Translations of Abstracts

Biometric Methodology

La technologie de séquençage de seconde génération peut opérer le séquençage de millions de petits fragments d'ADN en parallèle, et est capable d'assembler des génômes complexes pour seulement une fraction du prix et du temps des précédentes technologies. Dans les faits, un consortium récemment formé, le Projet 1000 Génômes, planifie de séquencer entièrement les génômes d'environ 1200 personnes. La perspective d'une analyse comparative à l'échelle de la séquence, d'un grand nombre d'échantillons au sein de multiples populations pourrait être atteinte dans les cinq prochaines années. Ces données présentent des défis inconnus jusque là dans l'analyse statistique. Par exemple, l'analyse doit procéder sur des millions de séquences nucléotidiques courtes, des lectures (reads) – séquences de A, C, G, ou T de longueur comprise entre 30 et 100 – qui sont le résultat d'un processus complexe de mesure d'intensité de fluorescence continue bruitée connue sous le nom de programme d'identification de base (base-programme).
La complexité du processus de discrétisation du programme d'identification de base provient de lectures de qualité largement variable entre échantillons de séquences comme au sein d'une même séquence. Cette variabilité dans la qualité du processus entraîne des erreurs peu fréquentes mais systématiques que nous avons observé créer des erreurs dans l'analyse d'aval des données de la séquence discrétisée. Ainsi, un objectif central du Projet 1000 Génômes est de quantifier la variation entre échantillons au niveau du simple nucléotide. A cette résolution, de petits taux d'erreur dans le séquençage s'avèrent significatifs, spécialement pour des variants rares. Le séquençage de seconde génération est une technologie relativement nouvelle pour laquelle les différents biais et sources cachées de variation ne sont pas encore bien connus. Par conséquent la modélisation et la quantification de l'incertitude inhérente à la génération des séquences pour les lectures est de la plus grande importance. Dans cet article nous présentons un modèle simple pour la prise en compte de l'incertitude provenant du programme d'identification de base Illumina/Solexa. Les paramètres du modèle ont une interprétation directe selon le programme d'identification de base, permettant des métriques facilement interprétables et informatives pour la variabilité de la qualité du séquençage. Notre modèle donne ces estimations informatives pour être directement utilisées avec les outils d'évaluation de la qualité, et améliorent significativement la performance du programme d'identification de base.

la transcription du gène. L'identification des sites de liaison des facteurs de transcription (TFBS) est une étape importante dans la compréhension de la régulation des gènes. Bien que sophistiquées pour la modélisation des TFBS et de leurs motifs combinatoires, les méthodes calculatoires de détection des TFBS et de recherche de motif procèdent souvent avec des hypothèses trop simplificatrices d'homogénéité de modèle. Etant donné que la composition en bases nucléotidiques varie entre les régions du génome, on s'attend à ce que la prise en compte de cette hétérogénéité dans la modélisation soit utile à l'identification des motifs. Lorsque des séquences de plusieurs espèces sont utilisées, l'hypothèse usuelle d'un même niveau de conservation dans les multiples alignements, est violée. Pour prendre en compte tous les types d'hétérogénéité, nous proposons un modèle dans lequel une chaîne de Markov segmentée est utilisée pour partitionner un alignement multiple en régions de compositions homogènes en bases nucléotidiques, et un modèle de chaîne de Markov cachée (HMM) est employé pour traiter différents niveaux de conservation. On développe une inférence bayésienne sur le modèle par échantillonneur de Gibbs avec récursivités par programmation dynamique. Des études de simulation et des constatations empiriques à partir d'ensemble de données biologiques montrent l'impact majeur de la modélisation du fond génétique sur l'identification de motifs, et démontrent que l'approche proposée est substantiellement plus performantes que les méthodes communément utilisées.

Les études d'association génétique s'intéressent souvent à l'effet d'haplotypes sur un critère d'intérêt. Les haplotypes ne sont pas observés directement, ce qui complique l'inclusion de tels effets dans les modèles de survie. Nous décrivons une nouvelle approche basée sur les équations d'estimation pour le modèle de régression de Cox afin de déterminer les effets d'haplotypes pour des données de survie. Ces équations d'estimation sont simples à mettre en œuvre et ne nécessitent pas l'utilisation de l'algorithme EM qui risque d'être lent dans le contexte du modèle de Cox semi-paramétrique avec information incomplète sur les covariables. Ces équations d'estimation conduisent aussi à des estimateurs d'écarts types directs, faciles à calculer et donc pallient à certaines des difficultés liées à l'utilisation de l'algorithme EM dans ce contexte pour obtenir des estimateurs de variances. Nous développons aussi une procédure, facile à mettre en œuvre, pour tester l'adéquation du modèle de Cox avec effets d'haplotypes. Enfin, nous appliquons les procédures présentées dans cet article pour étudier les effets possibles d'haplotypes du récepteur PAF sur les événements cardiovasculaires chez des patients atteints de pathologie coronaire, et comparons nos résultats à ceux basés sur l'algorithme EM.

Dans beaucoup de situations un sujet peut passer par à la fois un évènement terminal et un évènement non terminal où l'évènement terminal (par exemple la mort) censure l'évènement non terminal (par exemple la rechute), mais l'inverse n'est pas vrai.

Typiquement les deux évènements sont corrélés. Cette situation est appelée risques semi compétitifs (voir par exemple Fine, Jiang, et Chappell, 2001 ; Wang, 2003), et l'analyse est basée sur la fonction de survie jointe des deux évènements sur le quadrant positif mais avec des observations restreintes au triangle supérieur. Implicitement, cette approche entretient l'idée de variables de survie latentes et conduit à une discussion sur la distribution marginale de l'évènement non terminal qui n'est pas fondée en réalité. Nous soutenons que de façon similaire aux modèles à risque compétitifs, les temps de survie latents devraient être évités quand on modélise de telles données. Nous notons que les risques semi-compétitifs ont plus classiquement été décrits comme des modèles maladie-mort et cette approche évite une quelconque référence aux temps de survie latents. Nous considérons un modèle maladie mort avec fragilité partagée, qui dans sa forme la plus restrictive est identique au modèle à risques semi-compétitifs qui a été proposé et analysé mais qui permet beaucoup de généralisations et incorpore simplement des covariables. Une estimation de maximum de vraisemblance non paramétrique est utilisée pour l'inférence et l'estimation résultante pour le paramètre de corrélation est comparée avec les autres approches proposées. Les propriétés asymptotiques, les études de simulation et une application à un essai clinique randomisé dans le cancer nasopharyngé évaluent et illustrent les méthodes. Un algorithme simple et rapide est développé pour l'implémentation numérique..

Les essais cliniques recueillent souvent de manière périodique des informations sur la progression d'une maladie ainsi que les résultats d'analyses de laboratoire censés refléter les différentes étapes de la maladie. Un des objectifs premiers de ce type d'études est de déterminer la relation entre les analyses effectuées et la progression de la maladie. En l'absence de données manquantes ou de censure, ces analyses seraient simples. Cependant, les patients manquent souvent des visites et reviennent quand la maladie a progressé. Dans ce cas, non seulement on observe une censure sur les intervalles de temps de progression mais en outre les séries d'analyses sont incomplètes. Dans cet article, nous proposons un test simple de l'association entre un marqueur longitudinal et la durée d'un phénomène à partir de données incomplètes. Nous dérivons ce test selon une technique très intuitive qui consiste a calculer l'espérance du score des données complètes conditionnellement a l'observation des données incomplètes. Le problème est motivé par des données concernant le diabète.

On désigne par courbes de croissance des mesures répétées au cours du temps d'un processus continu de croissance sur une population d'individus. Ces données longitudinales sont classiquement analysées par des modèles non-linéaires à effets mixtes dont la fonction de régression impose une évolution monotone croissante du phénomène.

Ces modèles de croissance ne permettent pas de modéliser des modifications inattendues du taux de croissance. Nous proposons de prendre en compte ces éventuelles variations à l'aide d'équations différentielles stochastiques déduites du modèle de croissance standard par ajout d'une composante stochastique. Nous développons une méthode d'inférence Bayésienne de ces modèles reposant sur un algorithme de Gibbs. Dans le cas où la distribution conditionnelle du processus de diffusion n'est pas explicite, nous implémentons un schéma numérique d'Euler-Maruyama. Nous validons ce nouveau modèle en utilisant et en adaptant des critères basés sur la distribution prédictive a posteriori. Nous illustrons la pertinence de notre approche dans le cas d'un modèle de Gompertz sur des données de croissance de poulets.

Andrew C. Titman and Linda D. Sharples **742**
Semi-Markov Models with Phase-Type Sojourn Distributions

Les modèles multi-états à temps continu sont largement répandus pour des variables catégorielles, particulièrement pour la modélisation de maladies chroniques. Cependant l'inférence est difficile lorsque le processus n'est observé qu'à temps discret, sans information sur les types d'événements et leur occurrence entre les temps observés et sans qu'une hypothèse markovienne ne soit posée. Cette hypothèse peut être limitante lorsque les taux de transition entre les états de la maladie peuvent dépendre du temps passé depuis l'entrée dans l'état courant. Une telle formulation se résout par une modèle semi-markovien. Nous montrons que les problèmes de calcul associés à l'ajustement des modèles semi-markovien sur des données observées de panel peuvent être allégés en considérant une classe de modèles semi-markoviens avec des distributions de séjour phase-type. Cela permet d'appliquer des méthodes pour modèles de Markov cachés. D'autre part, des extensions sont données pour des modèles où les états observés sont sujets à des erreurs de classification. La méthodologie est démontrée sur un jeu de données relatif à la bronchiolite oblitérante après une transplantation du poumon chez des patients.

Florence Chaubert-Pereira, Yann Gu´edon, Christian Lavergne, and Catherine Trottier **753**
Markov and Semi-Markov Switching Linear Mixed Models Used to Identify Forest Tree Growth Components

La croissance des arbres est supposée être principalement le résultat de trois composantes : (i) une composante endogène supposée être structurée en une succession de phases approximativement stationnaires, séparées par des ruptures franches, asynchrones entre individus, (ii) une composante environnementale variant dans le temps supposée prendre la forme de fluctuations synchrones entre individus, (iii) une composante individuelle correspondant principalement à l'environnement local de chaque arbre. Dans le but d'identifier et de caractériser chacune de ces composantes, nous proposons d'utiliser des combinaisons semi-markoviennes de modèles linéaires mixtes, c'est-à-dire des mélanges finis de modèles linéaires mixtes avec dépendances semi-markoviennes. La semi-chaîne de Markov sous-jacente représente la succession de phases de croissance et leurs longueurs (composante endogène) alors que les modèles

linéaires mixtes attachés à chaque état de la semi-chaîne de Markov sous-jacente représentent – dans la phase de croissance correspondante – à la fois l'influence de covariables climatiques variant dans le temps (composante environnementale) comme effets fixes et l'hétérogénéité inter-individuelle (composante individuelle) comme effets aléatoires. Dans cet article, nous abordons l'estimation, dans un cadre général, des combinaisons markoviennes et semi-markoviennes de modèles linéaires mixtes. Nous proposons un algorithme de type MCEM dont les itérations se décomposent en trois étapes : (i) simulation des séquences d'états sachant les effets aléatoires, (ii) prédiction des effets aléatoires sachant les séquences d'états, (iii) maximisation. La modélisation statistique proposée est illustrée par l'analyse de pousses annuelles successives de troncs de pins laricio influencées par des covariables climatiques.

Ingelin Steinsland and Henrik Jensen **763**
Utilizing Gaussian Markov Random Field Properties of Bayesian Animal Models

Dans cet article nous montrons les avantages calculatoires et les nouvelles possibilités qu'offrent pour le modèle animal bayésien les champs aléatoires gaussiens de Markov. Nos inférences sont basées sur le calcul des distributions a posteriori des variables importantes en génétique quantitative. Pour le modèle animal unicaractère une approximation sans échantillonnage est présentée. Pour le modèle multicaractère nous établissons un modèle robuste et rapide de Chaîne de Markov par Monte Carlo (MCMC). La méthode proposée a été utilisée pour analyser les paramètres génétiques de caractères morphologiques d'une population de moineaux. Les résultats obtenus avec les modèles unicaractère et multicaractère ont été comparés.

Brian J. Reich, Montserrat Fuentes, Amy H. Herring, and Kelly R. Evenson **772**
Bayesian Variable Selection for Multivariate Spatially Varying Coefficient Regression

L'activité physique présente beaucoup de bénéfices bien documentés pour la santé cardiovasculaire et pour la maîtrise du poids. Pour les femmes enceintes, le Collège Américain des Obstétriciens et Gynécologues recommande couramment 30 minutes d'exercices modérés pour la plupart sinon tous les jours ; cependant peu de femmes enceintes atteignent ce niveau d'activité. Traditionnellement, les études se sont centrées sur l'examen des facteurs individuels et interpersonnels pour identifier des prédicteurs de l'activité physique. Il y a un renouveau d'intérêt pour savoir comment les caractéristiques de l'environnement physique dans lequel nous vivons et travaillons peuvent aussi influencer les niveaux d'activité physique. Nous considérons une des premières études sur les femmes enceintes qui examine l'impact des caractéristiques de l'environnement construit sur les niveaux d'activité physique. En utilisant un cadre sociologique, nous étudions les associations entre l'activité physique et différents autres facteurs incluant des caractéristiques personnelles, des variables de qualité météorologique/air, et des caractéristiques du voisinage pour des femmes enceintes dans quatre comtés de Caroline du Nord. Nous analysons simultanément six types d'activité physique et recherchons des dépendances croisées entre ces types d'activité. Une analyse exploratoire suggère que les associations sont différentes dans les différentes régions. Nous utilisons alors un modèle de régression multivariée avec des coefficients de régression variant dans l'espace. Ce

modèle comprend un paramètre de régression pour chaque covariable en chaque point de l'espace. Pour nos données comprenant beaucoup de prédicteurs, on a clairement besoin d'une forme de réduction de la dimension. Nous introduisons une procédure de sélection de variables bayesienne pour identifier des sous-ensembles de variables importantes. Notre algorithme stochastique de recherche détermine pour chaque effet d'une covariable d'être nul, non nul mais constant dans l'espace, et variant spatialement. Nous trouvons que les covariables de niveau individuel ont une plus grande influence sur les niveaux d'activité des femmes que les caractéristiques de voisinage environnemental, et certaines covariables de niveau individuel ont des associations variant spatialement avec les niveaux d'activité des femmes enceintes.

La détection de clusters spatiaux est une méthodologie importante permettant de détecter les clusters spatiaux dans des résultats sans faire d'hypothèses fortes sur la structure de dépendance spatiale. Pour des données sante concernant par exemple l'indice de masse corporelle ou l'obésité, la dépendance spatiale peut être difficile à modéliser puisque sa magnitude peut dépendre de grandeurs difficiles à quantifier. Cet article propose une approche robuste aux hypothèses de distributions pour des données ponctuelles ou agrégées et des variables aléatoires générales dont les deux premiers moments sont spécifiés, pouvant être des variables continues, binaires ou de comptage. Cette nouvelle méthodologie incorpore aisément des structures de pondération, telle que la population régionale, pour permettre de pondérer de manière inégale l'information des différentes régions dans le cas de données agrégées. La méthode proposée intègre également la possibilité d'ajouter des covariables d'ajustement au niveau individuel ou régional, ce qu'aucune méthode ne permettait jusqu'à maintenant pour des variables pondérées. La performance de la méthode est évaluée par simulation, puis la méthode est appliquée pour évaluer la classification spatiale d'indices de masse corporelle élevés dans une population de HMO (Health Maintenance Organization) dans les régions de Seattle et Washington aux Etats-Unis.

Quand on modélise des tableaux de grandes dimensions, la sélection des variables constitue un problème clé. Les méthodes de sélection disponibles pour une classification guidée par un modèle considèrent chaque variable indépendamment des autres, c'est-à-dire qu'une variable est retenue si elle permet de séparer au moins deux classes et rejetée si elle n'en sépare aucune. Mais dans de nombreuses applications, on aimerait aussi savoir précisément quelles sont les classes séparées par les variables informatives. Nous proposons une sélection de variables par couples de classes pour résoudre ce problème. L'analyse de données simulées et réelles montre que cette nouvelle approche fournit des résultats meilleurs et mieux interprétables que des méthodes alternatives fondées sur les pénalités l_1 et l_∞.

Nous proposons un modèle hiérarchique pour la probabilité de la toxicité limitant la dose (DLT) pour des combinaisons de doses de deux agents thérapeutiques. Nous appliquons ce modèle à un algorithme d'essai adaptatif bayésien dont le but est d'identifier des combinaisons avec taux de DLT proches d'un taux-cible préalablement spécifié. Nous décrivons des méthodes pour générer les distributions a priori des paramètres de notre modèle à partir d'un ensemble de base d'information explicité par les investigateurs cliniciens. Nous investiguons les performances de notre algorithme dans une série de simulations d'un essai hypothétique qui examine des combinaisons de quatre doses de deux agents. Nous comparons également la performance de notre approche à deux méthodes existantes et évaluons la sensibilité de notre approche au choix de la distribution a priori.

Ces dernières années, pour évaluer les stratégies thérapeutiques, les essais cliniques ont fréquemment recouru à des randomisations équilibrées sur de petits nombres de caractéristiques-patients. En général, de telles randomisations assurent un certain équilibre des facteurs pronostiques – connus et inconnus – entre les différents groupes de traitement ; en pratique, cependant, les investigateurs ne peuvent inoculer de la stratification qu'à petites doses, et l'on ne peut pas toujours obtenir simultanément un bon équilibre sur toutes les variables importantes, particulièrement quand ces variables sont nombreuses et que l'étude est de faible effectif. Ainsi en va-t-il de l'étude INSTINCT, conçue pour évaluer l'efficacité d'un programme éducatif destiné à améliorer l'utilisation du tPA chez les patients ayant subi un accident vasculaire cérébral. C'est pourquoi nous présentons ici une nouvelle méthode de randomisation qui consiste en un plan d'appariement équilibré avec pondération (« balance match weighted (BMW) design »), appliquant une technique d'appariement optimal avec contraintes à un essai prospectif randomisé : il s'agit, pour le plan BMW, de minimiser l'erreur quadratique moyenne de l'estimateur de l'effet traitement. De fait, une étude de simulation montre que, sous différents scénarios concernant les facteurs pronostiques, le plan BMW peut réduire substantiellement l'erreur quadratique moyenne de l'effet traitement, par comparaison à une randomisation simple ou à un plan apparié. Le plan BMW est également comparé, en termes d'efficacité et de robustesse de l'estimation de l'effet traitement, à une approche par modélisation ajustée sur un score de propension d'une part, et à la procédure d'E-estimation de Robins-Newey d'autre part. Il semble que le plan BMW soit plus robuste et souvent – mais pas toujours – plus efficace que chacune de ces deux approches. Le plan BMW s'avère également plus robuste en cas d'hétéroscédasticité. Pour illustration, nous appliquons notre méthode à l'étude INSTINCT.

Estimating Treatment Effects of Longitudinal Designs using Regression Models on Propensity Scores

Nous dérivons les estimateurs de régression qui peuvent comparer des traitements longitudinaux en utilisant uniquement les scores de propension longitudinaux comme régresseurs. Ces estimateurs, qui supposent des connaissances sur les variables utilisées dans l'attribution du traitement, sont importants pour réduire la grande dimension des covariables pour deux raisons. Premièrement, si les modèles de régression sur les scores de propension longitudinaux sont correctes, alors nos estimateurs partagent les avantages d'estimateurs basés sur les modèles et correctement spécifiés, un bénéfice non partagé par les estimateurs basés uniquement sur les poids. Deuxièmement, si les modèles ne sont pas correctes, la mauvaise spécification peut être plus facilement limitée par la vérification du modèle qu'avec les modèles complets. Ainsi, nos estimateurs peuvent être meilleurs quand ils sont utilisés à la place d'une régression sur l'ensemble des covariables. Nous utilisons nos méthodes pour comparer des traitements longitudinaux pour le diabète sucré de type 2.

Roula Tsonaka, Dimitris Rizopoulos, Geert Verbeke, and Emmanuel Lesaffre 834
Nonignorable Models for Intermittently Missing Categorical Longitudinal Responses

Une classe de modèles non ignorables est présentée pour prendre en compte des mécanismes de données manquantes non monotones pour des réponses catégorielles longitudinales. Cette classe de modèles inclut les traditionnels modèles de sélection et les modèles à paramètre partagé. Cela nous permet de réaliser une analyse de sensibilité plus large qu'usuellement. En particulier, au lieu de considérer les variations induites par un modèle non ignorable donné, nous étudions leur sensibilité aux différents mécanismes de données manquantes. Une caractéristique intéressante de la classe développée est que les paramètres sont obtenus avec une interprétation marginale, tandis que des modèles algébriquement simples sont considérés. Spécifiquement, des modèles à effets mixtes marginalisés (Heagerty, 1999) sont utilisés pour le processus longitudinal qui modélise séparément la moyenne marginale et la structure de corrélation. Pour la structure de corrélation, des effets aléatoires sont introduits et leur distribution est modélisée soit paramétriquement ou non paramétriquement pour éviter des mauvaises spécifications.

Mulugeta Gebregziabher and Bryan Langholz 845
A Semiparametric Missing-Data-Induced Intensity Method for Missing Covariate Data in Individually Matched Case Control Studies

Dans les études cas-témoins appariées individuelles dans lesquelles certaines variables explicatives sont manquantes, l'analyse basée sur les données complètes peut induire une grande perte d'information à la fois sur les variables manquantes et les variables complètement observées. Ce problème engendre généralement un biais et une perte d'efficacité. Pour traiter le problème de données manquantes sur les variables explicatives, nous proposons dans cet article une nouvelle méthode basée sur une approche d'intensité induite par les données manquantes dans le cas où le mécanisme de données manquantes ne dépend pas du statut cas-témoin, et nous montrons que cette

approche conduit à une généralisation de la méthode de l'indicateur manquant. Nous dérivons les propriétés asymptotiques des estimateurs issus de la méthode proposée, et évaluons les performances sur échantillon fini en terme de biais, d'efficacité et de couverture à 95% dans une étude de simulations approfondie utilisant plusieurs scénarios de données manquantes. Cette méthode est aussi comparée avec l'analyse sur données complètes et des méthodes déjà proposées pour traiter les données manquantes. Nos résultats montrent que, sous l'hypothèse de données manquantes prédictibles, la méthode proposée fournit une estimation valide des paramètres, est plus efficace que l'analyse sur données complètes, et reste compétitive par rapport à d'autres méthodes d'analyse plus complexes. Une étude cas-témoins sur le risque de myélome multiple et le polymorphisme du récepteur inter-leukine-6 (IL-6-□) est utilisée pour illustrer nos résultats.

Nous discutons l'identifiabilité de modèles pour des tests diagnostiques dichotomiques multiples en l'absence de méthode de référence (étalon-or). Les données se présentent sous la forme de comptages multinomiaux ou de produits de multinomiales, selon le nombre de populations échantillonnées. Les modèles sont généralement paramétrés en termes de prévalences, de sensibilités, de spécificités et de mesures de dépendance. On croit souvent que le modèle est identifiable dès que le nombre de degrés de liberté des données est supérieur ou égal au nombre de paramètres. Goodman (1974) a établi depuis longtemps que ce n'est pas une condition suffisante. Nous discutons les modèles actuellement disponibles et nous défendons une extension d'un modèle proposé par Dendukuri et Joseph (2001). Nous développons ensuite l'approche de Goodman et nous utilisons une présentation géométrique pour mieux appréhender la nature des modèles non identifiables. Nous illustrons ces méthodes à l'aide de données simulées et réelles.

L'estimation de l'incidence de l'infection par le VIH basée sur un échantillon transversal de sujets testés à la fois par un test de dépistage sensible et un test désensibilisé offre de nombreux avantages par rapport à l'estimation de l'incidence basée sur une étude de cohorte. Toutefois, la fiabilité de l'approche basée sur un échantillon transversal est remise en cause pour deux raisons principales. La première est la difficulté d'obtenir une approximation externe fiable de la moyenne de la période-fenêtre, définie comme le délai entre la détectabilité de l'infection par le test sensible et par le test désensibilisé, utilisée dans la procédure d'estimation de l'incidence du VIH par cette approche. La deuxième raison est la façon de traiter les faux négatifs avec le test de diagnostic désensibilisé ; c'est à dire les sujets qui sont testés négatifs, donc classés dans l'état infection-récente, alors qu'ils sont infectés depuis longtemps. Nous proposons et évaluons un schéma étendu d'estimation de l'incidence du VIH à partir d'enquêtes transversales dans lequel

Dans les essais randomisés, des mesures sont souvent recueillies sur chaque sujet initialement et après la randomisation à plusieurs reprises au cours de l'essai. L'analyse longitudinale de covariance dans laquelle les valeurs suivant la valeur initiale constituent le vecteur de réponse et la valeur initiale est traitée comme une covariable peut être utilisée pour évaluer les différences liées au traitement après la randomisation. Liang et Zeger (2000) proposent une analyse longitudinale contrainte des données dans laquelle la valeur initiale est incorporée dans le vecteur de réponse avec les mesures postérieures et une contrainte de moyenne initiale identique pour tous les bras est imposée en raison de la randomisation. Si la valeur initiale a des valeurs manquantes, l'analyse longitudinale contrainte s'avère plus efficace pour estimer les différences liées au traitement suivant la randomisation que l'analyse longitudinale de covariance. Le gain d'efficacité augmente avec le nombre de sujets avec valeur initiale manquante et avec le nombre de sujets avec valeurs ultérieures toutes manquantes et, pour le pre-post design, diminue avec la corrélation absolue entre valeurs initiale et ultérieures.

Josep L. Carrasco **897**
A Generalized Concordance Correlation Coefficient Based on the Variance Components Generalized Linear Mixed Models for Overdispersed Count Data

Le coefficient de corrélation de concordance classique (CCC) qui mesure l'accord pour un ensemble d'observateurs, suppose pour les données une distribution Normale et une relation linéaire entre la moyenne et les sujets et les effets observés. Ici, nous généralisons le CCC à n'importe quelle distribution de la famille exponentielle au moyen de la théorie des modèles mixtes linéaires généralisés et l'appliquons au cas de données de comptage surdispersées. Nous donnons un exemple de comptage de cellules CD34+ pour montrer l'applicabilité de la procédure. Dans le dernier cas nous définissons différents CCC et les appliquons aux données en changeant le modèle linéaire généralisé (GLMM) qui s'ajuste aux données. Une étude de simulation est menée pour explorer le comportement de la procédure avec des tailles d'échantillon petites et modérées.

Bo Cai, David B. Dunson, and Joseph B. Stanford **905**
Dynamic Model for Multivariate Markers of Fecundability

Les modèles à classes latentes dynamiques fournissent un cadre souple pour étudier des processus biologiques qui évoluent avec le temps. Motivés par l'étude des marqueurs de jours de fertilité pendant le cycle menstruel, nous proposons une approche à classes latentes dynamiques en temps discret qui permet aux changements de pente de dépendre du temps, de prédicteurs fixes et d'effets aléatoires. Les données observées sont des indicateurs multivariés ordinaux qui changent de façon dynamique et souple selon l'état de la classe latente. Etant donnée la flexibilité de cette approche qui inclut des composantes semi-paramétriques au travers de mélanges de Betas, des contraintes d'identifiabilité sont nécessaires dans la définition des classes latentes. Ces contraintes sont définies de manière appropriée à partir des connaissances biologiques du processus. La méthode bayésienne est spécifiquement développée pour analyser des données sur le mucus cervical dans une étude de femmes utilisant la planification familiale naturelle.

L'un des objectifs essentiels de la biologie évolutive est de comprendre la dynamique de la sélection naturelle au sein des populations. La force et la direction de cette sélection peuvent être décrites en régressant des mesures d'aptitude relatives sur des phénotypes écologiquement pertinents. Cependant, beaucoup de caractéristiques évolutives importantes des organismes sont complexes, et ont *de facto* des relations complexes avec les mesures d'aptitude. Les caractéristiques sexuelles secondaires comme les modes d'accouplement sont d'excellents exemples de traits complexes avec des conséquences importantes sur le succès de la reproduction. Classiquement, les chercheurs éclatent les traits sexuels comme les signaux d'accouplements en un ensemble de mesures incluant l'intensité et la durée afin de les inclure dans une analyse statistique. Cependant, ces mesures dérivées par les expérimentateurs ne capture vraisemblablement pas toute la variation phénotypique pertinente, en particulier lorsque les sources de sélection sont incomplètement connues. Afin de prendre en compte cette complexité, nous proposons un modèle linéaire généralisé bayésien de spectrogramme à dimension réduite, qui incorpore directement les représentations du phénotype dans son entier (signal acoustique à une dimension) dans le modèle comme une variable indicatrice tout en prenant en compte les multiples sources d'incertitude. La première étape de la réduction de dimension est obtenue en traitant le signal comme une 'image' et en identifiant ses fonctions orthogonales empiriques correspondantes. Ensuite, la seconde réduction de dimension est réalisée par une sélection de modèle utilisant une procédure stochastique de sélection de variables. De fait, le modèle que nous proposons caractérise les aspects clés du signal acoustique qui influence la sélection sexuelle tout en réduisant le besoin d'extraire a priori des traits correspondant à des signaux de plus haut niveau. Cet aspect de notre approche est fondamental et à le potentiel de fournir des informations biologiques supplémentaires comme illustré dans notre analyse.

Plusieurs méthodes statistiques pour détecter des associations entre des traits quantitatifs et des gènes candidats dans des populations structurées ont été développées pour des phénotypes complètement observés. Cependant, beaucoup d'études s'intéressent à évaluer le délai jusqu'à l'échec selon les phénotypes. Ce type de données est habituellement soumis à la censure. Dans cet article, nous proposons des méthodes statistiques pour détecter des associations entre un trait quantitatif censuré et des gènes candidats dans des populations structurées avec des niveaux multiples complexes de génétique chez les individus échantillonnés. Les méthodes proposées corrigent pour une stratification de population continue en utilisant à la fois des variables structures de la population comme des covariables et des termes de fragilités attribuables à la parenté. La

relation entre le délai jusqu'à l'échec et des scores génotypiques pour un marqueur candidat est modélisée via un modèle de survie accéléré avec une fragilité paramétrique de type Weibull (AFT) ainsi qu'un modèle AFT avec une fragilité semi-paramétrique, où la fonction de survie à l'inclusion est flexible pour être modéliser par un mélange d'arbres de Polya centré autour d'une famille de distributions de Weibull. Pour les modèles paramétrique et semi-paramétrique, les fragilités sont modélisées via une distribution a priori de type autorégressive conditionnelle Gaussienne intrinsèque, la matrice de parenté étant la matrice adjacente associée aux sujets. Des études de simulation et des applications à des jeux de données sur la ligne de temps de floraison de l' « *Arabidopsis thaliana* » ont démontré l'avantage de la nouvelle approche proposée par rapport aux approches existantes.

De plus en plus fréquemment, les études épidémiologiques s'intéressent à des hypothèses sur les interactions gène-environnement. Pour beaucoup de gènes largement étudiés et en présence d'un régime alimentaire et d'un comportement standard, des informations préalables importantes peuvent souvent être utilisées pour analyser les données présentes ou pour concevoir une nouvelle étude. Dans ce papier, nous proposons tout d'abord une approche complétement bayésienne pour l'analyse des études d'interactions gènes-environnement. L'approche bayésienne implique une incorporation naturelle de l'incertitude autour de l'hypothèse d'indépendance entre gènes et environnement souvent utilisée dans une telle analyse. Nous considérons ensuite un critère de détermination de la taille de l'échantillon à la fois pour l'estimation et le test d'hypothèse sur le paramètre d'interaction multiplicative gène-environnement. Nous illustrons les méthodes proposées à partir de données d'une grande étude cas-témoin en cours sur le cancer colorectal qui traite de l'interaction du N-acétyl transférase de type 2 (NAT2) avec le tabac et la consommation de viande rouge. Nous utilisons les données existantes pour extraire un prior et montrons comment cette information peut être employée pour répartir les cas et témoins lors de la plannification d'une future étude sur les mêmes paramètres d'interaction. Cette stratégie bayésienne pour le design et l'analyse est comparée à ses homologues fréquentistes.

L'imagerie par résonance magnétique fonctionnelle fournit des jeux de données de grande dimension caractérisés par des structures de dépendance complexes dues à des phénomènes neurophysiologiques sophistiqués et à certains aspects du plan d'expérience. Afin d'étudier les relations entre stimulation et modification de l'activité cérébrale, les méthodes classiques proposent une procédure en deux étapes, spécifiant tout d'abord un modèle au niveau individuel puis des paramètres de population (ou groupe). En règle générale, la corrélation temporelle entre les scans successifs acquis au cours d'une session est prise en compte dans la première étape. Cependant, les expériences IRMf

sont souvent constituées de plusieurs sessions, entraînant des dépendances temporelles entre les estimations successives de l'activité neuronale moyenne. En outre, les modèles et méthodes statistiques classiques négligent les corrélations spatiales des mesures de l'activité cérébrale.
Nous proposons un modèle spatio-temporel auto-régressif à deux niveaux prenant en compte simultanément les dépendances spatiales entre les voxels appartenant à une même région anatomique et les dépendances temporelles entre les estimations individuelles issues de plusieurs sessions.
Nous développons un algorithme exploitant la structure particulière de notre modèle de covariance pour fournir une méthode d'estimation rapide et efficace. La méthodologie proposée est appliquée sur des données IRMf issues d'une étude sur le contrôle inhibiteur de sujets dépendants à la cocaïne.

Martha Nason and Dean Follmann **958**
Design and Analysis of Crossover Trials for Absorbing Binary Endpoints

Le plan croisé « crossover » est un dispositif efficient et répandu dans le contexte de patients hétérogènes pour évaluer l'effet de traitements agissant relativement vite, et dont le bénéfice disparaît à l'interruption. Chaque patient peut servir comme son propre témoin puisque ce sont les réponses intra-individus au traitement et au placebo qui sont comparées. La prudence usuelle incite à considérer que ces dispositifs ne sont pas adaptés pour des issues binaires « absorbantes » telles que le décès ou l'infection par le virus HIV. Nous étudions l'utilisation des plans croisés dans le contexte de ces issues binaires absorbantes, et nous montrons qu'ils peuvent être plus efficients qu'un dispositif standard en groupes parallèles lorsqu'il y a une hétérogénéité dans les risques individuels. Nous introduisons également un nouveau dispositif en deux périodes dans lequel les « survivants » de la première période sont randomisés pour la seconde période.Ce dispositif combine un plan croisés avec une étude en groupes parallèles et présente certains avantages en terme d'efficience du plan croisé tout en garantissant que les groupes de la seconde période sont comparables en raison de la randomisation. Nous discutons la validité de ces nouveaux dispositifs et nous évaluons à la fois un modèle de mélange et un test de Mantel-Haenszel pour l'inférence. Le modèle de mélange suppose l'absence de report ou d'effets périodes, tandis que l'approche de Mantel-Haenszel conditionne les effets périodes. Nous utilisons des simulations pour comparer les différents dispositifs expérimentaux et nous donnons un exemple pour étudier les problèmes pratiques de l'implémentation.

L.G. Leon-Novelo, X. Zhou, B. Nebiyou Bekele, and P. Müller **966**
Assessing Toxicities in a Clinical Trial: Bayesian Inference for Ordinal Data Nested within Categories

Cet article traite de la modélisation et de l'inférence pour des données ordinales emboîtées dans des réponses catégorielles. Nous proposons un mélange de distributions normales pour des variables latentes associées aux données ordinales. Ce modèle de mélange nous permet de déterminer sans perte de généralité les paramètres de coupure qui relient la variable latente à la réponse ordinale observée. De plus nous montrons que

le modèle de mélange est plus souple pour estimer les probabilités des cellules, en comparaison du modèle de régression ordinale bayésien traditionnel avec des paramètres de coupure aléatoires. Nous étendons notre modèle pour prendre en compte de possibles dépendances entre les résultats de différentes catégories. Nous appliquons le modèle à une étude randomisée de phase III pour comparer les traitements sur la base des toxicités enregistrées par type de toxicité et grade par type. Les données sont constituées des différents types (catégories) de toxicités exprimées chez chaque patient. Chaque type de toxicité a un grade (ordinal) associé. La dépendance entre les différents types de toxicité exprimés par le même patient est modélisée par l'introduction d'effets aléatoires spécifiques du patient.

Souvent, les données d'essais cliniques nous arrivent sous la forme de tables de petites dimensions, où figure le modeste décompte d'événements peu pléthoriques. Du coup, des approximations classiques comme le test du score ou le test du maximum de vraisemblance s'avèrent imparfaites à bien des égards. Premièrement, à partir des mêmes données, elles peuvent aboutir à des résultats tout à fait différents. Deuxièmement, l'erreur réelle de première espèce peut différer significativement de l'erreur nominale que l'on souhaite contrôler, même lorsque les effectifs de l'essai commencent à être relativement importants. Troisièmement, les inférences exactes auxquelles on recourt alors sont parfois des fonctions fortement non monotones du paramètre d'intérêt, d'où des intervalles de confiance non contigus. Pour faire de l'inférence sur petits effectifs, nous disposons pourtant de deux approches novatrices. D'une part, on peut utiliser les méthodes asymptotiques dites d'ordres élevés, proposées par Reid (2003, *Annals of Statistics* **31**, 1695-1731), qui ajustent explicitement la statistique du rapport de vraisemblance – la théorie est compliquée, mais le calcul de la statistique est rapide. D'autre part, on peut effectuer des calculs exacts de significativité, en supposant que les paramètres de nuisance sont égaux à leur valeur sous l'hypothèse nulle : c'est l'approche de Lee et Young (2005, *Statistics and Probability Letters* **71**, 143-153), sorte de bootstrap paramétrique. Le but de cet article est d'expliquer et d'évaluer ces deux méthodes, dans le cas où l'on teste si une différence de proportions $p_2 - p_1$ excède une borne préfixée δ_0 de non-infériorité. En nous basant sur des calculs numériques extensifs, nous recommandons les *p*-values bootstrap de la deuxième méthode, lesquelles s'avèrent supérieures à toute autre alternative. Non seulement elles proposent des résultats quasiment identiques quelle que soit la statistique de test choisie, mais elles ont aussi une meilleure puissance, tout en contrôlant des risques réels de première espèce très proches des valeurs nominales. Qui plus est, leurs variations en fonction du paramètre δ_0 s'avèrent beaucoup moins erratiques que ce qu'on observe avec d'autres méthodes.

Reader Reaction

Ye et al. ont proposé en 2008 un modèle pour des mesures longitudinales et des données temps-évènement pour lesquelles les mesures longitudinales sont modélisées avec un modèle mixte semi paramétrique qui autorise des modèles complexes avec les bio markers longitudinaux. Ils ont proposé une approche régression calibration à deux étages, plus aisée à mettre en œuvre qu'une approche modélisation. A la première étape de leur approche le model mixte est ajusté sans tenir compte des données temps-évènement. Lors de la seconde étape l'espérance à postériori d'un effet aléatoire individuel du modèle mixte est pris en compte en tant que covariable dans un model de Cox. Bien que Ye et al aient reconnu que leur approche régression calibration pouvait produire un biais du au problème de l'erreur de mesure, ils répondent que ce biais est petit comparé à celui des méthodes alternatives. Dans cet article, nous montrons que ce biais peut se révéler conséquent. Nous montrons comment le réduire considérablement par une approche alternative de régression calibration qui peut être appliquée sur des données temps-évènement discrètes ou continues. Grâce à des simulations, il est montré que l'approche proposée conduit à un biais beaucoup moins conséquent qu'avec l'approche proposée par Ye et al. (2008). En accord avec la méthodologie proposé par Ye et al., un avantage de notre proposition de modélisation est qu'elle peut être mise en œuvre avec des logiciels statistiques standard et qu'elle n'exige pas de techniques d'estimation sophistiquée.