Translations of Abstracts

Biometric Methodology

La connaissance des taux d'incidence du VIH et d'autres maladies infectieuses est importante pour évaluer l'état d'une épidémie et pour concevoir des études interventionnelles. L'estimation de l'incidence d'une maladie à partir d'études longitudinales peut être coûteuse et longue. Comme alternative, Janssen et al., 1998 ont proposé une estimation de l'incidence du VIH à un temps donné en combinant un dosage standard et un dosage "déréglé" des anticorps. Ce papier replace le problème dans une perspective longitudinale à partir de laquelle l'estimateur du maximum de vraisemblance de l'incidence est déterminé et comparé à celui de Janssen. Cette formulation permet l'estimation dans des cas généraux, y compris des batteries de tests non identiques pour tous les sujets, l'inclusion de covariables et une évaluation comparative de différentes batteries de test en vue d'une future étude. Ces méthodes sont illustrées avec les données d'un essai interventionnel et d'une étude de séro-prévalence conduites au Botswana.

Des modèles à classes latentes ont été récemment développés pour l'analyse conjointe de variables quantitatives longitudinales et de temps d'événements. Ces modèles supposent que la population est divisée en G classes latentes caractérisées par différentes fonctions de risque pour l'événement et différents profils d'évolution pour le marqueur décrits par un modèle mixte dans chaque classe. Cependant, l'hypothèse clé d'indépendance conditionnelle entre le marqueur et l'événement sachant la classe est difficile à évaluer car les classes latentes ne sont pas observées. En utilisant un modèle conjoint à classes latentes et effets aléatoires partagés, nous proposons un test du score pour l'hypothèse nulle d'indépendance entre le marqueur et l'événement sachant la classe versus l'hypothèse alternative correspondant à un risque d'événement dépendant d'un ou plusieurs effets aléatoires du modèle mixte en complément des classes latentes. Une étude par simulations a été réalisée pour comparer le comportement du test du score à d'autres tests précédemment proposés, en incluant des situations où l'hypothèse alternative ou la fonction de risque de base sont mal spécifiées. Dans toutes les situations explorées, le test du score est le plus puissant. La méthode est appliquée au développement d'un modèle pronostique pour le risque de rechute du cancer de la prostate en fonction de l'évolution de l'antigène spécifique de la prostate dans une cohorte de patients traités par radiothérapie.

La littérature statistique traitant de la modélisation conjointe de données longitudinales et de temps jusqu'à événement s'est souvent concentrée sur le développement de modèles visant à rendre compte des spécificités des données étudiées. Mais on n'a guère porté d'attention, la plupart du temps, à l'établissement d'outils diagnostiques et de méthodes d'évaluation des modèles. Bien sûr, dans les modèles conjoints, la censure des données longitudinales par la survenue des événements n'est pas aléatoire, et c'est là le principal obstacle à l'utilisation des outils diagnostiques classiques. En particulier, en présence de données manquantes, la distribution de référence de certaines statistiques – par exemple, les résidus – n'est plus disponible directement, et il faut recourir à des calculs complexes pour la reconstituer. Cet article

présente une approche basée sur l'imputation multiple, permettant de simuler de multiples versions d'un jeu de données complet, toutes construites sous l'hypothèse du modèle conjoint mis en œuvre. On peut alors calculer des résidus et produire des représentations diagnostiques pour le modèle des données complètes. Nous illustrons ces propositions sur deux exemples réels.

S. Yang and R. Prentice **30**
Improved Logrank-Type Tests for Survival Data Using Adaptive Weights

Afin de tester l'effet d'un traitement avec un critère censuré, le test du logrank est très souvent utilisé et il possède des propriétés d'optimalité sous l'hypothèse des hasards proportionnels. Ce test peut également être associé à d'autres tests sous des hypothèses particulières de non proportionnalité. Nous définissons des tests universels qui utilisent des statistiques de logrank pondérés adaptatifs. Ces poids adaptatifs utilisent le risque relatif estimé à partir des modèles de Yang et Prentice. De nombreuses simulations ont été faites sous l'hypothèse des hasards proportionnels ou sous différentes alternatives dans un large éventail de risque relatifs. Ces simulations montrent que ces nouveaux tests améliorent les tests habituels en situation de non proportionnalité. En particulier, le logrank pondéré adaptatif reste optimal sous l'hypothèse de proportionnalité et augmente la puissance sous un large choix d'hypothèses de non proportionnalité. Ces tests sont utilisés sur de nombreux jeux de données pour illustrer leurs performances.

C.-Y. Huang, J. Qin, and M.-C. Wang **39**
Semiparametric Analysis for Recurrent Event Data with Time-Dependent Covariates and Informative Censoring

L'analyse d'évènements récurrents est conduite d'habitude avec la supposition que le temps de censure est indépendant du processus de récurrence des évènements. Dans beaucoup d'applications le temps de censure peut être informatif sur le processus de récurrence sous-jacent, spécialement dans les situations où un événement de défaillance corrélé pourrait potentiellement terminer l'observation des évènements récurrents. Dans ce papier, nous considérons un modèle semi-paramétrique d'évènements récurrents qui permet les corrélations entre les temps de censure et le processus d'évènements récurrents via une variable de fragilité. Ce dispositif flexible incorpore dans la formulation des covariables à la fois dépendantes et indépendantes du temps, en laissant les distributions de fragilité et de temps de censure non spécifiées. Nous proposons une nouvelle procédure d'inférence semi-paramétrique qui ne dépend ni de la distribution de la fragilité ni de celle des temps de censure. Les propriétés, lorsque la taille de l'échantillon est grande, de l'estimation des paramètres de régression et de la fonction d'intensité cumulée de base sont étudiées. Des études numériques démontrent que la méthodologie proposée marche bien pour des tailles d'échantillon réalistes. Une analyse de données d'hospitalisation pour des patients dans une étude de cohorte du sida est présentée pour illustrer la méthode proposée.

Y. Zheng, T. Cai, J. L. Stanford, and Z. Feng **50**
Semiparametric Models of Time-Dependent Predictive Values of Prognostic Biomarkers

L'évaluation statistique rigoureuse des valeurs prédictives de nouveaux biomarqueurs est un a priori critique avant d'appliquer ces nouveaux biomarqueurs dans des procédures de soins standards. Il est important d'identifier les facteurs qui influencent la performance du biomarqueur pour déterminer les conditions optimales d'exécution du test. Nous proposons une courbe des valeurs prédictives positives (PPV) avec covariables spécifiques et dépendant du temps pour quantifier la précision prédictive du marqueur pronostic, mesuré sur une échelle continue et avec des temps de défaillance censurés. L'effet de la covariable est représenté avec un modèle de régression semi-paramétrique. En particulier nous adoptons une technique de régression des temps de survie lissés (Dabrowska, 1997) pour prendre en compte la situation où le risque de l'occurence et de la progression de la maladie change avec le temps. De plus, nous fournissons une théorie asymptotique et des procédures basées sur des rééchantillonnages pour faire de l'inférence statistique sur les valeurs prédictives positives spécifiques de la covariable. Nous illustrons notre approche avec des études numériques et des données d'une étude sur le cancer de la prostate.

P. T. Reiss and R. T. Ogden **61**
Functional Generalized Linear Models with Images as Predictors

La régression sur composantes principales fonctionnelles (FPCR) est une nouvelle méthode prometteuse pour évaluer des observations scalaires à partir de prédicteurs fonctionnels. Dans cet article nous présentons une justification théorique de son usage. La FCPR est ensuite étendue dans deux directions, d'abord du modèle linéaire au modèle linéaire généralisé et ensuite depuis un prédicteur constitué par un signal unique jusqu'à des prédicteurs constituant une image à haute résolution. Nous montrons comment développer efficacement la méthode en adaptant la technologie des modèles additifs généralisés au contexte de la régression fonctionnelle. Une technique est proposée pour estimer simultanément les bandes de confiance des fonctions coefficients; dans le contexte de l'imagerie en neurologie cela donne de nouveaux moyens d'identification des régions associées à des événements cliniques. Nous décrivons une nouvelle application de tests du rapport de vraisemblance pour évaluer l'hypothèse nulle d'une fonction coefficient constante. Les performances de la méthodologie sont illustrées par des simulations et par l'analyse de données réelles avec des images de tomographie par émission de positrons comme prédicteurs.

Y. Li, X. Lin, and P. Müller **70**
Bayesian Inference in Semiparametric Mixed Models for Longitudinal Data

Nous nous intéressons à l'inférence bayésienne dans des modèles mixtes semi-paramétriques pour données longitudinales (« SPMMs »).
Les SPMMs sont une classe de modèles qui utilisent une fonction non-paramétrique pour modéliser l'effet du temps, une fonction paramétrique pour modéliser les effets des autres covariables, et des effets aléatoires paramétriques ou non-paramétriques pour prendre en compte la corrélation intra-sujet. Nous modélisons la fonction non-paramétrique en nous servant de la formulation bayésienne d'un ajustement par spline cubique, et la distribution de l'effet aléatoire à l'aide soit d'une distribution normale soit d'un processus a priori non-paramétrique de Dirichlet (« DP »). Lorsque la distribution de l'effet aléatoire est supposée être normale, nous proposons un prior de réduction uniforme (« USP ») pour la composante de variance et le paramètre de lissage. Quand l'effet aléatoire est modélisé de façon non-paramétrique nous utilisons un prior DP avec une distribution de base gaussienne et nous proposons un USP pour les hyper-paramètres de la distribution de base normale du DP. Nous affirmons que le prior DP habituellement utilisé implique une moyenne non nulle de la distribution des effets aléatoires, même lorsqu'une distribution de base normale de moyenne non nulle a été spécifiée. Ceci entraîne une identifiabilité faible des effets fixes et peut conduire à des estimateurs biaisés et à une faible inférence des coefficients de régression et de l'estimateur spline de la fonction non-paramétrique. Nous proposons un ajustement utilisant une technique post-traitement. Nous montrons que sous des conditions à faibles contraintes la loi a posteriori est propre conditionnellement à l'USP proposé, à un prior large pour les paramètres des effets fixes et à un prior impropre pour la variance résiduelle. Nous illustrons l'approche proposée à l'aide d'un jeu de données hormonales longitudinal et nous fournissons des résultats d'études de simulation pour comparer ses performances en échantillons finis à des méthodes existantes.

X. Ni, D. Zhang, and H. H. Zhang **79**
Variable Selection for Semiparametric Mixed Models in Longitudinal Studies

Nous proposons une approche de vraisemblance doublement pénalisée pour modéliser simultanément la sélection et l'estimation des modèles mixtes semi paramétriques dans un cadre longitudinal. Deux types de pénalités sont imposées à la log-vraisemblance classique : la pénalité la plus lourde sur la fonction de base non paramétrique et une pénalité de rétrécissement non concave sur les coefficients linéaires pour réaliser un modèle parcimonieux. Comparer aux approches basées sur les estimation d'équations, notre procédure fournit des inférences valides pour des données présentant des données manquantes aléatoirement, et serait plus efficace si le modèle spécifié est correcte. Un autre avantage de cette nouvelle procédure est sa facilité de calcul pour les paramètres de régressions et de variance. Nous montrons que le problème de pénalisation double peut être aisément reformulé dans le contexte des modèles linéaires mixtes, ainsi les logiciels existants peuvent être directement utilisés pour implémenter notre approche. Concernant l'inférence du modèle, nous dérivons dans le cadre fréquentiste et bayésien les estimations des variances des estimations

des composantes paramétriques et non paramétriques. Des simulation sont utilisées pour évaluer et comparer les performances de notre méthode par rapports aux autres. Nous appliquons la nouvelle méthode à un jeu de données réelle sur une étude de lactation.

K. Yu, W. Wheeler, Q. Li, A. W. Bergen, N. Caporaso, N. Chatterjee, and J. Chen **89**
A Partially Linear Tree-based Regression Model for Multivariate Outcomes

Dans les études génétiques des traits complexes, en particulier ceux liés au comportement, comme le tabagisme et l'alcoolisme, plusieurs mesures phénotypiques sont généralement obtenues pour le trait, mais aucune mesure unique ne peut représenter à elle seule les caractéristiques complexes du trait du fait de notre manque de compréhension de l'étiologie de la maladie. Si ces phénotypes partagent un mécanisme génétique commun, au lieu de les étudier séparément, il est plus avantageux de les analyser conjointement comme un trait multivarié de façon à améliorer la puissance de détection des gènes associés. Nous proposons un test d'association multipoint pour l'étude de traits multivariés. Le test s'appuie sur un modèle de régression incluant à la fois une partie linéaire (covariables) et une partie basée sur un arbre de classification (combinaison de marqueurs génétiques). Ce nouveau modèle fournit un cadre statistique formel pour évaluer l'association entre un trait multivarié et un ensemble de prédicteurs, comme des marqueurs dans un gène ou dans une voie métabolique, tout en ajustant pour d'autres covariables. Par simulations, nous montrons que la méthode proposée a une erreur de type 1 acceptable et une meilleure puissance qu'en utilisant un trait univarié (chaque trait est étudié séparément, puis ajustement pour comparaisons multiples). Nous illustrons notre approche avec une étude d'association sur un gène candidat pour des phénotypes liés au tabagisme. La méthode proposée est suffisamment générale pour être utilisée dans le cadre d'autres études biomédicales où l'on chercherait à évaluer l'effet conjoint d'un ensemble de facteurs de risque sur un trait multivarié.

R. I. Garcia, J. G. Ibrahim, and H. Zhu **97**
Variable Selection in the Cox Regression Model with Covariates Missing at Random

Nous considérons le problème de la sélection des variables dans un modèle des risques proportionnels (Cox, 1975) quand certaines covariables manquent au hasard. Nous envisageons la pénalité de la déviation absolue à découpage régulier, la pénalité LASSO adaptative puis nous proposons une procédure unifiée de sélection et d'estimation. Nous présentons un algorithme de calcul efficace qui optimise simultanément la fonction de vraisemblance pénalisée et les paramètres de pénalité. Nous optimisons également un critère de sélection du modèle, la statistique IC_Q (Ibrahim, Zhu et Tang, 2008), pour estimer les paramètres de pénalité et nous montrons qu'il sélectionne régulièrement toutes les variables importantes. Nous évaluons les propriétés à distance finie des estimateurs pénalisés à l'aide de simulations. En guise d'illustration, nous analysons deux jeux de données concernant le cancer du poumon.

Y. Yuan and G. Yin **105**
Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data

Nous étudions la technique des régressions de quantile (*quantile regression*) sur des mesures longitudinales avec données manquantes non ignorables, qu'elles soient intermittentes ou liées à des sorties d'essai. Contrairement à la régression classique, qui ne modélise que la moyenne, ces régressions sont capables de caractériser entièrement la distribution conditionnelle, et s'avèrent plus robustes à la présence de données atypiques ou à une mauvaise spécification de la distribution de l'erreur. Nous prenons en compte la corrélation intra-sujet en introduisant, dans la classique fonction de perte (*check function*) de la régression de quantile, une pénalité de type $\ell 2$. Cela permet, selon le principe de l'estimateur à rétrécissement, de rapprocher les valeurs des pentes et constantes individuelles des valeurs des pentes et constantes communes à l'ensemble de la population. On suppose que les données manquantes informatives sont liées au processus de la réponse longitudinale par le biais d'effets aléatoires latents communs à toutes les données, observées ou non. Le comportement de la méthode proposée est évalué à l'aide de simulations, puis illustré sur les données d'un essai clinique dans le SIDA chez l'enfant.

Dans cet article, nous cherchons à estimer les prédictions et les paramètres d'une régression semi-paramétrique, en présence de valeurs manquantes non aléatoires pour la variable d'intérêt. Nous nous plaçons dans le cas où que la valeur de la variable d'intérêt et son éventuelle absence sont indépendantes conditionnellement à une information auxiliaire de grande dimension. Une approche paramétrique peut alors conduire à une mauvaise modélisation de la relation entre les covariables et la réponse, alors qu'une modélisation non paramétrique est impossible à cause du fléau de la dimension. Pour dépasser cette situation, nous étudions une modélisation pour réduire la dimension de l'information auxiliaire afin d'estimer ensuite nos paramètres de façon non paramétriques sur l'espace des covariables de dimension réduite. Nos estimateurs sont alors robustes à double titre : pour qu'ils ne soient pas convergents, il faut que les deux modèles liant les covariables auxiliaires à la variable d'intérêt d'une part et à la non-réponse d'autre part soient tous les deux faux. Nous proposons de nombreuses simulations pour comparer les performances numériques de nos estimateurs à celles d'autres estimateurs disponibles dans la littérature sur les valeurs manquantes, dont l'analyse par score de propension et l'équation d'estimation pondérée par probabilités inverses. Nous illustrons notre approche à l'aide d'un jeu de données réelles.

La détection initiale des pneumonies acquises sous ventilation mécanique (PAVM) par des patients d'unités de soins intensifs (USI) nécessite d'évaluer des symptômes complexes à l'aide de critères cliniques tels que le score clinique d'infection pulmonaire (CPIS). Lorsque ce CPIS dépasse une valeur seuil, le diagnostique est confirmé par un lavage broncho-alvéolaire (LBA) qui permet de compter les bactéries pathogènes présentes. Le CPIS et le LBA sont donc deux indicateurs importants de la pneumonie fortement dépendants, et les échantillons sont incomplets pour le LBA. Pour comparer mes risques de pneumonies de plusieurs groupes de traitement à l'aide de ces échantillons incomplets, nous proposons une méthode qui combine des outils non paramétriques basés sur le rapport de vraisemblance empirique et des tests classiques pour les modèles paramétriques. Nous donnons des résultats théoriques sur les propriétés asymptotiques de la méthode proposée. Ces résultats sont confirmés par des simulations de Monte-Carlo, qui font également apparaître de bonnes propriétés de notre méthode en termes de puissance. La méthode est finalement appliquée aux données réelles issues des pratiques cliniques pour comparer les risques de PAVM des différents groupes de traitements.

Cet article s'intéresse au problème associé à l'évaluation d'un effet causal modéré dans les études longitudinales où le traitement (ou l'exposition) et ainsi les cofacteurs varient au cours du temps ce qui peut conduire à une modération dans son effet. Les *effets causaux intermédiaires* qui décrivent les effets causaux d'un traitement dépendant du temps conditionnellement au passé des covariables sont introduits et considérés comme une partie des modèles emboîtés structuraux moyens de Robin. Deux estimateurs des effets causaux intermédiaires, et leur écart-type, sont présentés et discutés. Le premier est un estimateur d'une régression en 2 étapes, le second est le G-estimateur de Robin. Les résultats d'une petite étude de simulation qui montre les performances des estimateurs sur des faibles échantillons par rapport à des grands, et sur le biais de variance entre les deux estimateurs sont présentés. La méthode est illustrée à partir d'une étude longitudinale sur la dépression.

Il est bien connu que les plans d'expérience optimaux sont fortement modèle-dépendants. Dans ce papier, nous appliquons l'approche du multiplicateur de Lagrange au choix du plan expérimental optimal, en

utilisant un modèle proposé récemment pour les effets résiduels. En général, les plans expérimentaux en croisé ne sont pas recommandés quand des effets résiduels sont présents et quand l'objectif principal est d'obtenir un estimateur non biaisé de l'effet traitement. Dans certains cas, des mesures de base sont supposées améliorer l'efficacité du plan expérimental. Ce papier examine l'impact des mesures de base sur des plans expérimentaux optimaux à partir de deux hypothèses différentes sur les effets résiduels pendant les mesures de base et en utilisant un plan expérimental en croisé non-conventionnel. Comme prévu, des mesures de base améliorent fortement l'efficacité du plan expérimental pour des plans à deux périodes, qui utilisent seulement les données de la première période pour obtenir des estimateurs non biaisés des effets traitement, tandis que l'amélioration reste modeste pour des plans à trois ou quatre périodes. En outre, nous trouvons peu de bénéfices supplémentaires à obtenir des mesures de base à chaque période de traitement par rapport à des mesures de base faites seulement en première période. Bien que notre étude de l'impact des mesures de base ne change pas les résultats sur des plans d'expérience optimaux rapportés dans la littérature, le problème de la dépendance au modèle est généralement reconnu. L'avantage de multiplier les périodes est évident, car nous avons trouvé qu'étendre des plans à deux périodes à trois ou quatre périodes réduit significativement la variabilité de l'estimation de l'effet direct du traitement.

On a recours à un échantillonnage biaisé par la taille lorsqu'une variable résultat est échantillonnée avec une probabilité de sélection proportionnelle à sa taille. Dans cet article, nous proposons un modèle de régression linéaire semi-paramétrique pour analyser les résultats biaisés par la taille. Dans le modèle que nous proposons, les paramètres de régression des covariables sont particulièrement importants, alors que la distribution des erreurs aléatoires est non spécifiée. Sous le modèle proposé, nous observons que les paramètres de régression sont invariants au regard de l'échantillonnage biaisé par la taille. En suivant cette propriété d'invariance, nous développons une procédure simple d'estimation pour l'inférence. Les méthodes que nous proposons sont évaluées par des études de simulation, et appliquées à deux analyses sur données réelles.

Nous décrivons une méthode stéréologique pour estimer la surface foliaire totale d'une plante in vivo et analysons le problème de l'estimation de la variance de l'estimateur correspondant. La méthode est basée sur un échantillonnage hiérarchique, avec trois niveaux d'échantillonnage aléatoire systématique uniforme : (I) sélection de plantes dans un couvert végétal en utilisant le disector optique lisse, (II) échantillonnage de feuilles de la plante sélectionnée en utilisant le disector optique, et (III) estimation de la surface des feuilles échantillonnées par comptage de points. Cette méthode est appliquée à l'estimation de la surface foliaire totale d'un couvert de chrysanthèmes (Chrysanthemum morifolium L.), ce qui nous permet d'évaluer le temps de calcul et la précision de l'estimateur. De plus, nous avons comparé la précision obtenue avec trois densités de comptage de points différentes.
Les résultats montrent que la précision de l'estimateur de la surface foliaire basée sur le comptage de points est élevée. En utilisant une intensité de 1.76 cm^2/point pour la grille de comptage, nous estimons des surfaces foliaires pour la plante et pour le couvert à des précisions similaires à ou meilleures que celles obtenues par analyse d'image ou avec un planimètre commercial. Pour une surface de couvert d'environ $1m^2$ (10 plantes), l'approche utilisant le disector optique avec une fraction d'échantillonnage de 1/11 suivie d'un comptage de points utilisant une grille de 4.3 cm^2/point a produit un coefficient d'erreur de moins de 7%. Le disector optique lisse peut être utilisé pour s'assurer que la contribution supplémentaire de la variabilité inter plante à la variance de l'estimateur est faible.

Les méthodes d'échantillonnage par transect avec double observation sont de plus en plus répandues, notamment pour l'estimation des abondances des mammifères marins à partir de bateaux et d'avions lorsque la détection des animaux sur une ligne est incertaine. Les données obtenues enrichissent les

données traditionnelles de distance par rapport au transect, de données de marquage-recapture à deux événements d'échantillonnage. Comme pour toutes les données de marquage recapture classiques, l'estimation de l'abondance pose problème en présence d'hétérogénéité. A la différence des méthodes de marquage-recapture, les méthodes d'échantillonnage sur transect utilisent pour l'inférence la distribution d'une covariable affectant la probabilité de détection – la distance à la ligne du transect. La connaissance de cette covariable peut être utilisée pour diagnostiquer une hétérogénéité non modélisée dans la partie marquage-recapture des données. En modélisant la covariance de la probabilité de détection avec la distance, nous démontrons que le problème d'estimation peut être formulé en terme de niveaux d'indépendance différents. D'un côté, une indépendance complète est supposée, comme dans l'estimateur de Petersen (qui n'utilise pas les données de distance); à l'opposé, l'indépendance n'est vérifiée que dans la limite où la probabilité de détection tend vers 1. Entre ces deux extrêmes existe une gamme de modèles, dont les plus couramment utilisés, qui présentent des niveaux intermédiaires d'indépendance. Nous démontrons que ce cadre conceptuel peut être utilisé pour améliorer l'analyse de données de transects avec double observation. Nous avons testé ces différentes méthodes par simulation et en analysant un jeu de données où les vraies abondances sont connues. Nous illustrons cette approche par l'analyse d'observations du petit rorqual en Mer du Nord et régions adjacentes.

En biologie des populations sauvages, l'utilisation de marqueurs individuels naturels par empreinte ADN ou autres caractéristiques biologiques se généralise. Cependant, les modèles classiques de capture-recapture n'autorisent pas d'erreur d'identification des animaux, ce qui peut poser des problèmes importants en particulier avec des marqueurs naturels. L'analyse statistique des mécanismes d'erreurs d'identification est très difficile par les méthodes statistiques classiques basées sur la vraisemblance, mais elle est facilitée par l'utilisation de méthodes bayésiennes. Nous présentons un cadre général pour l'analyse bayésienne de données catégorielles émanant d'une distribution multinomiale latente. Notre travail est motivé par un modèle spécifique, pour les analyses de capture-recapture en population fermée avec erreurs d'identification, avec des hypothèses fortes qui ne peuvent pas être généralisées ; cependant les méthodes développées ici peuvent très naturellement être adaptées à divers autres modèles de structure semblable. Supposons que les fréquences observées $\boldsymbol{f}$ sont une transformation linéaire connue $\boldsymbol{f = A'x}$ d'une variable multinomiale latente $\boldsymbol{x}$, associée à un vecteur de probabilités $\boldsymbol{\pi = \pi(\theta)}$. Sachant que les distributions conditionnelles complètes $\boldsymbol{[\theta|x]}$ peuvent être échantillonnées, l'implémentation de l'échantillonneur de Gibbs ne nécessite que de pouvoir échantillonner à partir de la distribution conditionnelle complète $\boldsymbol{[x \mid f,\theta]}$, ce qui est le cas si on connaît le noyau de A'. Nous illustrons cette approche sur deux jeux de données comportant des erreurs d'identification individuelle, l'un simulé, l'autre sur des salamandres identifiées par des marqueurs naturels.

La biodiversité est un sujet important de recherche en écologie. Une forme usuelle de données collectées pour investiguer les schémas de biodiversité est le nombre d'individus de chaque espèce dans des séries de lieux. Ces données contiennent de l'information sur le nombre d'individus (abondance), le nombre d'espèces (richesse) et la proportion relative de chaque espèce à l'intérieur d'un assemblage échantillonné (régularité). Si il y a assez de sites échantillonnés à l'intérieur d'un gradient environnemental alors les données doivent contenir de l'information sur comment ces trois attributs de biodiversité changent à travers les gradients. Nous montrons que la représentation de la distribution du rang d'abondance (RAD) des données fournit une méthode commode pour quantifier ces trois attributs constituant la biodiversité. Nous présentons un cadre statistique pour modéliser les RADs et permettons que leur distribution multivariée varie selon les gradients environnementaux. La méthode repose sur trois modèles : un modèle binomial négatif, un modèle binomial négatif tronqué, et un nouveau modèle basé sur une Dirichlet multinomiale modifiée qui tient compte du type particulier d'hétérogénéité observée sur les données RAD. La méthode est motivée par, et s'applique à, une étude marine à grande échelle au large de la côte d'Australie-Occidentale en Australie. Cela fournit une description riche de la biodiversité et de comment elle change avec les conditions environnementales.

G. Zheng, Z. Li, M. H. Gail, and J. L. Gastwirth **196**
Impact of Population Substructure on Trend Tests for Genetic Case–Control Association Studies

Les stratifications cryptiques de population dans les études d'association cas-témoins peuvent potentiellement impacter sur les performances des tests de tendances de Cochran-Armitage (CATTs). Trois scenarii sont envisagées ici : i) une hétérogénéité des fréquences génotypiques parmi les sous-populations cryptiques, i.e. non identifiées ; ii) une hétérogénéité des fréquences génotypiques et du risque de maladie parmi les sous-populations cryptiques et iii) des corrélations cryptiques au sein des populations cryptiques. Une approche unifiée est proposée pour dériver le biais et la distorsion de la variance sous les trois scenarii pour n'importe quel CATT dans une famille générale. A l'aide de ces formules analytiques, nous évaluons numériquement l'excès d'erreurs de type I des CATTs en présence de sous structure de population. Nos résultats montrent et expliquent pourquoi certaines corrections proposées pour ce type de biais et de distorsion de la variance ne sont pas optimales.

S. Basu, M. Stephens, J. S. Pankow, and E. A. Thompson **205**
A Likelihood-Based Trait-Model-Free Approach for Linkage Detection of Binary Trait

La cartographie génétique de caractères complexes fait souvent appel à des méthodes du type « allèles communs » (ou *ibd*), qui ont l'avantage de ne pas avoir à expliciter le déterminisme génétique sous-jacent. Le cadre du maximum de vraisemblance proposé par Kong et Cox (1997) permet de calculer des valeurs précises de *p* (la probabilité de l'hypothèse nulle) et du LOD pour tester les liaisons entre une région du génome et le caractère. La méthode spécifie un modèle de ségrégation d'allèles marqueurs dans une région du génome associée au caractère. Nous proposons ici un modèle de ce type modifié de manière à extraire le maximum d'information de pedigrees de grande taille couvrant plusieurs générations et contenant des données sur des individus apparentés. Mais notre modèle s'applique aussi à des pedigrees de taille réduite, et il présente des avantages par rapport aux modèles existants (Kong et Cox, 1997), incluant le fait qu'il intègre l'information sur des individus affectés et non affectés. Nous illustrons le modèle proposé sur des données réelles et simulées, et comparons ses résultats à ceux de l'approche existante (Kong et Cox, 1997). La méthode proposée est mise en œuvre dans le programme *lm.ibdtests* dans le cadre de MORGAN 2.8 (http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml).

K. N. Javaras, N. M. Laird, J. I. Hudson, and B. D. Ripley **214**
Estimating Disease Prevalence Using Relatives of Case and Control Probands

Nous montrons comment estimer une prévalence à partie d'une enquête familiale cas-témoins. Les enquêtes familiales cas-témoins permettent d'étudier l'agrégation familiale d'une maladie; les familles sont échantillonnées à partir de propositus cas ou témoins. Le statut et les covariables du propositus et de ses apparentés constituent les données de la famille. Nous présentons des estimateurs de la prévalence globale et de la prévalence spécifique de strates (par exemple de la prévalence par sexe). Ces estimateurs combinent les proportions d'atteints chez les apparentés de cas et chez les apparentés de témoins. Ces estimateurs sont approximativement sans biais si des hypothèses usuelles sont vérifiées. Nous montrons également comment construire des intervalles de confiance présentant de bonnes performances en termes de recouvrement, même pour des prévalences faibles. Nous présentons des simulations effectuées sous différents niveaux d'agrégation familiale. Dans tous les cas les estimations étaient proches de la valeur théorique et distribuées symétriquement autour de celle-ci, tandis que les pourcentages de recouvrement des intervalles de confiance étaient satisfaisants. Enfin, nous discutons les hypothèses sur lesquels reposent ces estimateurs et les situations où un estimateur alternatif, n'utilisant que les apparentés de témoins, serait préférable.

H. Huang, H. Zhou, F. Cheng, I. Hoeschele, and F. Zou **222**
Gaussian Process Based Bayesian Semiparametric Quantitative Trait Loci Interval Mapping

En cartographie génétique des caractères quantitatifs, il faut souvent inclure des covariables comme l'âge ou le poids pour augmenter la puissance des tests et éviter la production de résultats faussement positifs. Mais si la covariable du modèle n'est pas spécifiée correctement (par exemple si un effet quadratique est

spécifié par erreur comme étant linéaire) l'inclusion de la covariable peut réduire la puissance et la précision de l'identification du locus gouvernant le caractère quantitatif (ou QTL pour *quantitative trait locus*). De plus certaines covariables peuvent interagir entre elles d'une manière complexe. Nous proposons des modèles semi-paramétriques pour cartographier des QTL simples ou multiples. Les deux méthodes de cartographie offrent une fonction non spécifiée pour toute covariable étant ou soupçonnée d'être en relation avec le caractère étudié d'une manière non linéaire mais inconnue. Elles prennent en compte aussi les interactions entre différentes covariables. L'analyse est réalisée dans un contexte bayésien avec méthodes de Monte-Carlo par Chaînes de Markov (MCMC). Les avantages de notre méthode sont démontrés à l'aide de simulations à grande échelle et d'analyses de données réelles.

Dans le but de réunir les approches d'analyse des mélanges d'ADN, et de recherche dans les bases de données ADN, nous proposons une nouvelle approche pour évaluer l'évidence légale de mélanges d'ADN lorsqu'un suspect est identifié au moyen d'une recherche dans une base de données de profils ADN. Des formules générales sont développées pour le calcul du rapport des vraisemblances pour un mélange de deux individus dans différentes situations incluant les associations multiples et l'évidence imparfaite. L'influence des probabilités a priori sur le poids de l'évidence dans le scénario d'associations multiples est démontrée par un exemple numérique reposant sur des données provenant de Hong-Kong. Nous montrons que notre approche permet de présenter de manière très compréhensible l'évidence légale de mélanges d'ADN lorsqu'un suspect est identifié par une recherche dans une base de données.

En réponse à l'augmentation continue de la menace d'un terrorisme nucléaire ou par irradiations, des recherches actives sont réalisées pour développer de nouveaux traitements non toxiques ainsi que d'autres contre-mesures qui protègeraient ou limiteraient les effets indésirables des radiations. Bien que les études classiques pour d'identification de la dose limite 50 (LD_{50}) qui ont prévalu depuis des décades comme première étape des études précliniques aient été largement remplacées par des plans d'expériences requerrant moins d'animaux, elles restent toutefois nécessaires pour évaluer l'efficacité radio protecteur de nouveaux traitements (FDA, 2002). Pourtant, il n'existe pas de méthodes directement applicables pour déterminer le nombre de sujets nécessaire à la réalisation de ces études contrôlées d'efficacité. Ce papier présente un calcul de taille d'échantillon reposant sur la formule d'un test t de Student de comparaison d'effets. Ce travail répond à une demande de la FDA d'avoir des données d'efficacité robustes sur les effets des modulateurs de la réponse aux irradiations totales du corps lorsque les études cliniques ne sont pas faisables ou pas éthiques. Des simulations de Monte-Carlo prouvent les performances de la formule pour les tests t de Student, de Wald et du rapport de vraisemblance tant sous un modèle logistique que probit. Il faut noter que les résultats montrent de façon claire qu'il est justifié d'utiliser des tailles d'échantillon substantiellement plus faibles que celles communément utilisées pour ces études. Ce papier pourrait donc initier un dialogue entre les chercheurs qui étudient la survie d'animaux pour les études de radio-protection, les comités institutionnel de protection des animaux et les institutions réglementaires afin d'obtenir un consensus sur le nombre d'animaux nécessaire pour atteindre à une puissance statistique suffisante dans la démonstration de l'efficacité des traitements radio protecteurs.

Nous décrivons une approche bayésienne pour la calibration d'un modèle informatique stochastique de cinétique chimique. Comme dans de nombreuses applications dans le domaine des sciences biologiques, les données disponibles pour calibrer le modèle proviennent de différentes sources. De plus, ces données semblent fournir des informations discordantes sur les paramètres du modèle. Nous proposons un cadre de modélisation qui nous permet de synthétiser ces informations conflictuelles et d'inférer un consensus. En particulier, nous montrons comment les effets aléatoires peuvent être incorporés dans le modèle afin de prendre en compte l'hétérogénéité interindividuelle qui peut être la cause des contradictions observées.

Biometric Practice

Il est rare d'analyser des données génétiques sans avoir à tester l'équilibre de Hardy-Weinberg. Pour effectuer ce test, on a traditionnellement recours à des approches fréquentistes. Cependant, le caractère discret de l'espace d'échantillonnage interdit de fait de postuler une distribution uniforme des *p*-values sous l'hypothèse nulle, et l'énumération de tous les dénombrements possibles, conditionnels à celui de l'allèle secondaire, permet certes de calibrer les *p*-values mais en contrepartie d'un coût parfois élevé en temps de calcul. De plus, l'interprétation des *p*-values qui en résulteraient de même que le choix des seuils de significativité dépend de façon critique de la taille de l'échantillon, puisque l'hypothèse d'équilibre sera toujours rejetée avec de grands échantillons. Nous plaidons donc ici pour une approche bayésienne, à la fois basée sur les facteurs de Bayes et sur l'examen des distributions a posteriori. Nous décrivons des approches conjuguées simples, ainsi que des méthodes basées sur l'échantillonnage préférentiel (*important sampling*) de Monte Carlo. Les premières sont appréciables car elles aboutissent à des solutions analytiques pour l'expression du facteur de Bayes, facilitant leur application à un grand nombre de SNPs (polymorphismes pour un seul nucléotide), en particulier dans le contexte de la génomique. Nous décrivons également des méthodes directes d'échantillonnages pour examiner les distributions a posteriori des paramètres d'intérêt. Dans le cas où on s'intéresse à un grand nombre d'allèles à un locus, nous recourons à la méthode de Monte Carlo par chaîne de Markov. Nous discutons un certain nombre de possibilités en ce qui concerne la spécification de l'a priori, et appliquons les méthodes développées à plusieurs jeux réels de données.

Dans les études cas-témoins, pour détecter l'association entre un marqueur génétique et une maladie, on utilise classiquement le test de tendance de Cochran-Armitage. Ce test est localement optimal si le modèle génétique est correctement spécifié. Cependant, en pratique, le modèle génétique, et donc le test de tendance optimal, sont inconnus. Dans ce cas, les tests suivants sont utiles : le test du khi-2 de Pearson, le maximum des trois tests de tendance (optimal pour les modèles récessif, additif et dominant) et le test de tendance optimal pour le modèle génétique correspondant le mieux aux données. Dans cet article, nous avons d'abord proposé une modification de la méthode existant pour sélectionner le modèle génétique correspondant le mieux aux données dans le cas où l'allèle à risque est inconnu. Nous avons ensuite proposé une nouvelle approche de sélection qui exclut les modèles génétiques ne correspondant pas aux données. En utilisant, soit la sélection, soit l'exclusion de modèle, nous réduisons l'espace des modèles génétiques possibles conditionnellement aux données observées, ce qui permet d'augmenter la puissance de détecter une vraie association. Nous présentons des résultats de simulation basés sur les marqueurs génétiques identifiés par les études d'association du Wellcome Trust Case-Control Consortium. Nous montrons que l'approche utilisant l'exclusion de modèle génétique donne en général de meilleurs résultats que les méthodes existantes pour une diversité de modèles génétiques.

Pour la comparaison des formes des trains d'impulsions enregistrés dans deux situations expérimentales nous proposons une approche complètement inférentielle. La méthode est basée décrivant le déclenchement à l'aide de modèles de processus de Poisson non homogènes pour chaque situation associée à un modèle souple de mélange de lois a priori non paramétriques pour les intensités correspondantes des impulsions. Nous démontrons une inférence a posteriori par une analyse globale qui peut être utilisée aussi bien pour une comparaison couvrant la totalité de l'expérience que pour des comparaisons ponctuelles à des temps préétablis pour mettre en évidence des différences locales entre les deux processus de déclenchement. La méthode est appliquée à des enregistrements de deux neurones du cortex moteur

primaire d'un singe réalisant une succession de tâches de pointage.

L. Lin, D. Bandyopadhyay, S. R. Lipsitz, and D. Sinha **287**
Association Models for Clustered Data with Binary and Continuous Responses

Nous envisageons l'analyse de données bivariées groupées où chaque élément d'un groupe présente deux réponses, l'une binaire, l'autre continue. Nous proposons un nouveau modèle bivarié à effets aléatoires qui prend en compte l'association intra-groupe entre réponses binaires, l'association intra-groupe entre réponses continues et les associations entre réponses de types différents au niveau intra-groupe comme au niveau intra-sujet. Après intégration sur les effets aléatoires, le modèle marginal pour les réponses binaires conserve une forme logistique et le modèle marginal pour les réponses continues conserve une forme linéaire, ce qui facilite l'interprétation des coefficients. L'estimateur du maximum de vraisemblance des paramètres de notre modèle s'obtient à partir de logiciels standards comme la PROC NLMIXED de SAS. Des simulations établissent la robustesse de notre méthode vis-à-vis d'une mauvaise spécification de la composante fixe ou de la composante aléatoire du modèle. Un étude de la toxicité de l'éthylène glycol sur le développement de la souris illustre notre méthodologie.

B. Yu and P. Ghosh **294**
Joint Modeling for Cognitive Trajectory and Risk of Dementia in the Presence of Death

La démence est caractérisée par un déclin cognitif accéléré avant et après le diagnostic, par comparaison au vieillissement normal. Il est bien connu que la déficience cognitive survient longtemps avant le diagnostic de démence. Pour les individus développant une démence, il est important de déterminer la date où le taux de déclin cognitif a commencé à s'accélérer et le laps de temps qui a suivi jusqu'au diagnostic de démence. Pour les individus au vieillissement normal, il est également important de connaître l'évolution de la fonction cognitive jusqu'au décès. Un modèle bayésien de point de rupture est proposé pour l'ajustement de l'évolution de la fonction cognitive des individus qui développent une démence. Les individus âgés sont soumis à deux risques compétitifs, la démence et le décès hors-démence. Etant donné que la majorité des individus ne développent pas de démence, on propose un modèle mixte pour des données de survie avec risques compétitifs, reposant sur le temps d'apparition de la démence au-delà du point de rupture du déclin cognitif pour les sujets atteints de démence, et le temps de décès pour les individus sans démence. Les trajectoires de la fonction cognitive et les processus de survie sont modélisés conjointement, et les paramètres sont estimés par une méthode de Monte Carlo par Chaîne de Markov. A partir des données de l'Etude Honolulu Asie sur le Vieillissement nous montrons les trajectoires de la fonction cognitive et l'effet de l'éducation, du génotype de l'apolipoprotéine E4, et de l'hypertension sur le déclin cognitif et le risque de démence.

I. Ahmed, C. Dalmasso, F. Haramburu, F. Thiessard, P. Broët, and P. Tubert-Bitter **301**
False Discovery Rate Estimation for Frequentist Pharmacovigilance Signal Detection Methods

Les systèmes de pharmacovigilance ont pour objectif de détecter le plus précocement possible les effets indésirables des médicaments commercialisés. Ils maintiennent de grandes bases de notifications spontanées pour lesquelles ont été développées plusieurs méthodes de détection automatique de signaux. Une limite commune à l'ensemble de ces méthodes est que les règles de décision pour la génération de signaux sont fondées sur des seuils arbitraires. Dans cet article nous proposons une nouvelle procédure pour la génération de signaux. Le critère de décision est formulé sous la forme d'une région de rejet pour les degrés de signification issus de la méthode du rapport des chances rapporté ainsi que du test exact de Fisher. Pour ce dernier, nous étudions aussi l'utilisation des p-valeurs moyennes. La région de rejet est définie à partir du taux de faux positifs qui peut être estimé en adaptant les procédures basées sur des modèles de mélange pour la distribution des degrés de signification au cas des tests d'hypothèses unilatérales. La méthodologie proposée est principalement illustrée avec la procédure d'estimation position-dépendante. Elle est étudiée à travers une large étude par simulations et appliquée aux données françaises de pharmacovigilance.

Nous considérons une approche de l'analyse des données d'échantillonnage à distance entièrement basée sur un modèle. L'échantillonnage à distance a été largement utilisé pour estimer l'abondance (ou la densité) des animaux ou plantes dans un lieu d'étude spatialement explicite. Il n'existe, cependant, aucune méthode facilement disponible pour faire de l'inférence statistique sur les relations entre l'abondance et les covariables environnementales.

Les vraisemblances de processus de Poisson spatialisés peuvent être utilisés pour estimer simultanément la détection et les paramètres d'intensité en modélisant les données d'échantillonnage à distance comme un processus spatial ponctuel aminci. Une approche spatiale des données d'échantillonnage à distance basée sur un modèle possède trois avantages : elle permet l'utilisation de plans en transects complexes et opportunistes, elle permet l'estimation de l'abondance dans de petites sous régions, et elle fournit un cadre pour évaluer les effets de l'habitat ou la manipulation expérimentale sur la densité. Nous démontrons la méthodologie basée sur un modèle avec une petite étude de simulation et l'analyse du jeu de données des herbes de Dubbo. De plus, une simple méthode ad hoc pour prendre en compte la surdispersion est aussi proposée. L'étude de simulation montre que l'approche basée sur un modèle se compare favorablement par rapport aux méthodes conventionnelles d'échantillonnage à distance pour l'estimation d'abondance. De plus, la correction de la surdispersion a donné de bons résultats quand le nombre de transects était élevé. L'analyse du jeu de données de Dubbo a montré un effet transect sur l'abondance via une sélection de modèle par AIC. Une analyse plus approfondie de la qualité d'ajustement, cependant, a montré une potentielle confusion de l'intensité avec la fonction de détection.