport de vraisemblance. Nous considérons deux types de carte génétique : une carte dense et une carte générale et nous évaluons la puissance de la cartographie fonctionnelle non paramétrique au travers de simulations et d'exemples réels.

Y. C. Tai and T. P. Speed

40

On Gene Ranking Using Replicated Microarray Time Course Data

Supposons que l'on souhaite trier une liste de gènes issue de l'analyse temporelle de données microarray où plusieurs conditions biologiques sont étudiées et où les gènes d'intérêt sont ceux dont le profil au cours du temps varie entre conditions. Nous décrivons une statistique Bayésienne empirique et multivariée multi-échantillon pour trier les gènes en fonction de leur expression différentielle en utilisant l'information des données microarray longitudinales et transversales. Notre modèle longitudinal multi-échantillon suppose que les données temporelles sont indépendantes et identiquement distribuées selon des vecteurs multivariés normaux. D'autre part, nous construisons un modèle transversal basé une analyse de régression normale prenant en compte la structure de notre plan d'expérience. Dans les deux cas, nous utilisons dans notre modèle Bayésien les priors conjugués naturels pour nous approcher au mieux de la solution des odds à posteriori. Des simulations et deux études utilisant des données temporelles de microarray publiées sur le vers et la souris indiquent que les approches que nous proposons possèdent de bonnes propriétés.

G. Y. Qin and Z. Y. Zhu

52

Robustified Maximum Likelihood Estimation in Generalized Partial Linear Mixed Model for Longitudinal Data

Nous étudions, dans cet article, les estimations robustes simultanées de moyennes et de composantes de variance dans des modèles mixtes linéaires généralisés basées sur la construction d'une fonction de vraisemblance robuste. Les propriétés asymptotiques des estimateurs robustes proposés sont démontrées sous certaines conditions de régularité. Quelques simulations sont menées afin d'étudier la performance des estimateurs robustes proposés. Comme attendu ces estimateurs sont meilleurs que ceux résultant des estimations robustes impliquant des espérances conditionnelles tels ceux de Shina (2004) et Qin et Zhu (2007). En fin d'article la méthode robuste proposée est illustrée par l'analyse d'un jeu de données réel.

P. X.-K. Song, M. Li, and Y. Yuan

60

Joint Regression Analysis of Correlated Data Using Gaussian Copulas

Nous présentons une nouvelle approche pour la modélisation d'observations corrélées. Nous utilisons des copules gaussiennes pour concevoir un dispositif flexible qui permet d'associer des modèles linéaires généralisés unidimensionnels dans une analyse

conjointes de réponses corrélées qui peuvent être continues, discrètes ou mixtes. Cette démarche conduit à un analogue multivarié de la théorie des modèles linéaires généralisés univariés et à des estimateurs plus efficaces des coefficients de régression. L'existence des modèles de probabilité conjoints permet de construire des estimateurs du maximum de vraisemblance. Les exemples concernent des modèles de régression pour des observations corrélées discrètes, en particulier des modèles logistiques multidimensionnels et un modèle pour réponses mixtes, normale et binaire. Nous utilisons des simulations pour comparer ce modèle joint fondé sur les copules aux équations d'estimation généralisées (GEE), la méthode d'estimation utilisée classiquement pour associer des modèles univariés. Deux jeux de données réelles sont également présentés en illustration.

G. Marshall, R. De la Cruz-Mesía, F. A. Quintana, and A. E. Barón 69
Discriminant Analysis for Longitudinal Data with Multiple Continuous Responses and Possibly Missing Data

La caractérisation pertinente d'un effet nécessite souvent la prise en compte simultanée de différents aspects dudit effet, autrement dit la prise en compte de plusieurs critères. Cet article examine des modèles statistiques capables de reclasser correctement des unités dans deux ou plusieurs groupes prédéfinis, à partir de la mesure – répétée dans le temps – de plusieurs critères. Ces critères sont mis en relation dans un modèle linéaire mixte multivarié permettant de décrire les évolutions au sein des groupes. De par sa flexibilité, on peut utiliser l'approche à effets aléatoires pour modéliser conjointement les critères, et ainsi estimer les paramètres de population d'une analyse discriminante classifiant les unités dans des groupes (ou populations) distinct(e)s et prédéfini(e)s. Cette estimation des paramètres s'effectue par l'algorithme EM, avec une étape d'approximation linéaire. Des simulations éclairent les conséquences de cette approximation sur la classification. Nous présentons un exemple issu d'une étude portant sur 161 femmes enceintes à Santiago du Chili, dont l'objet était de prédire le bon déroulement (ou non) de la grossesse.

R. Tsonaka, G. Verkeke, and E. Lesaffre 81
A Semi-Parametric Shared Parameter Model to Handle Nonmonotone Nonignorable Missingness

Les études longitudinales génèrent souvent des réponses incomplètes selon un mécanisme manquant non aléatoire. Les modèles à paramètres communs représentent un cadre attrayant pour la modélisation jointe des processus de mesures et manquant, surtout dans le cas de manquement non-monotone, et supposent qu'un ensemble d'effets aléatoires détermine l'inter-dépendance. Des hypothèses paramétriques sont typiquement faites sur la distribution des effets aléatoires, dont la violation conduit à une mauvaise spécification du modèle avec un effet potentiel sur les estimations des paramètres et de leurs écart-types. Dans ce papier nous évitons toutes hypothèses paramétriques de la distribution des effets aléatoires et la laissons entièrement non spécifiée. L'estimation du modèle est alors réalisée par une méthode de maximum de vraisemblance semi-paramétrique. Notre proposition est illustrée sur une étude longitudinale randomisée portant sur des patients souffrant de polyarthrite rhumatoïde présentant des manquements non-monotones.

La capacité des essais cliniques randomisés à étudier les effets des traitements est souvent mise à mal par manque d'observance, ce que les analyses standards ignorent (analyse en intention de traiter) ou bien intègrent mais de telle sorte que l'estimateur obtenu ne décrit plus le seul lien de cause à effet (analyse selon le traitement réel). L'analyse du « complier average causal effect » (CACE) constitue une alternative à ces approches : il s'agit d'estimer l'effet traitement moyen sur la sous-population des patients qui seraient observants quel que soit le traitement qui leur serait alloué. Nous nous intéressons ici au cas d'un essai clinique où certains patients randomisés dans le bras contrôle prendraient en fait le traitement - et vice versa -, et où l'évaluation de la réponse serait parfois manquante. Nous développons des estimateurs du CACE issus de méthodes d'imputation multiple, lesquelles ont été appliquées avec succès à de nombreux problèmes de données manquantes mais jamais encore dans le cadre de l'inférence causale. Nous étudions par simulation les propriétés non asymptotiques (sur échantillon de taille finie) de ces estimateurs ainsi que celles de procédures concurrentes. Nous appliquons enfin nos méthodes, à titre d'illustration, à un essai réel de l'efficacité d'un vaccin anti-grippal (la randomisation portant sur un rappel effectué ou non à l'intention des médecins).

Cet article se penche sur l'analyse d'essais randomisés à deux bras, avec des défauts d'observance, dans le cas d'une distribution multinomiale de la variable-réponse. Dans de tels essais, nous définissons en premier lieu l'« effet causal » comme une fonction, estimée par la méthode du maximum de vraisemblance, qui va décrire et résumer la distribution de la réponse chez les patients observants, avec et sans traitement (un exemple d'une telle fonction est le CACE, « Complier Average Causal Effect », mesure stochastique de supériorité d'un traitement sur un autre chez les patients observants, également appelée ECMC, « Effet Causal Moyen du Conformant »). A l'aide de ladite fonction et de la statistique du rapport des vraisemblances, nous testons ensuite l'hypothèse de l'identité des distributions de la variable-réponse chez les patients observants, avec et sans traitement. Si la statistique du rapport des vraisemblances suit asymptotiquement une loi du Khi-deux lorsque les vraies valeurs des paramètres se situent à l'intérieur de l'espace des paramètres sous l'hypothèse nulle, cette propriété n'est plus vraie lorsque ces valeurs se trouvent sur la frontière de l'hypothèse nulle. C'est pourquoi, dans ces essais, nous proposons d'utiliser une version bootstrap (simple ou double) du test du rapport de vraisemblances. Ces méthodes sont illustrées par l'analyse de données issues d'un essai randomisé effectué dans un contexte de soins de santé primaires, analysant l'effet d'interventions éducatives encourageant la bonne observance de traitements antidépresseurs prescrits chez des patients âgés.

Les diagnostics classiques pour les modèles d'équations structurelles reposent sur des formes agrégées des données, et sont maladaptés pour la vérification d'hypothèse distributionnelles ou de linéarité. Nous étendons des tests d'ajustement récemment développés pour des données corrélées basées sur des résidus sujet-spécifiques, à des modèles d'équations structurelles avec variables latentes. Les tests proposés se prêtent eux-mêmes à des représentations graphiques et sont adaptés à la détection de spécification incorrecte de distribution ou de linéarité. Pour compléter les représentations graphiques, on définit des statistiques de test ; les distributions sous hypothèse nulle pour ces statistiques de test sont approximées par des techniques très efficaces de simulation. Les propriétés des tests proposés sont étudiées par des simulations. Nous illustrons les méthodes à l'aide de données provenant d'une étude sur l'exposition au plomb in-utero.

Nous proposons des outils de diagnostic de données influentes pour des modèles complexes de survie. Nous développons des outils de diagnostic de données influentes à la fois sur les distributions a posteriori jointe et marginale en se basant sur la divergence de Kullback-Leibler (divergence K-L). Nous présentons une expression simplifiée pour calculer la divergence K-L entre la distribution a posteriori basée sur l'ensemble des données et la distribution a posteriori basée sur l'exclusion d'un seul cas. De plus, nous examinons sa relation à l'ordonnée conditionnelle prédictive (CPO). Tous les calculs relatifs aux mesures de diagnostic proposées peuvent être faits facilement en utilisant des échantillons par méthodes de Monte-Carlo par chaînes de Markov (MCMC) de la distribution a posteriori sur les données complètes. Nous considérons le modèle de Cox avec un processus gamma comme distribution a priori sur le risque cumulé de base. Nous présentons aussi une relation théorique entre nos outils diagnostiques de suppression de cas et les outils diagnostiques basés sur la vraisemblance partielle de Cox. Un exemple basé sur des données simulées et deux exemples basés sur des données réelles sont présentés pour illustrer la méthode.

Mesurer la concordance est souvent intéressant dans les études cliniques afin d'évaluer la similarité des mesures produites par différents évaluateurs ou différentes méthodes sur les mêmes sujets. Nous proposons un coefficient kappa pondéré modifié pour mesurer la concordance de temps de survie bivariés discrets. Le coefficient kappa proposé prend en compte la censure en redistribuant les observations censurées dans une plage où les événements non observés pourraient survenir. Un kappa pondéré modifié généralisé est proposé pour analyser la concordance de temps de survie discrets multivariés. Nous estimons les coefficients kappa modifiés de façon non paramétrique par l'intermédiaire

de l'estimateur de la fonction de survie multivariée. Les propriétés asymptotiques des estimateurs kappa sont présentées et les performances de ces estimateurs analysés par des études de simulation de données de survie bivariées et trivariées. Nous illustrons l'application de ce coefficient kappa modifié en présence d'observations censurées à partir des données d'une étude portant sur le cancer de la prostate.

L. Kong and J. Cai

135

Case-Cohort Analysis with Accelerated Failure Time Model

Dans une étude cas-cohorte, les covariables sont seulement recueillies pour une sous-cohorte sélectionnée de manière aléatoire à partir de la cohorte entière, ainsi que pour des cas supplémentaires en dehors de la sous-cohorte. Cette stratégie est appropriée pour les grandes études de cohortes de maladies rares, en particulier, quand les expositions d'intérêt sont coûteuses à évaluer pour tous les sujets. Nous proposons des méthodes statistiques pour l'analyse de données issues d'études de type cas-cohorte avec un modèle semi-paramétrique à risque accéléré qui interprète les effets des covariables pour accélérer ou décélérer le délai jusqu'à l'échec. Les propriétés asymptotiques des estimateurs proposés sont développées. Les propriétés de l'estimateur cas-cohorte pour un échantillon fini et son efficacité relative par rapport à l'estimateur pour une cohorte entière sont évaluées via des études de simulations. Une étude de maladie cardiovasculaire est fournie comme exemple pour illustrer la procédure d'estimation.

H. Frydman and M. Szarek

143

Nonparametric Estimation in a Markov "Illness-Death" Process from Interval Censored Observations with Missing Intermediate Transition Status

Dans la plupart des essais cliniques, les patients sont suivis de manière intermittente pour évaluer la transition d'un état intermédiaire, tel que l'occurrence d'un événement non-létal lié à la maladie et le décès. L'estimation de la distribution du délai de survie sans événement non-létal, c'est à dire, le délai jusqu'à la première occurrence de cet événement non-létal ou le décès, est l'objectif principal de l'analyse des données. La difficulté de l'estimation de l'évaluation intermittente produit deux formes de données incomplètes : d'une part, les délais de l'occurrence des événements non-létals sont censurés par intervalle et d'autre part, quand un événement non-létal ne se produit pas avant la date de la dernière évaluation, le statut de l'événement non-létal pour le patient n'est pas connu entre la date de la dernière évaluation et la date de fin du suivi. Nous considérons ces deux formes de données incomplètes conjointement dans une structure de modèle « maladie-décès ». Nous développons une estimation non-paramétrique du maximum de vraisemblance (ml) dans un modèle « maladie-décès » à partir d'observations censurées par intervalle avec états de transition intermédiaire manquants. Nous montrons que les estimateurs ml sont consistants et nous proposons un algorithme pour les obtenir. Ce travail fournit alors une nouvelle méthodologie pour l'analyse de données de survie incomplètes des essais cliniques. Nous appliquons cette méthodologie à des données issues d'un essai clinique récent sur le cancer (*Bonner et al.*, 2006) et nous comparons nos résultats avec ceux obtenus selon les recommandations de la FDA.

Dans l'analyse de la courbe caractéristique de la performance d'un test (ROC) dépendant du temps en présence de plusieurs marqueurs, il est important de trouver une combinaison adéquate de marqueurs pour améliorer la précision des prédictions du statut vital dans le temps. En se basant sur des données de survie censurées, nous proposons une procédure d'estimation plus flexible pour la combinaison optimale des marqueurs dans le cadre d'un modèle linéaire généralisé pour le délai d'apparition de l'événement sans hypothèses restrictives sur le processus de censure. La convergence des estimateurs proposés est aussi établie dans cet article. En revanche, l'approche basée sur la pondération par l'inverse de la probabilité (IPW) peut introduire un biais quand les probabilités de sélection sont mal spécifiées dans les équations d'estimation. Les performances des deux procédures d'estimation sont examinées et comparées par une série de simulations. Il découle des simulations que les estimateurs proposés sont bien supérieurs aux estimateurs IPW. L'application de ces méthodes à une cohorte d'angiographies montre que notre procédure d'estimation se révèle utile pour prédire le délai jusqu'au décès de toute cause ou de cause liée à une maladie coronarienne.

Dans les méthodes de classification multi-catégorielles, les techniques standards traitent typiquement toutes les classes de façon égale. Ce traitement peut poser problème quand l'ensemble des données est déséquilibré au sens où certaines classes ont de très faibles proportions en comparaison avec les autres classes. Les classes minoritaires peuvent être ignorées ou ne pas être prises en compte durant le processus de classification à cause de leurs faibles proportions. Cela peut poser un sérieux problème si ces classes minoritaires sont importantes. Dans cet article nous étudions le problème des classifications déséquilibrées et proposons un nouveau critère pour mesurer la qualité de la classification. De plus nous proposons trois différentes procédures pondérées d'apprentissage, deux d'entre elles sont à une étape et la troisième est adaptative. Nous démontrons les avantages de ces nouvelles procédures en utilisant des SVM multi-catégorielles sur des données simulées ou réelles. Nos résultats indiquent que la méthodologie proposée peut effectivement traiter les problèmes de classification déséquilibrée.

Lors de la réalisation d'une analyse de la variance, l'investigateur a souvent deux objectifs principaux : déterminer les facteurs présentant un effet significatif sur la réponse, puis parmi ces facteurs trouver les différences entre les niveaux. Les comparaisons de niveaux sont faites à l'aide d'une analyse à posteriori basée sur des comparaisons deux à deux. Ce papier propose une approche de régression sous contrainte

particulière pour répondre simultanément aux deux objectifs à l'aide d'une restriction « shrinkage » dans une simple procédure automatisée. La configuration de la restriction possède la propriété de faire disparaître des niveaux dans un facteur en imposant à leurs effets d'être égaux pendant que dans le même temps on arrive à une sélection de facteurs par une mise à zéro de facteurs entiers. L'utilisation de cette approche conduit à l'identification d'une structure au sein de chaque facteur de telle façon que certains niveaux peuvent disparaître, pour constituer des groupes. En opposition avec les méthodes traditionnelles de comparaisons, ces groupes ne se recouvrent pas et ainsi les résultats sont interprétables en termes de sous-ensembles distincts de niveaux. La démarche proposée est montrée pour avoir la propriété « oracle » c'est-à-dire qu'asymptotiquement elle se comporte comme si l'exacte structure était connue au préalable. Une simulation et des données réelles montrent la puissance de cette méthode.

J. Y. Dai, M. LeBlanc, and C. Kooperberg

178

Semiparametric Estimation Exploiting Covariate Independence in Two-Phase Randomized Trials

Des résultats récents d'échantillonnage cas-témoin suggèrent que, sous l'hypothèse d'indépendance gène-environnement, une estimation semi-paramétrique utilisant cette indépendance aboutit à un gain important d'efficacité. Nous considérons l'estimation efficace de l'interaction traitement-biomarqueur dans un échantillonnage en deux étapes dans le cadre d'essais cliniques randomisés avec prise en compte de l'indépendance entre un traitement randomisé et les valeurs de départ des marqueurs. Nous développons un algorithme de Newton-Raphson appliqué à la vraisemblance pour dériver l'estimateur semi-paramétrique du maximum de vraisemblance (SPMLE). Notre algorithme s'applique à la fois à des réponses continues à la première étape et des valeurs continues des bio-marqueurs à la deuxième étape. La matrice d'information est établie explicitement par différentiation numérique. Dans certaines situations où le calcul du SPMLE est lent, nous proposons un estimateur du maximum de vraisemblance estimé (MELE) qui peut lui aussi prendre en compte l'indépendance des covariables. Cette approche de la vraisemblance estimée, qui utilise une distribution empirique des covariables, est facile à maximiser. Elle offre une estimation explicite de la variance dont la variance est peu augmentée par rapport à l'estimation efficace SPMLE. Nos résultats suggèrent que la prise en compte de l'indépendance des covariables dans un échantillonnage en deux étapes augmente sensiblement l'efficacité, en particulier pour estimer les interactions traitement-biomarqueur.

B. Rosner and R. J. Glynn

188

Power and Sample Size Estimation for the Wilcoxon Rank Sum Test with Application to Comparisons of C Statistics from Alternative Prediction Models

Le test U de Wilcoxon Mann-Whitney (WMW) est couramment utilisé pour faire une comparaison non paramétrique de deux groupes lorsque la Normalité de la distribution sous-jacente est discutable. Il y a eu déjà quelques travaux sur l'estimation de la puissance fondée sur cette procédure. Toutefois, l'erreur de type II du test WMW a été seulement

examinée pour certaines distributions (*Pettitt & Siskind, 1981*). Dans cet article, nous présentons une méthode d'estimation de l'erreur de type II qui est applicable à toute distribution continue, et qui peut également être étendue au traitement des données continues regroupées acceptant des ex aequo. Nous appliquons ces résultats pour obtenir les erreurs standard de l'aire sous la courbe caractéristique d'exploitation récepteur (AUROC) pour des règles de prédiction de risque sous H_1 et pour comparer des « AUROC » entre des règles de prédiction de risque appliquées aux mêmes ensembles de données. Ces résultats sont basés sur les fonctions que l'on peut appeler dans SAS qui évaluent l'intégrale Normale à deux dimensions et sont donc faciles à mettre en œuvre avec un logiciel standard.

B. Bornkamp and K. Ickstadt

198

Bayesian Nonparametric Estimation of Continuous Monotone Functions with Applications to Dose-Response Analysis

Dans cet article on considère la régression non paramétrique dans un cadre Bayésien. La fonction monotone à ajuster est modélisée comme un mélange de fonctions de répartition, différant par la position et l'échelle, une mesure de probabilité générale étant utilisée comme prior pour la fonction de mélange. On étudie le choix de la fonction de distribution paramétrique sous-jacente et on trouve qu'une fonction décrivant une puissance bilatérale est bien adaptée aussi bien du point de vue mathématique que calculatoire. Le modèle se rattache aux modèles non-linéaires traditionnels utilisés dans les études doses-réponses et fournit des possibilités de choix d'une prior informative sur certains des aspects de la courbe. Par une étude de simulation on compare la présente méthode avec une approche récente de régression non paramétrique monotone. On l'illustre avec une application à des données réelles d'une étude doses-réponses.

C. A. Calder, P. F. Craigmile, and J. Zhang

206

Regional Spatial Modeling of Topsoil Geochemistry

L'information géographique concernant les niveaux de toxiques dans le milieu environnemental est communément utilisée dans les études de santé régionale environnementale quand les mesures directes d'exposition personnelle sont limitées ou indisponibles. Dans ce papier, nous proposons un cadre statistique pour l'analyse de la distribution spatiale des propriétés géochimiques des couches arables, incluant les concentrations de produits toxiques variés. A cause d'une hétérogénéité de faible échelle des processus de la plupart des couches arables géochimiques, les mesures directes de ces processus eux-mêmes fournissent seulement une information hautement localisée ; il est alors financièrement prohibitif d'étudier les schémas spatiaux de ces processus à travers une large région en utilisant les analyses géostatistiques traditionnelles de ces données ponctuelles couche arable. Il est plutôt de pratique courante d'accéder à des schémas géochimiques à une échelle régionale en utilisant des mesures ponctuelles collectées en flux de sédiment depuis que, contrairement aux données couche arable, les mesures géochimiques de flux de sédiment individuelles sont représentatives du milieu environnant. Nous proposons un nouveau modèle sols multiéchelle (MSS) qui synthétise de façon formelle les données collectées dans la couche arable et par flux de sédiment et

qui permet l'utilisation d'une information en flux de sédiment plus riche pour informer sur le processus des couches arables qui est plus pertinente dans les études de santé environnementale. Notre modèle s'adapte à l'hétérogénéité de petite échelle des processus géochimique des couches arables en modélisant la dépendance spatiale à une résolution agrégée correspondant aux régions hydrologiquement similaires connus comme les lignes de partage des eaux. Nous présentons une analyse des niveaux d'arsenic, un métal hautement toxique, dans la couche arable dans le MidWest des Etats-Unis en utilisant le modèle MSS et montrons que ce modèle a de meilleures capacités prédictives que les approches alternatives utilisant les modèles statistiques plus conventionnels pour les données ponctuelles spatiales.

Z. Pawlas, J. R. Nyengaard, and E. B. Vedel Jensen

216

Particle Sizes from Sectional Data

Nous proposons une nouvelle approche statistique de recherche d'information concernant les distributions des tailles de particules à partir de données sectionnelles, sans hypothèse spécifique sur la forme des particules. La méthode se base sur des avancées récentes en stéréologie locale. Nous montrons comment estimer séparément à partir de données sectionnelles d'une part la variance due à la procédure d'estimation stéréologique locale et d'autre part la variance due à la variabilité des tailles de particules dans la population. On donne également des méthodes pour juger la différence entre la distribution des tailles estimées des particules et la distribution des tailles réelles des particules.

R. M. Fewster, S. T. Buckland, K. P. Burnham, D. L. Borchers, P. E. Jupp, J. L. Laake, and L. Thomas

225

Estimating the Encounter Rate Variance in Distance Sampling

La source de variance la plus importante dans l'échantillonnage par ligne-transect est habituellement la variance du taux d'observation. Des plans d'étude systématique sont souvent utilisés pour réduire la vraie variabilité entre les différentes réalisations du plan, mais l'estimation de la variance est difficile et les estimateurs approximent typiquement la variance en traitant le plan comme un simple échantillon aléatoire des lignes. Nous explorons les propriétés des différents estimateurs de la variance du taux d'observation sous les plans systématiques et aléatoires. Nous montrons qu'un estimateur de la variance basé sur le plan donne de meilleurs résultats que l'estimateur basé sur le modèle proposé par Buckland et al. (2001, p.79) quand les transects sont positionnés de façon aléatoire. Cependant, si les populations montrent clairement des tendances spatiales fortes, les deux estimateurs peuvent avoir un biais positif important sous les plans systématiques. Nous montrons que la post-stratification permet de réduire ce biais de manière efficace.

B. T. McClintock, G. C. White, M. F. Antolin, and D. W. Tripp

237

Estimating Abundance Using Mark-Resight When Sampling is With Replacement or the Number of Marked Individuals is Unknown

Bien que les méthodes de marquage-réexamen puissent souvent représenter des moyens peu coûteux et peu invasifs pour estimer l'abondance dans les programmes de

surveillance au long cours de population, les estimations produites ont deux limitations majeures car ils requièrent un échantillonnage sans remise et/ou la connaissance exacte du nombre d'individus marqués accessibles au réexamen. Ces contraintes peuvent être souvent difficiles à satisfaire. Nous nous y intéressons ici en introduisant les modèles à effets mixtes log-Poisson et lognormal-Poisson à troncature de zéro (respectivement PNE et ZPNE). Le cadre général de ces modèles permet l'utilisation efficiente de covariables dans la modélisation du taux de réexamen et des paramètres d'hétérogénéité individuelles, la sélection de modèles avec inférence multi-modèle, et l'incorporation de marquages individuels non identifiés. Les modèles peuvent être implémentés en utilisant les programmes habituels de calcul statistique, mais nous les avons aussi ajoutés au programme libre de marquage-recapture MARK. Nous illustrons l'utilisation et l'avantage de (Z)PNE en utilisant des données récemment collectées dans le Colorado sur le chien de prairie à queue noire. Nous étudions également la performance relative des modèles par des études de simulation. En comparaison aux autres estimateurs proposés, nous trouvons généralement que (Z)PNE est plus précis avec peu ou pas de perte dans la couverture de l'intervalle de confiance. Avec l'introduction récente du modèle à effets mixtes logit-normal et (Z)PNE, un cadre plus flexible et efficient pour l'estimation d'abondance par marquage-réexamen est maintenant disponible pour les échantillonnages les plus fréquents dans ces études.

A. I. Bandos, H.E. Rockette, T. Song, and D. Gur

247

Area under the Free-Response ROC Curve (FROC) and a Related Summary Index

L'évaluation des systèmes diagnostiques en « réponse-libre » est de plus en plus utilisée dans les domaines concernés par la détection, la localisation ou la classification d'une ou plusieurs anomalies chez un sujet. Une courbe ROC en réponse libre (FROC) permet d'évaluer la performance d'un système en réponse libre à tous les seuils de décision simultanément. En analyse des courbes ROC, il est classique de résumer l'ensemble de la courbe par un indice unique, comme l'aire sous la courbe. En revanche, il n'existe pas d'indice consensuel permettant de résumer la performance d'un système évalué sous le paradigme FROC. Nous proposons un nouvel indice résumant la performance en réponse libre simultanément à tous les seuils de décision et nous développons un modèle non paramétrique pour l'analyser. En termes algébriques, l'indice résumé que nous proposons correspond à l'aire sous la courbe FROC empirique, pénalisé en fonction du nombre de marques erronées, récompensé en fonction de la proportion d'anomalies décelées et ajusté sur la taille de la cible (ou rayon d'acceptation). En termes géométriques, l'indice peut s'interpréter comme une mesure de supériorité moyenne de la procédure par rapport à un processus artificiel de décision « au hasard » en réponse libre et il est analogue à l'aire comprise entre la courbe ROC et la diagonale qui correspond à un « tirage au hasard » de la décision. Nous développons un estimateur bootstrap idéal de la variance qui peut être utilisé pour construire des intervalles de confiance du bootstrap asymptotiques sans rééchantillonnage ou pour le calcul du nombre de sujets nécessaires. La procédure proposée ne dépend d'aucune hypothèse paramétrique et n'exige aucune hypothèse d'indépendance des réponses intra-sujet. Nous illustrons la méthode à l'aide de données d'imagerie médicale et nous présentons des simulations qui établissent l'intérêt de la procédure pour les tailles d'échantillon et l'ordre de grandeur des paramètres

considérés.

A. Crépet, H. Harari-Kermadec, and J. Tressou

257

Using Empirical Likelihood to Combine Data: Application to Food Risk Assessment

Dans cet article, nous proposons une méthode construite à partir de la vraisemblance empirique qui permet de combiner différents jeux de données, des enquêtes de consommation et des mesures de contamination, dans le cadre de l'estimation d'un risque alimentaire. L'indice de risque pertinent dans ce cadre est la probabilité que l'exposition au contaminant considéré dépasse la dose établie comme tolérable. Cet indice est donné par une fonction non linéaire des distributions des consommations et des contaminations qui prend la forme d'une U-statistique. Ce problème de non linéarité, ajouté à la grande taille des jeux de données, rend impossible l'évaluation directe de l'indice de risque. Nous proposons de considérer un problème approché à l'aide de U-statistiques incomplètes et de techniques de linéarisation. On obtient alors des intervalles de confiance asymptotique pour l'indice de risque. Nous proposons également une méthode alternative basée sur la vraisemblance empirique Euclidienne, ce qui revient à remplacer la divergence de Kullback-Leibler utilisée par la vraisemblance empirique par la distance Euclidienne. Ces deux méthodes sont validées par des simulations puis appliquées pour évaluer le risque lié à la présence de méthyle mercure dans le poisson et les produits de la mer

Biometric Practice

J. A. Royle

267

Analysis of Capture-Recapture Models with Individual Covariates Using Data Augmentation

Cet article concerne l'analyse de données de capture-recapture avec des covariables individuelles ayant un effet sur la probabilité de détection. Nous développons une procédure bayésienne basée sur la vraisemblance conjointe, en utilisant un cadre flexible d'augmentation des données, qui facilite l'analyse par simulations de Monte-Carlo sur chaînes de Markov (MCMC), et s'implémente directement de manière simple dans un logiciel libre. Nous appliquons cette approche à l'étude d'une population de campagnols des champs (*Microtus Pennsylvanicus*), pour laquelle on dispose de données individuelles de masse corporelle, une variable qui peut influencer la probabilité de détection. Dans un deuxième exemple, nous appliquons ce modèle à un comptage aérien de gibier d'eau, selon un protocole de double-observateur. L'unité fondamentale d'observation est le groupe d'oiseaux, et la covariable expliquant la probabilité de détection est la taille du groupe (covariable quantitative discrète).

R. B. O'Hara, S. Lampila, and M. Orell

275

Estimation of Rates of Births, Deaths, and Immigration from Mark-Recapture Data

L'analyse de données de marquage-recapture est dans une période de développement et d'expansion. Nous y contribuons ici en présentant un modèle incluant les naissances et l'immigration aussi bien que les morts. Les données proviennent d'une étude au long-

terme sur la mésange boréale (*Parus montanus*), dans laquelle nous pouvons supposer que toutes les naissances sont enregistrées, et de la sorte que les immigrants peuvent aussi être identifiés comme les oiseaux capturés adultes pour la première fois. Nous modélisons le taux d'immigration, le taux de naissance par parent, et le taux de décès des jeunes et des adultes. L'utilisation d'un modèle hiérarchique nous permet d'incorporer une variation annuelle à ces paramètres. Le modèle est ajusté aux données en utilisant les MCMC comme analyse bayésienne. Au delà de l'ajustement du modèle, nous examinons également plusieurs aspects de la qualité d'ajustement, en particulier la question de la variation de la survie selon l'âge ou l'état d'immigrant, et celle de l'influence de l'historique de capture sur la probabilité de capture. Ce dernier point est important car l'indépendance des historiques de capture est une hypothèse clé qui simplifie considérablement le modèle. Nous trouvons que la probabilité de capture dépend fortement d'une éventuelle capture l'année antérieure.

L. J. Welty, R. D. Peng, S. L. Zeger, and F. Dominici

282

Bayesian Distributed Lag Models: Estimating Effects of Particulate Matter Air Pollution on Daily Mortality

Un modèle de retard distribué (DLagM) est un modèle de régression qui inclut comme covariables, des variables retardées d'exposition; sa fonction de retard distribué (DL) correspondante décrit la relation entre le décalage et le coefficient de la variable d'exposition décalée. Les DLagMs ont été récemment utilisés en épidémiologie environnementale pour quantifier les effets cumulatifs du temps et de la pollution de l'air sur la mortalité et la morbidité. Les méthodes standards de formulation des DLagMs comportent des DLagMs sans contrainte, avec contraintes polynomiales ou splines pénalisées. Ces méthodes risquent de ne pas bénéficier de l'information *a priori* sur la forme de la fonction DL des expositions environnementales, ou pour toute autre exposition avec des effets que l'on croit approcher lentement zéro quand le décalage augmente, et sont donc à risque de conduire à des estimations sous-optimales.

Dans ce papier nous proposons un DLagM bayésien (BDLagM) qui incorpore des connaissances *a priori* sur la forme de la fonction DL et permet aussi d'estimer à partir des données, le degré de lissage de la fonction DL. Nous appliquons notre BDLagM aux données, qui l'ont motivé, issues de l'Etude Nationale de la Morbidité, Mortalité et de la Pollution de l'Air (NMMAPS) pour estimer les effets à court terme sur la santé de particule de pollution atmosphérique sur la mortalité de 1987-2000 à Chicago, Illinois. Dans une étude de simulation, nous comparons notre approche bayésienne avec les méthodes alternatives qui utilisent des DLagMs sans contrainte, avec contraintes polynomiales ou splines pénalisées. Nous illustrons aussi la relation entre BDLagMs et DLagMs avec splines pénalisées. Un logiciel pour estimer des modèles BDLagM et les données utilisées dans ce papier sont disponibles en ligne.

H. Miao, C. Dykes, L. M. Demeter, and H. Wu

292

Differential Equation Modeling of HIV Viral Fitness Experiments: Model Identification, Model Selection, and Multi-Model Inference

De nombreux processus et systèmes biologiques peuvent être décrits par un ensemble de

modèles à équations différentielles. Cependant la littérature pour les modèles à équations différentielles en statistique inférentielle est peu fournie. Nous proposons des méthodes d'estimation statistique, de sélection de modèle, et de moyenne multi-modèles pour des expériences *in vitro* de résistance virale HIV, pouvant être décrites par un ensemble d'équations différentielles ordinaires (ODE). Nous explorons aussi l'identifiabilité des paramètres dans les modèles à ODE. Nous appliquons les méthodes et techniques proposées à des données expérimentales de résistance virale pour la mutation 103N du virus HIV-1. Nous espérons que la modélisation proposée et les approches inférentielles pour les modèles à équations différentielles puissent être largement utilisés dans les études biomédicales.

M.-W. An, C. E. Frangakis, B. S. Musick, and C. T. Yiannoutsos **301**
The Need for Double-Sampling Designs in Survival Studies: An Application to Monitor PEPFAR

En 2007 il y avait 33,2 million de gens dans le monde vivants avec un traitement contre le VIH (UNAIDS/WHO, 2007). En mai 2003, le président des États-Unis annonçait un programme global connu sous le nom de President's Emergency Plan for AIDS Relief (PEPFAR) pour lutter contre cette épidémie. Nous cherchons à estimer la mortalité des malades suivis dans PEPFAR pour surveiller et évaluer ce programme. Cet effort est cependant gêné par les abandons en cours qui interviennent à un taux élevé. Cela a pour conséquence que les analyses standards de données de survie sur les données des non-abandons sont généralement biaisées et ne mettent pas en évidence un moyen objectif de corriger ce biais. Dans cet article, nous utilisons des plans d'échantillonnage double et leur méthodologie aux données de PEPFAR, et nous obtenons des estimations plus plausibles et considérablement différentes de celles obtenues avec les méthodes standards (mortalité à 1 an estimée à 9,6% comparée à 1,7%). Les résultats indiquent qu'un plan d'échantillonnage double est crucial pour mettre en évidence de façon objective les abandons possibles que l'on ne peut négliger, et pour obtenir alors des données fiables dans PEPFAR. Nous montrons de plus la nécessité de méthodes d'analyses appropriées couplées aux plans d'échantillonnage double.

A. Ivanova **307**
Dose Finding for Continuous and Ordinal Outcomes with a Monotone Objective Function: A Unified Approach

Dans beaucoup d'essais de phase I, l'objectif est de trouver la dose correspondante à un taux de toxicité donné. Dans certains essais, l'objectif est de trouver la dose associée à une somme pondérée de taux de divers grades de toxicités. Pour d'autres, l'objectif est de trouver la dose liée à une valeur moyenne donnée d'une réponse continue. Dans cet article, nous présentons un schéma de recherche de dose qui peut être utilisé dans n'importe lequel des essais de recherche de dose décrits ci-dessus, essais où la dose ciblée est définie comme une dose à laquelle une fonction monotone donnée de la dose atteint une valeur établie a priori. A chaque étape du schéma proposé, la différence normalisée entre la dose en cours et la dose ciblée est calculée. Si cette différence est proche de zéro, la dose est répétée. Sinon, la dose est augmentée ou diminuée, selon le

signe de cette différence.

M. A. Proschan and M. Nason

316

Conditioning in 2x2 Tables

Les données résultant d'une recherche biomédicale se présentent parfois sous la forme d'un tableau 2x2, du cadre très général d'un essai clinique avec réponse binaire au cadre très spécifique des méthodes de triage en cytométrie de flux, où les réponses immunes spécifiques à un antigène sont distinguées des réponses immunes générales. Ce type d'application présente des défis intéressants quant à ce sur quoi l'analyse devrait être conditionnée (nombre total d'événements, nombre d'événements dans le groupe contrôle, etc.). Nous donnons ici plusieurs exemples biostatistiques illustrant la complexité de l'analyse de données pourtant simples à première vue.