Translations of Abstracts

BIOMETRIC METHODOLOGY

Le mélange introductif de population peut être un facteur de confusion dans les études d'association génétique . Les méthodes basées sur les familles (Rabinowitz & Laird 2000) ont été proposées à la fois en estimation et en test pour ajuster sur ce facteur, particulièrement dans les études d'association cas-simple. Les méthodes basées sur les familles reposent sur le conditionnement par les génotypes parentaux observés ou sur la statistique suffisante minimale pour le modèle génétique sous l'hypothèse nulle. Dans certains cas ces méthodes n'appréhendent pas toute l'information disponible en raison d'une stratégie de conditionnement trop restrictive. Des méthodes générales efficientes pour ajuster sur un mélange introductif utilisant toute l'information disponible ont été proposées (Rabinowitz 2002). Cependant ces approches peuvent ne pas être toujours aisées à implémenter. Une approche précédente, à calculs simples, a été développée précédemment pour ajuster sur le mélange introductif par addition de covariables supplémentaires dans des modèles linéaires (Yang et al. 2000). Nous montrons ici que cette stratégie de modèle linéaire élargi de covariables appropriées peut être combinée avec les méthodes générales efficientes présentées par Rabinowitz (2000) pour obtenir un ajustement localement efficace et de calculs accessibles. Après obtention des covariables optimales, l'analyse ajustée peut être développée en utilisant des packages statistiques standard tels que SAS ou R. Les méthodes proposées ont une efficacité locale au voisinage du modèle exact. **Les études de simulation montrent que des améliorations non triviales en efficacité peuvent être obtenues en utilisant l'information non accessible aux méthodes qui reposent sur le conditionnement par les statistiques suffisantes minimales**. Les approches sont illustrées par une analyse de l'influence du génotype de l'apolipoprotéine E (APOE) sur la concentration plasmatique des lipoprotéines de basse densité (LDL) chez les enfants.

Les données à haute dimension telles que les données d'expression issues de puces à ADN ou d'un séquençage de nouvelle génération consistent en un grand nombre de variables dépendantes mesurées sur un petit échantillon. Un problème crucial en génomique est l'identification et l'estimation de facteurs associés à de nombreuses variables simultanément. L'identification du nombre de facteurs est également importante pour les analyses statistiques non-supervisées telles que la classification hiérarchique. Le modèle factoriel conditionnel est le plus fréquemment utilisé pour beaucoup de types de données génomiques allant de l'expression de gènes aux polymorphismes de nucléotides uniques en passant par la méthylation. Nous montrons ici que sous un modèle factoriel conditionnel pour données génomiques avec une taille d'échantillon fixée, les vecteurs singuliers à droite sont asymptotiquement consistants pour les facteurs latents non-observables lorsque le nombre de variables diverge. Nous proposons aussi un estimateur consistant de la dimension du modèle factoriel conditionnel sous-jacent pour une taille d'échantillon finie et un nombre infini de variables basé sur une décomposition en valeurs propres normée. Nous proposons une approche pratique pour la sélection du nombre de facteurs dans des jeux de données réels et nous illustrons l'utilité de ces résultats pour capturer les effets batch ou d'autres effets non-modélisés dans une expérience de puces à ADN à l'aide de l'approche noyaux de Leek et Storey (2008).

La plupart des méthodes existantes d'identification des régions aberrantes à l'aide de données CGH sont

limitées à une seule cible. En nous intéressant particulièrement à la comparaison d'échantillons multiples de deux groupes différents, nous développons une nouvelle approche de régression pénalisée avec une pénalisation adaptative lasso pour s'adapter à la dépendance spatiale des clones. Les segments génomiques non aléatoires aberrants sont déterminés en évaluant la différence entre clones voisins et segments voisins. L'algorithme proposé dans cet article est une première tentative de détection simultanée des régions aberrantes communes dans chaque groupe, et des régions où les deux groupes diffèrent sur les nombres de changement de copie. L'étude de simulation suggère que la procédure proposée surpasse les méthodes de segmentation usuellement utilisées pour la détection d'aberration sur simple échantillon, aussi bien en termes de faux positifs qu'en termes de faux négatifs. Pour évaluer plus avant la valeur de la méthode présentée, nous analysons un ensemble de données d'une étude d'identification des régions génomiques aberrantes associées aux sous-groupes définis par les grades des tumeurs malignes du sein.

D. R. Bickel **363**
Estimating the Null Distribution to Adjust Observed Confidence Levels for Genome-Scale Screening

Dans une nouvelle approche au problème des tests multiples, Efron (2004; 2007a; 2007b) a formulé des estimateurs de la distribution des statistiques de test ou des p-values nominales sous une hypothèse nulle appropriée à la modélisation de données de haute dimension issues des études d'association pangénomiques ou d'autres études biologiques. Les estimateurs de cette distribution nulle peuvent améliorer non seulement la procédure empirique de Bayes pour laquelle ils ont été originellement développés mais de nombreuses autres procédures de prise en compte des comparaisons multiples.
De tels estimateurs peuvent dans certains cas améliorer la procédure de comparaisons multiples (MCP) récemment proposée qui repose sur un cadre non bayésien de minimisation de la perte attendue par rapport à une confiance a posteriori, i.e. une distribution de probabilité des niveaux de confiance. La souplesse de cette méthode MCP est illustrée avec une fonction de perte non additive développée pour le criblage génomique plutôt que pour les études de validation.
L'intérêt d'estimer la distribution nulle est étudiée du point de vue de la confiance a posteriori de la méthode MCP (CPMCP). Dans la situation générique d'un problème de tests multiples à l'échelle du génome, conditionner le niveau de confiance observé sur la distribution nulle estimée comme une statistique ancillaire approximé améliore de façon importante l'inférence conditionnelle. Cependant, dans un contexte de données d'expression spécifiquement simulées, on observe que l'estimation de la distribution nulle tend à augmenter le biais conservatif qui résulte de la modélisation de distributions présentant des queues épaisses avec la famille normale.
Afin de permettre aux chercheurs de déterminer s'ils peuvent utiliser une estimation particulière de la distribution nulle pour l'inférence ou la prise de décision, un score d'information théorique est proposé. Défini comme la somme du degré d'ancilarité et du degré de pertinence inférentielle, ce score reflète l'équilibre que le conditionnement génère entre ces deux termes opposés.
La méthode CPMCP ainsi que les autres méthodes proposées sont appliquées à des données d'expression.

M. Capanu and C. B. Begg **371**
Hierarchical Modeling for Estimating Relative Risks of Rare Genetic Variants: Properties of the Pseudo-Likelihood Method

De nombreux gènes ont été identifiés comme influençant fortement les risques de cancer. Cependant, en général, le gène peut subir différentes mutations engendrant ou non un risque accru. Il est crucial d'identifier les mutations dangereuses et les mutations sans risque pour que les individus chez lesquels on diagnostique une mutation puissent être conseillés de manière appropriée. C'est une tache difficile puisque de nouvelles mutations sont continuellement découvertes et qu'il y a typiquement peu d'information sur chaque mutation. Dans un article précédent, nous avons utilise un modèle hiérarchique (Capanu et al. 2008) faisant appel à des méthodes de pseudo vraisemblance et d'échantillonnage de Gibbs pour estimer les risques relatifs de gènes mutants à partir de données d'une étude cas-témoins et avons montré que ce type de modèle permet de distinguer les mutations contribuant au risque de cancer. Cependant, d'autres recherches sont nécessaires pour valider l'utilisation de méthodes asymptotiques sur ces données éparses, i.e. contenant un grand nombre de zéros. Dans cet article, nous étudions en détail les propriétés de la pseudo vraisemblance par des techniques de simulation. Nous explorons aussi deux approches alternatives : la pseudo-vraisemblance avec correction de l'estimateur de variance proposée par Lin and Breslow (1996) et une méthode de pseudo-vraisemblance hybride avec une estimation Bayésienne de la variance. La validité de ces modèles hiérarchique est étudiée en analysant le biais et les propriétés de couverture des

estimateurs ainsi que l'efficacité des estimateurs du modèle hiérarchique par rapport au maximum de vraisemblance. Les résultats indiquent que les estimateurs de risques relatifs des mutations les plus rares ont des biais petits et que les intervalles de confiances a 95% obtenus ne sont en général pas conservateurs bien que les probabilités de couverture soient le plus souvent au delà de 90%. La largeur des intervalles de confiance diminue avec la variance résiduelle du modèle de second niveau hiérarchique. Les résultats montrent aussi que les estimateurs du modèle hiérarchique ont des intervalles de confiance plus étroits que ceux obtenus par une régression logistique simple et que cette amélioration est plus nette quand la mutation est plus rare.

L'identification de groupes homogènes d'individus est un problème important en génétique des populations. Récemment, plusieurs méthodes ont été proposées qui exploitent l'information spatiale pour améliorer les algorithmes de classification. Dans cet article nous développons un algorithme bayésien de classification, reposant sur un processus *a priori* de Dirichlet et qui utilise à la fois l'information génétique et l'information spatiale pour classer les individus en groupes homogènes pour des études ultérieures. Nous étudions la performance de notre méthode par une étude de simulation, et nous utilisons notre modèle pour classer les gloutons de l'Ouest du Montana en utilisant des données microsatellite.

En réponse à la multiplication de données de survie (ou plus généralement de données de délai de survenue d'un évènement) spatialement dépendantes, leur modélisation a fait l'objet d'une attention accrue. Dans les modèles classiques (modèles semi paramétriques du type modèles à hasard proportionnel ou à taux de défaillance accéléré), les motifs spatiaux sont introduits au travers de termes de fragilité spatialement dépendants dans le prédicteur. Dans le cadre des modèles à hasards proportionnels, nous proposons d'introduire la dépendance spatiale au niveau de la fonction de survie de base : plus précisément, nous proposons comme loi a priori sur cette fonction un mélange d'arbres de Polya spatialement dépendants.
Grâce aux techniques modernes de méthodes de Monte-Carlo par Chaînes de Markov (MCMC), cette approche reste raisonnable d'un point de vue computationnel dans un contexte bayésien hiérarchique. Nous comparons l'approche par mélange d'arbres de Polya spatialement dépendants avec l'approche classique par fragilité spatiale et nous illustrons l'intérêt de la méthode sur une analyse de données de survie au cancer du sein en Iowa collectées par le programme « Surveillance, Epidemiology, and End Results » de l'Institut National du Cancer.
Notre méthode démontre de meilleures performances que les méthodes alternatives classiques en terme de qualité d'adéquation, comme le montrent les calculs de la pseudo log-vraisemblance marginale (LPML), du critère de déviance (DIC) ou de la densité prédictive évaluée à partir de la totalité de l'échantillon observé (autrement appelée « full sample score »).

Dans cet article, nous proposons l'utilisation d'une famille de modèles de transformations semi-paramétriques à coefficients dépendants du temps, pour des données d'événement récurrent en présence d'un événement terminal tel que la mort. Ces nouveaux modèles offrent une grande flexibilité dans la formulation des effets de covariables sur les fonctions moyennes des événements récurrents au sein des survivants à un temps donné. Pour l'inférence sur les modèles proposés, on développe une classe d'équations d'estimation et on établit les propriétés asymptotiques des estimateurs correspondants. De plus, un test de défaut d'ajustement est proposé pour établir la validité du modèle, et quelques tests sont aussi proposés pour étudier la dépendance au temps des effets des covariables. Le comportement de la méthode proposée pour des échantillons de taille limitée est examiné par des études de simulation de Monte-Carlo, et on illustre également par une application à des données de cancer de la vessie.

Dans ce travail, les auteurs fournissent une nouvelle classe de modèles de fragilité à risques compétitifs pour des données de survie groupées. Cette classe se fonde sur une extension du modèle à risques compétitifs de Prentice et al (1978) permettant d'inclure des termes de fragilité, avec l'utilisation de modèles cause-spécifiques avec fragilité et risques proportionnels pour toutes les causes. Des estimateurs du maximum de vraisemblance paramétriques et non paramétriques sont proposés. Les principaux avantages de la classe de modèles proposés, par rapport aux modèles existants, sont : (1) l'inclusion de covariables ; (2) la structure flexible de la dépendance entre les différents types de délais de survie au sein d'un groupe ; et (3) la structure de dépendance intra-sujet non spécifiée. Les procédures d'estimation proposées produisent les estimateurs paramétriques et semi-paramétriques les plus efficaces et sont faciles à implémenter. Des études de simulation montrent que les méthodes proposées sont très performantes dans des situations pratiques.

Y. Li, L. Tian, and L.-J. Wei **427**
Estimating Subject-Specific Dependent Competing Risk Profile with Censored Event Time Observations

Dans une étude longitudinale, supposons que le critère principal soit le temps écoulé jusqu'à la survenue d'un certain type d'évènement. Cette variable réponse, cependant, peut être censurée par une variable de censure indépendante ou bien par la survenue d'un parmi plusieurs évènements dépendants en compétition. La question est de savoir comment construire une règle de prédiction fiable du profil des risques compétitifs d'intérêt à un moment donné pour un futur sujet dans une perspective décisionnelle en termes de bénéfice et de risque. Dans cet article nous proposons une procédure en deux étapes pour une inférence sur les profils spécifiques des sujets. Nous estimons ensuite de façon convergente les risques compétitifs moyens pour les sujets qui ont le même score d'indice paramétrique au moyen d'une procédure non-paramétrique d'estimation fonctionnelle. Nous illustrons cette nouvelle proposition avec des données d'un essai clinique randomisé pour évaluer l'efficacité d'un traitement du cancer de la prostate. Le critère principal pour cette étude était le temps écoulé jusqu'au décès par cancer de la prostate, mais il y avait deux types d'évènements dépendants en compétition à savoir le décès d'origine cardio-vasculaire et le décès pour toute autre cause.

T. Cai, T. A Gerds, Y. Zheng, and J. Chen **436**
Robust Prediction of t-Year Survival with Data from Multiple Studies

Pour évaluer un effet commun sur la base d'informations recueillies dans plusieurs études, la technique de méta-analyse, ces dernières années, a largement été utilisée. Regrouper des données issues d'études similaires est particulièrement intéressant en génomique, où les effectifs par étude sont plutôt faibles en comparaison du nombre de paramètres d'intérêt. Nous nous penchons ici sur l'établissement de règles pronostiques robustes pour la prédiction de la survie à t années sur la base de plusieurs études. Nous proposons de construire un score prédictif composite en ajustant un modèle stratifié de transformation semi-paramétrique ; ce modèle admet que les critères pronostiques recueillis dans les différentes études puissent ne pas être identiques, mais seulement reliés entre eux. Afin d'évaluer la pertinence du score obtenu, nous fournissons les estimateurs et intervalles de confiance des mesures usuelles d'adéquation d'un modèle, en particulier l'adaptation aux données de survie des courbes ROC ainsi que les valeurs prédictives positive et négative. Ces procédures sont appliquées à l'établissement de règles pronostiques de la survie à 5 ans de patientes atteintes du cancer du sein, sur la base de cinq études génomiques dans cette indication.

E. F. Acar, R. V. Craiu, and F. Yao **445**
Dependence Calibration in Conditional Copulas: A Nonparametric Approach

L'étude des dépendances entre variables aléatoires est un sujet récurrent en statistiques. En général, le degré de dépendance entre deux (ou plus) variables aléatoires est fonction d'une covariable observée. Dans ce cadre, nous proposons une méthode d'inférence statistique reposant sur un modèle de copule conditionnel dans lequel la fonction de copule appartient à une famille paramétrique et le paramètre de copule est fonction de la covariable. Afin d'estimer la relation fonctionnelle entre le paramètre de copule et la covariable, nous proposons une approche non-paramétrique fondée sur la vraisemblance locale. Pour un jeu de données fixé, le choix de la famille de copule le mieux adapté à ce jeu est une question délicate. Le cadre de travail proposé conduit naturellement à une nouvelle méthode de sélection de modèle de copule reposant sur les erreurs de prédiction estimées par validation croisée. Nous donnons les biais et variance asymptotiques de l'estimateur polynomial local et proposons une méthode de construction d'intervalles de confiance. Sur un

échantillon de taille finie, les performances de notre méthode sont étudiées sur données simulées et illustrées sur un jeu de données réel de naissances multiples.

D. Ryu, E. Li, and B. K. Mallick **454**
Bayesian Nonparametric Regression Analysis of Data with Random Effects Covariates from Longitudinal Measurements

Nous nous intéressons à la régression non paramétrique dans le contexte du modèle linéaire généralisé (GLM) pour des données avec des covariables qui sont les effets aléatoires sujet-dépendants de mesures longitudinales. L'hypothèse habituelle de linéarité des processus sous-tendant les effets des covariables longitudinales peut s'avérer peu réaliste, et dans ce cas crée un doute sur l'inférence faite pour ces effets de covariables. Supposant inconnues les fonctions de régression, nous proposons d'appliquer des méthodes bayésiennes nonparamétriques, avec des splines de lissage cubiques ou P-splines pour la possible nonlinéarité, et d'utiliser un modèle additif dans ce dispositif complexe. Nous proposons l'utilisation de schémas d'augmentation de données pour améliorer l'efficience des calculs. Cette approche permet des structures de covariances flexibles pour les effets aléatoires et les erreurs de mesure intra-sujet du processus longitudinal. L'espace des modèles a posteriori est exploré par un échantillonnage MCMC. Les méthodes proposées sont illustrées et comparées à d'autres approches, l'approche naïve et la calibration par régression, au travers de simulations et par une application qui étudie les relations entre l'obésité à l'âge adulte et les courbes de croissance de l'enfance.

C. Sommen, D. Commenges, S. L. Vu, L. Meyer, and A. Alioum **467**
Estimation of the Distribution of Infection Times Using Longitudinal Serological Markers of HIV: Implications for the Estimation of HIV Incidence

Pendant la dernière décennie, l'intérêt s'est porté sur l'analyse des anticorps anti-VIH et sur de nouvelles stratégies de tests qui permettent de distinguer les infections récentes des infections déjà établies à partir d'un seul échantillon de sérum. Des estimations de l'incidence sont obtenues en utilisant la relation entre prévalence, incidence et durée de l'infection récente (« période fenêtre »). Cependant, des travaux récents ont montré les limites de cette approche dues à l'utilisation d'une estimation de la « période fenêtre » moyenne. Nous proposons une approche alternative qui consiste à estimer la distribution des temps d'infection à partir des valeurs de marqueurs sérologiques obtenues au moment de la découverte de la séropositivité. Nous proposons un modèle pour la dynamique des marqueurs basé sur les mesures répétées des marqueurs virologiques de la séroconversion. Les paramètres du modèle sont estimés à partir des données d'une cohorte de sujets infectés par le VIH et inclus au cours de la primo-infection. Ce modèle peut être utilisé pour estimer la distribution des temps d'infection pour des individus nouvellement diagnostiqués et reportés dans le système de surveillance du VIH. Une nouvelle méthode est proposée pour estimer l'incidence de l'infection à VIH à partir de ces résultats.

D. Liu, T. Lu, X.-F. Niu, and H. Wu **476**
Mixed-Effects State-Space Models for Analysis of Longitudinal Dynamic Systems

Le rapide développement de nouvelles biotechnologies permet de comprendre en profondeur
les systèmes dynamiques biomédicaux au niveau cellulaire, mais également à un niveau plus fin. De nombreux systèmes biomédicaux, prenant en compte une composante individuelle, peuvent être décrit par un ensemble d'équations différentielles qui sont similaires à celles des systèmes dynamiques, utilisées en ingénierie. Dans cet article, motivé par une étude de la dynamique du VIH, nous proposons une classe de modèles à espace d'état qui inclue des effets mixtes, et qui se fonde sur la caractéristique longitudinale des systèmes dynamiques; ces modèles permettant de modéliser, avec une grande flexibilité, les corrélations intra et inter individuelles. Pour pouvoir estimer les paramètres des modèles proposés, nous modifions les approches bayésiennes et par maximum de vraisemblance utilisées dans les modèles à espace d'état et les modèles à effets mixtes standards. S'agissant de l'approche bayésienne, nous donnons les lois conditionnelles intervenant dans l'algorithme de Gibbs ; ce qui nous permet d'explorer la loi a posteriori. S'agissant de l'approche par maximum de vraisemblance, nous mettons en oeuvre un algorithme de type EM qui inclut une étape de Gibbs pour évaluer l'espérance conditionnelle qui apparaît dans l'étape E de l'algorithme EM. Des études par simulation sont réalisées pour comparer ces deux approches. Nous appliquons les modèles à espace d'état avec effets mixtes à un jeu de données issu d'un essai clinique relatif au sida, ce qui nous permet d'illustrer les méthodologies proposées. Les modèles et méthodes proposés dans

cet article ont potentiellement des applications dans l'analyse d'autres systèmes biomédicaux, tel que la dynamique d'une tumeur (dans l'étude du cancer), ou la modélisation de réseaux régulés en génétique.

Dans beaucoup d'investigations biomédicales, l'objectif principal est l'identification des sujets qui sont susceptibles à une un certain type d'exposition ou un traitement d'intérêt. Nous nous concentrons sur les méthodes répondant à cette question dans les études longitudinales quand l'intérêt porte sur la susceptibilité en rapport avec la mesure initiale d'un sujet ou le niveau moyen. Dans ce contexte, nous proposons un modèle à ordonnée à l'origine aléatoire et pente fonctionnelle ne nécessitant pas d'hypothèse d'association linéaire entre les coefficients aléatoires du modèle mixte et fournissant une estimation de la forme fonctionnelle de cette relation. Nous proposons une formulation de la fonction non paramétrique par spline pénalisé pour représenter cette relation, et implémentons une approche bayesienne complète pour estimer le modèle. Nous explorons la performance fréquentiste de notre approche par simulation, et appliquons le modèle à des données portant sur le débit sanguin coronaire dans une étude toxicologique animale. Les principes généraux introduits ici s'appliquent plus largement à des contextes dans lesquels un intérêt est porté à la relation entre mesure initiale et changement au cours du temps.

Nous nous intéressons à la sélection simultanée d'effets fixes et d'effets aléatoires dans une classe générale de modèles mixtes en utilisant l'estimation du maximum de vraisemblance pénalisé (MPL) à l'aide de fonctions de pénalités de la déviation absolue (SCAD) et LASSO- adaptative (ALASSO). On montre que les estimations du maximum de vraisemblance pénalisé possèdent des propriétés de convergence, de recouvrement, et ont des distributions asymptotiquement normales. Un critère de sélection de modèle, appelé la statistique IC_Q, est proposée pour la sélection des paramètres de pénalité (Ibrahim, Zhu et Tang, 2008). Il est montré que la procédure de sélection de variable basée sur IC_Q sélectionne les effets fixes et aléatoires de façon consistante. La méthode est très générale et peut s'appliquer à de nombreuses situations impliquant des effets aléatoires, incluant les modèles linéaires généralisés mixtes. Des études de simulations et sur un jeu de données réelles provenant d'une étude de Yale sur la croissance infantile ont été utilisées pour illustrer la méthode proposée.

Cet article considère le problème de la sélection de prédicteurs de délais de survie à partir d'un ensemble de très nombreux prédicteurs candidats en utilisant les données de multiples études. Alternativement aux approches de tests multi-étapes actuelles, nous proposons de modéliser explicitement l'hétérogénéité entre études à l'aide d'un modèle hiérarchique pour gagner en force. Notre méthode incorpore les données censurées par un modèle à vie accélérée. En utilisant une spécification de l'*a priori* soigneusement formulée, nous développons une approche rapide pour la sélection de prédicteurs et l'estimation de la réduction pour des prédicteurs de dimension élevée. Pour l'ajustement du modèle, nous développons un algorithme Monte Carlo EM (MC-EM) afin de concilier les données censurées. L'approche proposée, qui s'apparente à la machine à vecteur de pertinence (*relevance vector machine*, RVM), s'appuie sur l'estimation du maximum *a posteriori* (MAP) pour obtenir rapidement un estimateur dispersé. Comme pour la RVM typique, il y a une propriété de seuil intrinsèque selon laquelle les prédicteurs sans importance ont tendance à avoir leur coefficient réduit à zéro. Nous comparons notre méthode avec quelques procédures utilisées usuellement au moyen d'études de simulation. Nous illustrons aussi la méthode à l'aide des données sur les codes-barres d'expression de gènes provenant de trois études sur le cancer du sein.

La méthodologie de la réduction de dimension suffisante (SDR) a permis de faciliter l'analyse de

régression de données de grande dimension. Lorsque la réponse est censurée cependant, la plupart des estimateurs SDR existants ne peuvent être utilisés, ou imposent des conditions restrictives. Dans cet article nous proposons une nouvelle classe d'estimateurs SDR pondérés par l'inverse de la probabilité de censure, pour des régressions censurées. De plus, nous introduisons une régularisation pour aboutir à une sélection de variables et une réduction de dimension simultanées. Nous examinons les propriétés asymptotiques et les performances empiriques des méthodes proposées.

La portée clinique des conclusions issues de modèles de survie se juge à l'aune de l'erreur de prédiction. Plusieurs approches permettant d'évaluer la performance prédictive des modèles de survie ont été publiées. Nous analysons ici les propriétés de l'estimateur de l'erreur de prédiction développé par Schemper et Henderson (2000, *Biometrics* **56**, 249-255), qui quantifie la distance absolue entre les fonctions de survie prédite et observée. Nous apportons la preuve formelle que l'estimateur proposé par Schemper et Henderson n'est pas robuste à une mauvaise spécification du modèle de survie, c'est-à-dire que cet estimateur n'est pertinent que si la famille de modèles utilisée pour produire les prédictions a été correctement spécifiée. Pour remédier à ce problème, nous construisons un nouvel estimateur de la distance absolue entre les fonctions de survie prédite et observée. Nous montrons que cet estimateur modifié de Schemper-Henderson est robuste à une mauvaise spécification du modèle, permettant de l'appliquer à un large ensemble de modèles de survie. Les propriétés de l'estimateur de Schemper-Henderson et de sa version ainsi modifiée sont illustrées au moyen d'une étude de simulation et de l'analyse de deux jeux de données cliniques.

Un défi fréquent consiste à faire de l'inférence sur les paramètres d'un modèle statistique d'intérêt à partir de données longitudinales sujettes à la sortie d'étude. Ces dernières sont un cas particulier du cadre général des données dégradées monotones. Récemment, une attention considérable s'est portée sur les estimateurs doublement robustes qui font intervenir dans ce contexte des modèles présupposés pour le mécanisme de données manquantes (ou plus généralement de données dégradées) et pour des aspects de la distribution des données complètes. Ces estimateurs ont la propriété intéressante de fournir des inférences consistantes si un seul des modèles est correctement spécifié. Les estimateurs doublement robustes ont été critiqués à cause de performances potentiellement catastrophiques lorsque les deux modèles sont même très légèrement mal spécifiés. Nous proposons un estimateur doublement robuste qui s'applique à tous les cas de données dégradées monotones, et qui donne des performances comparables ou meilleures par rapport aux méthodes doublement robustes existantes. Nous démontrons cela par des études de simulation et par une application aux données d'un essai clinique sur le SIDA.

Avec les progrès de la médecine moderne et du diagnostic clinique il est souvent possible de caractériser plus finement des sous-types de cas avec les données d'études cas-témoin. Dans les études associant cas et témoins, l'absence de données sur les niveaux d'exposition entraîne souvent la suppression de tout l'ensemble de données concerné, avec une perte substantielle de l'information. Quand les sous-types de cas sont traités comme des données catégorielles, les données sont stratifiées et la suppression d'une observation est encore plus couteuse en terme de précision des odds-ratios relatifs à ces catégories, particulièrement pour le modèle logistique multinomial. Le modèle de régression par stéréotype pour réponse catégorielle est intermédiaire entre le modèle d'odds proportionnels et le modèle complètement multinomial ou modèle logistique avec baseline. L'utilisation de cette classe de modèles a jusqu'ici été limitée car sa structure implique des difficultés dans l'inférence dues à la non-linéarité et la non-identifiabilité des paramètres. Avec la régression par stéréotype, nous illustrons une façon de traiter le cas de données manquantes dans des études cas-témoins avec une subdivision fine des cas. Nous présentons deux méthodes, l'une est une méthode de Monte Carlo complètement bayésienne, l'autre un algorithme espérance/maximisation conditionnelle, pour l'estimation

des paramètres en présence de mécanismes de données manquantes complètement généraux. Nos méthodes sont illustrées par une application à une étude cas-témoins en cours sur le cancer colorectal. Des résultats de simulation sont présentés pour divers mécanismes de données manquantes et écarts par rapport aux hypothèses impliquées par la modélisation.

En recherche médicale, les courbes ROC sont utilisées pour évaluer la performance de biomarqueurs pour le diagnostic des pathologies ou la prédiction du risque de développer une maladie. L'aire sous la courbe ROC (AUC), comme mesure résumée de courbes ROC, est largement utilisée, particulièrement pour comparer plusieurs courbes ROC. Dans les études observationnelles, l'estimation de cette AUC est souvent compliquée du fait de valeurs manquantes des biomarqueurs, les estimateurs existants de l'AUC étant alors potentiellement biaisés. Dans cet article, nous développons des méthodes statistiques robustes pour l'estimation de l'aire sous la courbe ROC, et les méthodes proposées utilisent l'information de variables auxiliaires potentiellement prédictives de l'absence de biomarqueurs ou de valeurs des biomarqueurs. Nous nous sommes particulièrement intéressés aux variables auxiliaires qui permettent de prédire les valeurs manquantes des biomarqueurs. Dans le cas de valeurs manquantes aléatoire (MAR), c'est-à-dire lorsque l'absence des valeurs des biomarqueurs ne dépend que des des données observées, nos estimateurs possèdent la propriété appréciable d'être consistants en cas de spécification correcte, conditionnellement aux variables auxiliaires et à l'état pathologique, soit du modèle pour la probabilité d'absence soit du modèle des valeurs des biomarqueurs. Dans le cas d'absences non purement aléatoires (MNAR), c'est-à-dire où l'absence peut dépendre des valeurs non observées des biomarqueurs, nous proposons une étude de sensibilité pour évaluer l'impact de ces absences non purement aléatoires sur l'estimation de l'aire sous la courbe ROC. On étudie les propriétés asymptotiques des estimateurs proposés et des études de simulations évaluent le comportement à distance finie. Les méthodes sont enfin illustrées à partir de données provenant d'une étude sur la dépression de la future mère pendant la grossesse.

L'exactitude d'un diagnostic peut être améliorée par la combinaison de plusieurs tests diagnostiques. Nous considérons la combinaison linéaire optimale qui maximise l'aire sous la courbe ROC (Receiver Operating Characteristic). On peut estimer les coefficients de cette combinaison par une procédure non paramétrique. Cependant, pour estimer l'aire sous la courbe associée à ces coefficients, la solution simpliste remplaçant les coefficients par leur estimation est trop optimiste. Nous proposons plusieurs méthodes pour redresser ce biais positif, dont la validation croisée, particulièrement recommandée, et nous développons une approximation de la validation croisée qui réduit la charge de calcul. Ces méthodes peuvent aussi être utilisées dans une démarche de sélection de variables afin d'effectuer un choix parmi les tests disponibles. Nous les validons par simulations et nous les appliquons à trois jeux de données réelles.

En utilisant uniquement des comptages avec réplication spatiale et temporelle, Royle (2004b, *Biometrics* **760**, 108-115) a développé un modèle de *N*-mélange pour estimer l'abondance d'une population animale lorsque la probabilité de individuelle de détection d'un animal est inconnue. Une hypothèse inhérente à ce modèle est que les populations animales, pour chaque position échantillonnée, sont fermées relativement aux migrations, naissances et morts tout au long de l'étude. Dans le passé, ceci a pu être vérifié seulement par des arguments biologiques lis au schéma de l'étude car aucune vérification statistique n'était disponible. Dans cet article, nous proposons une généralisation du modèle de *N*-mélange pouvant être utilisé pour tester formellement l'hypothèse de fermeture. De plus, lorsqu'il est appliqué à une métapopulation ouverte le modèle généralisé fournit des estimateurs des paramètres de dynamique de population et donne des estimateurs d'abondance qui tiennent compte des probabilités imparfaites de détection et qui ne nécessitent pas l'hypothèse de fermeture. Une étude de simulation montre que ces estimateurs de l'abondance sont moins biaisés que les estimateurs correspondants obtenus à partir du modèle original de *N*-mélange. Le modèle proposé est ensuite appliqué deux ensembles de données de

comptage. Le premier exemple applique le test de fermeture que une étude à saison unique de canards colverts (*Anas platyrhynchos*), et la seconde utilise le modèle proposé pour estimer les paramètres de dynamique de population et l'abondance annuelle de merles d'Amérique (*Turdus migratorius*) à partir d'une étude sur plusieurs saisons.

Nous présentons une procédure quasi automatique pour localiser et dénombrer des nano-cristaux dans des images de dispositifs à kinésine motrice. Notre procédure utilise un estimateur de vraisemblance approché basé sur un mélange de modèles à deux composantes ; la première composante a une distribution normale et l'autre composante est distribuée selon la somme d'une variable aléatoire normale et d'une variable aléatoire exponentielle. La composante normale a une variance inconnue que nous modélisons comme une fonction de la moyenne. Nous utilisons des B-Splines pour estimer la fonction de variance sur des jeux d'apprentissage à partir d'une image appropriée et l'estimation est utilisée sur des images ultérieures. Les estimations des paramètres sont générées pour chaque image ainsi que des estimations de leurs écarts-types, le nombre de cristaux dans l'image est déterminé à l'aide d'un critère d'information et de tests du rapport des vraisemblances. Des simulations réalistes montrent que notre procédure est robuste qu'elle fournit des estimations justes, à la fois sur les paramètres et leurs écarts-types.

On propose un cadre général pour les schémas bayésiens, établis sur modèles, des essais de phase I en cancérologie, dans lequel un critère général de cohérence (Cheung, 2005) de schéma est également développé. Ce cadre peut incorporer des aspects d'éthique aussi bien individuelle que collective dans le schéma de l'essai. Nous proposons un nouveau principe qui minimise une fonction de risque comprenant deux termes, l'un représentant le risque individuel de la dose courante, et l'autre représentant le risque collectif. On étudie la performance de cette approche, mesurée en termes de précision de dose à la fin l'essai, de toxicité et de taux de surdosage, et de fonctions de pertes reflétant les éthiques individuelles et collectives, et la comparaison avec les schémas bayésiens existants montrent sa supériorité.

Supposons que nous soyons intéressés par faire de l'inférence sur un ensemble de paramètres contraints. Les régions de confiance pour ces paramètres sont souvent construits à partir d'une approximation normale de la loi d'un estimateur (convergent) d'une transformation de ces paramètres. Dans cet article, nous utilisons la notion de loi de confiance, le pendant fréquentiste de la loi a posteriori en statistique bayésienne, pour obtenir des régions de confiance optimales. Les membres d'une telle région peuvent être générés, de façon efficace, par des méthodes MCMC standards. Nous utilisons cette technique pour produire une inférence sur le profil temporel d'une fonction de survie, et ce en présence d'observations censurées. Nous illustrons cette nouvelle approche avec des données de survie provenant d'une étude bien connue, dite étude Mayo, qui porte sur la cirrhose biliaire primaire, et nous montrons que le volume de la région de confiance que nous proposons est 1.34 fois plus petite que celle de la bande de confiance conventionnelle.

L'un des éléments clefs de la découverte de nouveaux médicaments est la déclinaison chimique des pharmacophores, les supports conceptuels de l'activité. Le « patron » du pharmacophore est caractérisé par les agencements spatiaux relatifs et les propriétés physico-chimiques communes à toutes les molécules actives, appelées ligands, susceptibles de se lier à un récepteur particulier sur une protéine. Avec cette importante application en vue nous avons développé un modèle Bayésien hiérarchique pour la dérivation de patrons de pharmacophore à partir de multiples configurations d'ensembles de points, partiellement indexés par le type d'atome en ces points. Le modèle est développé au moyen d'un algorithme de recherche à

plusieurs niveaux qui produit une série de patrons respectant les relations géométriques entre les atomes qui se correspondent dans les différentes configurations. L'information chimique est prise en compte en distinguant entre les atomes d'éléments similaires ou différents, des espèces chimiques différentes ayant des probabilités moindres d'être mises en correspondance que les mêmes espèces. Nous illustrons notre méthode par des exemples de détermination de patrons de pharmacophores à partir d'ensembles de ligands tous susceptibles de liaison structurelle avec des sites actifs de protéines et nous montrons que notre modèle bayésien est capable de retrouver des caractéristiques clefs de pharmacophores dans deux cas tests.

L'Enquête Nationale sur le VIH/SIDA et la Santé Reproductive (NARHS) de 2005 au Nigeria a montré que le multipartenariat sexuel augmente le risque de contracter le VIH et les autres maladies sexuellement transmissibles. Par conséquent, la réduction du nombre de partenaires sexuels est une des stratégies de prévention pour atteindre l'objectif du Millénaire pour le développement de maîtriser et de stopper la propagation du VIH/SIDA. Nous considérons les nombres de petites amies, de partenaires occasionnels, et de prostituées fréquentés par des hommes hétérosexuels, rapportés dans l'enquête NARHS, comme indicateurs observés de leur attitude latente envers le multipartenariat sexuel. Pour explorer l'effet des facteurs de risque sur cette variable latente, nous proposons une extension d'une approche semi-paramétrique pour les modèles à variables latentes avec les indicateurs continus et catégoriels pour inclure les indicateurs de dénombrement. Ceci nous permet d'analyser simultanément des effets linéaires et non linéaires des covariables, telles que des facteurs sociodémographiques et la connaissance sur le VIH/SIDA, sur l'attitude envers le multipartenariat sexuel, qui sont censées influencer les indicateurs observés de dénombrement. Les résultats fournissent des indicateurs pertinents pour les décideurs politiques qui cherchent à réduire la propagation du VIH/SIDA au sein de la populace Nigériane à travers la réduction du nombre de partenaires sexuels.

Dans certaines études biomédicales impliquant des réponses binaires en grappe (disons, le statut d'une maladie), les tailles des grappes peuvent varier parce que certains éléments des grappes sont absents. Lorsqu'à la fois la présence d'une grappe et le statut binaire de la maladie d'un élément présent sont traités comme les réponses d'intérêt, nous proposons un nouveau cadre à deux degrés d'une régression logistique à effets aléatoires. Pour la simplicité d'interprétation des effets des régressions, on préserve les formes approximatives de la régression logistique à la fois pour la probabilité marginale de la présence/absence d'un élément et pour la probabilité conditionnelle du statut de la maladie d'un élément présent. Nous présentons une méthode d'estimation par maximum de vraisemblance qui puisse être implémentée en utilisant des logiciels statistiques standards. Nous comparons nos modèles et l'interprétation physique des effets de régression avec d'autres méthodes de la littérature. Nous présentons également une étude de simulation pour mesure la robustesse de notre procédure à une mauvaise spécification de la distribution des effets aléatoires et pour comparer les performances des estimations pour échantillons finis avec des méthodes existantes. La méthodologie est illustrée en analysant une étude sur le statut de santé parodontale d'une population de diabétiques chez les Gullah.

En épidémiologie, il est de plus en plus fréquent de doser des marqueurs biologiques ou environnementaux sur des regroupements de prélèvements individuels. Dans cet article, nous nous intéressons à l'ajustement d'un modèle à une réponse binaire quand la mesure d'une exposition importante a été effectuée sur des prélèvements regroupés. Nous utilisons une approche par calibration de la régression et nous proposons plusieurs méthodes dont des méthodes de remplacement qui prédisent l'exposition individuelle d'un sujet à partir de la mesure globale à laquelle il a contribué et de l'information apportée par d'autres covariables ainsi que des méthodes gaussiennes qui effectuent des ajustements supplémentaires en supposant la normalité des erreurs de calibration. Dans chacune de ces classes, nous envisageons deux approches de la calibration (par augmentation de covariable et par imputation). Des simulations montrent que ces approches réduisent le biais associé à la méthode naïve qui remplace les mesures individuelles par la mesure globale à

laquelle chaque sujet a contribué. En particulier l'approche par imputation gaussienne conduit à des résultats raisonnablement satisfaisants sous diverses conditions, y compris en présence d'une distribution dissymétrique des erreurs de calibration. Nous utilisons des données du Projet Collaboratif Périnatal pour illustrer ces méthodes.

Le test de la somme des rangs de Wilcoxon est largement utilisé pour les comparaisons de deux groupes de données non gaussiennes. Une hypothèse de ce test est l'indépendance des unités d'échantillonnage à la fois inter- et intra-groupes qui peut être violée dans un contexte de données corrélées comme pour des essais cliniques ophtalmologiques, où l'unité d'échantillonnage est l'individu mais l'unité d'analyse est l'œil. Dans ce but, nous avons proposé le test de Wilcoxon pour grappes afin de prendre en compte la corrélation entre les multiples sous-unités du même groupe (Rosner, Glynn et Lee, *Biometrics* 2003, 2006). Cependant une estimation de la puissance est nécessaire pour planifier les études qui utilisent cette approche analytique. Nous avons récemment publié des méthodes pour estimer la puissance et la taille d'échantillon pour un test ordinaire de la somme des rangs de Wilcoxon (Rosner et Glynn, *Biometrics*, 2009). Dans ce papier nous présentons des extensions de cette approche pour estimer la puissance pour un test de Wilcoxon groupé. Des études de simulations montrent une bonne concordance entre la puissance estimée et empirique. Ces méthodes sont illustrées avec des exemples d'essais randomisés en ophtalmologie. Une puissance accrue est obtenue en utilisant les sous-unités comme unités d'analyse au lieu de la grappe lorsqu'on utilise le test de la somme des rangs de Wilcoxon.

Litière & al (2007) ont présenté les résultats d'études de simulation qui, d'après eux, montraient que la mauvaise spécification de la forme de la distribution des effets aléatoires pouvaient produire des augmentations marquées de l'erreur de type II (baisse de la puissance) des tests basés sur les ajustements de modèles linéaires généralisés mixtes. Cependant, le papier contient une faille conceptuelle qui invalide leur conclusion. Nous présentons des études de simulation logiquement correctes qui démontrent un petit accroissement de l'erreur de type II, consistant avec les travaux précédents qui montrent un effet minime entraînée par une mauvaise spécification.

Le modèle à risque proportionnel de Fine-Gray est couramment utilisé pour estimer l'effet de facteurs pronostiques sur l'incidence cumulée d'un évènement d'intérêt en présence d'évènements concurrents. Cependant il est fréquent que l'hypothèse de proportionnalité des risques ne soit vérifiée, notamment dans des études multicentriques où le risque de base peut varier entre les centres. Une extension permettant la stratification du risque de base est présentée en considérant deux types de stratifications en fonction du rapport entre nombre de strates et effectifs des strates. Des estimateurs consistants des effets de facteurs pronostiques sont proposés ainsi que leur comportement asymptotiques pour des données faiblement stratifiées comportant un petit nombre de grandes strates, comme les bras de traitements d' un essai randomisé ou des données fortement stratifiées comportant un grand nombre de petites strates p.ex. les centres de greffes au sein d'un registre international.