Modèles de copules Archimédiennes pour données de Bernoulli corrélées

Thèse

Fodé Tounkara

Doctorat en Mathématiques
Philosophiæ doctor (Ph.D.)

Québec, Canada

Résumé

Cette thèse introduit et explore une nouvelle classe de modèles probabilistes pour des données de Bernoulli échangeables en forme de grappe. Dans ces modèles, la probabilité conditionnelle de succès est une fonction de la probabilité marginale de succès et d'un effet aléatoire positif spécifique à chaque grappe. La distribution de l'effet aléatoire contient un paramètre d'association qui est estimé pour donner une mesure de la force de la dépendance résiduelle ignorée par les marges. Nous montrons que la transformée de Laplace de l'effet aléatoire est liée au générateur des modèles de copules Archimédiennes, ce qui nous permet d'avoir un nouvel aperçu de ces modèles.

L'approche que nous proposons offre de nombreux avantages. En effet, la famille de copules Archimédiennes fournit une large classe de modèles pour la sur-dispersion dans une expérience de Bernoulli. D'un point de vue statistique, la fonction de vraisemblance marginale pour les données de l'échantillon a une expression explicite, les méthodes du maximum de vraisemblance sont alors faciles à mettre en œuvre.

Nous avons considéré quatre applications de nos modèles. Premièrement, nous construisons un intervalle de confiance par vraisemblance profilée pour le coefficient de corrélation intra-grappe (ICC). La deuxième application concerne l'estimation de la taille d'une population en présence d'hétérogénéité observée et non observée (résiduelle) dans une expérience de capture-recapture. Le troisième problème traite de l'estimation dans de petites régions, et enfin le quatrième indépendant des trois premiers, analyse les caractéristiques socio-économiques des hommes qui ont une préférence à épouser des jeunes filles de moins de 18 ans.

Dans la première application, nous considérons le cas le plus simple de nos modèles où aucune covariable n'est disponible puis proposons la méthode du maximum de vraisemblance pour l'estimation du coefficient de corrélation intra-grappe (ICC) à l'aide de plusieurs spécifications de copules Archimédiennes. La sélection d'un modèle particulier est effectuée en utilisant le critère d'information d'Akaike (AIC). La procédure comprend l'estimation du maximum de vraisemblance et la méthode du profil de vraisemblance (ou vraisemblance profilée). Nous avons fait des études de simulation pour mesurer la performance de la méthode d'intervalle par vraisemblance profilée sous nos modèles en termes de taux de couverture et de longueur d'intervalle de confiance, et la sensibilité de notre approche à la spécification d'un modèle de copule. La procédure que nous proposons a aussi été appliquée à des données réelles. Nous comparons notre méthode à celle proposée sous le modèle Béta-binomial, et la méthode d'intervalle de type Wald modifié proposée par Zou and Donner (2004). L'une des conclusions importantes de ces études est que l'intervalle de confiance par vraisemblance profilée obtenu sous nos modèles présente de belles propriétés en termes de taux couver-

ture et de longueur d'intervalle de confiance, même lorsque le nombre de grappes est petit. La sélection de modèle est une étape importante : si le modèle est mal spécifié, alors cela pourrait conduire à des résultats erronés.

La seconde application, une extension de la première pour accommoder des covariables au niveau des grappes, concerne la modélisation de l'hétérogénéité dans les probabilités de capture lors d'une expérience de capture-recapture dans une population fermée. Dans ce contexte, nos modèles sont utilisés pour modéliser l'hétérogénéité résiduelle qui n'est pas prise en compte par les covariables mesurées sur des unités capturées. Plusieurs modèles sont disponibles pour l'hétérogénéité non observée et la probabilité de capture marginale est modélisée en utilisant les fonctions de liens Logit et Log-Log complémentaire. Les paramètres sont estimés en utilisant la vraisemblance conditionnelle construite à partir des observations collectées sur les unités capturées au moins une fois. Ceci généralise le modèle de Huggins (1991) qui ne tient pas compte de l'hétérogénéité résiduelle. La sensibilité de l'inférence à la spécification d'un modèle est également étudiée par des simulations. Un exemple numérique est présenté.

La troisième application traite de la prédiction dans de petites régions. Nous proposons des techniques de Bayes basées sur nos modèles pour estimer des proportions régionales. L'inférence Bayésienne que nous proposons consiste à trouver la distribution a posteriori de l'effet aléatoire et sa transformée de Laplace sachant les données et les paramètres du modèle. Cette transformée de Laplace est ensuite utilisée pour trouver des estimateurs de Bayes et leurs variances a posteriori pour les vraies proportions. Nous développons une étude de comparaison entre le meilleur prédicteur de Bayes (BP) et le meilleur prédicteur linéaire sans biais (BLUP). Nous avons également étudié l'efficacité du BP obtenu sous nos modèles relativement au BLUP. Les paramètres du modèle sont estimés en utilisant la méthode du maximum de vraisemblance. L'avantage de notre approche est que la fonction de vraisemblance et l'expression du meilleur prédicteur (BP) ont une forme explicite, ce qui facilite la mise en oeuvre de leur évaluation sur le plan numérique. Nous obtenons un prédicteur empirique de Bayes (EBP) en remplaçant les paramètres par leurs estimateurs dans l'expression du BP. Nous utilisons le critère d'information d'Akaike (AIC) pour la selection d'un modèle. Nous utilisons la méthode du jackknife pour estimer l'erreur quadratique moyenne des prédicteurs empiriques. Des résultats empiriques obtenus à partir de données simulées et réelles sont également présentés.

Enfin, le quatrième problème traité dans cette thèse, qui est indépendant des trois premiers, concerne l'analyse des caractéristiques socio-économiques des hommes qui ont une préférence à épouser des jeunes filles de moins de 18 ans. Dans ce contexte, nous considérons les données de l'EDS 2006 du Niger et utilisons les copules Archimédiennes bidimentionelles pour modéliser l'association entre le niveau d'éducation (variable discrète) des hommes et leur revenu pré-marital (variable continue). Nous construisons la vraisemblance pour un échantillon issu de ce couple de variables aléatoires mixtes, et déduisons une estimation du paramètre de dépendance en utilisant une procédure semi-paramétrique où les marges sont estimées par leurs équivalents empiriques. Nous utilisons la méthode du jackknife pour estimer l'erreur type. Nous utilisons la méthode de Wald pour tester l'égalité entre l'association des caractéristiques socio-économiques des hommes qui épousent des jeunes filles mineures et celle des hommes qui se marient avec des femmes âgées. Les résultats du test contribuent à la validité de notre théorie selon laquelle les hommes qui épousent

des jeunes filles de moins de 18 ans ont un niveau d'éducation et un revenu pré-marital faibles, lorsqu'on les compare aux hommes qui ne le font pas.

Abstract

This thesis introduces and explores a new class of probability models for exchangeable clustered binary data. In these models, the conditional probability of success is characterized by a function of the marginal probability of success and a positive cluster-specific random effect. The marginal probabilities are modeled using the logit and complementary log-log link functions. The distribution of the random effect contains an association parameter that is estimated to give a measure of the strength of the within-cluster residual dependence that is not accounted for by the margins. We show that the random effect distributions can be related to exchangeable Archimedean copula models, thus giving new insights on such models.

The copula approach offers many advantages. Indeed, the family of Archimedean copulas provides a large class of models for over-dispersion in a Bernoulli experiment. From a statistical perspective, the marginal likelihood function for the sample data has an explicit expression, the maximum likelihood methods are then easy to implement and computationally straightforward.

Based on the proposed models, four applications are considered. First, we investigate the construction of profile likelihood confidence interval (PLCI) for the intra-cluster correlation coefficient (ICC). The second application is concerned with an heterogeneity in capture probabilities in a mark-recapture study for estimating the size of a closed population. The third contribution deals with the estimation in small areas, the fourth and final, independent of the other three, analyzes the socioeconomic characteristics of men who prefer to marry girls under 18 years old.

In the first application, we consider a simple case, without covariates and construct maximum likelihood inference procedures for the intra-cluster correlation using several specifications of Archimedean copulas. The selection of a particular model is carried out using the Akaike information criterion (AIC). Profile likelihood confidence intervals for the ICC are constructed and their performance are assessed in a simulation experiment. The sensitivity of the inference to the specification of the copula family is also investigated through simulations. Numerical examples are presented. We compare our approach with that proposed under the Beta-binomial model and with the modified Wald interval method proposed by Zou and Donner (2004). One of the important findings of these studies is that the profile confidence interval obtained under our models presents nice properties, even when the number of clusters is small. Model selection is an important step: if the model is poorly specified, then this could lead to erroneous results.

The second application, an extension of the first one to accommodate cluster level covariates, is concerned with an heterogeneity in capture probabilities in a capture-recapture study for estimating the size of a closed population. Unit level covariates are recorded on the units that are captured and copulas are used to model the residual heterogeneity that is not accounted for by covariates. Several models for the unobserved heterogeneity are available and the marginal capture probability is expressed using the Logit and the complementary Log-Log link functions. The parameters are estimated using a conditional likelihood constructed with the data obtained on the units caught at least once. The population size is estimated using a Horvitz-Thompson estimator constructed using the estimated probabilities that a unit is caught at least once. This generalizes the model of Huggins (1991) that does not account for a residual heterogeneity. The sensitivity of the inference to the specification of a model is also investigated through simulations. A numerical example is presented.

The third application uses the models of the first two in order to estimate small area proportions. We apply Bayes techniques using a new class of probability models, to estimate small area proportions. The Bayesian inference under the proposed models consists in obtaining the posterior distribution of the random effect and its Laplace transform. This posterior Laplace transform is then used to find Bayes estimates of small area proportions. We develop a comparison between the Best Predictor (BP) and the Best Linear Unbiased Predictor (BLUP). The model parameters are estimated using the maximum likelihood (ML) method. Under the proposed model, the likelihood function and the best predictor (BP) of small area proportion have closed form expressions. Model parameters are replaced by their ML estimates in the BP to obtain the empirical best predictor (EBP). We use the Akaike information criterion (AIC) for selecting a particular model. We propose the jackknife method to estimate the mean square error of the empirical Bayes predictor. Empirical results obtained from simulated and real data are also presented.

The fourth and last problem addressed in this thesis, independently of the others three, investigates socio-economic characteristics of men who prefer to marry girls under 18 years. We consider the data from the 2006 DHS Niger and use a bivariate Archimedean copula to model the association between education level (discrete) of men and their pre-marital income (continuous). We present the likelihood function for a sample from this pair of mixed random variables, and derive an estimate of the dependence parameter using a semi-parametric procedure where margins are estimated by their empirical equivalents. We use the jackknife method to estimate the standard error. We use a Wald-type procedure, to perform a parametric hypothesis test of equality between the association of the socio economic characteristics of men who marry underage girls and that of men who marry older women instead. These test results contribute to the validity of our theory that men who marry girls under 18 years old have a low level of education and income pre-marital, when compared to men who did not.

Table des matières

Liste des tableaux

Liste des figures

Cette thèse est dédiée à ma soeur Awa Tounkara et à ma tente Penda Coulibaly qui sont décédées à mon absence du pays, à mes parents, Fily Tounkara et Fanta Diouara, à ma tente Moussoukoro Kéita, à ma femme, Rakiatou Souley Harouna, à ma fille, Naïma Nabintou Tounkara, et à tous mes frères et soeurs.

Remerciements

La présente thèse à été réalisée que grâce à la Miséricorde d'Allah, et la collaboration de plusieurs personnes que je tiens à remercier.

En premier lieu, je remercie mes deux géniteurs, mon père et ma mère, pour leurs soutien et accompagnement, ainsi que leurs prières durant tout le long de mes études.

Ensuite, j'exprime mes profonds remerciements à mon superviseur de thèse, le professeur Louis-Paul Rivest, pour son aide précieuse et son soutien qu'il m'a apportés durant cette épreuve qu'est le doctorat. Je le remercie pour m'avoir attribuer une bourse d'étude de troisième cycle qui était pour moi indispensable pour la réalisation de cette thèse. Je dois également lui être reconnaissant pour sa patience, mais surtout sa grande sagesse, qui m'ont été d'un grand support face aux épreuves du doctorat. Son oeil critique et ses judicieuses suggestions m'ont été très précieux pour structurer et améliorer la qualité des différents chapitres de cette thèse. Ce fut un réel plaisir, sans aucun regres, de l'avoir eu comme directeur de thèse.

Je remercie ma chère épouse pour sa patience, sa compréhension et son soutien quotidien indéfectible.

Ma gratitude s'adresse aussi aux membres du jury, Dr Taoufik Bouezmarni (Département de mathématiques, Université de Sherbrooke), Dr Lajmi Lakhal Chaieb et Dr Anne-Sophie Charest (Département de mathématiques et de statistiques, Université Laval). L'examen d'une thèse est loin d'être une tâche facile, donc je suis reconnaissant pour leurs commentaires sérieux et détaillés qui ont contribué à améliorer cette thèse.

Je suis particulièment reconnaissant au Département de mathématiques et de statistique de la faculté des science et génie de l'Université Laval, pour son généreux soutien financier à travers l'assistanat d'études supérieures et des bourses de recherche de troisième cycle. J'ai eû la chance et l'honneur d'enseigner des cours dans ce département, et je tiens à remercier le professeur Frédéric Gourdeau, directeur de ce département, pour m'avoir confier des charges qui ont beaucoup contribuées à améliorer mon expérience dans ce domaine. Merci également aux agents de la gestion des études, en particulier, Madame Sylvie Drolet pour leur disponibilté à répondre à toute question d'ordre administratif.

Je tiens également à exprimer mes sincères remerciements et ma gratitude aux personnes suivantes qui m'ont aidé, ont travaillé avec moi, et ont partagé des idées inestimables avec moi pendant mes études à l'Université Laval :

Julie et Pierre-Louis, des amis québécois que j'ai connus en France, et par le biais de qui, j'ai eu la motivation de venir effectuer mon stage de fin d'étude de master à Québec.

Dr Belkacem Abdous, pour avoir accepter ma candidature aux études de doctorat. Je le remercie également pour son soutien financier d'une session au cours de mes études de doctorat.

Sophie Baillargeon, l'assistante de mon superviseur, pour tout son soutien à la réalisation des mes programmes R. Son assistance en programation a été crutiale et très appréciée.

Les consultants professionnels du Service de Consultation Statistique (SCS), Hélène Crépeau et Gaétan Daigle, pour m'avoir permis d'effectuer deux stages au SCS où j'ai vécu sous leur supervision une expérience enrichissante.

Mes amis et compatriodes sénégalais, Dr. Ibrahima Dione et Dr. Adama Souleymane Kamara, Madame Sénaba Sambe, Monsieur Moussa Diéne, et Monsieur Malick Diouf qui m'ont beaucoup encouragé à finir ce travail par des gestes d'amitié dont je suis reconnaissant. Je leurs souhaite une carrière professionnelle plein de succès.

Mr Mouhamed Ali Cissé et sa femme Nabintou Touré, pour tous les soutiens et services qu'ils m'ont apportés durant ces trois dernières années. Leurs conseils m'ont été très utils, je leurs exprime toute ma reconaissance pour ces gestes, et je souhaite à leurs deux aimables enfants Lugman et Ayemane beaucoup de succès dans leurs carrières.

Enfin, je suis redevable à mes frères Salifou Tounkara, Mamadou Diaby Tounkara, et mon oncle Souleymane Fofana pour tout le soutien qu'ils m'ont apporté, afin que je puisse aller en France poursuivre mes études. Je les remercie pour ces gestes, et leurs souhaite longue vie et plein de succès dans leurs carrières.

Avant-propos

Le présent document s'inscrit dans le cadre de l'obtention du diplôme de doctorat en mathématique (avec concentration statistique) à l'Université Laval à Québec.

Dans de nombreuses études statistiques, les unités échantillonnées sont des groupes de sujets, par exemple, les membres d'une même famille, les patients d'un même centre de santé, les élèves d'une même école ou les personnes d'une même communauté. Dans cette structure, les unités à l'intérieur d'un même groupe ont tendance à fournir des réponses similaires quand on les compare aux unités de groupes différents, ce qui entraine ainsi une corrélation entre les observations à l'intérieur d'un même groupe. Ce type de données est appelé des données en forme de grappes (ou en anglais clustered data). Les méthodes statistiques standards telles que le modèle linéaire généralisé qui ignorent une telle corrélation sont inadéquates pour analyser ce type de données et peuvent conduire à des conclusions erronées. Dans cette thèse, nous présentons une nouvelle classe de modèles basée sur la famille de copules Archimédiennes multidimensionnelles pour analyser des données de Bernoulli corrélées en forme de grappe. Les applications de ces modèles ont fait l'objet de quatre articles dont les deux premiers sont publiés dans des revues scientifiques, le troisième soumis, et le quatrième prêt à être soumis. Le premier article, présenté au chapitre II, a pour titre **Some new random effect models for correlated binary responses**. Cet article, écrit en collaboration avec Louis-Paul Rivest, mon directeur de recherche, a été publié à la revue *Dependence Modeling*, en 2014. Le deuxième article est présenté au Chapitre III. Son titre s'intitule **Mixture regression models for closed population capture-recapture data**, et ses auteurs sont Fodé Tounkara et Louis-Paul Rivest. Cet article a été publié dans *Biometrics*, en mai 2015. Au chapitre IV, nous présentons le troisième article, dont le titre est **Empirical Bayes predictions of small area proportions with Archimedean copulas**, et dont les auteurs sont les mêmes que ceux des deux premiers articles. Cet article va être soumis à la revue *Journal of Multivariate Analysis (JMVA)*. Le quatrième article, présenté au chapitre V, et dont le titre s'intitule **On the socioeconomic characteristics of men who marry underage girls** est soumis à la revue *The Bank World*. Ce dernier article est écrit en collabaration avec Sylvain Dessy, professeur au département d'économie de l'Université Laval, et son étudiante au doctorat Setou Mamadou Diarra.

Introduction

Motivation

Le modèle linéaire généralisé (GLM) est souvent utilisé pour analyser des données statistiques ayant une structure de grappe. Bien que très flexible et apte à modéliser plusieurs types de variables réponses, le GLM est basé sur l'hypothèse d'indépendance des observations, ce qui le rend inapproprié dans de nombreuses situations. En effet, dans plusieurs études pratiques, il existe des groupes d'observations où les variables réponses sont probablement corrélées entre elles. Dans certains cas, les observations sont collectées sur les mêmes individus à travers le temps. Par exemple, en épidémiologie il est fréquent que l'on veuille tester l'effet d'un traitement où son efficacité sur des sujets, ou bien voir les facteurs de risque pour des maladies comme par exemple, le cancer et les maladies cardiovasculaires. Dans ces études, les participants sont suivis pendant plusieurs semaines, plusieurs mois voire même plusieurs années, et les mesures collectées sont souvent corrélées. Ces types de données sont appelés des données longitudinales. Dans d'autres applications, les unités échantillonnées sont des groupes de sujets, par exemple, les membres d'une même famille, ou les patients d'un même centre de santé. La particularité de cette structure est que les unités à l'intérieur d'un même groupe sont susceptibles d'être plus proches et par conséquent de fournir des réponses similaires quand on les compare aux unités dans des groupes différents. Ce phénomène connu sous le nom de "l'effet de groupe", entraine ainsi une corrélation entre les observations à l'intérieur d'un même groupe. Ce type de données est appelé des données en forme de grappes (ou en anglais clustered data). Les méthodes statistiques standards telles que le GLM qui ignorent une telle corrélation fournissent un pauvre ajustement des données et peuvent conduire à des conclusions erronées.

Dans cette thèse, nous introduisons et explorons une nouvelle classe de modèles pour des données de Bernoulli corrélées échangeables en forme de grappe. L'hypothèse d'échangeabilité est raisonable dans de nombreuses applications notamment en tératologie animale, en toxicologie et en biologie. Pour rappel, un vecteur aléatoire est dit échangeable si sa loi jointe est invariante par permutation des composantes. Notre approche consiste à supposer que la corrélation intra-grappe est induite par un effet aléatoire associé aux unités à l'intérieur de la même grappe. Cet effet aléatoire peut être vu comme une combinaison de tous les facteurs aussi bien génétiques qu'environnementaux qui sont liés à la spécificité de la grappe. Nous admettons que cet effet aléatoire est non observé et provient d'une distribution de probabilité connue à support positif. De plus conditionnellement aux effets aléatoires, les observations à l'intérieur de chaque grappe sont supposées indépendantes et suivent la loi de Bernoulli. Les modèles sont enfin contruits en exprimant le paramètre de

Bernoulli comme une fonction de cette variable aléatoire positive et de sa transformée de Laplace.

Ces modèles contiennent des paramètres associés aux probabilités marginales de succès, et un paramètre de dépendance lié à la distribution de l'effet aléatoire. Le paramètre de dépendance est estimé afin de fournir une mesure du degré de dépendance résiduelle qui n'est pas prise en compte par les marges. Les paramètres sont estimés par la méthode du maximum de vraisemblance en utilisant la fonction de vraisemblance marginale. Cette fonction est obtenue en intégrant (ou en sommant) la vraisemblance conditionnelle des données de l'échantillon par rapport à l'effet aléatoire.

En utilisant la théorie des copules (Nelson, 2006), on montre par construction de la fonction de répartition conjointe ou de la fonction de survie conjointe des données de Bernoulli que la classe de modèles que nous proposons est associée à la famille de copules Archimédiennes (Mai and Scherer, 2012), ce qui nous permet d'avoir un nouvel aperçu de nos modèles.

Les copules sont des outils qui sont populaires pour la modélisation de distributions multivariées. Une fonction de copule modélise seulement la dépendance entre les variables dans une distribution multivariée, et elle peut être combinée avec tout un ensemble de distributions univariées pour les distributions marginales. Elles permettent aux distributions conjointes des observations d'avoir un paramètre qui mesure le niveau de dépendance entre les observations. L'une des propriétés dont jouissent également les modèles de copules est cette capacité de séparer l'effet des marges de celui de la dépendance.

Nous comptons plusieurs classes paramétriques de copules dépendant de la façon dont le paramètre de copule modélise la dépendance. Pour plus de détails concernant ces classes de copules, nous suggèrons de consulter les livres de (Joe, 1997; Nelson, 2006) et plus récemment de (Mai and Scherer, 2012; Joe, 2015).

Dans cette thèse nous nous sommes intéressés particulièrement à la classe des copules Archimédiennes. Cette classe de copules permet de modéliser une structure de dépendance échangeable : la valeur de la fonction de copule reste inchangée par permutation de ses composantes. Les copules Archimédiennes sont une classe de copules qui est facile à construire et qui jouit de très belles propriétés mathématiques.

L'utilisation des modèles de copules Archimédiennes, comme une formulation alternative des modèles à effet aléatoire, va donc nous permettre d'utiliser leurs propriétés, mais aussi de profiter de la grande variété de modèles qui sont disponibles pour tenir compte de la structure de dépendance entre les observations. Dans cette thèse nous considérons quatre familles paramétriques de copules Archimédiennes pour la surdispersion, les copules Gumbel, Clayton, Frank et Joe.

Les modèles que nous présentons dans cette thèse pour modéliser des variables de Bernoulli corrélées, constituent une alternative au modèle Beta-Binomial (BB) (Williams, 1975; Prentice, 1986; Williams, 1988), et aux modèles linéaires généralisés mixtes (GLMM) en présence de variables auxiliaires (Williams, 1982; Ochi and Prentice, 1984; McCulloch and Searle, 2001; McCulloch et al., 2008). Cette approche peut également être vue comme une alternative à la méthode des estimations d'équation généralisé (GEE) (Liang and Zeger, 1986) lorsque la dépendance est échangeable (symétrique). En présence de covariables, nous utilisons les fonctions de lien Logit ou Log-Log complémentaire pour exprimer la probabilité marginale de succès (ou échec) en fonction du prédicteur linéaire correspondant.

Du point de vue de l'inférence statistique, l'approche que nous proposons offre de nombreux avantages. On peut montrer que lorsque la transformée de Laplace de la distribution de l'effet aléatoire et son inverse ont des expressions explicites, alors les méthodes du maximum de vraisemblance sont faciles à mettre en œuvre. En effet, dans ce cas la fonction de vraisemblance marginale du modèle a une forme explicite et donc facile à implémenter. Les modèles que nous considérons dans cette thèse satisfont à cette exigence. La section suivante présente une revue bibliographique des méthodes traditionnelles pour des données de Bernoulli corrélées.

Revue de la littérature

Considérons K le nombre de grappes et n_i le nombre d'individus dans la $i^{\text{ième}}$ grappe, pour $i = 1,\ldots,K$. Soit Y_{ij}, $i = 1,\ldots,K$, $j = 1,\ldots,n_i$, une séquence d'observations mesurées dans la i^{me} grappe. À chaque Y_{ij}, on associe $x_{ij1},\ldots,x_{ijq}$, q covariables mesurées sur le $j^{\text{ième}}$ individu de la $i^{\text{ième}}$ grappe.

Les GLMs sont décrits par trois éléments fondamentaux : la composante aléatoire, la composante systématique, et la fonction de lien. La composante aléatoire est la distribution de la variable réponse. Ici on suppose que Y_{ij} est dichotomique avec comme valeurs possibles, succès et échec codées respectivement par un (1) et zéro (0). Le deuxième élément est le prédicteur linéaire η_{ij} associé à l'individu j dans la $i^{\text{ième}}$ grappe défini par

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij1} + \ldots + \beta_q x_{ijq}, \qquad i = 1,\ldots,K; \qquad j = 1,\ldots,n_i, \tag{1}$$

où $\beta = (\beta_0,\beta_1,\ldots,\beta_q)^T$ est le vecteur de dimension $(q+1)\times 1$ des coefficients de régression pour les covariables. La fonction de lien $g(\cdot)$ établit une relation entre la moyenne de la variable réponse, $\pi_{ij} = E(Y_{ij})$, et le prédicteur linéaire

$$g(\pi_{ij}) = \eta_{ij}, \qquad i = 1,\ldots,K; \qquad j = 1,\ldots,n_i. \tag{2}$$

Pour des variables réponses binaires, il existe plusieurs fonctions de lien, mais la plus populaire est la fonction de lien Logit :

$$g(\pi) = logit(\pi) = log(\frac{\pi}{1-\pi}). \tag{3}$$

Les modèles GLMs avec cette fonction de lien sont appelés modèles de régression logistique. L'inverse de la fonction de répartition de la loi norme standard, Φ, est aussi un choix populaire

$$g(\pi) = \Phi^{-1}(\pi). \tag{4}$$

Les modèles GLMs avec cette fonction de lien sont appelés modèles probit. La fonction Log-Log complémentaire peut aussi être utilisée :

$$g(\pi) = \log\{-\log(1-\pi)\}. \tag{5}$$

Cette fonction de lien est asymétrique, et peut conduire à des résultats différents par rapport aux fonctions de lien probit et Logit.

Si on utilise les notations matricielles et vectorielles, le GLM peut s'écrire comme suit :

$$Y_i \sim Bernoulli(\pi_i) \qquad i = 1, \ldots, K, \tag{6}$$

$$g(\pi_i) = X_i\beta, \qquad i = 1, \ldots, K, \tag{7}$$

où $Y_i = (Y_{i1}, \ldots, Y_{in_i})^T$ dénote le vecteur des variables réponses, $\pi_i = (\pi_{i1}, \ldots, \pi_{in_i})^T$ le vecteur des moyennes et

$$X_i = \begin{pmatrix} x_{i11} & x_{i12} & \cdots & x_{i1q} \\ x_{i21} & x_{i22} & \cdots & x_{i2q} \\ \vdots & & \ddots & 0 \\ x_{in_i1} & x_{in_i2} & \cdots & x_{in_iq} \end{pmatrix}$$

dénote la matrice de dimension $n_i \times q$ des covariables pour la $i^{\text{ième}}$ grappe.

Le GLM suppose que les variables réponses $Y_{i1}, \ldots, Y_{in_i}$ sont indépendantes. Cependant, dans de nombreuses situations les observations à l'intérieur des grappes ont tendance à être corrélées. Dans ce cas, l'hypothèse d'indépendance ne sera pas raisonnable et le modèle GLM pourrait mener à une sous-estimation des variances des estimateurs des paramètres.

Plusieurs méthodes ont été développées pour tenir compte de cette corrélation lorsqu'on veut faire de l'inférence sur les coefficients de régression β d'un modèle linéaire généralisé. On distingue deux types d'approche : l'approche marginale et l'approche conditionnelle.

L'approche conditionnelle suppose la présence d'un effet aléatoire spécifique à chaque grappe et modélise la probabilité de succès conditionnellement à cet effet aléatoire. C'est cet effet aléatoire qui permet de tenir compte de la variabilité entre les grappes, et donc de la dépendance entre les observations à l'intérieur d'une même grappe. De plus conditionnellement aux effets aléatoires, les observations à l'intérieur de chaque grappe sont supposées indépendantes et suivent la loi de Bernoulli. Dans une approche marginale, on ne modélise pas directement la corrélation, mais on en tient compte en corrigeant l'estimation des matrices de variances covariances des estimateurs.

Ci-dessous, nous présentons une description de la méthode des équations d'estimations généralisées (GEE), du modèle Béta-Binomial (BB) et du modèle linéaire généralisé avec ordonnée à l'origine aléatoire qui sont souvent utilisés pour traiter la corrélation présente dans les données de Bernoulli.

Equations d'estimations généralisées

Dans cette partie, nous donnons une description générale de l'approche marginale pour des données de Bernoulli corrélées. Nous discutons de la méthode GEE qui est la méthode d'estimation d'un modèle marginal la plus populaire.

Considérons le modèle marginal défini par les deux composantes suivantes :

(i) La moyenne marginale de la variable réponse, $E\left(Y_{ij}|x_{ij}\right) = \pi_{ij}$, dépend des variables auxiliaires x_{ij}, à travers une fonction de lien $g(\cdot)$ connue. Cette composante est tout simplement le modèle GLM défini par les équations (6) et (7).

(ii) Une spécification d'une structure de corrélation pour toutes les paires d'observations à l'intérieur de chaque grappe. La corrélation entre toute paire d'observations Y_{ij} et Y_{ik}, pour $j \neq k$ dépend d'un certain paramètre que l'on note par ρ.

Comme précédemment, on suppose que la variable réponse suit une distribution de Bernoulli de moyenne $\pi_{ij} = g^{-1}(\eta_{ij})$ et de variance $V\left(Y_{ij}|x_{ij}\right) = \pi_{ij}(1-\pi_{ij})$.

On distingue plusieurs formes pour la structure de corrélation. Les choix les plus populaires sont :

• Indépendance : Ici la corrélation entre toutes les paires d'individus est nulle,
$\text{corr}(Y_{ij}, Y_{ik}) = 0$, ce qui revient à ajuster le GLM.

• Echangeable (Exchangeable) : Connue aussi sous le nom de composante symétrique, cette structure suppose la même corrélation pour chaque paire d'individu à l'intérieur d'une même grappe, $\text{corr}(Y_{ij}, Y_{ik}) = \rho$. Cette structure est très appropriée lorsqu'on ne sait a priori que toutes les différentes paires d'observations ont la même corrélation. Elle est équivalent à la structure de corrélation intra-groupe supposée dans un modèle à effet aléatoire.

Pour toutes ces structures de corrélation, il est important de noter que la corrélation intra-grappe est supposée être la même pour toutes les grappes.

Les modèles marginaux tels que décrits plus haut sont souvent ajustés en utilisant la méthode des Estimations d'Equations Généralisées (GEE) proposée par (Liang and Zeger, 1986). La méthode GEE est une méthode de quasi-vraisemblance qui dépend seulement du premier moment et de la façon dont la variance de la variable réponse dépend de la moyenne. Soient A_i une matrice diagonale de dimension $n_i \times n_i$ et dont les élèments diagonaux sont $Var(Y_{ij}) = \pi_{ij}(1-\pi_{ij})$, et $R_i(\rho)$ la matrice des corrélations marginales pour le vecteur de variables réponses Y_i qui dépend d'un paramètre ρ, et souvent nommée matrice de corrélation de travail où "Working correlation matrix" en anglais. Alors, la méthode GEE estime les paramètres de régression β en résolvant les " équations d'estimations généralisées" suivantes :

$$\text{U}(\beta) = \sum_{i=1}^{K} D_i^T \text{V}_i^{-1} S_i = 0, \tag{8}$$

où $D_i = \partial\pi_i/\partial\beta$, $S_i = Y_i - \pi_i$, $\pi_i = (\pi_{i1}, \ldots, \pi_{in_i})^T$ et V_i est la matrice de variance-covariance pour les Y_i définie par $\text{V}_i = \phi A_i^{1/2} R_i(\rho) A_i^{1/2}$, où ϕ est le paramètre de dispersion.

Les paramètres β et la corrélation intra-grappe ρ peuvent être estimés en utilisant une procédure d'estimation itérative à deux étapes : (i) Etant donné une valeur courante de l'estimation de ρ et de ϕ, une estimation de β est obtenue comme solution de l'équation (8). (ii) Etant donné une valeur courante de l'estimation de β, une estimation des paramètres ρ et ϕ est obtenue à partir des résidus standardisés suivant : $r_{ij} = (Y_{ij} - \hat{\pi}_{ij})/\sqrt{\upsilon(\hat{\pi}_{ij})}$. Liang and Zeger (1986) ont montré que $\hat{\beta}$, la solution de l'équation (8), est un estimateur consistent de β, et ont proposé un estimateur de sa matrice de covariance basé sur les moments et qui est défini par

$$\mathrm{V}_T = \left(\sum_{i=1}^{K} \Big(\frac{\partial \pi_i}{\partial \eta_i}\Big)^T \mathrm{V}_i^{-1} \Big(\frac{\partial \pi_i}{\partial \eta_i}\Big) \right)^T \Bigg|_{\beta=\hat{\beta},\rho=\hat{\rho},\phi=\hat{\phi}},$$

où, $\hat{\rho}$ et $\hat{\phi}$ sont respectivement les estimateurs de ρ et de ϕ.

Le calcul de l'estimateur V_T suppose que la structure de travail $R_i(\rho)$ est la vraie matrice de corrélation pour les observations Y_i, ce qui n'est pas forcément vrai. Alors, on estime la matrice de covariance des $\hat{\beta}$ par un estimateur de variance sandwich robuste :

$$\widehat{V}ar(\hat{\beta}) = \mathrm{V}_T \left(\sum_{i=1}^{K} D_i^T \mathrm{V}_i^{-1} S_i S_i^T \mathrm{V}_i^{-1} D_i \right) \Bigg|_{\beta=\hat{\beta},\rho=\hat{\rho},\phi=\hat{\phi}} \mathrm{V}_T.$$

On l'appelle estimateur sandwich, du fait que l'estimation de la matrice de corrélation est "prise en sandwich" entre deux expressions similaires de matrices algébriques. Pour plus de détails concernant les estimateurs robustes, voir Liang and Zeger (1986), et McCulloch et al. (2008).

L'avantage des GEE est que c'est une méthode facile à implémenter. De plus, les estimations des coefficients de régression et leurs estimations de variances sont robustes et ne sont pas sensibles au changement de structure de corrélation, donc le choix de la structure de corrélation de travail n'est pas crucial. Les coefficients de régression sont interprétés comme étant l'effet d'un changement dans les variables explicatives sur la valeur moyenne de la variable réponse dans l'ensemble de la population, donc on parle aussi d'une approche "moyennée sur la population" (population-averaged). Cependant, la méthode GEE présente quelques désavantages. Premièrement, lorsqu'on utilise cette approche la distribution conjointe des observations n'est pas spécifiée et donc la fonction de vraisemblance n'est pas disponible. Par conséquent, la fonction de déviance, les critères AIC et BIC, et le test de rapport de vraisemblance qui dépendent de la vraisemblance maximisée ne peuvent être utilisés pour faire des inférences. Dans cette approche, la dépendance entre les observations à l'intérieur d'une même grappe est traitée comme un paramètre de nuisance et il n'existe aucun paramètre qui représente la variance inter-grappe, et donc on ne peut pas estimer la corrélation intra-grappe, l'hétérogéneité et l'effet grappe.

Modèle Béta-Binomial

En l'absence de covariables, le modèle Béta-Binomial (BB) est sans nul doute le modèle à effet aléatoire le plus utilisé pour analyser des données de Bernoulli corrélées.

Afin de tenir compte de la corrélation intra-grappe, ce modèle suppose que la probabilité de succès varie d'une grappe à une autre et est identique pour tous les individus d'une même grappe, et nous notons par p_i la probabilité de succès pour la i^{me} grappe. De plus les observations Y_{ij} sont conditionnellement indépendantes avec probabilité $P(Y_{ij}=1|p_i)=p_i$. Soit $Y_i=\sum Y_{ij}$ le nombre total de succès enregistrés dans la grappe i. Alors la distribution conditionnelle des Y_i pour $i=1,\ldots,K$ est binomiale, $Y_i|p_i$ ∼Binomial (p_i). On suppose enfin que les p_i sont indépendantes et indentiquement distribuées selon la loi Beta(a,b), de moyenne $E(p_i)=a/(a+b)$ et de variance $Var(p_i)=ab/\left[(a+b)^2(a+b+1)\right]$. Alors, la distribution inconditionnelle de Y_i est la distribution connue sous le nom de Beta-Binomial et est donnée par

$$P(Y_i=y_i)=\frac{B(a+y_i,b+n_i-y_i)}{B(a,b)},\qquad y_i=0,\ldots,n_i, \tag{9}$$

où $B(\cdot,\cdot)$ est la fonction Beta définie par

$$B(a,b)=\int_0^{+\infty}x^{a-1}(1-x)^{b-1}dx.$$

En utilisant ce modèle, on peut facilement calculer les moments des Y_{ij} en utilisant la méthode des espérances et des variances imbriquées. Ainsi, nous avons pour la moyenne $E(y_{ij})=a/(a+b)$ et pour la variance $Var(y_{ij})=ab/(a+b)^2$. En utilisant les résultats tels que

$$B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$$

et

$$\frac{\Gamma(s+h)}{\Gamma(s)}=(s+h-1)(s+h-2)\cdots(s+1)s=\prod_{t=0}^{h-1}(s+t),$$

on peut exprimer la probabilité définie en (9) en terme de produit, ce qui rend l'expression de la log-vraisemblance du modèle plus simple, voir (McCulloch et al., 2008) pour plus de détails.

Ce modèle est souvent utilisé pour estimer le coefficient de corrélation intra-grappe (ICC). Ce coefficient, que l'on note par ρ, mesure la corrélation entre toute paire d'individus à l'intérieur d'une même grappe. Pour deux individus j et l de la grappe i, tels que $j\neq l$, l'expression de ρ sous le modèle BB est donnée par

$$\rho=corr(Y_{ij},Y_{il})=\frac{1}{a+b+1},$$

où $corr(\cdot,\cdot)$ dénote la corrélation. Pour estimer ρ, on peut paramétriser la distribution des p_i en fonction de ρ et de la moyenne $\pi=a/(a+b)$, i.e $p_i\sim Beta(\pi,\rho)$, comme certains auteurs (McCulloch et al., 2008; Ahmed and Shoukri, 2010) l'ont suggèré. Une valeur de ρ positive signifie une similarité des réponses au sein d'une même grappe. Un exemple de corrélation positive est rencontré dans les expériences toxicologiques où l'on cherche à étudier les effets tératogènes des composés chimiques sur les animaux. En effet les foetus dans une portée ont tendance à réagir de façon plus similaire que les foetus de différentes portées, ce phénomène est connu sous le nom d'effet de la portée. En revanche une valeur de ρ négative traduit que les observations dans un groupe sont corrélées négativement. Ce dernier cas, moins fréquent en pratique, est impossible avec le modèle Béta-binomial. Cela se produit par exemple dans les études familiales lorsque les enfants peuvent être en compétition pour les soins maternels.

Sous le modèle Beta-Binomial, on peut montrer facilement que la variance des Y_i peut s'exprimer en fonction de ρ et de π comme suit $Var(Y_i)=n_i\pi(1-\pi)\{1+(n_i-1)\rho\}$. Puisque $\rho>0$, alors cette variance est supérieure à $n_i\pi(1-\pi)$ la variance de Y_i sous le modèle binomial, où l'on suppose l'indépendance des observations, on dit alors qu'il y a sur-dispersion.

On peut aussi utiliser le modèle Beta-Binomiale pour prédire la probabilité de succès conditionnelle, p_i. En effet, on peut déduire de la distribution marginale des observations et de la distribution a priori de p_i, que la distribution a posteriori de p_i est aussi une distribution Beta, $p_i|y_i,a,b \sim \text{Beta}(y_i+a, n_i-y_i+b)$. Donc, un meilleur prédicteur empirique pour p_i est donnée par

$$\begin{aligned}\hat{p}_i &= E(p_i|y_i,\hat{a},\hat{b}) \\ &= \frac{\hat{a}+y_i}{\hat{a}+\hat{b}+n_i} \\ &= \hat{\pi}\left(\frac{\hat{a}+\hat{b}}{\hat{a}+\hat{b}+n_i}\right)+\hat{y}_i\left(\frac{n_i}{\hat{a}+\hat{b}+n_i}\right),\end{aligned} \tag{10}$$

où $\hat{a}$, $\hat{b}$, $\hat{\pi}$ sont les estimations (obtenues soit par la méthode des moments où la méthode du maximum de vraisemblance) respectives des paramètres a, b et π, et $\hat{y}_i$ est la moyenne échantillonalle de Y dans la grappe i. Cet estimateur est équivalent au meilleur predicteur linéaire sans biais (BLUP). On remarque que cet estimateur est une combinaison linéaire d'une estimation de la moyenne de la population $\hat{\pi}$ et de l'estimateur direct $\hat{y}_i$. On dit que cet estimateur est ce qu'on appelle un "shrinkage estimator". Le lecteur intéressé à en savoir plus sur le modèle Beta-Binomial pourra consulter le livre de (McCulloch et al., 2008, Chapter 2).

Modèles linéaires généralisés avec une ordonnée à l'origine aléatoire

Les modèles linéaires généralisés avec une ordonnée à l'origine aléatoire sont une extension des GLMs où l'on introduit des effets aléatoires. Soit a_i l'effet aléatoire associé à la grappe i, donc pour une variable réponse Y_{ij}, et un vecteur de variables auxiliaires x_{ij} associés à l'individu j de la grappe i. Pour Y_{ij} binaire, le modèle linéaire généralisé avec effet aléatoire peut être défini par les deux composantes suivantes :

(i) Etant donné a_i, la distribution conditionnelle de Y_{ij} est Bernoulli de paramètre $p_{ij} = Pr(Y_{ij}=1|a_i)$, avec moyenne conditionnelle

$$g\{E(y_{ij}|x_{ij},a_i)\} = g(p_{ij}) = x_{ij}\beta + a_i,$$

où g est une fonction de lien connue.

(ii) On suppose que les effets aléatoires a_i sont indépendants identiquement distribués selon une distribution avec densité connue $f_a(;\alpha)$.

Il faut noter qu'il y a une hypothèse additionnelle d'indépendance conditionnelle, c'est à dire, étant donné a_i, les observations $Y_{i1},\ldots,Y_{in_i}$ sont mutuellement indépendantes. Le modèle logistique avec effet aléatoire défini par :

$$\log(\frac{p_{ij}}{1-p_{ij}}) = x_{ij}\beta + a_i, \quad a_i \sim N(0,\sigma^2)$$

est sans nul doute l'exemple de modèle linéaire généralisé avec effet aléatoire le plus populaire pour des données de Bernoulli. La méthode du maximum de vraisemblance est souvent utilisée pour faire de l'inférence sur les paramètres. La procédure d'estimation par la méthode du maximum de vraisemblance peut se résumer en ces deux étapes :

(i) L'estimation du maximum de vraisemblance des paramètres est basée sur la fonction de vraisemblance marginale obtenue en intégrant (où en sommant) la vraisemblance des données par rapport à la distribution des effets aléatoires.

$$\prod_{i=1}^{K}\int_{-\infty}^{-\infty} f(Y_i|X_i,a_i)f_a(a_i;\alpha)da_i,$$

où $X_i = (x_{i1},\ldots,x_{in_i})$.

(ii) Etant donné une estimation des paramètres β et α, un prédicteur empirique pour l'effet aléatoire a_i est obtenu en prenant l'espérence de la distribution a posteriori de a_i :

$$\hat{a}_i = E\big(a_i|Y_i,\hat{\beta},\hat{\alpha}\big).$$

Noter que la fonction de vraisemblance marginale ainsi que l'expression du prédicteur empirique $\hat{a}_i$ n'ont pas en général une expression analytique simple, et que les méthodes numériques ou les techniques de Monte Carlo sont souvent nécessaires pour les évaluer. Les modèles linéaires généralisés avec effets aléatoires sont des modèles conditionnels, où l'objectif est de faire de l'inférence sur les individus et de voir l'influence des covariables sur les individus. Dans les modèles linéaires généralisés avec effets aléatoires, les paramètres de régression mesurent l'effet des variables auxiliaires sur les individus. Ils ont un certain nombre d'avantages potentiels par rapport aux modèles GEEs. Plusieurs modèles, avec des niveaux de corrélation différents sont possibles. De plus l'estimation de ces modèles permet de corriger l'estimation du paramètre d'hétérogénéité. Cependant, le problème avec les modèles linéaires généralisés avec effets aléatoires est que les procédures d'estimation ne sont pas faciles à mettre en œuvre à cause de la complexité de la fonction de vraisemblance du modèle.

Organisation de la thèse

La suite de la thèse est organisée comme suit : Au chapitre 1, nous présentons les notions de copules ainsi que certaines de leurs propriétés, nous accordons une attention particulère à la famille des copules Archimédiennes. Nous discutons également de l'utilisation des copules pour des données discrètes, et des données mixtes continues et discrètes. Au chapitre 2, nous présentons le cas le plus simple de nos modèles, sans variable auxiliaire. Nous en déduisons une procédure d'inférence pour l'estimation du coefficient de corrélation intra-grappe. Le chapitre 3 propose ensuite une extension de ces modèles afin d'inclure des variables auxiliaires au niveau des grappes et d'estimer la taille de la population à partir des données d'une expérience de capture-recapture. L'utilisation des modèles pour faire de la prévision dans de petits domaines est présentée au chapitre 4. Le chapitre 5 présente l'utilisation des copules Archimédienne, pour modéliser la distribution d'un couple de variables aléatoires mixtes. Nous concluons la thèse et donnons des perspectives au chapitre 6. L'annexe A donne le matériel supplémentaire pour les chapitres 2 et 3. Dans l'annexe B, nous présentons la preuve du théorème du chapitre 3.

Chapitre 1

Théorie des copules

1.1 Introduction

La théorie des copules a connu ces dernières décennies un développement considérable avec ses applications dans de nombreux domaines, notamment en finance et en actuariat. L'introduction des copules dans la théorie des probabilités en vue d'applications statistiques est un phénomène relativement récent. Ce n'est que vers la fin des années 50 que les recherches dans ce domaine ont connu un essor considérable. En statistique, Sklar (1959) est le premier à utiliser le terme copule pour décrire les fonctions qui réunissent les fonctions de répartition unidimensionnelles pour former des fonctions de répartition multidimensionnelles. Schweizer and Wolff (1981) sont les auteurs du premier article associant les copules à l'étude de la dépendance entre des variables aléatoires. Une copule modélise la dépendance entre les variables dans une distribution multivariée, elle peut être combinée avec tout ensemble de distributions univariées pour les distributions marginales.

1.2 Définition et existence

Une copule est une fonction de répartition (FR) multivariée dont les distributions marginales sont toutes uniformes sur l'intervalle $(0,1)$. Supposons que le vecteur de variables aléatoires Y=$(Y_1,\ldots,Y_d)$, $d \geq 2$, a une FR multivariée F_Y avec des FRs marginales $F_{Y_1}(y_1),\ldots,F_{Y_d}(y_d)$. Alors, on peut montrer que chacune des $F_{Y_1}(Y_1),\ldots,F_{Y_d}(Y_d)$ a une distribution uniforme$(0,1)$ et par conséquent la FR du vecteur $\left(F_{Y_1}(Y_1),\ldots,F_{Y_d}(Y_d)\right)$ est une copule. Cette FR est appelée la copule de Y et on la note par C_Y. La fonction C_Y contient toute l'information concernant la dépendance entre les composantes de Y.

Il est facile de trouver une formule pour C_Y à partir de la distribution conjointe et des distributions marginales. Supposons que toutes les variables aléatoires ont une FR continue et strictement croissante, alors en utilisant la définition d'une FR nous avons

$$C_Y(u_1,\ldots,u_d) = F_Y\left(F_{Y_1}^{-1}(u_1),\ldots,F_{Y_d}^{-1}(u_d)\right) \tag{1.1}$$

En posant $u_j = F_{Y_j}(y_j)$, $j = 1,\dots,d$, dans (1.1), on obtient

$$F_Y(y_1,\dots,y_d) = C_Y\left\{F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\right\} \tag{1.2}$$

L'équation (1.2) fait partie du célèbre théorème de Sklar (1959) qui a donné les fondamments théoriques de l'application des copules. En effet, il résulte de ce théorème que la fonction de distribution conjointe d'un vecteur de variables aléatoires peut s'écrire en fonction de ses marges qui contiennent toute l'information concernant les distributions marginales univariées et d'une fonction de copule qui contient toute l'information relative à la structure de dépendance. De plus, ce théorème assure que si les FRs marginales sont continues, alors C_Y est unique.

Soit $c_{Y,\alpha}$ la fonction de densité de $C_{Y,\alpha}$ définie par :

$$c_Y(u_1,\dots,u_d) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C_Y(u_1,\dots,u_d). \tag{1.3}$$

En dérivant (1.2), et en utilisant (1.3), nous obtenons la fonction de densité du vecteur Y donnée par

$$f_Y(y_1,\dots,y_d) = c_Y\left(F_{Y_1}(y_1),\dots,F_{Y_d}(y_d)\right) f_{Y_1}(y_1)\cdots f_{Y_d}(y_d). \tag{1.4}$$

Notez que cette expression de la densité conjointe n'est valide que lorsqu'on utilise la copule pour la structure de dépendance entre des variables continues. Cependant, il y a un intérêt de plus en plus croissant dans l'application des copules pour des données binaires (Meester and Mackay, 1994; Nikolouplopoulos, 2009; Genest and Nešlehová, 2007). Dans ce dernier cas, Trégouët et al. (1999) donnent une expression de la fonction de probabilité conjointe en fonction de la copule comme suit :

$$P(Y_1 = y_1,\dots,Y_d = y_d|x) = \sum_{k_1=0,1} \cdots \sum_{k_d=0,1} (-1)^{\sum_{j=1}^d k_j} C_Y\left\{F_{Y_1}(y_1 - k_1|x_1),\dots,F_{Y_d}(y_d - k_d|x_d)\right\} \tag{1.5}$$

où $y_j = 0,1$ est la variable réponse associée à la composante j, $x = (x_1,\dots,x_d)$ où x_j est le vecteur des covariables pour la $j^{\text{ième}}$ composante, et F_{Y_j} est la fonction de répartition marginale associée au $j^{\text{ième}}$ individu définie par :

$$F_{Y_j}(u|x_j) = \begin{cases} 0 & \text{if} \quad u < 0 \\ 1 - \pi_j(x_j) & \text{if} \quad 0 \le u < 1 \\ 1 & \text{if} \quad u \ge 1. \end{cases}$$

où $\pi_j(x_j) = g^{-1}(x_j^T \beta)$ est la probabilité de succès marginale, g^{-1} est l'inverse d'une fonction de lien, x_j le vecteur des covariables associé à l'individu j, et β vecteur des paramètres de régression.

1.3 Propriétés des copules

Dans cette section, nous présentons quelques propriétés des copules. Le théorème suivant est une de leurs propriétés les plus importantes.

Théorème : Supposons que g_j est une fonction strictement croissante et $X_j = g_j(Y_j)$, pour $j = 1,\ldots,d$. Alors, nous avons

$$C_X(u_1,\ldots,u_d) = C_Y(u_1,\ldots,u_d)$$

où C_X est la copule associée au vecteur de variable aléatoires $X = (X_1,\ldots,X_d)$. En d'autres termes, les copules sont invariantes par transformation strictement croissante.

Preuve : En notant que la FR conjointe de $X = (X_1,\ldots,X_d)$ est,

$$\begin{aligned} F_X(x_1,\ldots,x_d) &= P\big\{g_1(Y_1) \leq x_1,\ldots,g_d(Y_d) \leq x_d\big\} \\ &= P\big\{Y_1 \leq g_1^{-1}(x_1),\ldots,Y_d \leq g_d^{-1}(x_d)\big\} \\ &= F_Y\big\{g_1^{-1}(x_1),\ldots,g_1^{-1}(x_1)\big\} \end{aligned} \tag{1.6}$$

Et donc la FR de X_j est

$$F_{X_j}(x_j) = F_{Y_j}\big\{g_j^{-1}(x_j)\big\}$$

et par conséquent,

$$F_{X_j}^{-1}(u) = g_j\big\{F_{Y_j}^{-1}(u)\big\}. \tag{1.7}$$

Il résulte ainsi, de l'application de (1.1) au vecteur X, des résultats (1.6) et (1.7), et enfin de l'application de (1.1) au vecteur Y, que la copule de X est donnée par.

$$\begin{aligned} C_X(u_1,\ldots,u_d) &= F_X\big\{F_{X_1}^{-1}(u_1),\ldots,F_{X_d}^{-1}(u_d)\big\} \\ &= F_Y\big[g_1^{-1}\{F_{X_1}^{-1}(u_1)\},\ldots,g_d^{-1}\{F_{X_d}^{-1}(u_d)\}\big] \\ &= F_Y\big\{F_{Y_1}^{-1}(u_1),\ldots,F_{Y_d}^{-1}(u_d)\big\} \\ &= C_Y(u_1,\ldots,u_d) \qquad \blacksquare \end{aligned}$$

Pour utiliser les copules afin de modéliser la dépendance multivariée, nous avons besoin des familles paramétriques de copules.

1.4 Exemples de copules

Cette section présente six exemples de copules. Les trois premiers sont d'intérêt spécial parce qu'elles représentent l'indépendance et les deux dépendances extrêmes.

• **La copule d'indépendance :** La copule d'indépendance de dimension d est la copule de d variables aléatoires Uniforme $(0,1)$. Elle est donnée par

$$C^{ind}(u_1,\ldots,u_d) = u_1\cdots u_d, \quad u_1,\ldots,u_d \quad \in \quad [0,1], \tag{1.8}$$

et sa fonction de densité est $c^{ind}(u_1,\ldots,u_d) = 1$.

• **La copule comonotone :** La copule comonotone de dimension d, notée par C^{Min}, a une dépendance positive parfaite. Soit U une variable aléatoire uniforme$(0,1)$. Alors, la copule comonotone est la FR de $\mathbf{U} = (U,\ldots,U)$; c'est à dire $\mathbf{U}$ contient d copies de U et donc toutes les composantes de $\mathbf{U}$ sont égales. Donc,

$$C^{Min}(u_1,\ldots,u_d) = P(U \leq u_1,\ldots,U \leq u_d) = P\{Y \leq min(u_1,\ldots,u_d)\} = min(u_1,\ldots,u_d).$$

• **La copule antimonotone :** La copule antimonotone de dimension 2 est la FR de $(U, 1-U)$, lequelle a une dépendance négative parfaite.

$$C^{Max}(u_1,u_2) = P(U \leq u_1, 1-U \leq u_2) = P(1-u_2 \leq U \leq u_1) = max(u_1+u_2-1,0).$$

Pour $d \geq 3$, il n'est pas possible d'avoir une copule antimonotone.

Notez que toute copule bivariée satisfait l'inégalité de Fréchet donnée par :

$$C^{Max}(u_1,u_2) \leq C(u_1,u_2) \leq C^{Min}(u_1,u_2).$$

• **La copule de survie :** Soit C une copule de dimension d et soit $U = (U_1,\ldots,U_d)$ un vecteur de variables aléatoires uniformes sur $[0,1]^d$ avec fonction de copule C. La copule de survie associée à la copule C, notée par $\bar{C}$, permet de connecter la fonction de survie conjointe S avec ses marges. La fonction de survie conjointe peut ainsi s'écrire comme suit :

$$\begin{aligned} S(u_1,\ldots,u_d) &= P(U_1 \geq u_1,\ldots,U_d \geq u_d), \quad \forall \quad (u_1,\ldots,u_d) \in [0,1]^d \\ &= \bar{C}\{S_1(u_1),\ldots,S_d(u_1)\}, \end{aligned}$$

où $S_j(u_j) = P(U_j \geq u_j) = S(0,\ldots,u_j,\ldots,0) = u_j$, $j = 1,\ldots,d$, est la fonction de survie univariée associée à la $j^{\text{ème}}$ composante du vecteur U.

Il existe une relation entre C et $\bar{C}$. Dans le cas bivarié, cette relation s'exprime comme suit :

$$\bar{C}(u_1,u_2) = u_1+u_2-1+C(1-u_1,1-u_2).$$

Ainsi, pour une copule C donnée, ce lien permet de construire la fonction de survie conjointe S en utilisant soit la copule C où la copule de survie $\bar{C}$.

Noter enfin que, si $(U_1,\ldots,U_d)$ est un vecteur de variables aléatoires uniformes sur $[0,1]^d$ avec copule C, alors $(1-U_1,\ldots,1-U_d)$ est un vecteur de variables aléatoires uniformes sur $[0,1]^d$ avec copule $\bar{C}$.

• **La copule normale** La copule normale est l'une des familles de copules les plus populaires en pratique. C'est une copule qui dépend seulement de la matrice de corrélation. La version bivariée de cette de copule est donnée par :

$$C^{Gauss}(u_1,u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\Big)dxdy,$$

où ρ, avec $\rho \in [-1,1]$, est le paramètre de la copule, et $\Phi^{-1}(\cdot)$ est l'inverse de la fonction de répartion de la distribution normale univariée standard.

Noter que la copule Gaussienne est symétrique dans le sens où elle est égale à sa copule de survie. Une copule Gaussienne dont la matrice de corrélation est la matrice identité, donc toutes les corrélations sont 0, correspond à la copule d'indépendance. Une copule Gaussienne converge vers la copule comonotone si toute les corrélations dans la matrice de corrélation convergent vers 1. Dans le cas bivarié, si la corrélation converge vers -1, alors la copule Gaussienne converge vers la copule antimonotone.

1.5 Mesures de dépendance

Nous avons vu à la section précédente que la copule d'une distribution multivariée permet de décrire la structure de dépendance. La présente section discute des mesures de concordance bivariées, à savoir, le Tau de Kendall, et les coefficients de dépendance caudales. Contrairement au coefficient de corélation linéaire, ces mesures de dépendance peuvent être exprimées en fonction de la copule sous-jacente. Les sections suivantes décrivent le tau de Kendall et le concept de dépendance caudale.

1.5.1 Tau de Kendall

Soit F la FR d'une distribution bivariée, et considérons (X_1,Y_1) et (X_2,Y_2) une paire de vecteurs aléatoires indépendants issus de cette distribution. On dit que les couples (X_1,Y_1) et (X_2,Y_2) sont concordants si $X_1 > X_2$ lorsque $Y_1 > Y_2$, et $X_1 < X_2$ lorsque $Y_1 < Y_2$, et on dit qu'ils sont discordants dans le cas contraire. Le tau de Kendall pour la distribution F est une mesure de concordance définie par

$$\begin{aligned}\tau &= Pr\big\{(X_1-X_2)(Y_1-Y_2)>0\big\} - Pr\big\{(X_1-X_2)(Y_1-Y_2)<0\big\}\\ &= E\big[signe\big\{(X_1-X_2)(Y_1-Y_2)\big\}\big],\end{aligned}$$

où signe(z) =-1, 0 où 1 selon que z =-1, 0 où 1, respectivement.

Il existe une relation entre le tau de Kendall et la copule associée à la distribution F. Si C représente cette copule, alors on peut montrer que :

$$\tau = 4E\big\{C(U,V)\big\} - 1 = 4\int_0^1\int_0^1 CdC - 1 = 4\int_0^1\int_0^1 C(u,v)c(u,v)dudv - 1,$$

où c est la densité de C.

Une estimation empirique du tau de Kendall, pour un échantillon $\{(x_i,y_i), i = 1,\ldots n\}$ de taille n, est le nombre paires concordantes moins le nombre de paires discordantes divisé par le nombre total de paires :

$$\hat{\tau}_n = \frac{1}{\binom{n}{2}} \sum_{1\leq i \leq j \leq n} signe\big\{(x_i-x_j)(y_i-y_j)\big\}.$$

1.5.2 Mesures de dépendance caudale

Dans cette section, nous discutons des coefficients de dépendance caudale à droite (et à gauche) qui ont été introduits par Embrechts et al. (1999) dans un contexte financier. Ces mesures de dépendance sont utilisées pour capturer la dépendance à la queue des distributions bivariées.

De façon plus précise, si X et Y sont des variables aléatoires de distributions respectives F_X et F_Y, alors, par définition, le coefficient de dépendance à la queue à droite quantifie la probabilité d'observer une grande valeur de Y, sachant qu'on a observé une grande valeur de X. Ce coefficient, que l'on note par λ_U, est donné par :

$$\lambda_U = lim_{v \to 1} Pr\big\{Y > F_Y^{-1}(v) | X > F_X^{-1}(v)\big\}.$$

De façon analogue, on définit le coefficient de dépendance caudale à gauche par :

$$\lambda_L = lim_{v \to 0} Pr\big\{Y \leq F_Y^{-1}(v) | X \leq F_X^{-1}(v)\big\}.$$

Il existe une relation entre ces coefficients et la copule associée à la distribution du couple (X,Y). Si C dénote cette copule, alors on peut montrer que :

$$\begin{aligned} \lambda_U &= lim_{v \to 1} \frac{1-2v+C(v,v)}{1-v}, \\ \lambda_L &= lim_{v \to 0} \frac{C(v,v)}{v}. \end{aligned}$$

Ces mesures sont indépendantes des distributions marginales, F_X et F_Y. De plus, elles sont invariantes par transformation strictement croissante des variables X et Y. Si $\lambda_U(\lambda_L) \in (0,1]$, alors il existe une dépendance caudale à droite (à gauche), tandis que $\lambda_U = 0 \;\; (\lambda_L = 0)$ correspond à l'absence de dépendance caudale à droite (à gauche). Pour plus de détails, voir Kjersti (2004).

1.6 Les copules Archimédiennes

Les copules Archimédiennes sont sans nul doute la classe de copules paramétriques la plus populaire. Elles ont été largement explorées par les statisticiens canadiens Christian Genest et Louis-Paul Rivest. Ils ont fortement contribué, par le biais de leurs nombreuses publications, à faire connaître ces classes de copules. Parmi leurs publications qui ont rendu ces familles de copules faciles d'utilisation, on peut citer (Genest and MacKay, 1986; Genest and Rivest, 1993).

Les copules Archimédiennes sont une classe très importante de copules qui sont faciles à construire et qui possédent de très belles propriétés mathématiques et statistiques. Parmi les raisons qui justifient aussi l'utilisation de cette classe, il y a le fait qu'elle contient plusieurs familles paramétriques, et une grande variété de structures de dépendance.

1.6.1 Définitions et propriétés

Cette section commence par donner la définition d'une copule Archimédienne bidimensionnelle.

Définition 1 *Soit* $\psi : [0,\infty] \to [0,1]$*, une fonction continue, strictement décroissante, et convexe telle que* $\psi(0) = 1$*. Une copule bivariée C est appelée copule Archimédienne bivariée si elle admet la représentation suivante*

$$C(u,v) = \psi\{\psi^{-1}(u) + \psi^{-1}(v)\} \quad \forall\, u,v \,\in\, [0,1]^2, \tag{1.9}$$

où $\psi^{-1}(t)$ *est l'inverse de* $\psi(t)$*, telle que* $\psi^{-1}(1) = 0$ *.*

La fonction ψ est connue sous le nom de générateur de la copule. Noter que si $\lim_{u\to 0} \psi^{-1}(0) = \infty$, alors on dit que ψ est un générateur strict, et dans ce cas on dit que la fonction C définie en (1.9) est une copule Archimédienne stricte.

Ci-dessous, nous présentons quelques propriétés intéressantes de cette famille de copules.

• **La symétrie** : Les copules Archimédiennes sont symétriques dans le sens où la valeur de $C(u,v)$ ne change pas par permutation de ses composantes. On dit que les copules Archimédiennes sont **échangeables**.

• **L'association** : C est associative, c'est à dire pour tout u, v et w dans $[0,1]$, on a

$$C\{u, C(v,w)\} = C\{C(u,v), w\}.$$

En utilisant ce dernier résultat, on peut facilement construire les copules Archimédiennes multidimensionnelles par une extension des copules Archiméediennes bivariées. Pour cela, rappelons d'abord la définition d'une fonction complètement monotone.

Définition 2 *Une fonction g est complètement monotone dans un intervalle J si elle admet des dérivées de tout ordre dans J et*

$$(-1)^k \frac{d^k g(t)}{dt^k} \geq 0, \quad \forall\, t \,\in\, J, \text{ et } k \,\in\, \{1,2,3,\ldots\},$$

où $\dfrac{d^k g(t)}{dt^k}$ *est la dérivée d'ordre k de la fonction g par rapport à t.*

On peut maintenant donner la définition d'une copule Archimédienne multivariée.

Définition 3 *Si* ψ *est complètement monotone, alors la fonction définie par*

$$C_\psi(u_1,\ldots,u_d) = \psi\{\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)\} \quad \forall\, u_1,\ldots,u_d, \,\in\, [0,1]^d, \tag{1.10}$$

est une copule Archimédienne pour tout d($d > 2$*).*

Les copules Archimédiennes peuvent être générées en utilisant la transformée de Laplace d'une variable aléatoire positive. Le théorème suivant, connu sous le nom de *formulate Bernstein's*, établit une relation entre le générateur d'une copule Archimédienne et la transformée de Laplace d'une variable aléatoire positive.

Théorème 1 *Une fonction $\varphi : [0,\infty[\rightarrow [0,1]$ est continument dérivable si et seulement si elle est la transformée de Laplace d'une variable aléatoire positive M, c'est-à-dire $\varphi(t) = E\{\exp(-tM)\}$, et $P(M > 0) = 1$.*

Si le générateur ψ d'une copule Archimédienne est la transformée de Laplace d'une variable aléatoire M de distribution G à support positif, telle que $G(0) = 0$, alors la procédure suivante que l'on peut trouver dans Marshall and Olkin (1993) peut être utilisée pour générer une copule Archimédienne :

• Générer M telle que sa transformée de Laplace est ψ.

• Générer d variables aléatoires uniformes sur $[0,1]$, $U_1,\ldots,U_d$ de façon indépendante.

• Alors, le vecteur de variables aléatoires défini par $V_i = \psi^{-1}\Big\{-\dfrac{\log(U_i)}{M}\Big\}$, pour $i = 1,\ldots,d$, est un vecteur de variables aléatoires uniformes avec fonction de répartition C_ψ.

Les copules Archimédiennes sont plus flexibles que la copule normale et la copule de la distribution t dans le sens qu'elles peuvent être symétriques, ou asymétriques, et peuvent avoir de la dépendance caudale à gauche et à droite. Pour plus d'informations concernant la famille de copules Archimédiennes, voir Nelson (2006),Joe (1997), et Mai and Scherer (2012). Ci-dessous, nous donnons quatre exemples de copules Archimédiennes populaires.

1.6.2 Exemples de copules Archimediennes

Dans cette section, nous présentons la copule de Clayton, la copule de Gumbel, la copule de Frank et la copule de Joe. Ces quatre familles sont sans doute les copules Archimédiennes les plus populaires.

a. Copule de Clayton

La copule de Clayton multidimensionnelle est définie par :

$$C^{Cl}(u_1,\ldots,u_d) = (u_1^{-\alpha} + \cdots + u_d^{-\alpha} - d + 1)^{-1/\alpha} \quad \forall\, u_1,\ldots,u_d,\, \in\, [0,1]^d,$$

où $\alpha\ \in\,]-1,\infty[\backslash\{0\}$.

Pour cette copule, le générateur $\psi(\cdot)$ et son inverse $\psi^{-1}(\cdot)$ sont respectivement donnés par :

$$\psi(t) = (1+\alpha t)^{-1/\alpha}, \quad \psi^{-1}(t) = \frac{(t^{-\alpha}-1)}{\alpha}.$$

Le générateur de la copule correspond à la transformée de Laplace de la distribution Gamma avec paramèmetre de forme $1/\alpha$, et paramétre d'échelle α ; $M \sim$ Gamma(forme=$1/\alpha$,échelle=α).

La copule de Clayton est une copule asymétrique dont les cas limites sont la copule d'indépendance pour $\alpha \rightarrow 0$, et la copule comonotone pour $\alpha \rightarrow \infty$. La famille de Clayton a une dépendance de queue à gauche

$\lambda_L = 2^{-1/\alpha}$, mais n'admet pas de dépendance caudale à droite. Le tau de Kendall pour cette famille est $\tau = \alpha/(\alpha+2)$.

b. Copule de Gumbel

La copule de Gumbel est une copule asymétrique définie par :

$$C^{Gu}(u_1,\ldots,u_d) = \exp\left[-\left\{\{-\log(u_1)\}^{\alpha+1} + \cdots + \{-\log(u_d)\}^{\alpha+1}\right\}^{1/(\alpha+1)}\right] \quad \forall\, u_1,\ldots,u_d,\, \in\, [0,1]^d,$$

où $\alpha\, \in\, [0,\infty[$.

Le générateur de cette copule et son inverse sont respectivement définis par :

$$\psi(t) = \exp\{-t^{1/(1+\alpha)}\}, \quad \psi^{-1}(t) = \left\{\log(t)\right\}^{\alpha+1}.$$

Le générateur de la copule de Gumbel est la transformée de Laplace d'une distribution de loi stable, M ~St$(1/(1+\alpha)$,1,γ,0), avec $\gamma = \left[\cos\left\{\pi/\{2(1+\alpha)\}\right\}\right]^{1+\alpha}$, voir Frees and Valdez (1998). Chambers et al. (1976) ont proposé un algorithme simple pour générer une distribution de loi stable. Pour $\alpha = 0$, la copule de Gumbel est la copule d'indépendance et elle converge vers la copule de la dépendance parfaite lorsque $\alpha \to \infty$. Le tau de Kendall pour cette copule est donnée par $\tau = 1-(1+\alpha)^{-1}$. La copule de gumbel n'admet pas de dépendance caudale à gauche, mais a une dépendance caudale à droite, qui vaut $\lambda_U = 2 - 2^{1/(1+\alpha)}$.

c. Copule de Frank

La copule de Frank de paramètre α, avec $0 < \alpha < \infty$, est donné par :

$$C^{Fr}(u_1,\ldots,u_d) = -\frac{1}{\alpha}\log\left[1 + \frac{\prod_{i=1}^{d}\{\exp(-\alpha u_i)-1\}}{\exp(-\alpha)-1}\right], \quad \forall\, u_1,\ldots,u_d,\, \in\, [0,1]^d.$$

Le générateur et son inverse sont donnés par :

$$\psi(t) = -\frac{1}{\alpha}\log\{e^{-t}(e^{-\alpha}+1)+1\}, \quad \psi^{-1}(t) = -\log\left\{\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\right\}.$$

Le générateur de la copule de Frank correspond à la transformée de Laplace de la distribution logarithmique, M ~Log$(1-e^{-\alpha})$, avec fonction de masse donnée par

$$Pr(M=m) = \frac{(1-e^{-\alpha})^m}{\alpha m}, \quad m = 1,2,\ldots.$$

Contrairement aux copules de Clayton et de Gumbel, en dimension 2 la copule de Frank est symétrique, et n'admet pas de dépendance à la queue, ni à gauche ni à droite. Les cas limites de cette famille sont la copule d'indépendance pour $\alpha \to 0$, et la copule comonotone pour $\alpha \to \infty$. Pour cette famille, la relation entre le tau de Kendall et le paramètre de dépendance est donnée par

$$\tau = 1 - 4\frac{[D_1(\theta)-1]}{\theta},$$

où $D_1(\theta) = \frac{1}{\theta}\int_0^\theta x/(e^x-1)dx$ est une fonction de Debye de premier ordre.

d. Copule de Joe

La copule de Joe multidimensionnelle est définie par,

$$C^{Joe}(u_1,\dots,u_d) = 1 - \Big\{\sum_{i=1}^d (1-u_i)^{1+\alpha} - \prod_{i=1}^d (1-u_i)^{1+\alpha}\Big\}^{1/(1+\alpha)}, \quad \forall\, u_1,\dots,u_d, \in\ [0,1]^d,$$

où $\alpha\ \in\ [0,\infty[$.

Pour cette famille, le générateur et son inverse sont respectivement donnés par :

$$\psi(t) = 1-(1-e^{-t})^{1/(1+\alpha)}, \quad \psi^{-1}(t) = -\log\big\{1-(1-t)^{\alpha+1}\big\}.$$

Le générateur de la copule de Joe est la transformée de Laplace de la distribution de Sibuya de fonction de masse donnée par, $Pr(M=m) = \binom{1/(1+\alpha)}{m}(-1)^{m-1}$, pour $m \in \mathscr{N}^+$, voir Mai and Scherer (2012, chap. 2).

Lorsque $\alpha = 0$ cette famille correspond à la copule d'indépendance, et lorsque $\alpha \to \infty$, elle correspond à la dépendance parfaite. Le tau de Kendall n'a pas une expression explicite et est donné comme une série infinie, voir Mai and Scherer (2012, chap. 2). Cette copule a une dépendance caudale à droite qui vaut $\lambda_U = 2-2^{1/(1+\alpha)}$, mais n'a pas de de dépendance caudale à gauche.

Pour plus de discussion concernant ces quatre familles et leurs algorithmes de simulation, voir (Mai and Scherer, 2012, chap. 2).

1.7 Copules Archimédiennes et données de Bernoulli

Dans cette section, nous présentons la classe de modèles à effet aléatoire que nous proposons pour des données de Bernoulli corrélées. Ces modèles traitent la corrélation en exprimant la probabilité de succès comme fonction d'une variable aléatoire positive. Sachant cette variable aléatoire, les observations sont supposées indépendantes et identiquement distribuées selon la loi de Bernoulli. Nous obtenons ainsi la distribution conjointe de survie (et de répartion) des observations en intégrant la distribution conjointe de survie (et de répartition) conditionnelle par rapport à la densité de l'effet aléatoire. Nous montrons que cette distribution conjointe a la forme d'une copule Archimédienne. Ce qui nous permet d'établir le lien entre la classe de modèles à effet aléatoire et la famille des copules Archimédiennes.

Soit Y_j, $j=1,\dots,d$, une séquence de d variables aléatoires de Bernoulli corrélées, où $Y_j = 1$ est un succès et 0 est un échec. Le modèle à effet aléatoire que nous considérons dans cette thèse suppose que la dépendance intra-grappe est induite par une variable aléatoire a à support positif $[0,+\infty[$ dont la fonction de distribution, notée par $F_\alpha(\cdot)$ dépend d'un paramètre de dépendance positif α. Nous notons par ψ_α, la transformée de Laplace associée à la distribution $F_\alpha(\cdot)$,

$$\psi_\alpha(t) = E(e^{-ta}) = \int_0^{+\infty} e^{-at}dF_\alpha(a).$$

Le modèle est défini en supposant que la probabilité de succès varie d'une unité à une autre selon la relation suivante :

$$p_j = e^{-a\psi_\alpha^{-1}(\pi_j)}, \quad j = 1, \ldots, d, \tag{1.11}$$

où $\pi_j = E(p_j)$ est la probabilité de succès marginale associée à l'observation j, et $\psi_\alpha^{-1}(\pi_j)$ dénote l'inverse de la transformée de Laplace $\psi_\alpha(\cdot)$ évaluée à π_j.

Sachant l'effet aléatoire a, les observations sont conditionnellement indépendantes. Ainsi pour le modèle (1.11), la distribution conjointe de survie conditionnelle des observations est :

$$\begin{aligned} P(Y_1 \geq y_1, \ldots, Y_d \geq y_d|a) &= \prod_{j=1}^{d} P(Y_j \geq y_j|a) \\ &= \prod_{j=1}^{d} p_j^{y_j} \\ &= \exp\{-a\sum_{j=1}^{d} y_j\psi_\alpha^{-1}(\pi_j)\}, \end{aligned} \tag{1.12}$$

avec y_j =0, 1, pour $j = 1, \ldots, d$. La distribution de survie conjointe marginale peut ainsi être obtenue en intégrant (1.12) par rapport à la densité de l'effet aléatoire a :

$$\begin{aligned} P(Y_1 \geq y_1, \ldots, Y_d \geq y_d) &= E\{P(Y_1 \geq y_1, \ldots, Y_n \geq y_n|a)\} \\ &= \int_0^{+\infty} \exp\{-a\sum_{j=1}^{d} y_j\psi_\alpha^{-1}(\pi_j)\}f_\alpha(a)da. \end{aligned} \tag{1.13}$$

L'expression en (1.13) n'est rien d'autre que la transformée de Laplace ψ_α évaluée à $\sum_{j=1}^{d} y_j\psi_\alpha^{-1}(\pi_j)$, donc on a

$$P(Y_1 \geq y_1, \ldots, Y_d \geq y_d) = \psi_\alpha\{\sum_{j=1}^{d} y_j\psi_\alpha^{-1}(\pi_j)\}. \tag{1.14}$$

La distribution marginale de survie pour chaque Y_j peut être obtenue en mettant toutes les autres composantes à zéro dans (1.14)

$$S_j(y_j) = P(Y_j \geq y_j) = P(Y_1 \geq 0, \ldots, Y_j \geq y_j, \ldots, Y_d \geq 0) = \psi_\alpha\{y_j\psi_\alpha^{-1}(\pi_j)\}, \tag{1.15}$$

où $S_j(y_j)$ vaut 1 si $y_j = 0$ et π_j si $y_j = 1$.

En utilisant l'approche par les copules, la distribution conjointe de survie est donnée par :

$$P(Y_1 \geq y_1, \ldots, Y_d \geq y_d) = C_\alpha\{S_1(y_1), \ldots, S_d(y_d)\}, \tag{1.16}$$

où C_α est une fonction de copule paramétrisée par le paramètre de dépendance α, et S_j est la fonction de survie marginale de Y_j, $S_j(y_j) = P(Y_j \geq y_j) = \pi_j^{y_j}$, pour $y_j = 0, 1$.

Pour comparer le modèle de copule et le modèle à effet aléatoire défini en (1.11), nous considérons la famille des copules Archimédiennes définie par :

$$C_\alpha(u_1, \ldots, u_d) = \phi_\alpha\{\phi_\alpha^{-1}(u_1) + \ldots + \phi_\alpha^{-1}(u_d)\}, \tag{1.17}$$

où ϕ_α est une fonction décroissante à support positif telle que $\phi_\alpha(0) = 1$, et ϕ_α^{-1} est son inverse telle que $\phi_\alpha^{-1}(1) = 0$.

En choisissant le générateur ϕ_α comme étant égal à la transformée de Laplace, ψ_α alors la copule en (1.17) devient :

$$C_\alpha(u_1, \ldots, u_d) = \psi_\alpha\{\psi_\alpha^{-1}(u_1) + \ldots + \psi_\alpha^{-1}(u_d)\}. \tag{1.18}$$

Il résulte ainsi des relations (5.10) et (1.18) que sous le modèle de copule, la distribution conjointe de survie est

$$\begin{aligned} P(Y_1 \geq y_1, \ldots, Y_d \geq y_d) &= \psi_\alpha\big[\psi_\alpha^{-1}\{S_1(y_1)\} + \ldots + \psi_\alpha^{-1}\{S_d(y_d)\}\big] \\ &= \psi_\alpha\big[\sum_{j=1}^{d} \psi_\alpha^{-1}\{S_j(y_j)\}\big]. \end{aligned} \tag{1.19}$$

En remplaçant enfin $S_j(y_j)$ par son expression dans (1.15), on peut remarquer que les distributions conjointes en (1.14) et (1.19) sont les mêmes, ce qui prouve que le modèle à effet aléatoire défini en (1.11) n'est rien d'autre qu'un modèle de copule Archimédienne particulier.

Il est important de noter que l'utilisation des copules pour des données discrètes n'est pas nouveau. Ce phénomène a fait l'objet d'études et de discussions dans la littérature, voir par exemple, Meester and Mackay (1994); Nikoloulopoulos and Karlis (2008); Nikolouplopoulos (2009); Nikoloulopoulos and Karlis (2010). Cependant, Genest and Nešlehová (2007) note un certain nombre de complications liées à l'application directe des modèles de copule à des données discrètes, en utilisant des marges discrètes dans une copule. Ces difficultés sont liées au fait que pour ces données, la copule n'est pas unique. On note également un problème d'interprétation du paramétre de dépendance. Noter cependant que du point de vue de la modélisation, la non unicité de la copule n'est pas un problème, car les paramètres du modèle sont identifiables, et donc peuvent être estimés.

1.8 Copules et données mixtes

Les données en forme d'une combinaison de variables aléatoires continues et discrètes (nominale, ordinale, ou binaire) corrélées interviennent dans de nombreux domaines d'application, notamment en santé, en médecine, en économétrie, ou en actuariat. Ces données sont souvent appelées données multivariées mixtes corrélées. L'analyse du comportement conjoint de ces données nécessite la spécification de modèles flexibles qui tiennent compte de la dépendance. De tels modèles permettent de tenir compte des relations entre les variables, d'évaluer l'influence des covariables sur la distribution conjointe, et de caractériser la structure de dépendance.

Plusieurs méthodes sont disponibles pour analyser des données multivariées continues, et les approches traditionnelles comme les modèles linéaires généralisés mixtes peuvent être utilisés pour faire des inférences. Cependant, l'analyse conjointe de données mixtes, discrètes et continues, mène à des complications qui sont souvent dûes à des difficultés liées à la construction de modèles pour ce type de données.

Une approche consiste à transformer une catégorie de variables en l'autre type de variables (par exemple la discrétisation des variables continues), et ensuite choisir une méthode d'analyse appropriée. Bien que simple et facile à mettre en oeuvre, cette approche est souvent inefficace du fait que la conversion de variables n'est pas appropriée dans de nombreuses applications. Nous avons donc besoin de développer des méthodes alternatives pour construire de façon directe ou indirecte des modèles conjoints pour des données multivariées mixtes. Pour construire ces modèles, plusieurs approches ont été proposées dans la littérature.

Soient X un vecteur de variables aléatoires discrètes (nominale, binaire, ou ordinale) et Y une vecteur de variables aléatoires continues. Une analyse conjointe de ces types de variables aléatoires, nécessite une spécification de façon directe ou indirecte de leur densité conjointe, $f_{X,Y}(x,y)$.

La factorisation est souvent utilisée dans la littérature comme une méthode directe qui exprime la densité conjointe $f_{X,Y}(x,y)$ en une densité marginale pour une catégorie de variables et une densité conditionnelle de l'autre ensemble de variables. Cette approche suggère donc deux formulations de la distribution conjointe. Le premier exprime la distribution conjointe comme suit, $f_{X,Y}(x,y) = f_X(x)f_{Y|X}(y|x)$, où f_X est la densité marginale de X, et $f_{Y|X}$ dénote la densité conditionnelle de Y sachant X. Un exemple de cette approche est les modèles de localisation généraux (GLOMs) introduits par Olkin and Tate (1961). Ces modèles utilisent la distribution multinomiale pour les variables discrètes, et la distribution normale multivariée conditionnelle pour les variables continues. Des applications récentes des GLOMs peuvent être trouvées dans Fitzmaurice and Laud (1995); Fitzmaurice and Laird (1997); Hirakawa (2012); Kang and Yang (2013). La distribution conjointe peut aussi être formée à partir du produit de la densité marginale f_Y pour la variable continue Y, et la densité conditionnelle $f_{X|Y}$ de la variable discrète X sachant la variable continue Y. On peut ainsi écrire : $f_{X,Y}(x,y) = f_Y(y)f_{X|Y}(x|y)$. Un exemple de ce modèle est les modèles appelés modèles conditionnels continus groupés (CGCMs) dont on peut trouver des applications récentes dans Catalano and Ryan (1992); Catalano (1997); de Leon (2005). D'autres extensions des modèles GLOMs et CGCMs ont été considérées par de Leon and Carrière (2007), qui utilisent ces deux modèles pour contruire un modèle hybride.

Bien que les modèles obtenus par la factorisation soient simples et faciles à construire, ils sont souvent inadéquats dans plusieurs applications. En effet, différentes factorisations conduisent à différents modèles, qui peuvent conduire à différentes estimations. L'autre difficulté de cette approche est que la factorisation ne prend pas correctement en compte les différentes échelles de mesure des variables. Dans le cas des données de grande dimension, les modèles de factorisation peuvent être très difficiles à mettre en oeuvre. Pour plus de discussion concernant ces problèmes, voir Teixeira-Pinto and Normand (2009).

Hormis les modèles obtenus par les méthodes directes, les modèles conjoints construits à partir de méthodes indirectes pour des données mixtes corrélées ont aussi été étudiés par de nombreux auteurs, voir par exemple, Gueorguieva and Agresti (2001), et Faes (2013), pour ne citer que ceux-là. Les modèles linéaires généralisés mixtes (GLMMs) sont des exemples de modèles indirects les plus utilisés. Ces modèles introduisent des effets aléatoires, soit identiques ou corrélés, pour accommoder l'association entre les données mixtes. Contrairement à la factorisation, cette approche permet un traitement symétrique entre les observations. En utilisant des effets aléatoires, les GLMMs incorporent dans l'analyse les effets spécifiques aux individus, intègrent la structure de corrélation entre les mesures longitudinales pour un même individu, ou entre les différentes observations d'une même grappe. Soit A un vecteur d'effets aléatoires, et $f_{X,Y|A}(x,y|a)$ la densité conjointe conditionnelle de X et Y, sachant A. La densité conjointe $f_{X,Y}(x,y)$ peut donc s'écrire comme suit :

$$f_{X,Y}(x,y) = \int f_{X,Y|A}(x,y|a) f_A(a) da.$$

On suppose l'indépendance conditionnelle entre X et Y, c'est à dire $f_{X,Y|A}(x,y|a) = f_{X|A}(x|a) f_{Y|A}(y|a)$. Des applications des GLMMs pour des données mixtes peuvent être trouvées dans (Gueorguieva and Agresti, 2001; Lin et al., 2010; Gueorguieva, 2013). Bien que très attractifs en pratique, les modèles à effets aléatoires ont aussi leurs limites qui les rendent inappropriées dans certaines situations, voir de Leon and Carrière Chough (2010); McCulloch (2008) pour plus de détails. L'utilisation des copules comme une stratégie indirecte de construction de modèles pour des données mixtes a beaucoup attiré l'attention des chercheurs ces dernières années. L'application des copules pour des données mixtes a été étudiée et discutée, par exemple dans de Leon and Wu (2011), Dobra and Lenkoski (2011), Song et al. (2009), Zimmer and Trivedi (2006), pour ne citer que ceux-là. L'utilisation des copules comme outils de modélisation de la dépendance pour des données mixtes discrètes et continues est un phénoméne récent. Pour plus de détails, voir la thèse de Beilei (2013).

1.9 Transition

Au chapitre suivant, nous présentons le cas le plus simple des modèles à effet aléatoire où aucune covariable n'est disponible. Nous en déduisons une procédure d'inférence pour l'estimation du coefficient de corrélation intra-grappe (ICC). Nous présentons la méthode d'estimation du maximum de vraisemblance et la méthode d'intervalle de confiance par vraisemblance profilée. Nous présentons des études de simulation pour mesurer la performance de la procédure que nous proposons. La méthode que nous proposons a aussi été testée sur des données réelles.

Chapitre 2

Some new random effect models for correlated binary responses

F. Tounkara and L.-P. Rivest.

Dependence Modeling, 2, 73-87, 2014.
Keywords : Multivariate exchangeable copulas, Exchangeable binary data, Profile interval, Maximum likelihood.

2.1 Résumé

Des copules échangeables sont utilisées pour modéliser une variation extra-binomiale dans une expérience de Bernoulli avec un nombre variable d'essais. Une procédure d'inférence basée sur la méthode du maximum de vraisemblance pour la corrélation intra-grappe est construite pour plusieurs familles de copules. La sélection d'un modèle particulier est effectuée en utilisant le critère d'information d'Akaike (AIC). Des intervalles de confiance profils sont construits pour la corrélation intra-grappe, et leurs performances sont évaluées dans une étude de simulation. La sensibilité de l'inférence à la spécification de la famille de copules est également étudiée par des simulations. Des exemples numériques sont présentés.

2.2 Abstract

Exchangeable copulas are used to model an extra-binomial variation in Bernoulli experiments with a variable number of trials. Maximum likelihood inference procedures for the intra-cluster correlation are constructed for several copula families. The selection of a particular model is carried out using the Akaike information criterion (AIC). Profile likelihood confidence intervals for the intra-cluster correlation are constructed and their performance are assessed in a simulation experiment. The sensitivity of the inference to the specification of the copula family is also investigated through simulations. Numerical examples are presented.

2.3 Introduction

Data in the form of clustered binary responses arise in many fields of study. For example, in ophthalmology the two eyes of a patient are a cluster. In toxicological studies Kuk (2004), a litter is a cluster of newborn animals. In education, the school and the classroom are clusters of students. Units within cluster are more likely to be similar than units from different clusters and the study variables often exhibit an intra-cluster correlation. Standard statistical methods that ignore such a correlation provide a poor fit of data and can lead to erroneous inferences. Such an extra-binomial variation can be accommodated by specifying a random cluster effect : the responses in a cluster are conditionally independent given the cluster effect. A common model is the beta-binomial, see Williams (1975), where the conditional probability of success in a cluster has a beta distribution. This model has been criticized for its lack of flexibility when the cluster sizes are variable Feng and Crizzle (1992). Other random effects models for clustered data include the logistic-normal-binomial model described by Williams (1982), and the probit-normal-binomial, see Ochi and Prentice (1984).

Several measures of association for binary traits have been proposed in the literature. These include the Kappa coefficient (Eldridge and Kerry, 2012, Chap. 8), the odds ratio (Lipsitz et al., 1991; Ananth and Preisser, 1999). This work focusses on the estimation of the "uncorrected" intra-cluster correlation (ICC). The ICC is a measure of the similitude of observations taken in the same cluster. It can be interpreted as an index of familial aggregation and as measure of inter-rater agreement, depending on the situation. When planning a cluster sampling or a cluster randomization trial, the clusters are the statistical units and the precision of the results depend on the ICC. Nowadays cluster randomized trials are widely used to assess various modes of community interventions. They involve assigning randomly clusters of persons, often belonging to the same neighborhood or group, to the arms of a trial. In such situations the ICC plays a crucial role in the sample size determination, as it measures the homogeneity within clusters Eldridge and Kerry (2012). ICC estimates are now published in the epidemiological literature for various types of intervention, see Pals et al. (2009) for a recent example. It is therefore important to have precise ICC estimators with reliable methods for constructing confidence intervals.

The estimation of the ICC for binary data has been considered by many authors. Ridout et al. (1999) compare more than 20 estimators of the ICC, derived using either maximum likelihood or a modified method of moments, in a Monte Carlo study. They conclude that the estimator of Fleiss and Cuzick (1979) can be used as an omnibus estimator as it generally has good sampling properties. Recently, the construction of confidence intervals for the ICC has been investigated by (Zou and Donner, 2004; Chakraborty et al., 2009; Saha, 2012). The first modeled the extra binomial variation using Madsen (1993) generalized binomial distribution, Chakraborty et al. (2009) use a z-transform and a distribution-free large sample variance to construct an ICC confidence interval while Saha (2012) assumes a beta-binomial distribution. Saha (2012) also compared six methods for constructing confidence intervals for the ICC and found that, for beta-binomial data, the best procedure was a profile likelihood confidence interval (PLCI). Note that inference about the ICC using Bayesian framework have also been investigated, see (Turner et al., 2001, 2006; Ahmed and Shoukri, 2010).

This work investigates the construction of PLCI for the ICC using a new class of probability models for exchangeable clustered binary data, associated with exchangeable Archimedean copulas (Mai and Scherer, 2012, chap. 2). This family contains several specifications for the extra binomial variation. The procedure investigated in this work consists in (i) selecting a model for the random cluster effect and (ii) constructing a profile confidence interval for the ICC using the model selected at step 1.

Section 2.4 introduces a class of model for an extra binomial variation associated with multivariate Archimedean copulas. Graphical models' comparisons are provided. Section 2.5 discusses model selection, the calculation of maximum likelihood estimates for the ICC, and the construction of profile likelihood confidence intervals. These new statistical methods are investigated through simulations in Section 2.6. Two numerical examples are used to illustrate the proposed approach in Section 2.7.

2.4 Within cluster dependency modeling for binary exchangeable data

2.4.1 Model formulation

Consider a random sample of K clusters of size $n_i,\ i = 1,2,\ldots,K$. Let $Y_{ij}, j = 1,2,\ldots,n_i$, be a set of n_i exchangeable Bernoulli random variables in cluster i, where $Y_{ij} = 1$ is a success and 0 is a failure. These random variables are identically distributed and possibly dependent. The total number of successes in the i^{th} cluster is $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. A standard model for the within cluster dependency assumes that the probability of success p_i is random and varies from one cluster to the next. The models considered here associate to cluster i a positive random variable θ_i whose Laplace transform, $\psi_\alpha(t) = E(e^{-t\theta_i})$, depends on a positive dependency parameter α is a such way that $P(\theta_i = 1) = 1$ when $\alpha = 0$. Let $\pi \in (0,1)$ stands for the marginal probability of success ; the proposed model is

$$p_i = e^{-\theta_i \psi_\alpha^{-1}(\pi)} \quad i = 1,\ldots,K, \tag{2.1}$$

where $\psi_\alpha^{-1}(\pi)$ denotes the inverse function of the Laplace transform $\psi_\alpha(\cdot)$ evaluated at π. Under model (2.1), $p_i \in (0,1)$ and the marginal probability of success is $E(p_i) = \pi$. This model has a clear separation between π and the random variable for the dependency, θ_i. Indeed, $\psi_\alpha^{-1}(\pi)$ is a mere scale parameter for the distribution of $-\log p_i$. Model (2.1) assumes exchangeability within each cluster. This assumption could be questionable for familial data where the mother-child association could be stronger than that between two siblings.

Table 2.1 gives the inverse of the Laplace transforms $\psi_\alpha^{-1}(t)$ and the densities $f_\alpha(\theta)$ of the random effects θ_i for four models commonly used in the copula literature. For Gumbel (G) and Clayton (C) families, f_α is a density defined on $\mathscr{R}^+$ while, for the models of Frank (F) and Joe (J), f_α is a probability mass function defined on the positive integers $\mathscr{N}^+$. For the models in (2.1), the k-moments $E(p_i^k)$ have the following form, $\lambda_k = \psi_\alpha\{k\psi_\alpha^{-1}(\pi)\}$; explicit expressions for these moments are given in Table 2.1. It follows that, given π and the association parameter α, the moment based definition of the intra-cluster correlation (ICC) ρ under models (2.1) is given by

$$\rho = \frac{E(p_i^2) - \{E(p_i)\}^2}{E(p_i)\{1 - E(p_i)\}} = \frac{\psi_\alpha(2\psi_\alpha^{-1}(\pi)) - \pi^2}{\pi(1-\pi)}. \tag{2.2}$$

TABLE 2.1 – Inverse of Laplace transforms, the densities of the random effects θ_i, and moments for four copula families, Clayton (C), Frank (F), Gumbel (G) and Joe (J).

Families	$\psi_\alpha^{-1}(t)$	$f_\alpha(\theta)$	λ_k
C	$(t^{-\alpha}-1)/\alpha$	$\frac{\theta^{1/\alpha-1}\exp(-\theta/\alpha)}{\alpha^{1/\alpha}\Gamma(1/\alpha)}$	$\left\{k\pi^{-\alpha}-k+1\right\}^{-\frac{1}{\alpha}}$
F	$-\log\left(\dfrac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$	$(1-e^{-\alpha})^\theta/(\theta\alpha)$	$-\frac{1}{\alpha}\log\left\{1-\frac{(1-e^{-\alpha\pi})^k}{(1-e^{-\alpha})^{k-1}}\right\}$
G	$\left\{-\log(t)\right\}^{\alpha+1}$	Positive Stable	$\pi^{k^{\frac{1}{1+\alpha}}}$
J	$-\log\{1-(1-t)^{\alpha+1}\}$	$\binom{\frac{1}{\alpha}}{\theta}(-1)^{\theta-1}$	$1-\{1-(1-(1-\pi)^{1+\alpha})^k\}^{\frac{1}{1+\alpha}}$

The ICC is the correlation between different Bernoulli variables, Y_{is} and $Y_{i\ell}$, of cluster i. It is an increasing function of α, with $\rho=0$ when $\alpha=0$ for all models in Table 2.1. Further interpretations of this coefficient can be found in (Eldridge and Kerry, 2012, Chap.8).

For Gumbel's model, the random effect θ_i has a positive stable distribution with parameter $1/(\alpha+1)$ and $f_\alpha(\theta)$ does not have a closed form expression. For this model, the moments λ_k are equal to those for the q model of Kuk (2004). This provides an additional motivation for Kuk's proposal as it can be derived using stable random effects in (2.1). Consider now Clayton's family. When $\alpha=1$, θ_i has a negative exponential distribution with parameter 1, and p_i has a beta distribution with parameters $(\psi_\alpha^{-1}(\pi),1)$. Therefore Clayton's model with $\alpha=1$ gives the beta-binomial distribution with parameters $(\psi_\alpha^{-1}(\pi),1)$. For Frank's and Joe's model, θ_i is an integer valued random variable defined on $\mathscr{N}^+$. For Frank's copula, θ_i follows a logarithmic distribution, while for Joe's family it is the Sibuya distribution. The later distribution is heavy tailed as no moment exists. See Mai and Scherer (2012) for more information about these distributions. It might seem unrealistic to have latent cluster effects distributed according to a positive stable law or to a discrete distribution. These assumptions are made to get, within (2.1), tractable models with probability mass function (pmf) having closed form expressions.

Swapping the successes and the failures, (2.1) gives new models for an extra binomial variation. The probability of success is then expressed as

$$p_i = 1-e^{-\theta_i\psi_\alpha^{-1}(1-\pi)} \quad i=1,\ldots,K. \tag{2.3}$$

This duality was noted by Kuk (2004) whose p model can be expressed as (2.3) where θ_i has a positive stable distribution. In the sequel, models obtained with (2.3) will be called dmodel since, as will be seen in the next section, they arise when the joint distribution of the Bernoulli variables Y_{ij} is expressed using a copula. In this paper, Kuk's p model is labeled the dGumbel model. For (2.3), the intra cluster correlation is $\rho=\{\psi_\alpha(2\psi_\alpha^{-1}(1-\pi))-(1-\pi)^2\}/\{\pi(1-\pi)\}$.

Two approaches are available to calculate the pmf of Y_i . The first method evaluates alternating sums directly using the moments λ_k given in Table 2.1 :

$$P\big(Y_i = k\big) = \binom{n_i}{k} E\{p_i^k (1-p_i)^{n_i-k}\} \tag{2.4}$$

$$= \binom{n_i}{k} \sum_{j=0}^{n_i-k} (-1)^j \binom{n_i-k}{j} \lambda_{k+j}. \tag{2.5}$$

For large clusters, say larger than 30, the evaluation of the alternating sums (2.5) is numerically unstable and yields negative probabilities. The second method to calculate the pmf of Y_i is to evaluate (2.4) directly by integrating over the random effects distribution. For the random effects having gamma distributions of Clayton's family, we tried using the Gauss-Laguerre quadrature method. It gave poor approximations for large values of the dependence parameter α. Integrating over the random effect is possible for the models of Joe and Frank as they have discrete distributions. For these models,

$$P(Y_i = k) \approx \begin{cases} \left[1 - \sum_{\theta=1}^{N} \{1 - (1 - e^{-\theta\psi_\alpha^{-1}(\pi)})^{n_i}\} f_\alpha(\theta)\right] & \text{if } k = 0 \\ \\ \binom{n_i}{k} \sum_{\theta=1}^{N} e^{-\theta\psi_\alpha^{-1}(\pi)k} \left\{1 - e^{-\theta\psi_\alpha^{-1}(\pi)}\right\}^{n_i-k} f_\alpha(\theta) & \text{if } 0 < k \leq n_i \end{cases} \tag{2.6}$$

where N is a positive integer large enough for these approximations to be accurate. The values $N = 100$ and $N = 400$ yielded nearly identical results in the small cluster simulation that are presented in Section 4. We used $N = 400$ in the large cluster simulations presented in the next section.

Note that binomial mixture models constructed in terms of Laplace transforms can be found in (Alanko and Duffy, 1996; Kuk, 2004). The innovative aspect of our paper is to parameterize these models in terms of a marginal probability π and the ICC ρ and to investigate this large class of models in a systematic way.

2.4.2 A copula formulation of the model

This section derives an alternative formulation for (2.1), in terms of the joint survival distribution of the Bernoulli random variables $Y_{ij}, j = 1, 2, \ldots, n_i$, for the units within a cluster. Dropping subscript i, let $\{y_j\}$ be a sequence of n zeros and ones. Under (2.1) the joint survival distribution of the Y_j's within a cluster is

$$P(Y_1 \geq y_1, \ldots, Y_n \geq y_n) = C_{\alpha,n}\{\bar{F}(y_1), \ldots, \bar{F}(y_n)\}, \tag{2.7}$$

where $\bar{F}(y)$ is the marginal survival function of Y_j, $\bar{F}(y) = P(Y \geq y) = \pi^y$ for $y = 0, 1$, and $C_{\alpha,n}(u_1, \ldots, u_n) = \psi_\alpha\{\sum_{j=1}^{n} \psi_\alpha^{-1}(u_j)\}$ is, for $u_j \in (0,1)$ $j = 1, \ldots, n$, an n dimensional Archimedean copula for the dependency between the variables, see (Mai and Scherer, 2012, Chap. 2) for more discussion. Note that the inverse $\psi_\alpha^{-1}(\cdot)$ of ψ_α is a decreasing function defined on $(0,1)$ such that $\psi_\alpha^{-1}(1) = 0$. Under (2.7), $\lambda_k = P(Y_1 \geq 1, \ldots, Y_k \geq 1) = C_{\alpha,k}\{\pi, \ldots, \pi\} = \psi_\alpha\{k\psi_\alpha^{-1}(\pi)\}$.

Equation (2.7) is proved by noticing that, assuming (2.1),

$$\begin{aligned}P(Y_1 \geq y_1,\ldots,Y_n \geq y_n) &= E\{P(Y_1 \geq y_1,\ldots,Y_n \geq y_n|\theta)\}\\ &= E[\{\exp\{-\theta \sum y_j \psi_\alpha^{-1}(\pi)\}]\\ &= \psi_\alpha\{\sum y_j \psi_\alpha^{-1}(\pi)\}\\ &= \psi_\alpha[\sum \psi_\alpha^{-1}\{\bar{F}(y_j)\}]\end{aligned}$$

since, by construction, $y_j \psi_\alpha^{-1}(\pi) = \psi_\alpha^{-1}\{\bar{F}(y_j)\}$.

The value $\alpha = 0$ gives the independence copula, $C_{0,n}(u_1,\ldots,u_n) = u_1 \times \ldots \times u_n$. As α goes to ∞ the copulas for all the models of Table 1 converge to what is known as the Fréchet upper bound,

$$\lim_{\alpha\to\infty} \psi_\alpha\{\sum \psi_\alpha^{-1}(u_j)\} = \min_j(u_j).$$

This corresponds to the perfect correlation where all the Bernoulli variables within a cluster takes the same value, e. g. $P(Y_1 = \ldots = Y_n) = 1$. An attractive feature of all models in Table 1 is that (π,ρ) has the full range $]0,1[\times]0,1[$, as $\pi \in (0,1)$ and $\alpha \in (0,\infty)$. Moreover, ρ is a monotone function of α satisfying the following properties

$$\alpha \longrightarrow \infty \Longleftrightarrow \rho \longrightarrow 1 \ \text{ and } \ \alpha = 0 \Longleftrightarrow \rho = 0.$$

There is a copula behind most models for clustered binary data. Consider for instance the generalized binomial distribution, see Madsen (1993). Under this model the joint survival distribution within a cluster can be expressed as (5.9) with

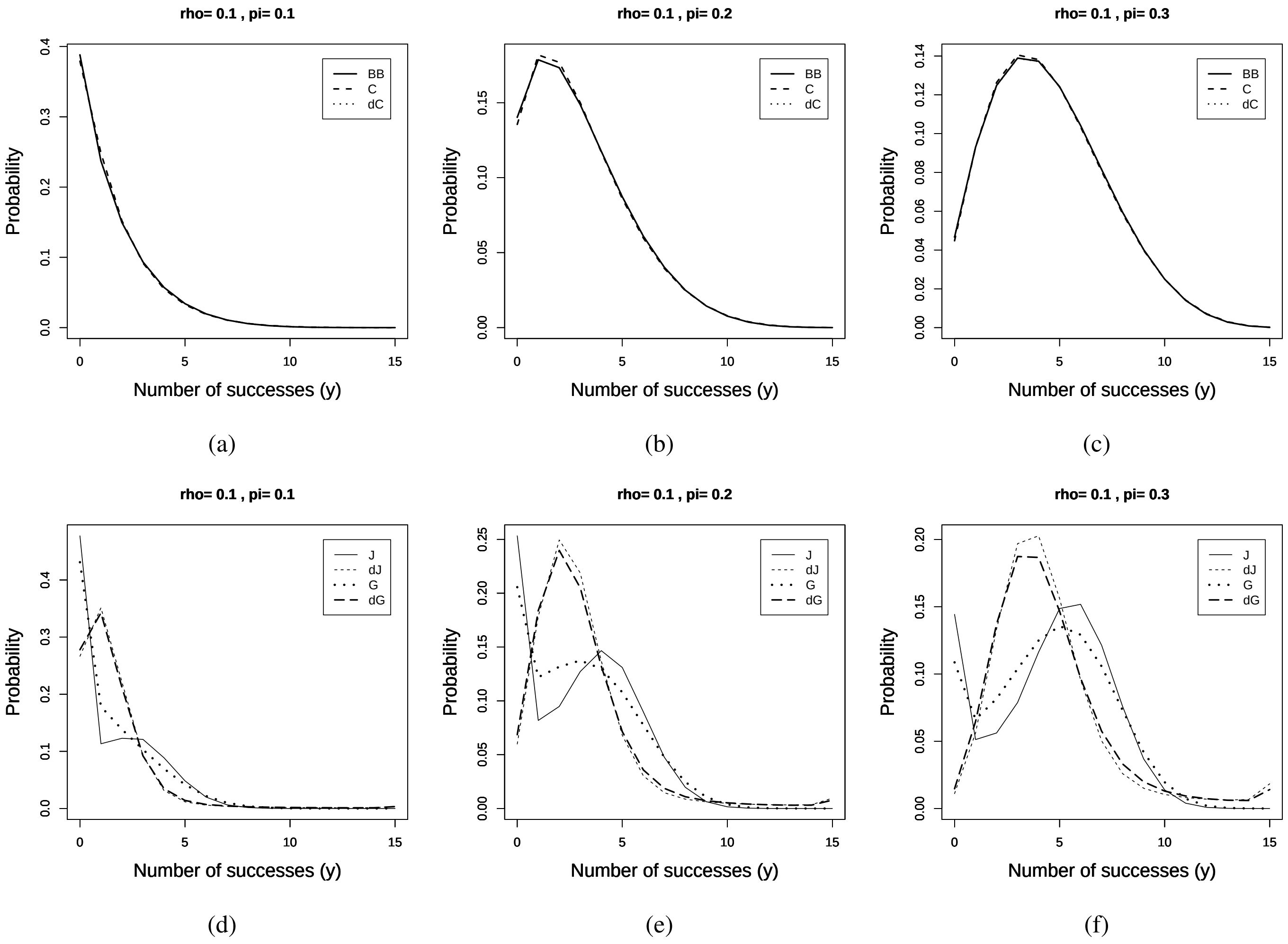
$$C_\alpha(u_1,\ldots,u_n) = (1-\alpha)\prod_{j=1}^{n} u_j + \alpha \min(u_j), \quad \alpha \in (0,1).$$

This expression highlights that the generalized binomial distribution is a discrete mixture model featuring two latent classes with respective probabilities $(1-\alpha)$ and α. The clusters in the first class are made of independent units while in the second class all the units of a cluster have the same Y values. Such a model is rather unlikely in practice and one would expect (2.1) to represent better the within cluster association. Despite its unrealistic description of the dependency, the generalized binomial has been widely used to investigate statistical procedures for cluster binary data, see (Ridout et al., 1999; Zou and Donner, 2004).

Finally note that the estimation of the intra cluster correlation for discrete data using copulas was investigated by Shoukri et al. (2011) in a different context. They were concerned with split-cluster designs, where individuals within a cluster are randomized to one of the two arms of a trial. They use bivariate copulas to model the association between the mean responses for the treated and for the untreated individuals in a cluster.

2.4.3 Graphical comparisons

A total of 8 Archimedean copulas models are available to account for an extra binomial variation. They were compared using exploratory analysis techniques such as correspondence analysis. The findings of these comparisons are now briefly presented. First the beta-binomial, the Clayton and the dClayton (dC) models are, for all practical purposes, equal. This is illustrated in (a), (b), and (c) of Figure 2.1 where the pmf for these three models obtained when $n = 15$, $\pi = 0.1, 0.2, 0.3$ and $\rho = 0.1$ are presented. Similar results are found for $\rho = 0.3$; they are presented in the supplementary material section. This means that when the beta-binomial is a candidate model for a data set, there is no need to include the Clayton and dClayton model as they provide fits that are very similar.



(a)

(b)

(c)

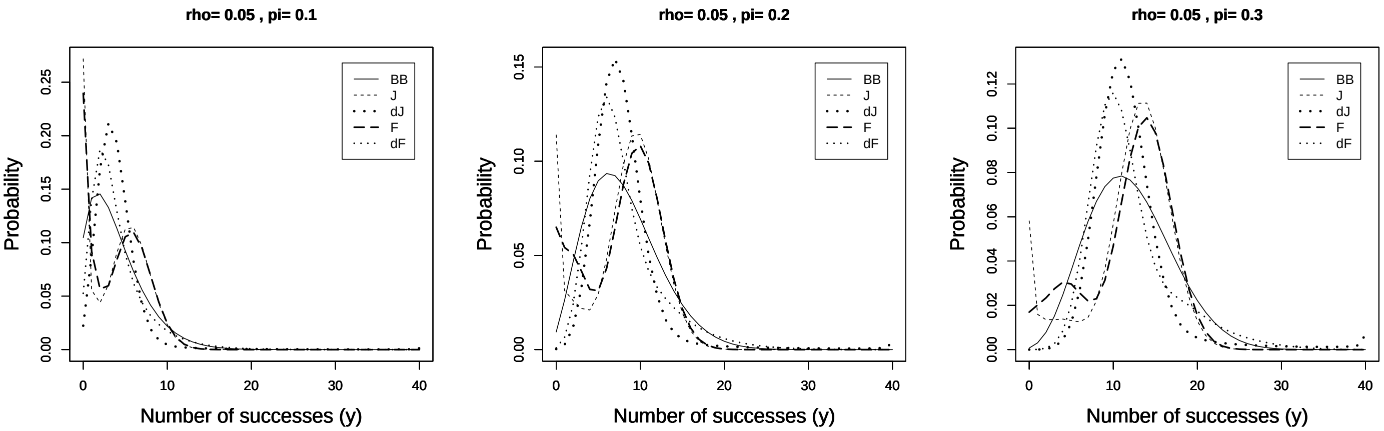
(d)

(e)

(f)

FIGURE 2.1 – A comparison of the probability mass functions for cluster size 15 under some small clusters models, for $\rho = 0.1$ and $\pi = 0.1, 0.2, 0.3$

Graphs (d), (e) and (f) of Figure 2.1 compare the pmf for Joe (J), dJoe (dJ), Gumbel (G) and dGumbel (dG) models. The dJoe and dGumbel pmfs have nearly identical unimodal shapes while of the Gumbel and the Joe pmfs are similar with a bimodal structure. The pmfs of Figure 2.1 have a wide range of values for



(a) (b) (c)

FIGURE 2.2 – A comparison of the probability mass functions for clusters of size 40 under the beta-binomial (BB), Joe (J), dJoe (dJ), Frank (F) and d-Frank (dIG), for $\rho = 0.05$ and $\pi = 0.1(a), 0.2(b), 0.3(b)$

$Pr(Y_i = 0)$, when $\pi = 0.2$ it goes from 0.05 to 0.25. Additional comparisons, with $\rho = 0.3$ are presented in Web Appendix A in the supplementary material of Chapter 2. In all comparisons, the Joe and the dJoe pmf presented the most extreme shapes, while the beta-binomial pmf has an average shape, between these two extremes. Therefore, in the small cluster simulations presented in the next section, only these three models are considered as being representative of all the models of Table 2.1.

Figure 2.2 presents some pmf comparisons for clusters of size $n = 40$, with $\rho = 0.05$ and $\pi = 0.1, 0.2, 0.3$. Such small values of ρ are typical in community intervention studies, see (Légaré et al., 2011). The models presented are the beta-binomial (BB), Frank (F), dFrank (dF), Joe (J) and dJoe (dJ). Figures 2.2 shows that the pmf for all the models are different, except possibly for the dFrank and dJoe models, whose shape are very similar. The pmf shows a unimodal structure for the models of dFrank and dJoe, and a bimodal structure under F's and J's models. Joe's pmf assigns a large probability to zero as compared with the others models and the beta-binomial pmf has an average shape between that of the Joe and the dJoe models. Graphs for $\rho = 0.1$ are presented in the supplementary material chapter.

2.5 Maximum likelihood inference for ICC(ρ)

This section shows how to calculate maximum likelihood estimates for ρ and π using data $\{(n_i, y_i) : i = 1, \ldots, K\}$. It also discusses the construction of profile likelihood confidence intervals for ρ. Inference on ρ is carried out in a frequentist rather than a Bayesian set-up as this is commonly used in research on clustered randomized trials (Eldridge & Kerry, 2012, p.193).

2.5.1 Maximun Likelihood Estimates

The parameters of interest are ρ and π, thus α needs to be expressed in terms of these two parameters. The only model for which equation (2.2) leads to an explicit expression for α is Gumbel's families for which

$$\alpha = \frac{\log(2)}{\log\left[1+\frac{\log\{\rho(1-\pi)+\pi\}}{\log\pi}\right]} - 1.$$

For all the other models of Table 1, solving (2.2) numerically is needed to calculate α ; this can be done using the `R` function `uniroot`. The log-likelihood is

$$\log L(\pi,\rho) = \log\left\{\prod_{i=1}^{K}\binom{n_i}{y_i}E\{p_i^{y_i}(1-p_i)^{n_i-y_i}\}\right\}. \tag{2.8}$$

It can be evaluated using one of the two methods presented in Section 2.4.1.

The maximum likelihood estimates are the values $\hat{\pi}$ and $\hat{\rho}$ that maximize the log-likelihood defined in (2.8). They were calculated using `R`. First an R function evaluating (2.8) in terms of ρ and π was written where α was evaluated in terms of ρ and π by solving (2.2) using `uniroot`. The search interval for α was set to $(10^{-5}, 10^{10})$; for some copulas (2.2) cannot be evaluated for α values as large as 10^{10} and this generates a `uniroot` warning. This had no impact on the result since `uniroot` automatically lowers the upper bound of the search interval when this happens. Then the values of π and ρ that maximized the likelihood were obtained using the function `optim` of `R`. We used as starting values, the sample mean $\hat{\pi}_0 = \sum y_i / \sum n_i$ and $\hat{\rho}_{FC}$, the Fleiss Cuzick moment based estimate of ρ defined in (2.12), see the next section. The log-likelihood is maximized in the square $[1e^{-5}, .99999] \times [1e^{-5}, .99999]$. The Fisher information matrix was then estimated as minus the Hessian of the likelihood function ; this yields standard errors, *s.e.*, for the two estimates. Negative estimates of ρ are not possible. Negative estimates of ρ have little practical interest, indeed (Eldridge and Kerry, 2012, Chapter 8) recommend setting them to zero and proceeding as the data with cluster was independent.

2.5.2 Model selection criteria

A model's fit can be assessed using a deviance, equal to the likelihood ratio statistics comparing the copula model to that of a saturated model where each cluster has its own fixed probability of success p_i,

$$Dev = 2[\log\{\sum_{i=1}^{K} \binom{n_i}{y_i} y_i^{y_i}(n_i - y_i)^{n_i - y_i}/n_i^{n_i}\} - \log L(\hat{\pi}, \hat{\rho})]. \tag{2.9}$$

Note, however, that this deviance does not have an asymptotic χ^2_{K-2} distribution when the copula model fits well as the null copula model is not a special case of the saturated model with cluster specific p_i when K is fixed.

To select a specific model to work with, we used the Akaike Information Criterion (AIC), defined by

$$AIC = -2\log L(\hat{\pi}, \hat{\rho}) + 2m, \tag{2.10}$$

where m, the number of parameters, is equal to 2 for all the copula models considered in this work. The preferred model is the one with the minimum AIC value. We did not use the model averaging approach of Nikoloulopoulos and Karlis (2008) as this does not permit the calculation of profile confidence intervals.

2.5.3 Profile Likelihood Confidence Interval (PLCI)

Wald confidence intervals for ρ, with a $100(1-\tau)$ confidence level, are given by $\hat{\rho} \pm z_{1-\tau/2} s.e.(\hat{\rho})$, where $z_{1-\tau/2}$ is a normal quantile. This procedure may perform poorly for small to moderate sample sizes when estimating ρ, see Donner and Eliasziw (1992). This section reviews the construction of PLCI for ρ when using copula models associated to (2.1). Such intervals are generally considered to be more accurate than Wald confidence intervals. The profile likelihood for ρ is given by $L_P(\rho) = \max_\pi L(\pi, \rho)$, $\rho \in (1e^{-5}, .99999)$. We used the R function `optimize` for its evaluation, it involved a `uniroot` function to evaluate α as a function of ρ and π. It may happen that $L_P(\rho)$ cannot be numerically evaluated for large values of ρ ; to calculate the profile confidence interval it is important to identify a value ρ_M such that $L_P(\rho)$ can be evaluated for $\rho \in (1e^{-5}, \rho_M)$. The $100(1-\tau)\%$ PLCI for ρ is defined as the set

$$CI = \Big\{\rho : \quad \log L_P(\rho) > \log L(\hat{\rho}, \hat{\pi}) - \frac{1}{2}\chi^2_1(1-\tau)\Big\}, \tag{2.11}$$

where $\chi^2_1(1-\tau)$ is the $(1-\tau)$ quantile of a chi-squared distribution with one degree of freedom. The set CI is typically an interval and is denoted $(\hat{\rho}_L, \hat{\rho}_U)$. We computed the two limits by finding two solutions to the equation $\log L_P(\rho) = \log L(\hat{\rho}, \hat{\pi}) - \frac{1}{2}\chi^2_1(1-\tau)$, one in the interval $(0, \hat{\rho})$ and the other in the interval $(\hat{\rho}, \rho_M)$, using the R function `uniroot`. When $\log L(\hat{\rho}, \hat{\pi}) - \frac{1}{2}\chi^2_1(1-\tau) < \log L_P(1e^{-5})$, the interval lower bound was set at $\hat{\rho}_L = 0$. If ρ_M is not specified correctly then $\log L_P(\rho_M)$ cannot be evaluated ; in such cases `uniroot` returns $\hat{\rho}$ as the value of the confidence interval upper bound. This is not a legitimate upper bound and samples for which $\hat{\rho}_U = \hat{\rho}$ were rejected in the simulations, such unacceptable upper bounds occurred only in the large cluster simulations. For a quick evaluation of $(\hat{\rho}_L, \hat{\rho}_U)$ the R function `plkhci` could also be used. When the true value of ρ is close to 0, the lower bound $\hat{\rho}_L$ is likely to be 0 and in this case the true coverage of the confidence might be slightly above its nominal level.

PLCI for ρ, constructed using the beta-binomial distribution, were investigated by Saha (2012). He concluded through Monte Carlo simulations that the profile confidence interval performs well over a wide range of parameter values. A goal of this work is to investigate whether the beta-binomial profile confidence interval is an omnibus procedure applicable in all circumstances or whether the fit of the beta-binomial model should first be ascertained.

2.6 Simulation study

The purpose of this simulation study is to assess whether the AIC criterion (2.10) provides a reliable model selection criterion. Its goal is also to assess the performance of the profile likelihood confidence intervals for ρ constructed for the models of Table 1 . Two sets of simulation were carried out. The first featured small clusters, with sizes ranging between 1 and 15 with a mean of 4. The second one had large clusters with n_i between 27 to 65 with a mean of 38. Data was generated from the beta-binomial distribution, the Joe, dJoe, frank and dFrank models. For J's and dJ's models, cluster specific θ_i were generated from the Sibuya distribution while for the Frank and dFrank model, the logarithmic distribution was used. The accuracy of the AIC model selection criterion was measured using the proportions of times that model M_0 is selected when the data was simulated from model M_1 :

$$minAIC(M_0, M_1) = \frac{\sum_{i=1}^{1000} I_i}{1000} \times 100,$$

where the indicator I_i equal to 1 if model M_0 provides the smallest AIC for sample i and 0 otherwise.
The simulations also investigated 95% two-sided confidence intervals for ρ. Their performance was measured using the coverage (cov) and the average length (CIL) defined as

$$Cov = \frac{\sum_{i=1}^{1000} A_i}{1000} \times 100; \qquad CIL = \frac{\sum_{i=1}^{1000} (\hat{\rho}_{U_i} - \hat{\rho}_{L_i})}{1000},$$

where A_i equal to 1 if the true value ρ is in $(\hat{\rho}_{L_i}, \hat{\rho}_{U_i})$, and 0 otherwise. All programming was done in R. Samples with $y_i = 0$ or $y_i = n_i$ for all clusters were rejected.

2.6.1 Small cluster simulations

Three models are investigated in the small cluster simulations, beta-binomial, Joe and dJoe. Two values of ρ and π, namely 0.1 and 0.3, and three values of K, 100, 250, and 500, were considered. Table 2.2 summarizes the proportion of times that a model is selected as a function of the model from which the data is simulated. The relatively high values on the diagonals of each quadrant of Table 2.2 show that the models are identifiable. This is more pronounced, as either K, π or ρ increases. When a data set is generated from Joe's model, dJoe is nearly never selected and vice-versa. These results are in line with the findings of the graphical comparisons of Figures 2.1 where these two models were very different. Table 2.2 also confirms the middle ground nature of the beta-binomial since this model has the smallest number of correct identifications.

Table 2.3 reports results on PLCI that assume that the model is selected correctly. It evaluates PLCI for ρ under the beta-binomial, the Joe and the dJoe models. The PLCI has generally good statistical properties ; its

TABLE 2.2 – Small cluster simulations : The proportion of times that a model is selected by the AIC criterion as a function of the model used to simulate the data.

			$\pi = 0.1$			π =0.3		
				Data			Data	
ρ	K	Models	BB	J	dJ	BB	J	dJ
0.1	100	BB	42.50	21.30	20.90	55.50	18.40	15.60
		J	43.50	74.20	14.70	27.90	79.80	5.90
		dJ	14.00	4.50	64.40	16.60	1.80	78.50
	250	BB	58.60	16.80	13.90	79.50	11.10	7.00
		J	33.60	82.80	4.90	15.00	88.90	1.00
		dJ	7.80	0.40	81.20	5.50	0.00	92.00
	500	BB	75.50	13.20	7.70	92.40	4.80	2.00
		J	22.20	86.80	0.80	5.40	95.20	0.00
		dJ	2.30	0.00	91.50	2.20	0.00	98.00
0.3	100	BB	54.70	16.60	15.70	79.00	8.50	9.50
		J	33.30	83.00	3.40	14.40	91.50	0.20
		dJ	12.00	0.40	80.90	6.60	0.00	90.30
	250	BB	79.10	10.10	6.70	94.80	2.60	0.80
		J	16.60	89.90	0.10	3.80	97.40	0.00
		dJ	4.30	0.00	93.20	1.40	0.00	99.20
	500	BB	93.20	4.70	1.30	99.80	0.10	0.00
		J	5.90	95.30	0.00	0.20	99.90	0.00
		dJ	0.90	0.00	98.70	0.00	0.00	100.00

coverage is close to nominal 95% level and its average length is relatively small. As expected, the expected confidence length decreases as the number of clusters increases. The PLCI for Joe's model is always shorter than the other two, with a very good empirical coverage. Thus PLCI constructed using the three models investigated in this section are reliable statistical techniques.

Additional simulation results are presented in Web Appendix B in the Supplementary Material of Chapter 2. First the coverage and the average length of Zou and Donner (2004) ICC confidence interval constructed using the moment estimator of Fleiss and Cuzick (1979),

$$\hat{\rho}_{FC} = 1 - \frac{\sum y_i(n_i - y_i)/n_i}{(\sum n_i - K)\hat{\pi}(1-\hat{\pi})} \tag{2.12}$$

are reported together with the performance of the beta-binomial PLCI when the model is misspecified, that is when the data comes from the Joe and the dJoe model (see Web Table 2.1 in the "Web Appendix B", in the Supplementary Material of Chapitre 2). The Zou & Donner (2004) confidence interval is very conservative as it is always much wider than the PLCI. These simulations also show that the PLCI for the beta-binomial is sensitive to a model misidentification. For instance for dJ data, $K = 250$, $\pi = 0.1$, and $\rho = 0.1$, the estimated coverage of the beta-binomial PLCI, 76.70%, is much smaller than the nominal value of 95%. However, when data are from J's model this method in general has a tendency to provide the observed coverage probabilities that are slightly larger than the nominal level.

TABLE 2.3 – Small cluster simulations : Confidence interval length (CIL) and empirical coverage (COV) of two-sided PLCI for ρ with a nominal 95% confidence level under beta-binomial, Joe, and dJoe models

			$\rho = 0.1$				ρ =0.3			
			$\pi = 0.1$		π =0.3		$\pi = 0.1$		π =0.3	
	Data	Models	CIL	Cov	CIL	Cov	CIL	Cov	CIL	Cov
$K = 100$	BB	BB	0.22	95.30	0.17	93.90	0.33	95.20	0.23	95.60
	J	J	0.18	93.80	0.15	95.40	0.27	93.90	0.21	94.00
	dJ	dJ	0.27	95.90	0.19	93.70	0.38	95.29	0.26	95.40
$K = 250$	BB	BB	0.14	95.90	0.12	95.00	0.21	94.20	0.15	93.00
	J	J	0.11	94.50	0.10	94.40	0.17	95.10	0.13	94.10
	dJ	dJ	0.19	94.60	0.13	94.30	0.26	94.70	0.17	96.20

2.6.2 Large cluster simulations

We ran large cluster simulations with K =10, 25, and 50 clusters. Two values of ρ, 0.05 and 0.1, and of π, 0.1 and 0.3, were investigated. Besides the beta-binomial model, the models of Frank, dFrank, Joe, and dJoe were considered. Table 2.4 shows that the AIC has a hard time identifying the models when $K = 10$. The proximity, in Figure 2, of the pmf for Joe's and Frank's model shows up in Table 2.4 where the AIC has difficulties distinguishing one from the other. This improves as K increases and at $K = 50$ most models are identifiable. The identifiability increases with π. Results in Tables 2.5 show that, when the model is correctly identified, the PLCI constructed with the beta-binomial, the Joe, the dJoe, the Frank and the dFrank models have good coverage properties, even with a relatively low ($K = 10$) number of clusters. As expected, the average confidence interval length decreases with the number of clusters. In general, the Joe and the Frank PLCI provide smaller expected lengths and an empirical coverage close to the nominal values

In Web Table 2.2 of "Web Appendix B", in the Supplementary Material of chapitre 2, we reported confidence interval length (CIL) and empirical coverage (COV) of Zou and Donner's interval and of the beta-binomial PLCI, when the data comes from J's, dJ's F's and dF's distributions. We can see that Zou and Donner's method is very conservative. On the other hand, the PLCI based on beta-binomial model provides coverage below the nominal level when the model is misspecified. Thus using an omnibus PLCI based on the beta-binomial distribution is not a robust statistical procedure.

2.7 Examples

In this section, we present two real life examples to illustrate the performance of estimators obtained under (2.1) and (2.3).

2.7.1 A pilot trial about shared decision making

This example is concerned with the impact of training physicians in shared decision making on their patients' involvement in the decision making process, see Légaré et al. (2011) for more details. The data in Table 2.7 presents the number of patients y_i reporting an active role in the decision about taking an antibiotics treatment for an acute respiratory infection, and the total patient enrollment n_i in $K = 9$ community health services of

TABLE 2.4 – Large cluster simulations : The proportion of times that a model is selected by the AIC criterion as a function of the model used to simulate the data.

			$\pi = 0.1$					π =0.3				
			Data					Data				
ρ	K	Models	BB	J	dJ	F	dF	BB	J	dJ	F	dF
0.05	10	BB	40.85	5.90	32.86	8.77	34.38	54.25	11.58	28.93	19.22	27.78
		J	10.58	63.07	11.43	43.93	6.16	0.94	50.60	0.00	14.41	0.23
		dJ	1.48	0.00	20.20	0.10	8.01	1.15	0.00	30.91	0.00	6.94
		F	19.68	30.62	4.69	45.67	6.97	17.94	36.16	4.46	64.92	3.12
		dF	27.41	0.41	30.82	1.53	44.48	25.71	1.67	35.70	1.45	61.92
	25	BB	62.53	1.50	20.76	4.70	33.40	74.15	2.13	14.80	4.74	15.92
		J	3.71	68.20	4.89	36.64	1.42	0.40	72.44	0.00	19.17	0.00
		dJ	0.60	0.00	39.24	0.00	9.24	0.50	0.00	54.50	0.00	8.16
		F	11.02	30.30	4.58	58.66	1.93	9.62	25.33	1.60	75.98	0.41
		dF	22.14	0.00	30.53	0.00	54.01	15.33	0.10	29.10	0.10	75.51
	50	BB	80.10	0.00	12.77	1.00	24.95	90.40	0.10	5.41	1.30	8.50
		J	0.20	72.70	2.81	28.30	0.00	0.20	85.90	0.11	9.10	0.00
		dJ	0.10	0.00	55.81	0.00	5.01	0.10	0.00	73.41	0.00	5.40
		F	3.40	27.30	3.07	70.70	0.10	2.20	14.00	0.97	89.60	0.00
		dF	16.20	0.00	25.54	0.00	69.94	7.10	0.00	20.11	0.00	86.10
0.1	10	BB	38.60	4.10	23.95	7.11	32.42	48.64	4.66	18.16	12.29	24.01
		J	11.45	63.20	8.41	42.28	4.00	1.51	60.10	0.00	22.46	0.00
		dJ	3.14	0.00	30.10	0.30	11.05	1.01	0.00	41.84	0.00	8.73
		F	23.51	31.90	4.85	48.00	11.05	20.95	34.72	2.37	63.24	2.81
		dF	23.30	0.80	32.69	2.30	41.47	27.90	0.52	37.63	2.01	64.45
	25	BB	67.20	0.90	13.83	3.70	30.33	79.70	0.50	5.24	3.60	12.14
		J	1.20	67.00	2.15	35.60	0.10	0.00	80.10	0.00	13.70	0.00
		dJ	0.60	0.00	51.97	0.00	9.01	0.00	0.00	67.43	0.00	6.32
		F	10.80	32.10	2.15	60.40	1.00	7.30	19.40	0.21	82.70	0.10
		dF	20.20	0.00	29.92	0.30	59.56	13.00	0.00	27.12	0.00	81.44
	50	BB	82.20	0.00	5.31	0.90	19.80	93.50	0.00	0.81	0.80	4.60
		J	0.10	70.90	1.06	28.80	0.10	0.00	87.90	0.00	7.50	0.00
		dJ	0.10	0.00	71.44	0.00	3.80	0.00	0.00	83.35	0.00	2.40
		F	3.00	29.10	0.32	70.30	0.00	0.80	12.10	0.10	91.70	0.00
		dF	14.60	0.00	21.87	0.00	76.30	5.70	0.00	15.74	0.00	93.00

the Quebec region.

The ICC ρ enters the sample size calculation in these trials. Getting a reliable estimate is needed for planning purposes. Its value is typically small in community intervention trials and the upper limit of a confidence interval is useful to calculate a conservative sample size.

The deviance of the independence model is 19.366 for 9 degrees of freedom ; the p-value of the independence test is $\Pr(\chi^2_9 > 19.366) = 1.3\%$. This suggests that $\rho > 0$. This is a large cluster data set and only six sets of estimates are available. The values of $\hat{\rho}$ reported in Table 2.7 differ little, except for those obtained with the dJ and dF models which are null. The corresponding AIC are larger than that for the independence model, suggesting a poor fit for these models. The best fitting model is that of Frank . Thus Frank's model is plausible for this data and the 95% profile confidence interval for ρ of (0.001,0.097) appears to be reliable.

As an application of this result, suppose that one were to design a cluster randomized trial to compare the efficiency of a method for training physicians in share decision making, with respect to no training, on the proportion of patients engaging in share decision making. The test level is set at 5%, the power at 80% and

TABLE 2.5 – Larger cluster simulations : Confidence interval length (CIL) and empirical coverage (COV) of two-sided PLCI for ρ with a nominal 95 confidence level for data generated from 5 models

			$\rho = 0.05$				ρ =0.1			
			$\pi = 0.1$		π =0.3		$\pi = 0.1$		π =0.3	
	Data	Models	CIL	Cov	CIL	Cov	CIL	Cov	CIL	Cov
$K = 10$	BB	BB	0.19	98.48	0.15	97.81	0.27	97.87	0.21	98.03
	J	J	0.11	96.60	0.17	92.92	0.14	95.97	0.19	97.24
	dJ	dJ	0.46	93.96	0.31	96.30	0.49	97.18	0.33	98.28
	F	F	0.10	96.24	0.14	96.08	0.14	94.81	0.17	97.36
	dF	dF	0.25	97.82	0.20	96.47	0.32	98.57	0.24	98.60
$K = 25$	BB	BB	0.10	96.03	0.08	97.74	0.15	95.67	0.12	96.46
	J	J	0.06	96.43	0.10	94.22	0.08	95.19	0.12	96.64
	dJ	dJ	0.29	95.20	0.18	97.04	0.31	97.06	0.21	97.39
	F	F	0.06	96.80	0.09	95.63	0.08	93.59	0.10	95.97
	dF	dF	0.13	97.22	0.11	96.79	0.19	95.26	0.15	95.51
$K = 50$	BB	BB	0.06	94.16	0.06	94.93	0.11	94.10	0.09	95.60
	J	J	0.04	95.60	0.07	95.12	0.06	94.50	0.09	96.76
	dJ	dJ	0.19	97.54	0.12	97.53	0.22	98.22	0.15	94.86
	F	F	0.04	95.63	0.06	95.30	0.06	94.50	0.08	97.15
	dF	dF	0.08	96.45	0.07	94.74	0.13	95.90	0.10	94.78

TABLE 2.6 – Number of patients reporting an active role in the decision making process for 9 community health services

y_i	3	8	10	7	2	3	13	8	8
n_i	38	39	49	29	49	35	51	42	27

TABLE 2.7 – Six set of estimators for a community intervention trial.

	MLEs		95% CI for ρ		Model selection
Models	$\hat{\pi}$	$\hat{\rho}$	Lower	Upper	AIC
ZD	0.17	0.023	0.000	0.541	-
Binomial	0.17	-	-	-	52.36
BB	0.17	0.029	0.000	0.128	51.32
F	0.18	0.033	0.001	0.097	50.08
dF	0.17	0.000	0.000	0.277	54.36
J	0.18	0.027	0.000	0.104	52.22
dJ	0.17	0.000	0.000	0.293	54.36

the two proportions of patients engaging in shared decision making are 50% and 70% for the untreated and the treated community health services. The unit are patients and suppose that a community health service contributes 10 patients to the trial. How many community health services are needed, for each arm of the trial, to achieve the planned power of 80% ? The result critically depends on the value of ρ. Using the standard sample size formula ((Chakraborty et al., 2009, section 2.2), (Eldridge and Kerry, 2012, Chap. 7)) with $\rho = 0.097$, 17 health community services are needed while this number drops to 12 if the calculations are done with the mle $\hat{\rho} = 0.033$.

2.7.2 Chronic Obstructive Pulmonary Disease (COPD) data

This data is presented in Example 3 of Liang et al. (1992), and is considered by (Stefanescu and Turnbull, 2003; Zou and Donner, 2004). In this data the familial aggregation of the COPD is used as a measure of how genetic and environmental factors may contribute to disease etiology. It involves 203 siblings from 100 families with sizes ranging from 1 to 6. The binary response here indicates whether a given sibling has impaired pulmonary functions.

In the Table 2.8 , we report estimates of π and ρ, profile confidence intervals for ρ, and deviances obtained for all the models considered in this work. Results derived with the nonparametric (NP) model of Stefanescu and Turnbull (2003), with six parameters $\lambda_k, k = 1, \ldots, 6$, are also provided. As expected from graphical comparisons, the dG and dJ models have very similar fits. These models provide the best fits to the COPD data in terms of the AIC value. Moreover, the likelihood ratio test for comparing the fit of dGumbel model with that of the nonparametric model has $\chi^2_4 = 1.75$ p-value=0.78 ; this is not significant and the dGumbel fits well. The dGumbel 95% profile confidence interval for ρ is $(0.036, 0.395)$; it is very close to the dJoe interval that has been shown to be reliable in the small cluster simulations. Their lengths are 10% smaller than that of the Zou & Donner interval. As expected, the beta-binomial distribution, the Clayton and dClayton models give similar fits. Joe's and Frank's model do not fit. This agrees with graphical comparisons of Figure 1 where these two pmfs give much larger probabilities to 0 than the other model. Globally all profile likelihood confidence intervals are similar in this example, with differences less than 10% in confidence interval length. Given the small family size, the fact that the selection of a particular model has a limited impact on the analysis does not come as a surprise. The statistical methods proposed in this work are useful when dealing with larger clusters.

TABLE 2.8 – MLEs of π and ρ, and two-sided 95% confidence interval for ρ under 11 models for the (COPD) data. Degrees of freedom (*DF*), Deviance (*Dev*) and , AIC criterion of models are also reported.

	MLEs		95% CI for ρ			Goodness of fit		Model selection
Models	$\hat{\pi}$	$\hat{\rho}$	Lower	Upper	CL	*DF*	*Dev*	*AIC*
ZD	0.296	0.180	0.008	0.402	0.394	-	-	-
B	0.296	-	-	-	-	99	146.232	188.065
BB	0.282	0.213	0.059	0.401	0.342	98	137.406	181.239
G	0.283	0.207	0.052	0.388	0.336	98	138.760	182.593
dG	0.280	0.190	0.036	0.395	0.359	98	136.018	179.851
C	0.282	0.212	0.059	0.400	0.341	98	137.344	181.177
dC	0.282	0.213	0.059	0.403	0.344	98	137.407	181.240
J	0.286	0.191	0.038	0.368	0.329	98	139.848	183.681
dJ	0.281	0.180	0.031	0.390	0.359	98	136.101	179.934
F	0.286	0.201	0.045	0.382	0.336	98	139.342	183.175
dF	0.282	0.206	0.051	0.402	0.352	98	136.826	180.659
NP	0.284	0.200	-	-	-	94	134.266	186.099

2.8 Discussion

This paper has suggested to model the intra cluster association for binary data using multivariate Archimedean copulas. This family contains a wide range of distributional shape for the extra binomial variation. It has been demonstrated through a simulation study that the AIC is a useful criterion for selecting a particular model for the extra binomial variation. An important conclusion of our study is that profile likelihood confidence intervals for Archimedean copulas models have good coverage properties, even when the number of clusters is small. The model selection step is important. An omnibus method, such as using the beta-binomial profile likelihood confidence interval for the ICC regardless of whether this model fits well, may lead to confidence interval with poor coverage properties.

2.9 Transition

Dans le chapitre suivant, nous présentons une extension des modèles de la section 2.4 pour accommoder des covariables au niveau des grappes. Dans ce chapitre, nous nous intéressons à l'hétérogéneité dans les probabilités de capture lors d'une expérience de capture-recapture dans une population fermée. Dans ce contexte, nos modèles sont utilisés pour modéliser l'hétérogéneité résiduelle qui n'est pas prise en compte par les covariables mesurées sur des unités capturées. Plusieurs modèles sont disponibles pour l'hétérogénéité non observée et la probabilité de capture marginale est modélisée en utilisant les fonctions de liens Logit et Log-Log complémentaire. Les paramètres sont estimés en utilisant la vraisemblance conditionnelle construite à partir des observations collectées sur les unités capturées au moins une fois. Ceci généralise le modèle de Huggins (1991) qui ne tient pas compte de l'hétérogénéité résiduelle. La sensibilité de l'inférence à la spécification d'un modèle est également étudiée par des simulations. Un exemple numérique est présenté.

Chapitre 3

Mixture regression models for closed population capture-recapture data

Tounkara and Rivest (2015).
Biometrics, 71, 721-730, doi : 10.1111/biom.12325, 2015.
Keywords : Archimedean copulas, Closed population, Maximum likelihood, Random unit effect, Residual heterogeneity, Unit level covariates.

3.1 Résumé

Dans les études de capture-recapture, l'utilisation de covariables associées aux probabilités de capture a été recommandée pour obtenir des estimations de population stables. Toutefois, une certaine hétérogénéité résiduelle pourrait encore exister et ignorer une telle hétérogénéité pourrait conduire à sous-estimer la taille de la population (N). Dans ce travail, nous explorons deux nouveaux modèles avec des probabilités de capture en fonction de deux variables et les effets aléatoires non observés, pour estimer la taille d'une population. Une procédure d'inférence, comprenant une estimation de Horvitz Thompson et un intervalle de confiance pour la taille de la population, est ensuite déduite. La sélection d'un modèle particulier est effectuée en utilisant le critère d'information d'Akaike (AIC).

Premièrement, nous généralisons le modèle à effet aléatoire de Darroch et al. (1993) pour accommoder les covariables individuelles et discutons de ses limites. La seconde approche est une généralisation du modèle binomial traditionnel tronqué à zéro qui utilise un effet aléatoire pour tenir compte de l'hétérogénéité non observée. Cette approche fournit des utiles pour faire de l'inférence sur N. En effet les quantités clés telles que les moments, les fonctions de vraisemblance et les estimations de N et leurs erreurs standard ont des expressions explicites. Plusieurs modèles pour l'hétérogénéité non observée sont disponibles et la probabilité de capture marginale est exprimée en utilisant les fonctions de lien Logit et Log-Log complémentaire. La sensibilité de l'inférence à la spécification d'un modèle est également étudiée par des simulations. Un exemple numérique est présenté. Nous comparons la performance de l'estimateur proposé à celui obtenu sous de modèle M_h de Huggins (1991).

3.2 Abstract

In capture-recapture studies, the use of individual covariates has been recommended to get stable population estimates. However, some residual heterogeneity might still exist and ignoring such heterogeneity could lead to underestimating the population size (N). In this work, we explore two new models with capture probabilities depending on both covariates and unobserved random effects, to estimate the size of a population. Inference techniques including Horvitz-Thompson estimate and confidence intervals for the population size, are derived. The selection of a particular model is carried out using the Akaike information criterion (AIC).

First, we extend the random effect model of Darroch et al. (1993) to handle unit level covariates and discuss its limitations. The second approach is a generalization of the traditional zero-truncated binomial model that includes a random effect to account for an unobserved heterogeneity. This approach provides useful tools for inference about N, since key quantities such as moments, likelihood functions and estimates of N and their standard errors have closed form expressions. Several models for the unobserved heterogeneity are available and the marginal capture probability is expressed using the Logit and the complementary Log-Log link functions. The sensitivity of the inference to the specification of a model is also investigated through simulations. A numerical example is presented. We compare the performance of the proposed estimator with that obtained under model M_h of Huggins (1991).

3.3 Introduction

Capture-recapture methods are commonly used to estimate the size of closed populations. Units in the population are captured, marked, and released before being re-captured. This operation is repeated on t $(t > 1)$ occasions of capture. We suppose that the probability of capture for each unit does not change during the experiment. In such a case, the number of captures per unit is a sufficient statistic and is reported for all units. Note that no information about units that were never captured are available. The objective is to estimate the number of such units to construct an estimate for the population size (N).

A variety of statistical models is available to analyze capture-recapture data. The homogeneity model, denoted by M_0, supposes that the number of captures has a zero-truncated binomial distribution with the same capture probability for all units. However, this estimator can be highly biased when the capture probability changes from one unit to the next, see for instance (Otis et al., 1978; Burnham, 1987; Hwang and R. Huggins, 2005).

The capture probabilities are said to be variable or heterogeneous when some units are easier to catch than others. This occurs in many applications. For example, in ecological studies, some animals might be more active and easier to catch than others. In the species richness problem, common species are easier to detect than rare ones. The acronym M_h denotes models that handle such heterogeneity. These models can be classified into two groups depending on the availability of individual covariates. The first group relates the capture probability to some unobserved random variable associated to the catchability of a unit. For example, Darroch et al. (1993) suggested a random effect model to examine the coverage of the census. The heterogeneity can be explained by the latent classes models of Norris and Pollock (1996) and Pledger

(2000). They assume that capture probabilities vary between groups of units and are homogeneous within groups. The heterogeneity can also be modeled using a continuous latent variable having a normal (Coull and Agresti, 1999) or a beta (Burnham, 1972; Dorazio and Royle, 2003) distribution. The difficulty with this approach is that given a data set, several models may provide a good fit and lead to very different estimates of population size. Thus, there is a serious identifiability problem for N (Huggins, 2001; Link, 2003).

One solution to the non estimability of N is to collect auxiliary covariates measured on each unit caught in the study and to use them to model heterogeneity in capture probabilities. For example, Pollock et al. (1984) introduced models for categorical covariates, and Huggins (1989) proposed an extension to handle continuous covariables. In these models, the Logit and Log-Log link functions are used to express capture probabilities in term of covariates, and a conditional likelihood function is used to estimate model parameters. The population size is estimated by using the Horvitz-Thompson (H-T) estimator. A generalization that allows a nonparametric specification of the capture probabilities in terms of the covariates has been investigated, see for example Stoklosa and Huggins (2012). These models are adequate as long as covariates explain all of the heterogeneity. There may however still remain unexplained heterogeneity and these models could then lead to biased estimators of N ; indeed Huggins (1991) proposed a score test to assess such a residual heterogeneity. Thus alternatives that account simultaneously for observed and unobserved heterogeneity are needed. Up to date, this issue has not been addressed in the literature, see however the discussion in Rivest and Baillargeon (2014).

This work explores two new specifications of M_h, featuring both covariates and random effects, to model capture probabilities when estimating a population size. These models take into account observed as well as unobserved heterogeneity. The paper is organized as follows : Section 3.4 presents a general formulation of zero-truncated binomial mixture regression models. Maximum likelihood estimators for the parameters and the Horvitz-Thompson estimator for the population size are then derived. Section 3.4.1 introduces an extension of the random effect model of Darroch et al. (1993) to handle covariates. In Section 3.4.2 we describe a new class of models for residual heterogeneity. In Section 3.5, simulation studies are carried out to assess the performance of the proposed models and to compare them with the model M_h in Huggins (1991). To illustrate the methods in real life, we use the Harvest Mouse Data set analyzed in (Stoklosa and Huggins, 2012). Conclusions are drawn in Section 3.6.

3.4 Random effect models for residual heterogeneous

In this section, we present a general formulation of random effect models with covariates for heterogeneity in capture probabilities. We then propose an extension of the random effect model of Darroch et al. (1993) to handle covariates, and a new class of models.

3.4.1 General model formulation

Consider a closed population of size N and a capture-recapture experiment conducted over t capture occasions, where it is assumed that the units are captured independently one from the other. Let Y_i be a random variable, denoting the number of captures for population member i, and x_i a vector of covariates for this unit. The binomial model can be written as

$$P^{(B)}\big(Y_i = y_i|p_i\big) = \binom{t}{y_i} p_i^{y_i}(1-p_i)^{t-y_i} \qquad y_i = 0,\ldots,t, \tag{3.1}$$

where p_i, for $0 \leq p_i \leq 1$ is the capture probability for unit i. The standard homogeneity model M_0 corresponds to the case where p_i is the same for all units in the population. In the presence of heterogeneity, p_i varies between units. We consider models that express p_i in terms of x_i a vector of covariates and a_i a random variable associated with the catchability of unit i,

$$p_i = p(x_i^\top \beta; a_i), \qquad i = 1,\ldots,N, \tag{3.2}$$

where β is a vector of regression parameters for the covariables.

Under (3.2), model (3.1) gives a binomial mixture model

$$P^{(BM)}\big(Y_i = y_i|\alpha,\beta\big) = \binom{t}{y_i} E\{p_i^{y_i}(1-p_i)^{t-y_i}\}, \qquad y_i = 0,\ldots,t, \tag{3.3}$$

where the distribution of a_i is parametrized by α and $E(\cdot)$ gives an average value over the unobserved a_i.

In capture-recapture experiments, only units captured at least once are observed. The data set is $\{(y_i,x_i), i = 1,\ldots,n\}$, where n is the number of units caught at least once, y_i is the number of captures and x_i the vector of covariates for unit i. An important quantity is the probability that unit i is captured at least once,

$$\begin{aligned}\pi_i(\alpha,\beta) &= 1 - P^{(BM)}\big(Y_i = 0|\alpha,\beta\big)\\ &= 1 - E\{(1-p_i)^t\}.\end{aligned} \tag{3.4}$$

Because the event $y_i = 0$ cannot be observed, the data set comes from a zero-truncated binomial mixture regression model with a probability function

$$P^{(BM)}\big(Y_i = y_i|y_i > 0,\alpha,\beta\big) = \frac{P^{(BM)}\big(Y_i = y_i|\alpha,\beta\big)}{\pi_i(\alpha,\beta)} \qquad y_i = 1,\ldots,t. \tag{3.5}$$

In the absence of individual covariates, model (3.5) leads to standard binomial mixture models such as the beta-binomial, the logistic-normal and the latent class models.

The zero-truncated binomial regression model corresponds to model (3.5) where a_i is constant and p_i depends only on covariates through some link function. Assuming a Logit link function, (3.5) corresponds to the model M_h of Huggins (1991).

The log-likelihood function of model (3.5) is given by

$$L(\alpha,\beta) = \sum_{i=1}^{n} \log\Big\{P^{(BM)}\big(Y_i = y_i|y_i > 0,\alpha,\beta\big)\Big\}. \tag{3.6}$$

The model parameters can be estimated by maximizing (3.6) using a general maximization routine, which enables parameter estimation via an iterative procedure. We denote by $(\hat{\alpha},\hat{\beta})$ the estimates of (α,β). The standard error estimates are obtained from the Fisher information matrix, which can be evaluated numerically. The Horvitz-Thompson estimator of population size is given by

$$\widehat{N} = \sum_{1}^{n} 1/\hat{\pi}_i,$$

where $\hat{\pi}_i = \pi_i(\hat{\alpha}, \hat{\beta})$. The variance of $\widehat{N}$ is calculated as in Huggins (1989) and is given by

$$\widehat{Var}(\widehat{N}) = \sum_{i=1}^{n} (1-\hat{\pi}_i)/\hat{\pi}_i^2 + \left(\frac{\partial \widehat{N}(\hat{\alpha}, \hat{\beta})}{\partial(\hat{\alpha}, \hat{\beta})}\right)^{\top} \widehat{\mathscr{I}}^{-1} \left(\frac{\partial \widehat{N}(\hat{\alpha}, \hat{\beta})}{\partial(\hat{\alpha}, \hat{\beta})}\right),$$

where $\widehat{\mathscr{I}}$ is the estimated Fisher information matrix for (α, β), and $\left(\frac{\partial \widehat{N}(\hat{\alpha}, \hat{\beta})}{\partial(\hat{\alpha}, \hat{\beta})}\right)$ is the vector of partial derivatives of $\widehat{N}$ with respect to $\hat{\alpha}$ and $\hat{\beta}$,

$$\frac{\partial \widehat{N}(\hat{\alpha}, \hat{\beta})}{\partial(\hat{\alpha}, \hat{\beta})} = -\sum_{i=1}^{n} \frac{1}{\hat{\pi}_i^2} \frac{\partial \pi_i(\hat{\alpha}, \hat{\beta})}{\partial(\hat{\alpha}, \hat{\beta})}. \tag{3.7}$$

As $\widehat{N}$ is likely to have a skewed distribution, we suggest using the log-transform method of Burnham (1987) to obtain a confidence interval (CI) for N. The $100(1-\gamma)\%$ confidence interval for N is given by

$$IC_{100(1-\gamma)\%}(N) = \left(\widehat{N}_L, \widehat{N}_U\right) = \left[n + (\widehat{N} - n)/c, n + (\widehat{N} - n)c\right], \tag{3.8}$$

where $c = \exp\left\{z_{\gamma/2}\sqrt{[\log(1 + \widehat{Var}(\widehat{N})/(\widehat{N} - y)^2)]}\right\}$, and $z_{\gamma/2}$ is a $N(0,1)$ critical value.

We use the Akaike Information Criterion (AIC) for selecting a model given by

$$AIC = -2\log L(\hat{\alpha}, \hat{\beta}) + 2k, \tag{3.9}$$

where k is the number model parameters.

3.4.2 The model of Darroch et al. (1993)

A method to account for unmeasured heterogeneity is to include random individual effects in the model for the capture probability. Using a Logit link function leads to the following mixed model,

$$p_i = \frac{\exp(x_i^{\top}\beta + a_i)}{1 + \exp(x_i^{\top}\beta + a_i)}, \tag{3.10}$$

where a_i is a $N(0, \alpha)$ random effect. This model is not practical because it does not lead to closed form expressions for the capture probabilities $\pi(\alpha, \beta)$, see (3.4), entering in the formula for $\hat{N}$. This can be circumvented by assuming that the density of a_i is proportional to $(1 + e^{x_i^T\beta + a})^t h_\alpha(a)$, where $h_\alpha(.)$ denotes the density of $N(0, \alpha)$, with moment generating function given by $\exp\left(s^2\alpha/2\right)$. This proposal leads to the following probability function

$$P^{(BM)}\left(Y_i = y_i | \alpha, \beta\right) = \frac{\binom{t}{y_i}\exp(y_i x_i^T\beta + y_i^2\alpha/2)}{\sum_{\ell=0}^{t}\binom{t}{\ell}\exp(\ell x_i^T\beta + \ell^2\alpha/2)}, \qquad y_i = 0, \ldots, t.$$

It generalizes a model proposed by Darroch et al. (1993) to unit level covariates. The probability to be captured at least once in (3.4) is

$$\pi_i(\alpha,\beta) = 1 - \frac{1}{\sum_{\ell=0}^{t}\binom{t}{\ell}\exp(\ell x_i^T\beta + \ell^2\alpha/2)},$$

and then the conditional distribution of Y_i in (3.5) has a simple log-linear form. The vector of the partial derivatives defined in (3.7) is written as

$$\frac{\partial \hat{N}(\hat{\alpha},\hat{\beta})}{\partial(\hat{\alpha},\hat{\beta})} = -\sum_{i=1}^{n}\frac{1}{\pi_i^2}\left\{\sum_{\ell=0}^{t}\binom{t}{\ell}\exp(\ell x_i^T\beta + \ell^2\alpha/2)\begin{pmatrix}\ell^2/2\\ \ell x_i\end{pmatrix}\right\}.$$

Since the model (3.10) has an exponential form, the parameters are easily estimated. The population size and its variance have closed form expressions. Note that this approach is not a standard way to model heterogeneity since the density of the random effects depends on both, the number of capture occasions t, and the covariates x_i. The peculiar nature of the random effect means that in covariate free applications, this model usually gives a poor M_h fit when t is large, say larger than 10. Other M_h models, such as the Logit normal or the beta binomial, are then superior to Darroch's. Thus the intuitive way to combine covariates and random effects (i.e. adding a random component to the linear part of the model) leads to estimates that are either intractable or depend on a peculiar distribution for the random effects. It does not give a widely applicable estimation method for N.

3.4.3 A population based model for residual heterogeneity

The population based approach to model construction first specifies the average probability, ρ_i, for a unit with covariate x_i to be captured ; ρ_i is related to the covariates x_i through some link function g, $g(\rho_i) = x_i^\top\beta$. The Logit and the complementary Log-Log link functions are considered here. To account for the dependency between the captures for the same animal, we use a_i, a positive random variable with Laplace transform, $\psi_\alpha(t) = E(e^{-ta_i})$ where α is a positive dependence parameter. The model is defined in terms of the conditional probability of capture,

$$p_i = 1 - e^{-a_i\psi_\alpha^{-1}\{1-\rho_i\}}, \qquad i = 1,\dots,n, \tag{3.11}$$

where $\psi_\alpha^{-1}(\cdot)$ denotes the function inverse of $\psi_\alpha(\cdot)$. Under model (3.11), $p_i \in (0,1)$ and $E(p_i) = \rho_i = g^{-1}(x_i^T\beta)$, where g^{-1} is the inverse link function. Thus the requirement of having ρ_i as the average capture probability for covariate x_i is met with (3.11).

Under model (3.11), the Horvitz Thompson estimator for N and its variance have closed form expressions if the Laplace transform $\psi_\alpha(t)$ and its inverse have explicit analytical expressions. Table 3.1 gives four families of Laplace transforms that satisfy this condition. The four possible distributions for a_i are the gamma distribution and the positive stable distribution which are defined on $\mathscr{R}^+$ and two distributions, the logarithmic and the Sibuya, whose support is the set of positive integers. In Table 3.1, these models are represented using their names in the copula literature (Mai and Scherer, 2012). The model in which a_i has a logarithmic series distribution is called Frank's model. See Tounkara and Rivest (2014) for more discussion about the Archimedean copulas underlying the models considered in this section.

TABLE 3.1 – Laplace transforms, the distribution F_α of the random effects a_i and moments for four copula families, Clayton (C), Frank (F), Gumbel (G), Joe (J).

Families	$\psi_\alpha^{-1}(u)$	$a_i \sim F_\alpha$	$\lambda_k(\alpha,\rho)$
C	$(u^{-\alpha}-1)/\alpha$	$\Gamma\big(1/\alpha, 1/\alpha\big)$	$\big\{k(1-\rho)^\alpha - k + 1\big\}^{-\frac{1}{\alpha}}$
F	$-\log\left(\dfrac{e^{-\alpha u}-1}{e^{-\alpha}-1}\right)$	$Log(1-e^{-\alpha})$	$-\frac{1}{\alpha}\log\left\{1-\frac{(1-e^{\alpha(1-\rho)})^k}{(1-e^{-\alpha})^{k-1}}\right\}$
G	$\big\{-\log(u)\big\}^{\alpha+1}$	Positive Stable	$(1-\rho)^{k^{\frac{1}{1+\alpha}}}$
J	$-\log\{1-(1-u)^{\alpha+1}\}$	Sibuya $\big(1/(1+\alpha)\big)$	$1-\{1-(1-\rho^{1+\alpha})^k\}^{\frac{1}{1+\alpha}}$

In (3.11), p_i increases with a_i, thus larger values of a_i are associated to units that are captured often, for all the models in Table 3.1. The case $\alpha = 0$ corresponds to the absence of residual heterogeneity. In fact, for all models of Table 3.1, as α goes to 0, a_i converges to 1 in probability since $\psi_\alpha(t)$ tends to e^{-t}. In that case, model (3.11) gives, $p_i = \rho_i$ for all $i = 1,\cdots,N$; leading to the standard zero-truncated binomial regression models corresponding to model M_h in (Huggins, 1991). Note that $\psi_\alpha^{-1}(\cdot)$ is a decreasing function defined on $(0,1)$ with $\psi_\alpha^{-1}(1)$=0.

In Table 3.1, $\Gamma\big(1/\alpha, 1/\alpha\big)$ denotes the gamma distribution with shape $1/\alpha$ and rate $1/\alpha$. For this model, when $\alpha = 1$, a_i has a negative exponential distribution with parameter 1 and p_i has a beta distribution with parameters 1 and $1/\psi_\alpha^{-1}(1-\rho_i)$. Thus Clayton's model with $\alpha = 1$ and the beta-binomial distribution with parameters 1 and $1/\psi_1^{-1}(1-\rho_i)$ coincide. In the capture recapture literature, this model is considered by (Dorazio and Royle, 2003). In some sense, Clayton's model generalizes that of Dorazio and Royle (2003) to handle unit level covariates.

For Frank's model, $Log(\cdot)$ denotes the logarithmic distribution with parameter $1-e^{-\alpha}$ and probability mass function $Pr(a = k) = (1-e^{-\alpha})^k/(k\alpha)$, at $k \in \mathscr{N}^+$. For Joe family, Sibuya $\big(1/(1+\alpha)\big)$ denotes the Sibuya distribution with parameter $1/(1+\alpha)$. Its probability mass function is given by $Pr(a = k) = \binom{1/(1+\alpha)}{k}(-1)^{k-1}$, for $k \in \mathscr{N}^+$. Under Frank's and Joe's models, the smallest possible value of a_i is 1 and these models have a lower bound for the conditional probability of capture p_i. Consider now Gumbel's family. For this model, F_α is the positive stable distribution with parameter $1/(\alpha+1)$ and its density does not have a closed form expression. Chambers et al. (1976) propose a simple simulation algorithm for these variables. For more discussion about these families and their sampling algorithms, see (Mai and Scherer, 2012, Chap. 2).

Under model (3.11), α is the parameter that measures the level of residual heterogeneity. However this parameter does not have the same interpretation for all models. As an alternative we use Kendall's τ, a semi-parametric correlation measure for the underlying copula (see (Nelson, 2006)), which is a one-one a function of α. Kendall's τ belongs to $(0,1)$ and a value close to 0 is associated to the absence of residual heterogeneity, while large τ corresponds to strong heterogeneity in capture probabilities. For all models given in Table 3.1, Kendall's tau can be evaluated in term of ψ_α^{-1} and its derivative, see for example Chapter 2 of Mai and Scherer (2012). It has an explicit form for Gumbel's family, $\tau = \alpha/(\alpha+2)$ and for Clayton's,

$\tau = \alpha/(\alpha+1)$. For Frank's family, its evaluation involves the Debye function of order one, i.e., $D_1(\theta) = \frac{1}{\theta}\int_0^1 x/(e^x-1)dx$ and for Joe's family it is given as infinite series, see (Mai and Scherer, 2012, Chap. 2,p. 66-67) for further discussions.

For the models in (3.11), the moments $E\{(1-p_i)^k\}$ have the following form, $\lambda_k(\alpha,\rho_i) = \psi_\alpha\{k\psi_\alpha^{-1}(1-\rho_i)\}$; explicit expressions are given in Table 3.1. The probability that unit i is captured at least once is $\pi_i(\alpha,\beta) = 1-\lambda_t(\alpha,\rho_i), i = 1,\ldots,n$. The following theorem whose proof is given in the appendix, shows that for models in (3.11) the probability of being captured at least once decreases with α.

Theorem 1 *For the four models in Table 3.1, if $\alpha_1 \leq \alpha_2$ then*

$$\pi_i(\alpha_1,\beta) \geq \pi_i(\alpha_2,\beta)$$

for all β.

Thus, for a given marginal probability of capture ρ, units becomes, on average, more difficult to capture as the residual heterogeneity increases. For Gumbel's model, the complementary Log-Log-link deserves a special mention because it leads to a very simple expression for $\pi_i(\alpha,\beta)$. Under this model,

$$\pi_i(\alpha,\beta) = 1-\exp\left\{\log(1-\rho_i)\cdot t\,\frac{1}{1+\alpha}\right\},$$

Taking a Log-Log transform and using $\log\{-\log(1-\rho_i)\} = x_i^T\beta$ leads to

$$\begin{aligned}\log[-\log\{1-\pi_i(\alpha,\beta)\}] &= \frac{1}{1+\alpha}\log(t) + \log\{-\log(1-\rho_i)\}\\ &= \frac{1}{1+\alpha}\log(t) + x_i^T\beta\end{aligned}$$

Thus both ρ_i and $\pi_i(\alpha,\beta)$ have a linear form when expressed on the Log-Log scale.

The marginal distribution of Y_i in (3.3) can be written in terms of the moments $\lambda_k(\alpha,\rho_i)$ as

$$P^{(BM)}(Y_i = y_i|,\alpha,\beta) = \binom{t}{y_i}\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\lambda_{t-y_i+j}(\alpha,\rho_i) \quad y_i = 1,\ldots,t \tag{3.12}$$

Figure 3.1 presents graphs of the probability mass functions of the 4 models in Table 3.1 for ρ=0.1, 0.2, 0.3 ; and two heterogeneity levels τ=0.1, 0.2. It can be seen that the probability of zero capture is in general larger under Clayton (C) and Frank (F) models. The Clayton and Joe (J) models are the two extremes and the models of Gumbel (G) and Frank are intermediate. Figure 3.1 suggests the ordering C-F-G-J in terms of model shape.

3.5 Model evaluation and data analysis

In this section, we evaluate the sampling properties of the estimator $\widehat{N}$ obtained with the models depicted in Section 2.3 using both simulated and real data.

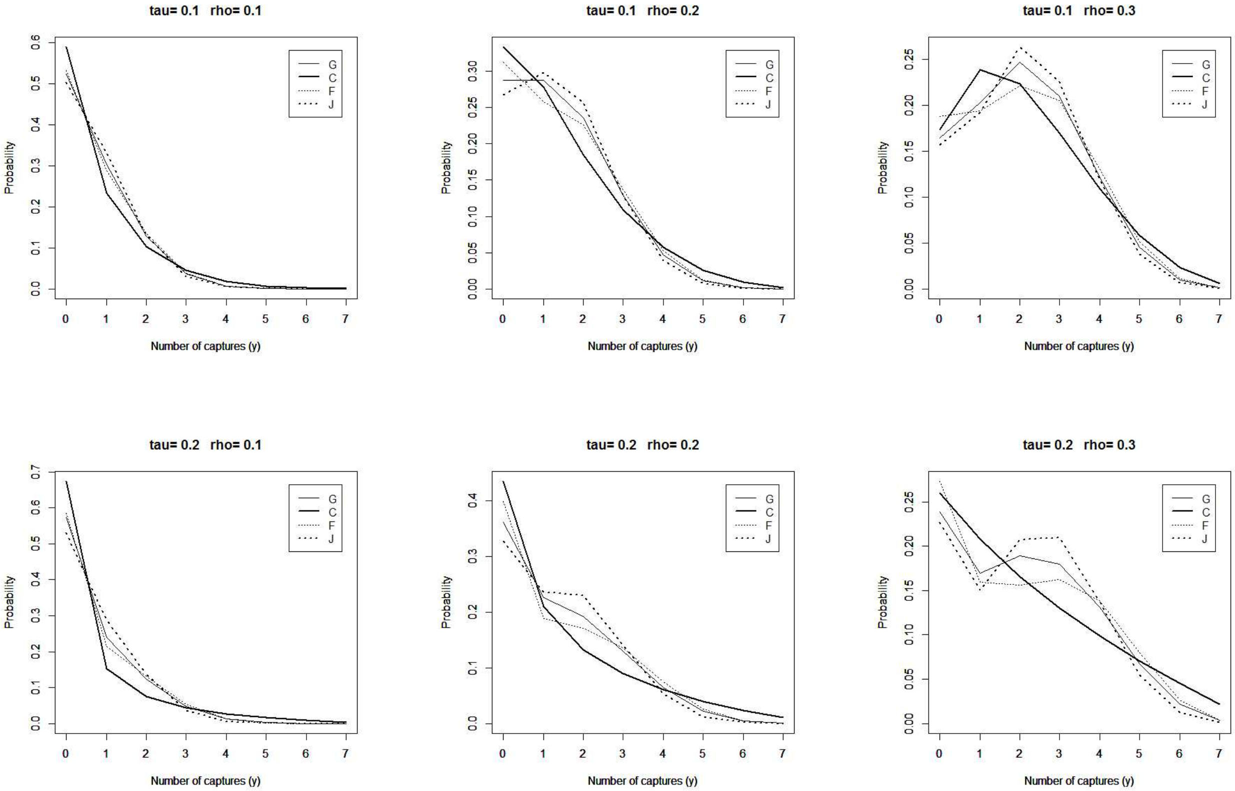


FIGURE 3.1 – Comparison of Clayton (C), Gumbel (G), Frank (F) and Joe (J) probability mass functions for t =7 capture occasions , τ =0.1, 0.2 and marginal capture probabilities ρ_i=0.1 , 0.2 , 0.3

3.5.1 Monte Carlo simulations

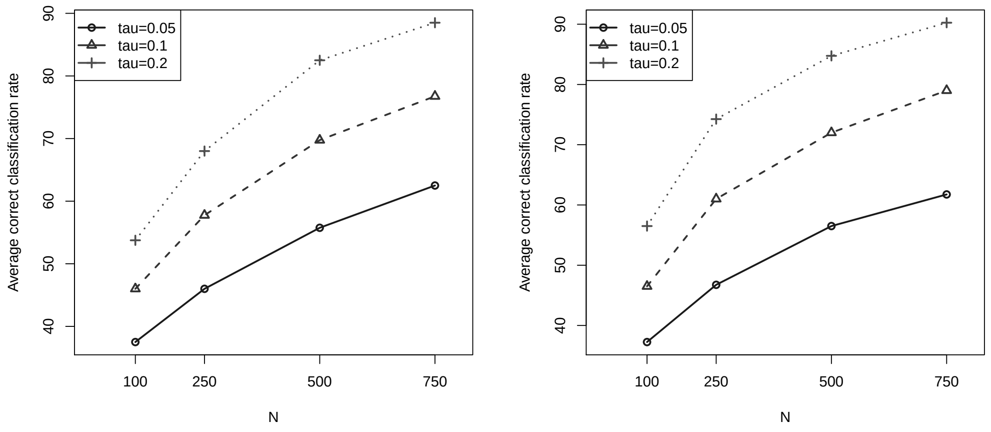
The aim of this study was threefold. First we investigated whether the AIC defined in (**??**) allows to identify the right model for the residual heterogeneity, among all those presented in Table 3.1. Secondly, we focused on the sampling properties of $\widehat{N}$ and investigate whether adding unit level covariates to the model improved the inference for N. Finally, we evaluated the performance of the confidence interval (CI) given in (3.8) for the models of Table 3.1. We compared these results with those obtained under the model M_h of Huggins (1991). The study used $t = 5$, four values of N, 100, 250, 500 and 750, and three levels of residual heterogeneity τ, 0.05, 0.1, 0.2. The model for ρ_i had a single continuous explanatory variable x with a $N(0,1)$ distribution and $\beta_0 = -0.5$, $\beta_1 = 1$. Data was generated from (3.11), using Gumbel (G)'s, Clayton (C) 's, Frank (F)'s and Joe (J)'s models with both Logit and complementary Log-Log link functions for the marginal capture probabilities. For each parameter combination, 1000 zero-truncated samples were generated. These were obtained by deleting the zero counts.

All programming was done in R. For the models of Clayton and Frank, in Table 3.1, $\psi_\alpha^{-1}(t)$ has an indeterminate form at $\alpha = 0$ and an evaluation of log-likelihood (3.6) at $\alpha = 0$ gives a numerical error. Thus only values of α larger than 10^{-5} were considered when maximizing (5.11) for these two models. The estimates $\hat{\beta}$ and $\hat{\alpha}$ were obtained using the function `optim()` of R. For the starting values, we used Huggins

(1991) estimates for the regression parameters and $\theta = \log(\alpha) = 1$; parameter α was log-transformed to avoid numerical boundary problems. In the simulations, the average proportion of units captured was nearly identical for all distributions. It decreased as τ increases, in agreement with Theorem 1. For the Logit link these proportions were 82 %, 80 % and 76 % for τ =0.05, 0.1, 0.2, respectively ; they were about 2 % larger with the Log-Log link function.

Objective 1 : Model selection

The model selection accuracy was measured using the proportions of times that a model was selected when the data was simulated from another model. In Figure 3.2, we plot the average correct classification rates in term of the population size. We can see that this rate increases with N and τ. Detailed results are presented in Web Table 3.1 of Web Appendix of the Supplementary material of Chapter 3. Web Table 3.1 highlights that the extreme models in Figure 3.1, Joe and Clayton, are more easily identifiable than Frank's and Gumbel's. The two link functions give similar results so we focus on the Logit case. The small population ($N = 100$) and small heterogeneity ($\tau = 0.05$) scenario gives the smallest proportion, 37 %, of correct classification which is defined as the mean diagonal percentage. It is slightly larger than 25%, the result that would be obtained if the models were indistinguishable. Correct classifications increase to 54% for $N = 100$ and $\tau = 0.2$ and to 82% for $N = 500$ and $\tau = 0.2$.



(a) Logit link

(b) c-Log-Log link

FIGURE 3.2 – Average correct classification rate in terms of τ and N for models constructed using the Logit (a) and the Log-Log (b) link functions.

Objective 2 : Sampling Distributions

Figure 3.3 presents results from simulations carried out at N =500 with unit level covariates x where a_i is distributed according to a Gamma distribution with 2 levels of heterogeneity τ =0.05 and 0.1. Simulations for the two link functions, Logistic and Log-Log, were run. Figure 3.3 give boxplots for the sampling distributions of population size estimators obtained under the four models of Table 3.1 and the zero-truncated binomial of Huggins (1991), labeled H. Boxplots for models fitted with covariates appear on the lower row while those without covariate are on the upper row.

As the simulation model is Clayton's (C) with covariates, the estimators obtained with that model are generally unbiased. When the covariates are in the model and the heterogeneity is misspecified, the estimates of N have small biases that increase with τ. All biases are negative ; which is to be expected since, in Figure 3.1, Clayton's model has, in most cases, the largest 0 probability. In the boxplots with covariates, the largest biases are for Huggins (1991) (H) model. Accounting for the heterogeneity appears to be useful, even even if the heterogeneity model is misspecified. The specification of a model for the heterogeneity have a relatively small impact on the sampling distribution of $\hat{N}$. In the "no covariate" part of Figure 3.3, (H) gives the estimates obtained with the homogeneity model M_0. The no covariate results are very variable. All the estimators are biased and have heavily skewed distributions. The selection of a particular heterogeneity model has a large impact on the sampling distribution of $\hat{N}$. Thus using unit level covariates in the model is useful even if they do not account for all the heterogeneity in capture probabilities. It reduces the differences between the estimates for N. The supplementary material section presents sampling distributions similar to those displayed in Figure 3.3, obtained with data sets simulated using Gumbel's, Frank's and Joe's models. The results are similar and lead to the same conclusions.

Objective 3 : the performance of confidence interval

The simulations also investigated 95% two-sided confidence intervals for N defined in (3.8) under the proposed heterogeneity models for N. Their performance was measured using the coverage (COV) and the average length (CIL) defined as

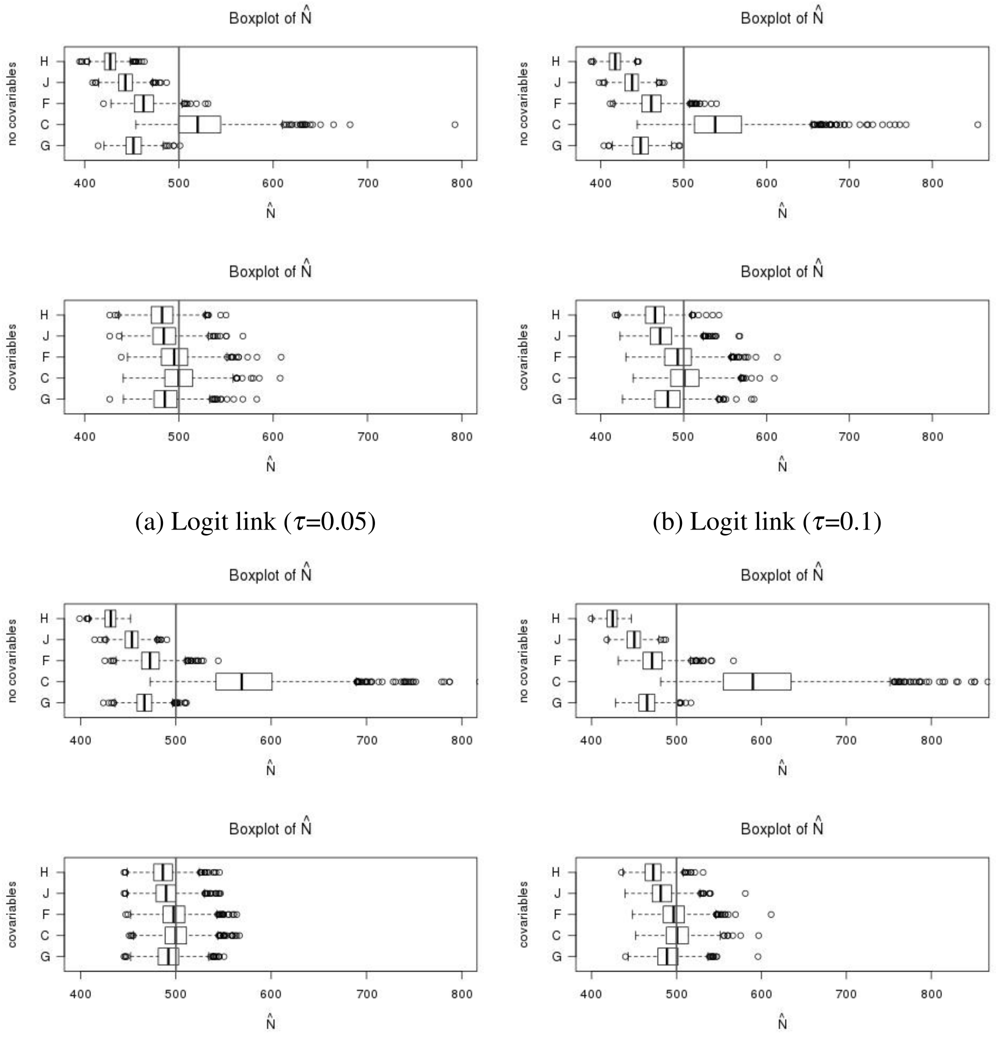
$$Cov = \frac{\sum_{i=1}^{1000} A_i}{1000} \times 100; \qquad\qquad CIL = \frac{\sum_{i=1}^{1000} (\hat{N}_{U_i} - \hat{N}_{L_i})}{1000},$$

where A_i equal to 1 if the true value N is in $(\hat{N}_{L_i}, \hat{N}_{U_i})$ defined in (3.8), and 0 otherwise.

Figure 3.4 summarizes the CIL as a function of N and τ. As expected CIL increases with N and τ. It is shorter with the Log-Log link function as N is on average slightly larger for the Log-Log simulations. Detailed simulation results are presented in Web Table 3.2 of the SM of Chapter 3 report. We investigated the variations of the coverage rate CV using an ANOVA model. The only significant factor was the population size N : the average coverage increased with N, from 93.92% at N=100 to 94.67% at N=750.

When residual heterogeneity is present, Huggins (2001) estimator is biased and the associated confidence interval shows a serious under-coverage. Its coverage goes to zero as N or τ increases, see Web Table 3.2 in the "Web Appendix B" of SM of Chapter 3. Additional simulation results, not provided here, show that estimation procedures based on Darroch model are biased for the models of Table 3.1, except when $\tau = 0.05$, when it provides estimates of N with small bias and confidence intervals with good coverage properties.

In summary, the simulation study shows that in the presence of residual heterogeneity the Huggins (1991) M_h models can seriously underestimate the population size. The Laplace transform models of Table 3.1 correct this bias. Model selection is an issue, especially for small populations where the likelihood based selection procedure can be erratic. Still, even when the wrong model is selected, estimators for N that account for a residual heterogeneity have relatively good sampling properties.



(a) Logit link (τ=0.05)

(b) Logit link (τ=0.1)

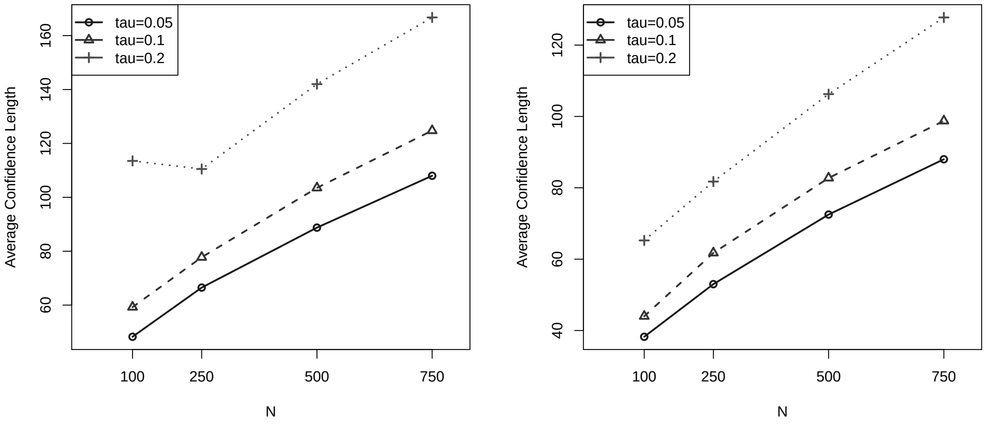
(c) c-Lg-Lg link (τ=0.05)

(d) c-Lg-Lg link (τ=0.1)

FIGURE 3.3 – Boxplots of estimates of N obtained with the models of Gumbel (G), Clayton (C), Frank (F), Joe (J) and Huggins (H), calculated without covariates (above) and with covariates (below). The data is simulated using Clayton's model with $N = 500$, τ=0.05, 0.1 and the marginal probability of capture modeled using Logit (graphs (a)-(b)) and complementary-Log-Log (c-Log-Log) (graphs (c)-(d)) link functions.

3.5.2 Example : the Harvest Mouse Data of Stoklosa and Huggins (2012)

To illustrate the performance of the proposed models, we used the Harvest Mouse Data analyzed in Stoklosa and Huggins (2012). This data set involves a total of $n = 142$ mice. For each mouse, the number of captures, over $t = 14$ occasions, and the mice weight were collected. Table 3.2 gives the estimates of dependence parameter τ, the estimate of N and its standard error under various models. We also report the lower and upper bounds and the length of 95% confidence interval for N and the AIC of all the models. The model of Huggins has a much larger AIC than the others and is rejected. This data set has unexplained heterogeneity in capture probability. The estimates of τ are positive for all models, indicating the presence of residual heterogeneity in the data. According to the AIC, Frank's model is preferred with Clayton's model a close



(a) Logit link

(b) c-Log-Log link

FIGURE 3.4 – The average confidence interval length (CIL) in terms of τ and N for the Logit and the Log-Log link functions.

second. Note that Huggins M_h estimate for N is not in Frank's 95% confidence interval for N. Not accounting for the residual heterogeneity yields biased results. In this example the two link functions, Logit and Log-Log, give similar results.

The fits of no covariate models are also provided in Table 3.2. They have large AIC and lead to drastically different values for $\hat{N}$. This is in agreement with the findings presented in figure 3.3 ; using covariates reduces the differences between the estimates of N obtained under various models. The Table 3.2 also includes results obtained with the two support points latent-class model of Norris and Pollock (1996). The beta-binomial model of Dorazio and Royle (2003) was also fitted to this data and gave essentially the same estimate as Clayton's model.

Stoklosa and Huggins (2012) found that the relationship between ρ and the body weight x was not linear on the Logit scale and constructed estimators that used non parametric link functions. They obtained estimates around the Huggins M_h estimate of 176. It would be interesting to investigate whether improving the link functions for the models considered in Table 3.2 has an impact on the model selected and on the final estimate for this example.

3.6 Discussion

This paper introduced zero-truncated binomial mixture regression models to account for a residual heterogeneity in capture probability, when estimating the size of a closed population. The standard technique of adding a random effect to the model for the capture probabilities does not yield a tractable estimator for N. A new class of models with a wide range of distributions for heterogeneity in capture probability has been proposed. Simple inference procedures based on Horvitz-Thompson estimator and confidence interval for the population size have been derived.

TABLE 3.2 – The results obtained when fitting the Laplace transform models of Table 3.1 to the Harvest Mouse Data.

		$\hat{\tau}$	AIC	$\hat{N}$	s.e.	95% CI for N	
Covariate Models							
Logit Link	G	0.12	440.9	211	24.7	177.0	278.4
	C	0.32	426.5	264	55.0	194.5	425
	F	0.16	426.0	239	35.0	191.0	334.5
	J	0.08	447.2	202	21.9	172.0	261.8
	H	-	455.9	176	9	162	200
c-Lg-Lg Link	G	0.12	440.7	210	23.7	176.7	273.9
	C	0.32	426.6	264	55.2	194.0	426.0
	F	0.16	425.8	238	34.6	190.5	332.6
	J	0.09	447.0	201	21.1	171.9	258.6
	H	-	456.01	175	9	161	198
Models without covariate							
	G	0.18	470.7	204	20.8	174.6	259.3
	C	0.81	455.5	847	729.8	274.4	3896.9
	F	0.37	455.7	326	260.2	165.7	1576.7
	J	0.14	475.9	196	14.1	174.4	231.3
	M_0		529.2	158	5.0	150.8	171.0
	Latent-class		456.9	211	22.2	178.9	269.3
	BB		455.5	880	1489.9	202.6	9116.5

The simulation study reveals that the AIC is a legitimate criterion for model selection. We have also demonstrated that adding unit level covariates to models is useful, as it reduces the discrepancy between the estimators obtained with the different models. It provides a partial solution to the identifiability problem as the selection of a particular model for the residual heterogeneity has a smaller impact on $\hat{N}$ than in the no covariate analysis. Of course, if the covariates explain all of the heterogeneity, then the identifiability problem disappears and Huggins (1991) M_h model should be used.

When the copula model is correctly specified then the analysis gives a confidence interval for N that has good coverage properties with a reasonable average length. Generalization of these estimators to other models in Huggins (1991) classification is an area that warrants additional investigations.

3.7 Transition

Au chapitre suivant, nous considérons les modèles des deux derniers chapitres pour faire de la prédiction dans de petites régions. Nous proposons les méthodes empiriques de Bayes basées sur nos modèles pour estimer des proportions. Nous obtenons une expression explicite des prédicteurs empiriques de Bayes et de leurs variances a posteriori pour les vraies proportions régionales. Nous utilisons la méthode du jackknife proposée par Lohr et Rao (2009) pour estimer l'erreur quadratique moyenne des prédicteurs empiriques. Nous évaluons par des études de simulation le gain d'efficacité du prédicteur empirique relativement au prédicteur linéaire (EBLUP). Nous présentons un exemple basé sur des données réelles.

Chapitre 4

Empirical Bayes predictions of small area proportions with Archimedean copulas

F. Tounkara and L.-P. Rivest.

Journal of multivariate analysis , under preparation for submission.
Keywords : Multivariate exchangeable copulas, Small area proportion, Empirical Bayes Techniques.

4.1 Résumé

Les données d'enquête sont utilisées pour obtenir des informations concernant les caractéristiques telles que la moyenne et la proportion relative non seulement à une population, mais aussi à des domaines. Cependant, la taille de l'échantillon dans certaines régions peut être faible, conduisant ainsi à des estimateurs directs de mauvaise qualité. Des estimateurs indirects basés sur l'information auxiliaire et construits à partir de méthodes basées sur des modèles définis de façon implicite ou explicite sont recommandés. Sur ce point, la méthodologie est très riche, mais pour le cas de données binaires la théorie est limitée. Ici, nous proposons une approche basée sur les modèles de copules Archimédiennes pour l'estimation régionale pour les données binaires. Des techniques empiriques de Bayes sont utilisées pour obtenir une forme explicite pour des estimations de proportions. En utilisant une étude de simulation, nous avons évalué la performance des estimateurs. Enfin, nous présentons un exemple numérique.

4.2 Abstract

Survey data are used to obtain information regarding characteristics such as mean and proportion relating not only to a population, but also to areas. However, the sample size in some areas may be small, leading to direct estimators of poor quality. Indirect estimators based on auxiliary information and constructed from implicitly and explicitly models-based methods are recommended. On this point the methodology is very rich, however for the case of binary data the theory is limited. Here, we propose a copula-based approach of small area estimation for binary data. Empirical Bayes techniques are used to obtain explicit form for

estimates of proportions. Using simulation study, we evaluated the performance of estimators. Finally, we present a numerical example.

4.3 Introduction

Survey data are used to obtain estimates of characteristics such as means and proportions relating not only to a population, but also to population subgroups. Standard survey estimates for subgroups usually termed direct estimates, are based only on the data and the sampling weights for the sampled units in the area of interest. However, the sample size in some areas may be small or even zero, leading to design-based estimators of poor quality. Therefore, alternative approaches have to be used in order to produce reliable estimates for small areas. The use of models that borrow strength from neighboring areas may produce such reliable estimates. Here, we consider the situation where the dependent variable of interest is dichotomous and where the small area parameters of interests is a proportion. The estimation of proportions for small area has been considered in the literature, see for example Wong and Mason (1985).

In the absence of covariates, the empirical best prediction (EBP) approach based on a Beta-Binomial (BB) model has been used to estimate small area proportions, see for example Rao (2003). In this case, the EBP is equivalent to the empirical best linear unbiased predictor (EBLUP). This EBLUP is a weighted average of an area-specific observed proportion and the mean proportion across all areas. Alternatives to the BB model, for instance the logit-normal and the probit-normal models, are also used to estimate small area proportions, see (Rao, 2003, Chap 9). A major problem of these approaches is that the EBP is difficult to be implemented, because it has no closed-form expression. For the logit-normal model, the EBP involves single-dimensional integrals over normal densities. An alternative to the EBP approach, the Hierarchical Bayes (HB) method, has been used to provide inference about small area proportions using mixed models and the Beta-Binomial models, see for instance Malec et al. (1997) and Aitkin et al. (2009). The HB approach is straightforward in the sense that the posterior distributions, once computed, can be used for all inferential purposes ; but this method is computationally intensive.

When auxiliary variables are available, both EBP and HB approaches based on generalized linear mixed models (GLMM) that include both fixed effects and random area-specific effects are often used for small area estimation (McCulloch, 2003; Jiang and Lahiri, 2006a). For example, the EBP method based on a logistic regression model containing both fixed and random effects is commonly used to estimate small-area proportions. This approach was originally proposed by Dempster and Tomberlin (1980) for the local estimation of census undercounts. Using a Bayesian method, MacGibbon and Tomberlin (1997) developed a methodology based on random effects, a multiple logistic regression model and empirical Bayes techniques. Others applications of the EBP method based on logistic regression with mixed effect can be found in Wong and Mason (1985),Tomberlin (1988) and Jiang and Lahiri (2006b). Applications of the HB method to provide inference about small area proportions using a logistic regression model has also been considered, see for instance Malec et al. (1997), and Liu (2009). The HB approach is straightforward in the sense that, posterior distributions, once computed, can be used for all inferential purposes, but this method is computationally intensive.

In this work, we apply Bayes techniques using a new class of probability models, to estimate small area proportions. These models have parameters for the marginal probability of success and a random effect that takes into account variations between small areas. The distribution of the random effect is associated with exchangeable Archimedean copulas (Mai and Scherer, 2012, Chapter 2). This family contains several specifications for the extra binomial variation. The Bayesian inference under the proposed models consists in obtaining the posterior distribution of the random effect and its Laplace transform. This posterior Laplace transform is then used to find Bayes estimates of small area proportions. The model parameters are estimated using the maximum likelihood (ML) method. Under the proposed model, the likelihood function and the best predictor (BP) of small area proportion have closed form expressions. Model parameters are replaced by their ML estimates in the BP to obtain the empirical best predictor (EBP). We use the Akaike information criterion (AIC) for selecting a particular model. Early applications of these models include the construction of a profile likelihood confidence interval for the intra-cluster correlation by Tounkara and Rivest (2014).

The estimation of the prediction mean squared error (PMSE) of the EBP under the proposed models is also of interest. When the small area estimators are considered under a frequentist approach, the EBP is used to estimate the true proportion and the estimated posterior variance is used as a measure of variability. However, the use of this measure of variability leads to an underestimation of PMSE of the EBP, because they do not take into account the uncertainty from having to estimate the model parameters, (Pfeffermann, 2013). In addition to using a naive approach, this work utilizes the jackknife techniques proposed by Jiang and Lahiri (2002) to arrive at estimates of the uncertainty associated with empirical Bayes estimates of small area proportions. It estimates the marginal prediction mean square error by integrating over the joint distribution of the random effects and the responses variable.

The rest of the paper is structured as follows. In Section 2, the basic model, where all parameters are assumed to be known is described. Bayesian inference for random effects, including the posterior distribution and its Laplace transform are developed in Section 3. This posterior Laplace transform is then used to find Best Predictor (BP) of small area proportions. Section 4 describes the Empirical Best Linear Unbiased Predictor (EBLUP) for proportions. In Section 5, we develop a comparison between the BP and the BLUP. Section 6 develops the ML inference, including estimation of parameters and model selection. The EBP and the jackknife estimator of its PMSE are then derived. Section 7 investigates these new statistical procedures through simulated and real data. We end with a discution.

4.4 Basic model

The proposed model involves two parameters : the marginal probability of success, $\pi \in (0,1)$, and $\alpha > 0$ a within area dependency parameter ; both are assumed to be known. The model is based on a positive, area specific, random effect a whose distribution depends on α. Given a, the probability of a success in a small area is

$$p = e^{-a\psi_\alpha^{-1}(\pi)}, \tag{4.1}$$

TABLE 4.1 – Inverse of Laplace transforms, the densities of the random effects a and moments for four copula families, Clayton (C), Frank (F), Gumbel (G), Joe (J).

Families	$\psi_\alpha^{-1}(t)$	$g_\alpha(a)$	$\lambda_k(\pi,\alpha)$
C	$(t^{-\alpha}-1)/\alpha$	$\frac{a^{1/\alpha-1}\exp(-a/\alpha)}{\alpha^{1/\alpha}\Gamma(1/\alpha)}$	$\left\{k\pi^{-\alpha}-k+1\right\}^{-\frac{1}{\alpha}}$
F	$-\log\left(\dfrac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$	$(1-e^{-\alpha})^a/(a\alpha)$	$-\frac{1}{\alpha}\log\left\{1-\frac{(1-e^{-\alpha\pi})^k}{(1-e^{-\alpha})^{k-1}}\right\}$
G	$\left\{-\log(t)\right\}^{\alpha+1}$	Positive Stable	$\pi^{k^{\frac{1}{1+\alpha}}}$
J	$-\log\{1-(1-t)^{\alpha+1}\}$	$\binom{\frac{1}{\alpha}}{a}(-1)^{a-1}$	$1-\{1-(1-(1-\pi)^{1+\alpha})^k\}^{\frac{1}{1+\alpha}}$

where ψ_α^{-1} is the inverse of ψ_α, the Laplace transform of a, $\psi_\alpha(t)=E\{\exp(-ta)\}$. Observe that $\pi=E(p)$ so that π is truly the marginal probability of success.

Table 4.1 gives the inverse of the Laplace transforms $\psi_\alpha^{-1}(t)$ and the densities $g_\alpha(a)$ of the random effects a for four models commonly used in the copula literature. For Gumbel (G) and Clayton (C) families, f_α is a density defined on $\mathscr{R}^+$ while, for the models of Frank (F) and Joe (J), f_α is a probability mass function defined on the positive integers $\mathscr{N}^+$, and the k-moments $E(p^k)$ have the following form, $\lambda_k(\alpha,\pi)=\psi_\alpha\{k\psi_\alpha^{-1}(\pi)\}$; explicit expressions for these moments are given in Table 4.1. See Tounkara and Rivest (2014) for more discussion about the distributions of the random effects of Table 4.1 and the Archimedean copulas underlying the models considered in this section. In all cases when $\alpha=0$, the Laplace transform is $\psi_0(t)=\exp(-t)$ and $P(a=1)=1$; this is the independence model where $p=\pi$. Swapping successes and failures, (4.1) gives new models for an extra binomial variation. The probability of success is then expressed as

$$p=1-e^{-a\psi_\alpha^{-1}(1-\pi)}, \tag{4.2}$$

where a is one of the random effects defined in Table 4.1. This gives a new class of models that are called dmodels in the sequel.

Consider $Y_1,\ldots,$ Y_n a set of n Bernoulli random variables coming from the same area with conditional probability of success p, see (4.1). Let $\rho=corr(Y_j,Y_{j'})$, $(j\neq j^{'})$ denotes the intra-cluster correlation, and we set $\phi=(\rho,\pi)^T$ the vector of model parameters. Under model (4.1), $E(Y_j)=\pi$ and $V(Y_j)=\pi(1-\pi)$. The responses Y_j and $Y_{j^{'}}$ of two units j and $j^{'}$ $(j\neq j^{'})$ are dependents ; their covariance is $cov(Y_j,Y_{j^{'}})=\lambda_2(\pi,\alpha)-\pi^2$, and the intra-area correlation is

$$\rho=\frac{\lambda_2(\pi,\alpha)-\pi^2}{\pi(1-\pi)}=\frac{\psi\{2\psi_\alpha^{-1}(\pi)\}-\pi^2}{\pi(1-\pi)}. \tag{4.3}$$

Let $Y=\sum_{j=1}^n Y_j$. The conditional distribution of Y, given a, is binomial and its conditional mass probability function (pmf) can be written as :

$$
\begin{aligned}
f(y|a,\alpha,\pi) &= \binom{y}{n} e^{-a\psi_\alpha^{-1}(\pi)y}\{1-e^{-a\psi_\alpha^{-1}(\pi)}\}^{n-y} \\
&= \binom{n}{y}\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell e^{-a\psi_\alpha^{-1}(\pi)(y+\ell)},
\end{aligned} \tag{4.4}
$$

where y is a realization of Y. The marginal pmf of Y is obtained by integrating $f(y|a,\alpha,\pi)$ in (4.4) over the distribution of the random effect a. It is

$$
P(Y=y|\alpha,\pi) = \binom{n}{y}\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell \lambda_{y+\ell}(\pi,\alpha). \tag{4.5}
$$

This marginal pmf is used to develop Bayes inference for the conditional probability p.

4.5 Bayesian inference for p

In this section, we develop Bayesian inference for p. As seen in section 4.4, given p in (4.1), Y be a binomial random variable with parameters n and p. First the posterior distribution of a in (4.1) given that $Y=y$ is derived and the conditional Laplace transform is obtained. This posterior Laplace transform is thus used to develop Bayes inference for the conditional probability of success.

The following theorem gives the posterior distribution of the random effect a.

Theorem 1 *The posterior density (pmf) of a given $Y=y$ can be expressed as*

$$
g(a|y,\alpha,\pi) = \frac{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell e^{-a(y+\ell)\psi_\alpha^{-1}(\pi)} g_\alpha(a)}{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell \lambda_{y+\ell}(\pi,\alpha)}, \tag{4.6}
$$

where $g_\alpha(\cdot)$ denotes the density (pmf) of a.

The proof of Theorem 1 combines equations (4.4) and (4.5) using Bayes formula. The support of the conditional distribution of a is contained in $(0,\infty)$; thus it has a Laplace transform that is given in the next proposition,

Proposition 1 *The posterior Laplace transform of a given $Y=y$ is given by*

$$
\psi_{\alpha,y}(t) = \frac{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell \psi_\alpha\{t+(y+\ell)\psi_\alpha^{-1}(\pi)\}}{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^\ell \lambda_{y+\ell}(\pi,\alpha)} \tag{4.7}
$$

Proof of Proposition 1
By definition, the posterior Laplace transform is

$$
\psi_{\alpha,y}(t) = E\left(e^{-ta}|y,\alpha,\pi\right). \tag{4.8}
$$

Using the posterior density, equation (4.8) can be written as

$$\psi_{\alpha,y}(t) = \int_0^\infty e^{-at} g(a|y,\alpha,\pi)da,$$

$$= \frac{\sum_{\ell=0}^{n-y} \binom{n-y}{\ell} (-1)^\ell \int_0^\infty e^{-a[t+(y+\ell)\psi_\alpha^{-1}(\pi)]} g_\alpha(a)da}{\sum_{\ell=0}^{n-y} \binom{n-y}{\ell} (-1)^\ell \lambda_{y+\ell}(\pi,\alpha)}. \tag{4.9}$$

Noting that the integrals in (4.9) are the Laplace transforms of a evaluated at $\left(t + (y+\ell)\psi_\alpha^{-1}(\pi)\right)$ ends the proof.

Now, consider $X_1,\ldots,X_q$ a sequence of q non-observed Bernoulli random variables coming from the same area as $\{Y_i : i = 1,\ldots,n\}$. Given that $Y = y$, the posterior Laplace transform $\psi_{\alpha,y}(\cdot)$ in (4.7) characterizes the conditional dependence through X_j, $j = 1,\ldots,q$. This is presented formally in the next proposition,

Proposition 2 *Let $X_1,\ldots,X_q$ denote a sequence of q non-observed Bernoulli random variables. Given that $Y = y$, their dependency can be modeled by the random effect model having the same form as (4.1), with the following characteristics (we use the subscript (y) to denote updates parameters or functions)*

$$\pi_y = \psi_{\alpha,y}\{\psi_\alpha^{-1}(\pi)\},$$
$$p_y = e^{-a_y \psi_{\alpha,y}^{-1}(\pi_y)}, \tag{4.10}$$

where a_y denotes a positive random effect, with Laplace transform $\psi_{\alpha,y}(t)$, see (4.7), that accounts for the residual dependency given the data.

Given p_y, X_j, $j = 1,\ldots,q$, are independent and the conditional distribution of $X_j|p_y$ is $Bernoulli(p_y)$. Under model (4.10), the marginal mean of X_j is $E(X_j) = \pi_y$, and its marginal variance $V(X_j) = \pi_y(1-\pi_y)$. The intra-cluster correlation between X_j and $X_{j'}$, $(j \neq j')$, is given by

$$\rho_y = \frac{\psi_{\alpha,y}\{2\psi_{\alpha,y}^{-1}(\pi_y)\} - \pi_y^2}{\pi_y(1-\pi_y)}.$$

4.5.1 The BP of p and its MSE

Here, we apply results in Theorem 1 to the Bayes estimation of p in (4.1). Given $Y = y$, the Bayes estimate of p and its variance are $E(p|y)$ and $E(p^2|y) - E(p|y)^2$, respectively. These Bayes estimates can be obtained from the posterior Laplace transform.

Theorem 2 *Assuming that $\phi = (\alpha,\pi)$ is known, the best predictor (BP) of p under model (4.1) and its variance can be expressed as*

$$\hat{p}^{BP}(y,\phi) = \psi_{\alpha,y}\{\psi_\alpha^{-1}(\pi)\}, \tag{4.11}$$

and

$$MSE^c\{\hat{p}^{BP}(y,\phi)\} = \psi_{\alpha,y}\{2\psi_\alpha^{-1}(\pi)\} - \Big[\psi_{\alpha,y}\{\psi_\alpha^{-1}(\pi)\}\Big]^2, \tag{4.12}$$

respectively.

The best predictor $\hat{p}^{BP}(y,\phi)$ can be seen as the posterior estimation of the conditional probability of success p. The unconditional mean square error of $\hat{p}^{BP}(y,\phi)$ can be obtained by averaging $V(p|y,\phi)$ over all possible values of y. It is given by :

$$\begin{aligned} MSE\{\hat{p}^{BP}(y,\phi)\} &= E\big\{V(p|y,\phi)\big\} \\ &= \lambda_2(\phi) - \sum_{\ell=0}^{n} \big\{\hat{p}^{BP}(\ell,\phi)\big\}^2 f(\ell,\alpha,\pi) \end{aligned} \tag{4.13}$$

where $f(k,|\alpha,\pi) = P(Y_i = k)$ is the pmf of y defined in (4.5).

To end this section, note that for large sample sizes n, say larger than 30, the alternating sum in (4.7) is numerically unstable and may yield negative predictions. In such case, we evaluate the $\hat{p}_i^{BP}(y,\phi)$ using an approximate method proposed by Tounkara and Rivest (2014), which consists in calculating the numerator and the denominator directly by integrating over the random effect distribution. Note that this approach is only possible for the model of Joe and Frank as their random effects have discrete distributions. The following section presents an alternative to the BP of p, the best linear unbiased prediction (BLUP) of p.

4.6 The BLUP of p

To predict p, we can consider the Best linear unbiased prediction (BLUP) approach. Consider an arbitrary binomial mixture model, with $y|p \sim \text{Binomial}(n,p)$, and $E(p) = \pi$, where $\varphi = (\rho,\pi)$ is known and ρ is the intra-cluster correlation. A linear predictor for p is a function $b_0 + b_1 y$. Among all the b_0 and b_1 that give an unbiased prediction of p, the values that give the smallest prediction error variance are $b_0 = (1-\gamma)\pi$ and $b_1 = \gamma/n$ where $\gamma = n\rho/\{1+(n-1)\rho\}$. Thus the BLUP is given by

$$\hat{p}^{BLUP}(y,\varphi) = \gamma\frac{y}{n} + (1-\gamma)\pi. \tag{4.14}$$

Since γ belongs in $(0,1)$, it is worth mentioning that the BLUP estimator is the weighted combination of the direct estimator $p^d = y/n$ and the synthetic estimator π. The minimal value for the variance of the prediction error is given next.

Theorem 3 *The MSE of $\hat{p}^{BLUP}$ can be written as*

$$MSE\{\hat{p}^{BLUP}(y,\varphi)\} = \frac{\pi(1-\pi)\rho(1-\rho)}{1+(n-1)\rho}. \tag{4.15}$$

Proof of Theorem 3.

The result in (4.15) is obtained by minimizing,

$$\begin{aligned} f(\gamma) &= E\big[\{\hat{p}^{BLUP}(y,\varphi) - p\}^2\big] \\ &= E\big\{(\gamma\frac{y}{n} + (1-\gamma)\pi - p)^2\big\} \\ &= E\big\{\gamma^2(\frac{y}{n}-\pi)^2 + (p-\pi)^2 - 2\gamma(\frac{y}{n}-\pi)(p-\pi)\big\} \\ &= \gamma^2 E\big\{(\frac{y}{n}-\pi)^2\big\} + E\big\{(p-\pi)^2\big\} - 2\gamma E\big\{(\frac{y}{n}-\pi)(p-\pi)\big\} \end{aligned} \tag{4.16}$$

with respect to γ.

Under a binomial mixture model, we have $E(p) = \pi$, $V(p) = \rho\pi(1-\pi)$, $E(y) = \pi$, $V(y) = n\pi(1-\pi)[1+(n-1)\rho]$. Noting that $E\{(\frac{y}{n}-\pi)(p-\pi)\} = cov(p,p) = V(p)$, equation (4.16) can be written as

$$\begin{aligned} f(\gamma) &= \gamma^2 V(\frac{y}{n}) + V(p) - 2\gamma V(p) \\ &= \gamma^2 \frac{n\pi(1-\pi)[1+(n-1)\rho]}{n^2} + \rho\pi(1-\pi) - 2\gamma\rho\pi(1-\pi). \end{aligned} \tag{4.17}$$

Setting the partial derivative

$$\frac{\partial f(\gamma)}{\partial \gamma} = 2\gamma \frac{n\pi(1-\pi)[1+(n-1)\rho]}{n^2} - 2\rho\pi(1-\pi)$$

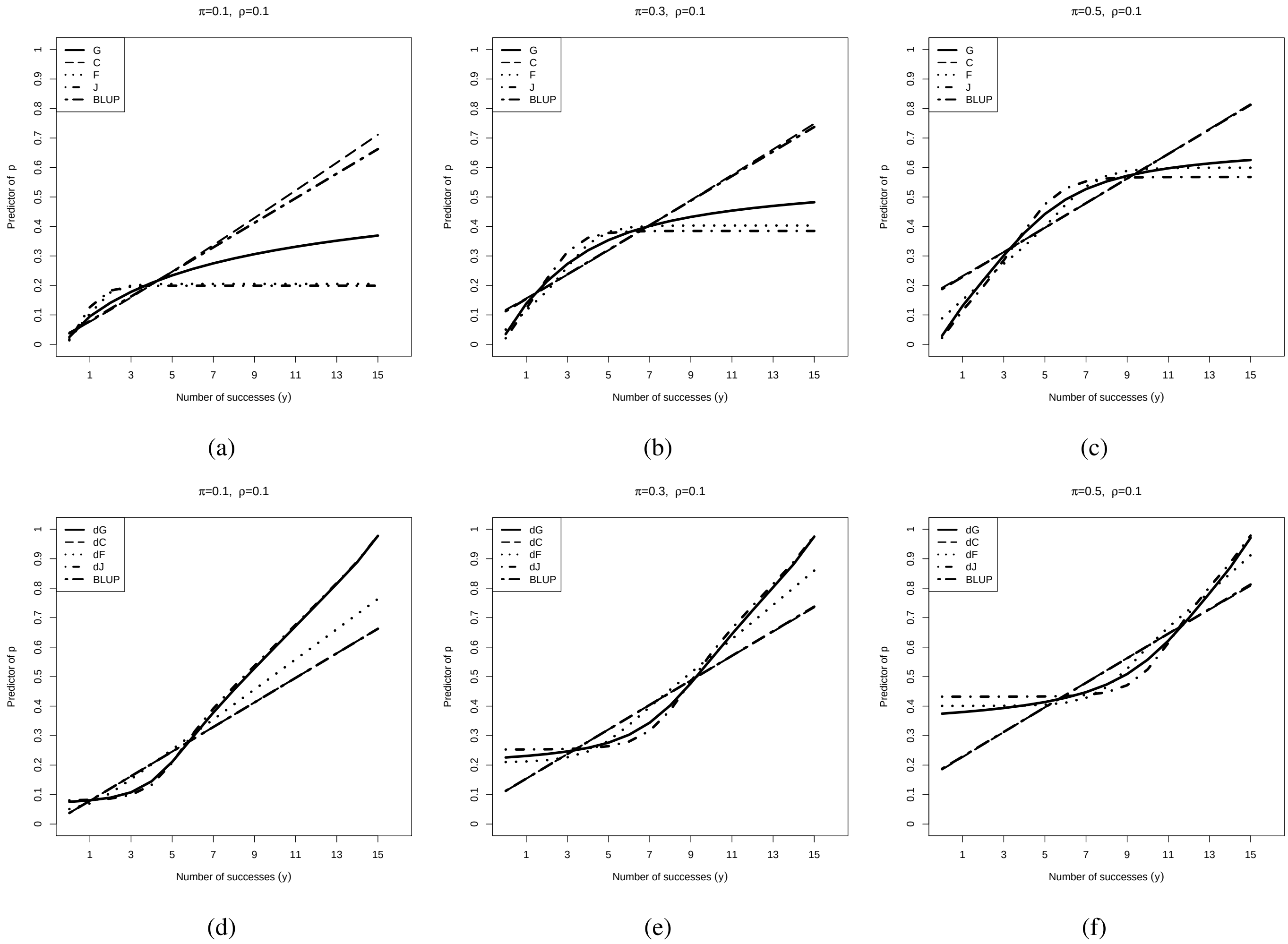
equal to zero, yields $\gamma_0 = n\rho/(1+(n-1)\rho)$ as the minimizer of $f(\gamma)$. Replacing γ by this value into (4.17) gives (4.15).

4.7 The comparison of BP and BLUP

This section gives the functional forms for the BP and the BLUP under the models defined by (4.1) and (4.2). It also evaluates the efficiency of BP relatively to the BLUP. To do so, it is convenient to parameterize all models in terms of ρ and π. The dependence parameter α in Table 1 can be expressed in terms of these two parameters by solving equation (4.3), see Tounkara and Rivest (2014).

4.7.1 The BP versus the BLUP

For fixed parameter values both predictor are functions of y, see (4.11) and (4.14), that are plotted in Figure 4.1. The findings of these comparisons are now briefly presented. First for Clayton's and the d-Clayton's models, the BLUP is equivalent to the BP. They are, for all practical purposes, equal. This is illustrated in Figure 4.1 where the EBUP and the BP for Clayton's and d-Clayton's models obtained when $n = 15$, $\pi = 0.1, 0.2, 0.3$ and $\rho = 0.1$ are presented. This means that when the BLUP is a candidate to produce a prediction for a proportion, there is no need to include the Clayton and dClayton model as they provide predictions that are very similar. We observe similar results between the BP obtained under Joe (J) and Gumbel (G), and also between d-Joe (dJ) and dGumbel (dG). In Figure 4.1, the Joe and the dJoe BP exhibits a high degree of nonlinearity as compared with the BLUP which is a linear function of y. Frank's BP is very similar to Joe's while Gumbel's is closer to the BLUP.



(a) (b) (c)

(d) (e) (f)

FIGURE 4.1 – A comparison of the BPs and BLUP for area size 15 under some area level models, for $\rho = 0.1$ and $\pi = 0.1, 0.3, 0.5$

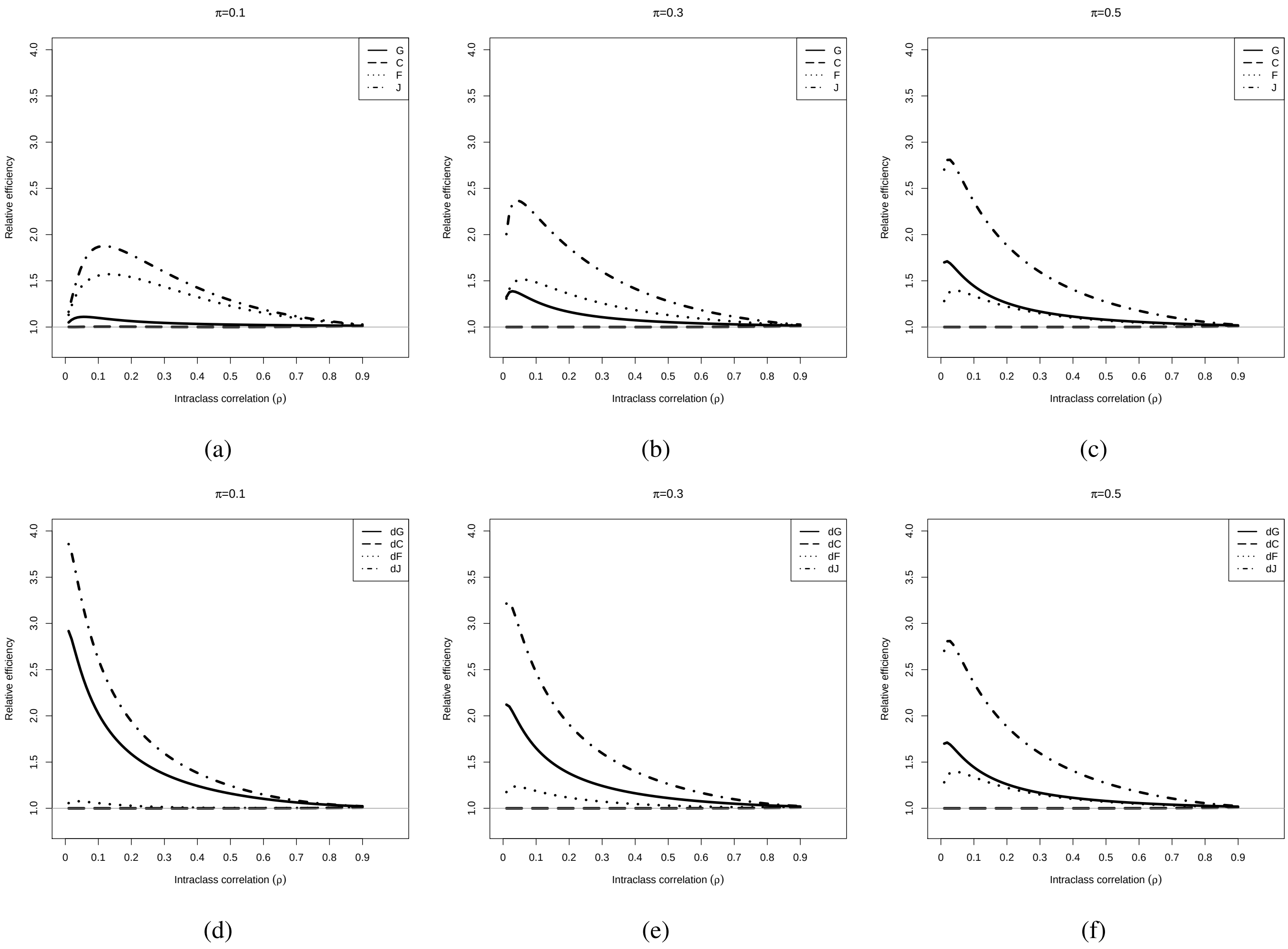
Suppose that in a cluster, $n = 15$ Bernoulli random variables have been recorded yielding $y = 3$ successes. Knowing that the model for the within cluster dependency has $\pi = 0.5$ and $\rho = 0.1$, what is the probability that an additional Bernoulli random variable from this cluster will give a success ? Considering either graph (c) or (f) of Figure 4.1, the BLUP gives a probability of 0.25 ; if a dmodel is used then it is around 0.4 while for model (4.1) it is closer to 0.22. Thus the predicted probability depends heavily on the underlying model.

4.7.2 The efficiency of BP relative to the BLUP

The MSE of the BP in (4.13) is used to evaluate the efficiency of the BP (p^{BP}) relatively to the BLUP (p^{BLUP}), using the formula

$$RE = \frac{MSE(p^{BLUP})}{MSE(p^{BP})},$$

where the MSEs are given in (4.15) and (4.12). Figure 4.2 shows, for $\pi = 0.1$, 0.3, and 0.5, how the relative efficiency of the BP depends on the underlying models and on the the intra-class correlation (ρ). Overall, except for the Clayton's and d-Clayton's model, the BP can offer substantial gains in precision. The most important gains occur for the Joe and the d-Joe model.



(a) (b) (c)

(d) (e) (f)

FIGURE 4.2 – The relative efficiency of the BP associated with 8 copula models relatively to the BLUP for $\pi = 0.1, 0.3, 0.5$

4.8 Estimation and empirical prediction

The BP and BLUP of p developed in the previous section assume that the parameters α, π, and ρ are known. In practice these parameters are unknown and they are estimated from sample data. This section presents the maximum likelihood procedure and suggests a model selection criterion to construct empirical predictions of proportion.

4.8.1 The maximum likelihood estimation

Let $\{(y_i, n_i), i = 1, \ldots, m\}$ be a sample from m small areas. The likelihood function is obtained by taking the product across the m small areas of the pmf in (4.5).

$$L(\alpha, \pi) = \prod_{i=1}^{m} f(y_i, \alpha, \pi). \tag{4.18}$$

To obtain the ML estimate of α and π, one can use an iterative algorithm implemented in the optim function of R, applied to the log of the likelihood function (4.18). One can use $\alpha = 1$ and $\pi = \sum_{i=1}^{m} y_i / \sum_{i=1}^{m} n_i$ as starting values. Let $\hat{\alpha}$ and $\hat{\pi}$ denote the ML estimors of α and π. Substituting $\hat{\alpha}$ and $\hat{\pi}$ in (4.3), we obtain $\hat{\rho}$, the ML estimator of ρ.

4.8.2 The model selection criterion

For selecting a particular model to work with, we used the estimate of Akaike Information Criterion (AIC), defined by

$$AIC = -2\log L(\hat{\alpha}, \hat{\pi}) + 2k,$$

where L is given in (4.18) and k, the number of parameters, is equal to 2 for all the copula models considered in this work. The model with the minimum AIC value is preferred. The AIC is an useful criterion for selecting a particular copula model for correlated binary data, see Tounkara and Rivest (2014).

4.8.3 The empirical prediction

We denote by p_i the conditional probability for the i^{th} area, $i = 1, \ldots, m$, and let p_i^{BP} and p_i^{BLUP} be the corresponding BP and BLUP, respectively. Here, we present the empirical BP (EBP) of p_i and its error estimation, and the empirical BLUP (EBLUP) of p_i. A method for estimating their mean square error is also given.

EBP and its MSE estimate

Let $\hat{\phi} = (\hat{\alpha}, \hat{\pi})$ be the estimate of $\phi = (\alpha, \pi)$. Replacing ϕ by $\hat{\phi}$ in (4.11), yields the empirical best predictor (EBP) of p_i, $\hat{p}_i^{EBP} = \hat{p}_i^{BP}(y_i, \hat{\phi})$, where y_i is the binomial observation for area i. A naive estimate of variability is the estimated posterior variance. However, this empirical variance induces an underestimation of the mean squared error (MSE) because it ignores the variability associated with the estimation of model parameters. Thus, an estimator of the mean square error, $mse_1 = E(\hat{p}_i^{EBP} - p_i)^2$ for $i = 1, \ldots, m$ is needed. Here, we use the jackknife method proposed by Jiang and Lahiri (2002) and described in Pfeffermann (2013). The MSE of $\hat{p}_i^{EBP}$ can be decomposed as

$$mse_1 = E\{\hat{p}_i^{BP}(y_i, \phi) - p_i\}^2 + E\{\hat{p}_i^{EBP} - \hat{p}_i^{BP}(y_i, \phi)\}^2 = \eta_{1i} + \eta_{2i}, \tag{4.19}$$

where η_{1i} is the unconditional MSE of the BP and η_{2i} denotes the contribution to the MSE from estimating the model parameters. We estimate them using the Jackknife. Let $\hat{\eta}_{1i}^{BP}(\hat{\phi})$ be the naive estimator of η_{1i}, obtained by substituting ϕ by the ML estimator $\hat{\phi}$ in (4.12). We also denote by $\hat{\eta}_{1i}^{BP}(\hat{\phi}_{-\ell})$ the naive estimator obtained by deleting area ℓ when estimating ϕ, and we set $\hat{p}_{i,-\ell}^{EBP} = \hat{p}_i^{BP}(\hat{\phi}_{-\ell})$ is the corresponding EBP. The Jackknife estimator of mse_1 is given by :

$$mse_1^J(\hat{p}_i^{EBP}) = \hat{\eta}_{1i} + \hat{\eta}_{2i}, \tag{4.20}$$

where

$$\hat{\eta}_{1i} = \hat{\eta}_i^{BP}(\hat{\phi}) - \frac{m-1}{m} \sum_{\ell=1}^{m} \{\hat{\eta}_i^{BP}(\hat{\phi}_{-\ell}) - \hat{\eta}_i^{BP}(\hat{\phi})\},$$

and

$$\hat{\eta}_{2i} = \frac{m-1}{m} \sum_{\ell=1}^{m} \left\{ \hat{p}_{i,-\ell}^{EBP} - \hat{p}_i^{EBP} \right\}^2.$$

EBLUP and its MSE estimate

Analogously, replacing $\varphi = (\rho, \pi)$ by $\hat{\varphi} = (\hat{\rho}, \hat{\pi})$ in (4.14), leads to the EBLUP, $\hat{p}_i^{EBLUP} = \hat{p}_i^{BLUP}(y_i, \hat{\varphi})$, for p_i. The naive measure of variability of the EBLUP, obtained by replacing φ by $\hat{\varphi}$ in (4.15), can lead to severe underestimation of the MSE because it ignores the variability associated with $\hat{\rho}$ and $\hat{\pi}$. We use the jackknife procedure of Jiang and Lahiri (2002) described above to obtain estimation of MSE($\hat{p}_i^{EBLUP}$). We consider the following decomposition of the MSE of $\hat{p}_i^{EBLUP}$

$$mse_2 = E\{\hat{p}_i^{BLUP}(y_i, \varphi) - p_i\}^2 + E\{\hat{p}_i^{EBLUP} - \hat{p}_i^{BLUP}(y_i, \varphi)\}^2 = M_{1i} + M_{2i}. \tag{4.21}$$

We have $\hat{p}_i^{EBLUP} = \hat{p}_i^{BLUP}(y_i, \hat{\varphi})$ and $\hat{p}_{i,-\ell}^{EPLUB} = \hat{p}_i^{BLUP}(y_i, \hat{\varphi}_{-\ell})$, where $\hat{\varphi}$ is the ML estimate of φ, and $\hat{\varphi}_{-\ell}$ the delete-ℓ ML estimatr obtained from $\{(y_i, n_i), i = 1, \ldots, m, i \neq \ell\}$. Let $\hat{M}_i^{BLUP}(y_i, \hat{\varphi})$ be the naive estimator of M_{1i}, obtained by substituting φ by the ML estimator $\hat{\varphi}$ in (4.15), and let $\hat{M}_i^{BLUP}(\hat{\varphi}_{-\ell})$ denotes the naive estimator when estimating parameters obtained by deleting the area ℓ to the data. The Jackknife estimator of mse_2 is given by :

$$mse_2^J(\hat{p}_i^{EBLUP}) = \hat{M}_{1i} + \hat{M}_{2i}, \tag{4.22}$$

where

$$\hat{M}_{1i} = \hat{M}_i^{BLUP}(\hat{\varphi}) - \frac{m-1}{m} \sum_{\ell=1}^{m} \{\hat{M}_i^{BLUP}(\hat{\varphi}_{-\ell}) - \hat{M}_i^{BLUP}(\hat{\varphi})\},$$

and

$$\hat{M}_{2i} = \frac{m-1}{m} \sum_{\ell=1}^{m} \left\{ \hat{p}_{i,-\ell}^{EBLUP} - \hat{p}_i^{EBLUP} \right\}^2.$$

4.9 Extension to unit level covariates

In this section we extend the results of sections 4.4-4.5 to handle covariates. Let N_i be the known population size of the area $i (i = 1, \ldots, m)$ and p_{ij} be the response probability of unit j in area i. This probability may depend on covariates $x_{ij} = (x_{ij1}, \ldots, x_{ij\ell})$ measured on that individual. The proposed model is

$$p_{ij} = e^{-a_i \psi_\alpha^{-1}(\pi_{ij})}, \quad g(\pi_{ij}) = x_{ij}^t \beta, \quad i = 1, \ldots, m, \quad j = 1, \ldots, N_i, \tag{4.23}$$

where a_i is a positive random effect belonging to one of the family of Table 4.1 whith Laplace transform $\psi_\alpha(t)$, g denotes some link function, and β is a vector of regression parameters. Let $\phi = (\alpha, \beta_1, \ldots, \beta_\ell)^T$, be the vector of parameters ; this section assumes that they are known. The goal is to develop Bayesian inference for p_{ij} given some observed data $\{Y_{i1}, \ldots, Y_{in_i}, n_i < N_i\}$ from the i^{th} area.

Let $\mathbf{x}_i$ denotes the matrix of covariates x_{ij} for the n_i observed units. Under models (4.23), the conditional joint probability mass function for the i^{th} area is given by :

$$\begin{aligned} f(\mathbf{y}_i|a_i,\mathbf{x}_i,\phi) &= Pr\big(Y_{i1}=y_{i1},\ldots,Y_{in_i}=y_{in_i}|a_i,x_i,\phi\big) \\ &= \prod_{j=1}^{n_i} e^{-a_i y_{ij}\psi_\alpha^{-1}(\pi_{ij})}\big\{1-e^{-a_i\psi_\alpha^{-1}(\pi_{ij})}\big\}^{1-y_{ij}} \\ &= \sum_v (-1)^{\sum_{j=1}^{n_i} v_j} e^{-a_i\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}}. \end{aligned} \tag{4.24}$$

Where the sum is over all $v=(v_1,\ldots,v_{n_i})\in\{0,1\}\times\cdots\times\{0,1\}$; $\bar{F}_{ij}(y_{ij}+v_j)$ takes the value 1 if $y_{ij}+v_j=0$, π_{ij} if $y_{ij}+v_j=1$ and 0 if $y_{ij}+v_j=2$. The marginal joint pmf of $(Y_{i1},\ldots,Y_{in_i})$ is obtained by integrating $f(\mathbf{y}_i|a_i\mathbf{x}_i,\phi)$ over the random effect a_i. It is given by,

$$l_i(\mathbf{y}_i|\mathbf{x}_i,\phi) = \sum_v (-1)^{\sum_{j=1}^{n_i} v_j}\psi_\alpha\Big\{\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}\Big\}. \tag{4.25}$$

Using equations (4.24)-(4.25), an application of Bayes 's formula leads to the following posterior distribution of the random effect a_i

$$g(a_i|\mathbf{y}_i,\mathbf{x}_i,\phi) = \frac{\sum_v(-1)^{\sum_{j=1}^{n_i} v_j} e^{-a_i\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}} g(a_i;\alpha)}{l_i(\mathbf{y}_i|\mathbf{x}_i,\phi)}. \tag{4.26}$$

Where $g(a_i;\alpha)$ is the marginal density (or marginal pmf) of a_i. The Laplace transform of this distribution is

$$\begin{aligned} \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(t) &= E\Big(e^{-ta_i}|\mathbf{y}_i,\mathbf{x}_i,\phi\Big) \\ &= \int_0^\infty e^{-at} g(a_i|\mathbf{y}_i,\mathbf{x}_i,\phi)da_i \\ &= \frac{\sum_v(-1)^{\sum_{j=1}^{n_i} v_j}\int_0^\infty e^{-a_i\left[t+\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}\right]} g(a_i;\alpha)da_i}{l_i(\mathbf{y}_i|\mathbf{x}_i,\phi)}, \end{aligned} \tag{4.27}$$

The integrals in (4.27) are the Laplace transforms of a_i evaluated at $t+\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}$. Therefore,

$$\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(t) = \frac{\sum_v(-1)^{\sum_{j=1}^{n_i} v_j}\psi_\alpha\big[t+\sum_{j=1}^{n_i}\psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij}+v_j)\}\big]}{l_i(\mathbf{y}_i|\mathbf{x}_i,\phi)}. \tag{4.28}$$

Given $\mathbf{y}_i$, this posterior Laplace transform characterizes the dependency between non- observed units in the i^{th} area.

Proposition 3 *Given $(Y_{i1},\ldots,Y_{in_i})=\mathbf{y}_i$, the dependence between non sampled units in area i, indexed by $j=n_i+1,\ldots,N_i$, is modeled by a random effect model defined by (4.23) with the following characteristics*

$$\begin{aligned} \pi_{ij\mathbf{y}_i} &= \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{\psi_\alpha^{-1}(\pi_{ij})\}, \\ p_{ij\mathbf{y}_i} &= e^{-a_{i\mathbf{y}_i}\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}^{-1}(\pi_{ij\mathbf{y}_i})}, \end{aligned} \tag{4.29}$$

where $a_{i\mathbf{y}_i}$ denotes the random effect that account for dependence given the data, with Laplace transform $\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(\cdot)$; $\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}^{-1}(\cdot)$ is the inverse of $\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(\cdot)$.

Consider now the Bayes estimation of p_{ij}. Results in the Proposition 3 can be used to find $E(p_{ij}|\phi,\mathbf{y}_i)$, the Bayes estimate of p_{ij}, and its posterior variance. These quantities can be expressed in terms of the posterior Laplace transform $\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(\cdot)$ as

$$\hat{p}^{BP}_{ij,\mathbf{y}_i,\mathbf{x}_i}(\phi) = E_{a_{i,\mathbf{y}_i}}[p_{ij,\mathbf{y}_i}] = \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{\psi_\alpha^{-1}(\pi_{ij})\}, \tag{4.30}$$

and its posterior variance is

$$MSE^c\{\hat{p}^{BP}_{ij,\mathbf{y}_i,\mathbf{x}_i}(\phi)\} = E_{a_{i,\mathbf{y}_i}}[p^2_{ij,\mathbf{y}_i}] - E_{a_{i,\mathbf{y}_i}}[p_{ij,\mathbf{y}_i}]^2 = \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{2\psi_\alpha^{-1}(\pi_{ij})\} - [\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{\psi_\alpha^{-1}(\pi_{ij})\}]^2. \tag{4.31}$$

4.10 Model evaluation and data analysis

In this section, we evaluate the performance of the EBP obtained with the models depicted in Section 2 using both simulated and real data. We only investigate estimators considered when there are no covariates.

4.10.1 Simulation study

A simulation study was conducted to examine the finite sample performance of the predictors, EBP, and EBLUP, of a proportion. This simulation exercise considers two cases : case I) $\pi = 0.1$ and $\rho = 0.1$; case II) $\pi = 0.1$ and $\rho = 0.3$. Data was generated from (4.1), using Joe (J)'s and dJoe (dJ)'s models, with m =20, 40 areas and $n_i = 5$ units in each area i, $i = 1,\dots,m$. The number of Monte Carlo simulations is $R = 1000$. For each of the four combinations of parameters, a set of true proportions $\{p_i^r, i = 1,\dots,m\}$, and Monte Carlo samples $\{y_i^r, i = 1,\dots,m\}$ were generated 1000 times.

Using the data from each sample r, we computed the EBLUP $\{\hat{p}^r_{1,i}, i = 1,\dots,m\}$, and the EBP $\{\hat{p}^r_{2,i}, i = 1,\dots,m\}$ of p_i^r. The EBP was calculated using Joe (J)'s and dJoe (dJ)'s models. To compare the EBLUP and EBP, we can calculate the Relative Bias (RB) and the Relative Root Mean Squares Error (RRMSE) to evaluate the accuracy of the small area estimates. We also calculate the efficiency (RE) of EBP relative to EBLUP. For each predictor k, these quantities are defined by :

- The relative bias

$$RB_i = \frac{\frac{1}{R}\sum_{r=1}^R(\hat{p}^r_{k,i} - p_i^r)}{\sum_{r=1}^R p_i^r/R}, \quad k = 1,2.$$

- The relative root mean square error

$$RRMSE_i = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^R(\hat{p}^r_{k,i} - p_i^r)^2}}{\sum_{r=1}^R p_i^r/R}, \quad k = 1,2.$$

- The relative accuracy, RE of the EBP, $\hat{p}^r_{2,i}$ with respect to the EBLUP, $\hat{p}^r_{1,i}$

$$RE_i = \frac{\sum_{r=1}^R(\hat{p}^r_{1,i} - p_i^r)^2}{\sum_{r=1}^R(\hat{p}^r_{2,i} - p_i^r)^2}.$$

Table 4.2 repports mean relative biases ($\overline{RB}$), and the mean relative root mean square error ($\overline{RRMSE}$) of the EBP and EBLUP that assume that the model is selected correctly. The mean relative accuracy ($\overline{RE}$) is also repported. All means are over small areas, and are expressed as $(\times 100)$. For Joe's data, Table 4.2 shows that the mean RBs of EBLUP and EBP very are similar, especially for $\rho = 0.1$. However, the mean RRMSE of the EBP is always smaller than that of the EBLUP. The efficiency of the EBP relative to EBLUP is always larger than 1.

On the other hand, when data are generated from dJoe's model, the EBLUP displays mean RB slightly smaller than that of EBP. However, it mean RRMSE is always above that of the EBP. The mean relative accuracy of EBP with respect to the EBLUP is always larger than 1. As expected, results are better, as the number of area increases, or as the intra-area correlation decreases. Comparing now the EBP under Joe's and dJoe's models, Table 4.2 shows that the accuracy of the EBP relative to the EBLUP under dJoe's model exceeds that for Joe's model.

Overall, these results show that when copula model is correctly specified, the EBP performs better than the EBLUP. the EBP constructed under models investigated in section 4.4 are reliable statistical techniques.

				EBLUP		EBP		Efficiency
	m	Data	Models	$\overline{RB}(\times 100)$	$\overline{RRMSE}(\times 100)$	$\overline{RB}(\times 100)$	$\overline{RRMSE}(\times 100)$	$\overline{RE}$
Case I	20	Joe	Joe	0.35	88.15	0.35	85.85	1.05
	40	Joe	Joe	0.84	83.21	0.85	79.89	1.09
	20	d-Joe	d-Joe	-0.96	73.83	-1.21	63.76	1.35
	40	d-Joe	d-Joe	-0.52	74.60	-0.62	60.97	1.51
Case II	20	Joe	Joe	1.72	101.72	1.80	93.91	1.17
	40	Joe	Joe	-0.25	98.06	-0.22	87.60	1.26
	20	d-Joe	d-Joe	-0.75	90.98	-0.92	80.75	1.27
	40	d-Joe	d-Joe	-0.83	90.75	-1.01	77.56	1.37

TABLE 4.2 – The mean relative bias ($\overline{RB}$) and the mean relative root mean square error ($\overline{RRMSE}$) of EBLUP and EBP, and the mean efficientcy ($\overline{RE}$) of EBP relative to EBLUP.

4.10.2 An example

The data analyzed here is presented and analyzed in Example 5.1 of Tounkara and Rivest (2014). It is concerned with the impact of training physicians in shared decision making on their patients involvement in the decision making process. It involves patients in $m = 9$ community health services of the Quebec region, and the binary response indicates whether a patient reports an active role in the decision about taking an antibiotics treatment for an acute respiratory infection. The goal is to estimate the proportion of such patients. The analysis of Tounkara and Rivest (2014) shows that the Frank model provided best fit of the data. Thus, we only consider the results for this model. Their result also reveal that the intra-cluster correlation for this data is significantly positive.

In Table 4.3, we report the estimates $(\hat{p}_i)$ and the square root of the estimated jackknife mean squared error (mse^J) for the EBP and the EBLUP. These results display the domination of the EBP over the EBLUP except for comminuties 1 and 5, where the EBLUP seems preferable.

			Direct		EBLUP		EBP	
	n	y	$\hat{p}_i$	s.e	$\hat{p}_i$	$\sqrt{mse^J}$	$\hat{p}_i$	$\sqrt{mse^J}$
1	38	3	0.08	0.04	0.12	0.09	0.11	0.10
2	39	8	0.20	0.07	0.20	0.06	0.22	0.05
3	49	10	0.20	0.06	0.20	0.06	0.22	0.05
4	29	7	0.24	0.08	0.21	0.07	0.22	0.06
5	49	2	0.04	0.03	0.10	0.11	0.05	0.14
6	35	3	0.09	0.05	0.13	0.09	0.13	0.08
7	51	13	0.26	0.06	0.23	0.07	0.22	0.05
8	42	8	0.19	0.06	0.19	0.06	0.22	0.05
9	27	8	0.30	0.09	0.24	0.08	0.22	0.06

TABLE 4.3 – Estimates and standard error for a community intervention trial.

4.11 Conclusion

This work investigated the empirical best method under Archimedean copulas to estimate small area proportions. A wide range of distributions for whitin area dependency is available. AIC can be used to select a particular copula model. We obtain an explicit expression for the EBP. The efficiency of the BP relative to BLUP is studied using graphical tools. We used the jackknife to obtain estimates of uncertainty associated with the empirical estimators. The empirical evaluations based on simulated and real data reveal good performance of the EBP under the proposed models. They show potential gains the EBP compared to the EBLUP. Further study could investigate the hierarchical Bayesian methods that specify distribution for parameters.

4.12 Transition

Le prochain et dernier chapitre de cette thèse est indépendant des chapitres précédents. Ce chapitre concerne l'analyse des caractéristiques socio-économiques des hommes qui ont une préférence à épouser des jeunes filles de moins de 18 ans. Dans ce contexte, nous utilisons les copules Archimédiennes bidimentionelles pour modéliser l'association entre le niveau d'éducation (variable discrète) des hommes et leur revenu prémarital (variable continue). Nous construisons la vraisemblance pour un échantillon issu de ce couple de variables aléatoires mixtes, et déduisons une estimation du paramètre de dépendance en utilisant une procédure semi-paramétrique où les marges sont estimées par leurs équivalents empiriques. Nous utilisons la méthode du jackknife pour estimer l'erreur type. Nous utilisons la méthode de Wald, pour tester l'égalité entre l'association des caractéristiques socio-économiques des hommes qui épousent des jeunes filles mineures et celle des hommes qui se marient avec des femmes âgées. Un exemple numérique, portant sur les données de l'EDS 2006 du Niger, sera présenté.

Chapitre 5

On the socioeconomic characteristics of men who marry underage girls

S. Dessy, S. Diarra and F. Tounkara.

The World Bank Economic , submitted.
Keywords : Child marriage, developing countries, men, education, income.

5.1 Résumé

La lutte contre le mariage des enfants dans les pays en développement est urgente, car elle a des retombées positives pour le développement humain. La littérature existante suggère que pour éliminer cette pratique néfaste, les facteurs de l'offre et de la demande doivent être ciblées. Cependant, alors que la pauvreté multiforme des parents est largement reconnue comme le principal moteur de la fourniture de filles mariées, sur le côté de la demande, il n' y a toujours pas d'explication claire de la raison pour laquelle une grande proportion d'hommes dans les pays en développement épousent des jeunes filles de moins de 18 ans. En se basant sur la théorie économique et des modèles empiriques avec l'application sur des données du Niger, nous constatons que les hommes qui ont un niveau de revenu bas et/ou une éducation basse ont tendance à épouser des filles mineures, tandis que ceux qui ont un niveau plus élevé de revenu et de l'éducation ont tendance à se marier à des femmes d'âge légal.

5.2 Abstract

Combating child marriage in developing countries is urgent because it has positive spillovers for human development. The existing literature suggests that to eliminate this harmful practice, both supply and demand factors must be targeted. However, while parents' multifaceted poverty is widely recognized as the main driver of the supply of child brides, on the demand side, there is still no clear explanation of why a large proportion of men in developing countries marry underage girls. Based on economic theory and empirical modeling with application to Niger, we find that men who have a low level of income and/or education tend

to marry underage girls, while those who have a higher level of both income and education tend to marry legal-age women.

5.3 Introduction

Like child labor— a well-documented barrier to development—, child marriage has almost disappeared in developed countries, but remains prevalent in many developing countries. However, unlike child labor which victimizes both boys and girls indiscriminately, and has comparable impacts on their respective lives, child marriage is predominantly a girls' phenomenon, and constitutes a far greater threat to young girls' lives and future prospects than to boys'(UNFPA 2012). According to a 2012 Report produced by The United Nation's Population Fund (UNFPA), 8 of the 10 countries with the highest prevalence rates of child marriage in the world are located in sub-Saharan Africa. Combatting child marriage in sub-Saharan Africa and elsewhere is thus urgent because it has positive spillovers for the achievement of the Millennium Development Goals (MDGs) : it can promote gender equality and empower women (MDG 3) by reducing the age gap between married women and their spouses ; it can reduce both child mortality (MDG 4) and maternal mortality (MDG 5) by increasing the average age at first birth for women ; it can also promote environmental sustainability (MDG6) by mitigating demographic explosion that exerts increased pressure on existing natural resources and tends to encourage the adoption of unsustainable production methods.

The development literature (Jensen and Thornton, 2003; Vogelstein, 2013) suggests that to eliminate this harmful practice, both supply and demand factors must be targeted. On the supply side, there is a consensus that public intervention must target multifaceted poverty in order to be effective at mitigating incentives for parents to marry off their underage girls. But on the demand side, effective action remains impeded by lack of understanding of the reasons why men marry underage girls. One possible reason is that men in countries reporting a high prevalence of child marriage may have no choice but to marry underage girls despite them not being emotionally and physically fit to reproduce, because legal age single women are in short supply. But if so, then one would expect grooms to be willing to wait a few years (when their brides reach the legal age of 18) before initiating the reproductive process. Yet the 2006 Niger's Demographic and Health Survey (DHS) reveals that the median age at first birth was 17 years for married women aged 20 – 24 at the time of the survey, implying that more than half of these women became pregnant before the legal age of 18.

Another possible reason is that a large cross-section of single men simply attaches an added value to marriage with underage girls. Indeed, using data from 2012 Niger's DHS, Cockburn and al. (2015) show that 27% of the observed prevalence of child marriage among married women aged 20 – 24 can be explained by the added value men, in average, attach to having a child bride. Motivated by this finding, this paper aims to explore both analytically and quantitatively the socioeconomic outcomes of men likely who attach an added value to having a child bride : what are the socioeconomic outcomes of men who prefer underage girls ? Do these outcomes differ from those of men who choose to have legal-age brides ?

To address these questions, our starting point is a descriptive overview of the distribution of married men in Niger, by bride's type (child bride of legal-age bride), level of education, and income. The data used in this descriptive overview is extracted from the 2006 DHS, and represent a sample of married men whose wives were aged 20 – 29 at the time of the survey. We identified each married man in this sample by his number of years of schooling completed, his household standard of living (which we take as a proxy of his pre-marital standard of living), and the age of his wife at the time their marital union was sealed.

Men's characteristics		% of married men by bride's type	
Education	**Wealth**	< 18	≥ 18
Less educated	poor	85.53	14.47
	rich	80.88	19.12
Educated	poor	72.89	27.11
	rich	48.66	51.34

TABLE 5.1 – Descriptive statistics for Niger

Table 5.1 above represents four different categories of married men, depending upon their pair of socioeconomic outcomes. The first category consists of married men who have a low level of education (i.e., have less than 6 years of education) and are relatively poor (i.e., they belong to the first, second, or third wealth quintiles) ; the second consists of those who have low levels of education, but have a relatively high standard of living (i.e., they belong to the fourth and fifth wealth quintiles) ; the third consists of those who are educated (i.e., have more than six years of education) but poor ; finally, the fourth category consists of married men who are educated and have a relatively high standard of living. Table 1 shows that nearly 86% of men who have a low level of education and are poor married underage girls. Among men who have a low level of education but have a relatively high standard of living, the percentage of those who married underage girls decreases slightly to 81%. It decreases further among the educated and poor, but still remains high at 73%. By contrast, this figure falls to 49% among men the educated and rich. Table 5.1 thus suggests that men who choose legal-age bride in Niger tend to have both a high level of education and a high income, while those who choose child brides tend to have a low level of education, or a low level income, or a low level of both.

The present work therefore aims to achieve two main objectives. First, we build a model of prospective grooms' decision on bride type that reproduces the stylized facts described in Table 1 above. Our theory has two main ingredients. The first consists of the desired characteristics of a bride, including reproductive fitness, capacity to contribute income to the newly formed household, and submission to husband's control of household fertility and economic resources. On the basis of these characteristics, we model a prospective groom's surplus from marriage as having two components, including a reproductive surplus and an economic surplus tied to his span of control over household resources. The second ingredient consists of the determinants of the prospective groom's ability to select a bride on the basis of these desired characteristics. There are two such determinants, namely the prospective groom's level of education and his level of income. Prospective grooms differ with respect to these two socioeconomic outcomes. The joint distribution of these outcomes interacts with the desired bride's characteristics to divide the population of prospective grooms between those who choose child brides and those who choose legal-age brides instead. Our model predicts that prospective grooms with a low level education and/or income choose child brides, while those with a higher level of both education and income choose legal-age brides.

The intuition behind this prediction is as follows. On the one hand, relative to a prospective groom with a low level of education, a prospective groom with high level of education may more clearly perceive the inferior reproductive fitness of a child bride. However, because reproductive fitness is not the only desired bride's characteristic, having a high level of education, though necessary, is not sufficient for a prospective groom to prefer a legal-age bride over a child bride. On the other hand, a prospective groom with a high level of income may be indifferent between a child bride and a legal-age bride from the viewpoint of both their capacity to contribute the economic prosperity of their future household and the extent to which they can be sexually controlled. However, he may still lean towards a child bride ; if he has too low a level of education to clearly perceive a child bride's inferior reproductive fitness. This explains why, in our model, prospective grooms with a high level of both education and pre-marital are those most likely to choose to have legal-age brides.

Second, using the women's questionnaire in the 2006 DHS, we build two bivariate samples of married men. One such sample consists of independent observations of education and pre-marital income among men whose wives were underage (i.e., aged less than 18) at the start of their marriage, and the other consists of similar bivariate data pertaining to men whose wives had the legal age (i.e., aged 18 or above). Then, for each of these two samples, we parameterize the joint distribution of education and income using a copula-based approach. The purpose of this empirical exercise is to examine the extent to which the structure of the association between education and income differs between Niger's men who marry underage girls and those who don't. There are two main steps in our empirical analysis.

In the first step, we develop a statistical model for the joint distribution of education and income among married men. Following Genest and Favre (2007) and Zimmer (2013), we adopt a copula-based approach for modeling this joint distribution. This approach allows us to parametrically specify the unknown joint bivariate distribution of these two socioeconomic outcomes in terms of their respective margins, and the copula that binds them together. Indeed, a copula is characterized by a dependence parameter measuring the strength of the association between two or more variables. Our main interest in this copula-based approach is therefore to estimate this dependence parameter. There are two technical issues which we must resolve in order to make effective use of this copula-based approach for the data at hand : first, how to construct a copula-based log-likelihood function underlying a maximum-likelihood-based estimation of the dependence parameter ; and second, what parametric model to choose for the copula.

The first of the two technical issues mentioned above occurs because copulas are parametric expressions of cumulative distribution functions (CDFs), which means that to apply the maximum likelihood estimation method they must be converted to probability density functions (pdfs). In the case of continuous outcomes, this conversion is obtained by straightforward cross-partial differentiation of the copula. The data underlying our empirical analysis, however, are mixed discrete and continuous outcomes, which do not lend themselves to cross-partial differentiation. To convert to pdfs, we follow Zimmer (2013), by combining differentiating, for the continuous outcome (income), and differencing, for the discrete outcome (level of education).

As stated above, the second technical issue is a model selection problem. We combine pragmatism and formal analytical tools to solve this problem. Since a positive dependence is expected between men's years of schooling completed and their (pre-marital) income, we focus our search for candidate copula models to those with parameter ranges that encompass both positive and negative dependence. This makes the class of Archimedean copulas an adequate source of candidate models, due to their flexibility and mathematical tractability. For the empirical analysis to be carried in this work, candidate copulas models include the Clayton, Frank, Gumbel, and Joe copulas. Given these candidates models, we then complete our selection of the most suitable copula model for the data set at hand by using the *Akaike Information Criterion* (AIC). The AIC provides a ranking of all the candidate copula models in terms of how small the information lost is when a given copula model is used to represent the process generating the data. Despite being conceived for use in parametric settings, its use in a semi-parametric setting like ours is still justified in the analysis of one-parameter bivariate copulas when the sample size and the Kendall tau measuring dependence in the data are sufficiently high $\Big($ Jordanger and Tjostheim (2014)$\Big)$. Our data indeed satisfy this requirement.

In the second step of our empirical analysis, we perform a hypothesis test to gauge the extent to which the socioeconomic outcomes of men who marry underage girls differ from those of men who marry older women instead. Our hypothesis test takes the form of a parametric test of equality between two dependence structures.

In line with our theory also outlined above, we perform a one-tailed hypothesis test under the null hypothesis of identical dependence structures for men who married underage girls and for those who married women of legal age instead. The alternative to this null hypothesis is that the dependence between education and pre-marital income is weaker for men who married underage girls than for those who married women of legal age.

We use a Wald-type test statistic constructed using the estimators of the dependence parameters from the first and second samples. Under the null hypothesis of identical dependence structures, the difference between the two estimators should be zero. Following Genest and Favre (2007), these two estimators were obtained by applying a maximum pseudo-likelihood estimation method (mple)—a semi-parametric estimation method for the dependence parameter. However, since this estimation method underestimates the variance of the estimator, we also apply a Jackknife resampling method to mitigate this efficiency loss.

We reject the null hypothesis of equality between the two dependence structures, in favor of the alternative that dependence between years of schooling completed and pre-marital income is weaker for men married to underage girls than for those married to older women. Indeed, estimation results show that dependence between years of schooling completed and pre-marital income is twice as strong for men who married older women (5.38) than for those who married underage girls (2.36). The interpretation of this result reads as follows. First, as we show in Table 3 further below, men whose wives were underage at the start of the marriage have in average a lower level of education and a lower level of pre-marital income than those whose wives were of legal age at the start of the marriage. Second, this feature of the data, combined with the fact that the positive dependence between education and pre-marital income is twice as strong for men who married legal-age women than for those who married underage girls, suggests that men who married legal-age women are more likely to have both a higher level of education and a higher level of pre-marital income. Third, a man with either a low level of education or a low level of pre-marital income is therefore more likely to marry an underage girl, because dependence between these two variables is weaker for men married to underage girls. These findings suggest that intervention aimed at discouraging the demand for child brides should target young men who have either a low level of education, or a low level of pre-marital income, a low level of both.

The economics literature on child marriage is still in its infancy, although several notable contributions exist, including Jensen and Thornton (2003), Field and Ambrus (2008), Nguyen and Quentin (2012), and Wahhaj (2014). Our paper is more closely related to Wahhaj (2014) who models the effects of demand-side factors on the prevalence of child marriage, with application to South Asia. We contribute to this literature by offering a theoretical and empirical exploration of the socioeconomic characteristics of men most likely to marry underage girls.

The rest of this study is structured as follows. In section 5.4, we develop a theoretical model of child marriage emphasizing the determinants of the demand for underage brides. Section 5.5 outlines the probability model that provides a basis for testing the predictions of our theory. Section 5.6 applies this probability model to Niger's data and presents the test results. Finally section 5.7 concludes.

5.4 Elements of the model

Consider a population of N prospective grooms who must decide on the type of bride they want to have. The choice is between a child bride (understood as a single female under 18 years of age) and a legal-age bride (understood as a single female aged 18 or above). Each prospective groom is characterized by a pair of socioeconomic outcomes (S_i, W_i), drawn from an unknown bivariate cumulative probability distribution function, $H(s, \omega)$, where s and ω are realizations of the random variables S and W, respectively. The variables S and W represent a prospective groom's level of education and pre-marital income, respectively.

Let $a_i \in \{0, 1\}$ denotes the binary decision on bride type : $a_i = 0$ means prospective groom i chooses to marry an underage girl, while $a_i = 1$ means he choose to marry a legal-age women. We assume that every prospective groom's choice of bride type is unconstrained by availability of the type. This allows us to focus on the role played by men's preferences for bride type. When a prospective groom makes his choice, marriage is arranged between the two parties, at an exogenous cost b to the groom. In the case of sub-Saharan Africa, one can think of b as the brideprice or a gift extended by the groom to the bride family, including livestock, poultry, jewelry, and/or cash (Anderson, 2007).

5.4.1 A groom's expected surplus from a marital match

Our demand-side theory of child marriage builds from a number of stylized facts related to men's desired characteristics of brides in developing countries. First, social norms in many developing countries prescribe virginity as a desired characteristic of the bride, particularly in the context where sexual promiscuity and sexually transmitted diseases such as HIV are prevalent Jensen and Thornton (2003).

Concerns about bride's virginity may thus tip the balance in favor of child brides. Second, child brides are perceived as more submissive, implying that they are less likely to challenge their spouse for equal control of household resources and of the couple's sexuality (Jensen and Thornton, 2003; Field and Ambrus, 2008). We formalize these three facts by relating them to the marital surplus P a prospective groom derives from a marital match. In particular, we assume that prospective groom i's marriage surplus depends on his marriage-posting strategy a_i, and has two components : a reproductive surplus $\Phi(a_i)$ and an economic surplus $\Omega(a_i)$. In other words, if $P(a_i)$ denotes prospective groom i's realized surplus from a marital match, then

$$P(a_i) := \Phi(a_i) + \gamma\Omega(a_i), \qquad i = 1, ..., N, \tag{5.1}$$

where γ is a positive scalar that converts units of economic surplus into equivalent units of reproductive surplus.

The reproductive surplus

Assume for the time being that there are no differences between child brides and legal-age brides in terms of the risk of death from maternal causes. If a prospective groom i posts his marriage offer in the child segment of the supply side of the market (i.e., $a_i = 0$), and thus marries an underage girl, he gets a reproductive surplus given by $\Phi(0) = k_0$. If he posts his offer in the older women segment instead, and thus marries an older woman, he gets a reproductive surplus given by $\Phi(0) = k_1$. The terms k_0 and k_1 summarize men's perceptions of the physical characteristics of underage, and full age brides respectively, including the length of their reproductive life, and their sexual status (virgin or not). We thus make the following assumption :

Assumption 1. $k_0 > k_1 > 0$.

Assumption 1 is a direct reflection of the facts that men perceive child brides as having potentially longer reproductive lives, are more likely to be virgin, and are also more sexually controllable by their husbands due to their younger age (Jensen and Thornton, 2003; Field and Ambrus, 2008). To the extent that these characteristics are desired by developing country's men, and in the absence of differential risks of death from maternal causes between the two groups of women, one would expect men who marry underage girls to receive a higher (expected) reproductive surplus than those who marry older women.

We can then generically express the realized reproductive surplus of a prospective groom who adopt the marriage-posting strategy $a_i \in \{0, 1\}$ as follows :

$$\Phi(a_i) = (1 - a_i) k_0 + a_i k_1, \tag{5.2}$$

all i.

The economic surplus

Once a match occurs, childbearing also begins. We assume that a bride's economic contribution starts after her childbearing years. Thus a child bride will contribute w_0 to the household economic resources through her labor, after her childbearing years, while a legal-age bride will contribute w_1.

Assumption 2. $w_0 \leq w_1$.

Assumption 2 states that the economic contribution of a child bride to her household's economic resources after completion of childbearing is at most as high as that of a legal-age bride. The legal-age bride being more mature and having a shorter reproductive life-span may be more productive than her younger counterpart.

The total level of resources generated by a marital match involving a child bride is given by $W_i + w_0 - b$, of which the husband controls a share $\lambda_0 \in (0,1)$. Likewise, the corresponding level of household resources for a marital match involving a legal-age bride is $W_i + w_1 - b$, of which the husband controls a share $\lambda_1 \in (0,1)$. We can thus obtain a generic expression for the realized economic surplus from the marital match accrued to the groom as follows : $\forall a_i \in \{0,1\}$,

$$\Omega(a_i) = \ln([W_i + a_i w_1 + (1-a_i) w_0 - b] \Lambda(a_i) - \underline{c}), \tag{5.3}$$

all i, where $\underline{c}$ denotes an exogenously given subsistence requirement for income, and

$$\Lambda(a_i) = \begin{cases} \lambda_0 & \text{if } a_i = 0 \\ \\ \lambda_1 & \text{if } a_i = 1 \end{cases} .$$

Assumption 3. $0 < \lambda_1 < \lambda_0 < 1$.

Assumption 3 reflects the fact that child brides tend to have a lower status in their husbands' families, and thus are more controllable, see Field and Ambrus (2008).

The risk of maternal mortality

The surpluses specified in (5.2) and (5.3) above are those that would have realized to a typical groom i in the absence of maternal mortality. However, empirical evidence from a 2014 *WHO* Report points to fact that maternal mortality rates are generally high in developing countries, with girls who marry early being overrepresented among the victims. [1] Denote as $\rho_0 \in (0,1)$ the probability of maternal mortality in a marital match where the bride was underage at the time she was married off, and by $\rho_1 \in (0,1)$ the corresponding figure forlegal-age bride.

Assumption 4. $0 < \rho_1 < \rho_0 < 1$.

As reproductive fitness is one of the desired characteristics of brides in developing countries, inferior fitness as implied by Assumption 4 should normally make potential child brides unattractive to men. Yet in developing countries such as Niger, high rates of prevalence of child marriage tend to coexists with high rates of maternal mortality for child brides, suggesting men's inability to factor in this information in their marriage decision. To capture this fact, we assume that a prospective groom's perception of his preferred bride's likelihood of death from maternal causes depends on his level of education, as measured by his number of years of schooling completed.

[1] WHO (2014). Maternal Mortality, *Fact Sheet* N^{o}348, available online at http ://www.who.int/mediacentre/factsheets/fs348/en/

It is well-known from anthropological and psychological studies (Van Dyk, 2001) that in sub-Saharan Africa, lack of education nurtures beliefs in witchcraft and sorcery as explanation to events such as illness, and death. It is therefore plausible that an individual who adheres to traditional beliefs in witchcraft and sorcery will not perceive the likelihood of maternal mortality to be different between a child bride and a bride of legal age. This fact highlights the role education may play in dispelling such erroneous perception.

Denote as $\rho_0(S_i)$ the subjective probability of bride's maternal mortality as perceived by a prospective groom with a level of education, $S_i \in \{0,....,\bar{s}\}$, when his proposed bride is underage. Likewise, denote as $\rho_1(S_i)$ the corresponding subjective probability when the proposed bride has legal age.

Assumption 5. Given $S_i \geq 0$,

$$\rho_0(S_i) = \begin{cases} \frac{S_i}{s^*}\rho_0 + \frac{(s^*-S_i)}{s^*}\bar{\rho} & S_i < s^* \\ \\ \rho_0 & S_i \geq s^* \end{cases} \tag{5.4}$$

$$\rho_1(S_i) = \begin{cases} \frac{S_i}{s^*}\rho_1 + \frac{(s^*-S_i)}{s^*}\bar{\rho} & S_i < s^* \\ \\ \rho_1 & S_i \geq s^* \end{cases}$$

all $i = 1,.....,N^m$, where $s^* \in \{0,\bar{s}\}$ denotes the threshold number of years of schooling completed from which a prospective groom is able to clearly perceive the true likelihood of maternal mortality of his chosen bride type, and $\bar{\rho} \in [0,1]$.

Assumption 5 implies that for a prospective groom with no formal education, i.e., $S_i = 0$, $\rho_0(0) = \rho_1(S_i) = \bar{\rho}$, implying that he sees no difference in the likelihood of maternal mortality between a child bride and a legal-age bride. By contrast, for a prospective groom with a level of education $S_i = s^*$, $\rho_0(s^*) = \rho_0$ and $\rho_1(s^*) = \rho_1$, implying that he can clealry perceives the difference in the likelihood of maternal mortality between an underage bride and a legal-age bride.

As a final feature of men's marriage surplus, we also assume that when a bride suffers maternal mortality, her husband loses the entire reproductive surplus, $\Phi(a_i)$, as well as her economic contribution, $w(a_i) \in \{w_0, w_1\}$. In that case, his marriage surplus reduces to $\gamma \ln(W_i - b - \underline{c})$. Therefore, combining (5.2), (5.3) and Assumption 5, we obtain the expected marital surplus of prospective groom i as follows :

$$\widetilde{P}(a_i, S_i, W_i) = \begin{cases} [1-\rho_0(S_i)] f_0(W_i) + \rho_0(S_i)\gamma \ln(W_i - b - \underline{c}) & \text{if } a_i = 0 \\ \\ [1-\rho_1(S_i)] f_1(W_i) + \rho_1(S_i)\gamma \ln(W_i - b - \underline{c}) & \text{if } a_i = 1 \end{cases}, \tag{5.5}$$

for all $i = 1,...,N^m$, where

$$f_0(W_i) \; : \; = k_0 + \gamma \ln[(W_i + w_0 - b)\lambda_0 - \underline{c}]$$

$$f_1(W_i) \; : \; = k_1 + \gamma \ln[(W_i + w_1 - b)\lambda_1 - \underline{c}]$$

and $\rho_0\left(s_i\right)$ is as defined in (5.4). The above specification of the expected marital surplus accrued to a groom implicitly assumes that when it occurs, maternal mortality strikes at first birth for all types of women, so that the marital match yields no reproductive surplus, and in addition, the widower loses the economic contribution of the deceased bride. This is just a simplifying assumption. Our results are qualitatively invariant to the relaxation of this assumption.

5.4.2 Who gains from marrying underage girls ?

In this sub-section, we discuss the socioeconomic outcomes of prospective grooms who gain from marrying underage girls. Consider a prospective groom i with socioeconomic outcomes (S_i, W_i). Suppose he chooses to have a child bride (i.e., $a_i = 0$). Because choosing the marriage strategy $a_i = 0$ has an opportunity cost given by the expected surplus from marrying a legal-age woman $\widetilde{P}(1, S_i, W_i)$, we define the net expected surplus associated with the marriage-posting strategy $a_i = 0$ as follows :

$$\pi(S_i, W_i) := \widetilde{P}(0, S_i, W_i) - \widetilde{P}(1, S_i, W_i), \tag{5.6}$$

where $\widetilde{P}(a_i, S_i, W_i)$ is as defined in (5) above.

A prospective groom with socioeconomic outcomes (S_i, W_i) such that $\pi(S_i, W_i) > 0$ derives a positive net expected surplus from marrying an underage girl, and so gains from pursuing that bride-choice strategy (i.e., $a_i = 0$). In contrast, a prospective groom with socioeconomic outcomes (S_i, W_i) such that $\pi(S_i, W_i) < 0$ gains from having a legal-age bride, and thus will choose $a_i = 1$.

Without loss of generality, suppose $s = \bar{s}$, implying that perfect information about the true likelihood of maternal mortality for both underage and legal-age brides requires that the groom has the maximum level of formal education. Then, combining (5.5) and (5.6) by way of substitution, using (5.4) and Assumption 5, rearranging terms, leads to

$$\pi\left(S_i, W_i\right) = \left(1-\bar{\rho}\right)\left[\left(k_0 - k_1\right) + \zeta\left(W_i\right)\right] - \left[\beta + \left(\rho_0 - \rho_1\right)\xi\left(W_i\right) + \chi\left(W_i\right)\right]\frac{S_i}{\bar{s}} \tag{5.7}$$

all $i = 1, ..., N$, where

$$\begin{aligned}
\zeta\left(W_i\right) \quad &: \quad = \gamma \ln\left[\frac{\left(W_i + w_0 - b\right)\lambda_0 - \underline{c}}{\left(W_i + w_1 - b\right)\lambda_1 - \underline{c}}\right] \\
\xi\left(W_i\right) \quad &: \quad = \gamma \ln\left(W_i - b - \underline{c}\right) \\
\chi\left(W_i\right) \quad &: \quad = \gamma \ln\left(\frac{\left[\left(W_i + w_1 - b\right)\lambda_1 - \underline{c}\right]^{\rho_1 - \bar{\rho}}}{\left[\left(W_i + w_0 - b\right)\lambda_0 - \underline{c}\right]^{\rho_0 - \bar{\rho}}}\right)
\end{aligned}$$

$$\beta \quad : \quad = \left(\rho_0 - \bar{\rho}\right)k_0 - \left(\rho_1 - \bar{\rho}\right)k_1 > 0.$$

Then, take a prospective groom i with socioeconomic outcomes $(S_i, W_i) \in \{0, \bar{s}\} \times [\underline{\omega}, \bar{\omega}]$. We know from our discussion above that he will choose $a_i = 0$ if and only if, $\pi(S_i, W_i) > 0$. Using (5.7), yields a reformulation of this condition as follows :

$$S_i < \frac{(1-\bar{\rho})\left[(k_0 - k_1) + \zeta(W_i)\right]}{\beta + (\rho_0 - \rho_1)\,\xi(W_i) + \chi(W_i)}\bar{s} \equiv \varphi(W_i)\,. \tag{5.8}$$

Given W_i, the terms $\varphi(W_i)$ denotes the threshold education level below which a prospective groom gains from marrying an underage girl.

Recall that the main goal of this work is to gauge the Economique outcomes of men who choose to have child brides. We modeled a prospective groom's level of education (S) and (pre-marital) income (W) as joint determinants of bride choice strategy. Therefore, define

$$\begin{aligned}\Omega_0 &= \{(S_i, W_i) \in \{0, \bar{s}\} \times [\underline{\omega}, \bar{\omega}] : S_i < \varphi(W_i)\} \\ \\ \Omega_1 &= \{(S_i, W_i) \in \{0, \bar{s}\} \times [\underline{\omega}, \bar{\omega}] : S_i > \varphi(W_i)\}\end{aligned}$$

to be the sets of prospective grooms who choose to have a child bride and a legal-age bride, respectively. Denote as M_0 and M_1, the number of prospective grooms who choose to have a child bride and a legal-age bride respectively. Then ,

$$M_0 = |\Omega_0| \qquad \text{and} \quad M_1 = |\Omega_1|\,,$$

where $\left|\Omega_j\right|$ denotes the cardinality of the set Ω_j, $j = 0, 1$.

A question of interest for this work is whether the structure of the association between S and W in the set Ω_0 is similar/different to that in the set Ω_1. In other words, do the socioeconomic outcomes of men who choose to have a child bride differ from that of those who choose to have a legal-age bride ? The answer to this question depends in large part on the properties of the threshold function, $\varphi(.)$. Unfortunately, expression (5.8) is too complex to allow for straightforward discussion of these properties. For this reason, in what follows we solve a numerical example to illustrate the properties of the function $\varphi(.)$ and discuss their implication for the socioeconomic outcomes of men who marry underage girls.

5.4.3 A numerical example

In this example, we assign numerical values to relevant parameters, and graphically represent the threshold level of education below which a prospective groom choose to make a marriage offer to an underage girl. Without loss of generality, we set $\bar{\rho} = 0$, and normalize the brideprice to $b = 0$. We set the subsistence income $\underline{c}$ equals to unity. Since in sub-Saharan Africa, women in general earn less than their husband, we set an underage bride's contribution to household wealth to $w_0 = 0$, and and the contribution of an adult bride to $w_1 = \underline{c} = 1$.

A 2007 UNICEF Report focusing on women and children lists the age-gap between the spouses as an important factor giving men uncontested control over household resources. [2] Echoeing these facts, we pick

[2] UNICEF, 2007. Women and Children : The Double Dividends of Gender Equality, *The State of the World's Children 2007*

$\lambda_0 = .90$ and $\lambda_1 = .65$ as the share of household resources controlled by the husband in a household where the wife was married off as a child and in one where she was married at a legal age (i.e., above 18 years of age) respectively. With respect to the reproductive surplus, we pick $k_0 = 10$ and $k_1 = 5$. According to a 2014 Report by the *World Health Organization (WHO)* the probability that a 15 year old woman will eventually suffer maternal mortality is 1 in 160 in developing countries.[3] The 2012 DHS for Niger set at 15, the mean age at first marriage child brides. Therefore we pick $\rho_0 = 1/160 = 0.00625$. Again according to the same 2014 Report by the *WHO*, the maternal mortality ratio in developing countries in 2013 is 230 per 100 000, or 0.0023. However, this probability includes underage women, who are reported to face a higher probability of death from a maternal cause than older women. Therefore to account for this fact, we pick $\rho_1 = 0.0015$, which is smaller than the reported 0.0023. Finally, we set the scalar γ that converts units of economic surplus into equivalent units of reproductive surplus equal to 10, and add another scaler $\mu = 1/45$, so as to contain the threshold $\varphi(W_i)$ within the close range, $[0, 21]$, for all levels of income, where $21 = \bar{s}$ is the maximum level of education. The following table summarizes parameters use to generate Figure 5.1 below.

Parameter values													
γ	λ_0	λ_1	ρ_0	ρ_1	$\bar{\rho}$	μ	b	$\underline{c}$	k_0	k_1	$\bar{s}$	w_0	w_1
10	0.90	0.65	0.00625	0.0015	0	1/45	0	1	10	5	21	0	1

TABLE 5.2 – Parameters values

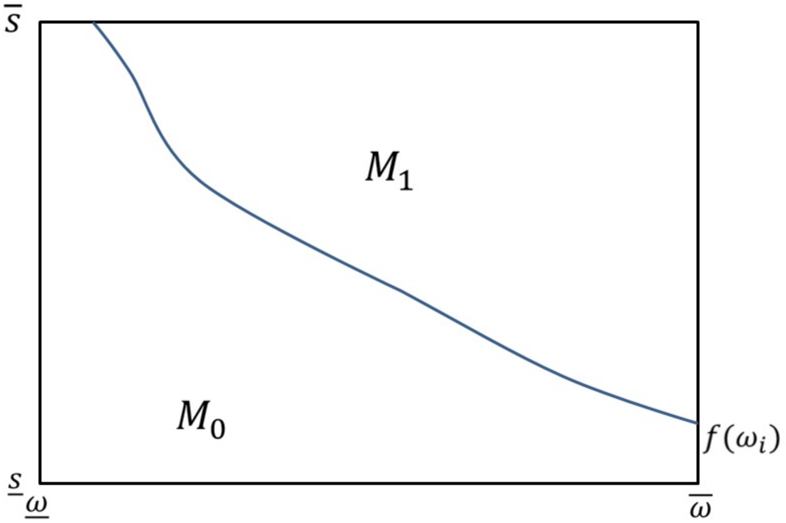


FIGURE 5.1 – Separating men who marry underage girls from those who do not

The area below the curve representing the function $\varphi(.)$ denotes the socioeconomic outcomes pairs (s, ω) such that all prospective grooms with such characteristics choose to marry underage girls. The area above this curve represents all the socioeconomic characteristics pairs (s, ω) such that all prospective grooms with such characteristics choose to marry older women. Figure 5.1 suggests that the higher the pre-marital income of a prospective groom, the more educated he will have to be in other to gain from marriage with a woman of legal age. This implies that the socioeconomic outcomes S and W display stronger positive dependence

[3] See WHO (2014) available online at http ://www.who.int/mediacentre/factsheets/fs348/en/

among men who marry legal-age women. In the next section, we specify a statistical model for structuring the association between education and pre-marital income and use it to fit our theory to Niger's DHS data.

5.5 A statistical model for the joint distribution of education and income

In this section we lay out the foundations of a statistical model for structuring the association between education and income among married men in Niger. As we mentioned above, our approach to constructing this statistical model is copula-based. The basic foundation of this approach is the application of Sklar's theorem, which ensures that for every bivariate joint cumulative distribution function (CDF) F and for any pair of outcome variables of X and Y, there exists a coupla $C : [0,1]^2 \to [0,1]$ such that :

$$F(x,y) = C[F_x(x), F_y(y); \theta] \equiv C(u,v;\theta), \tag{5.9}$$

where $u = F_x(x)$ and $v = F_y(y)$ are the realizations of the uniform random variables $U = F_x(X) \sim uniform[0,1]$ and $V = F_y(Y) \sim uniform[0,1]$ respectively, and θ denotes a parameter measuring the dependence between F_x and F_y. If X and Y are defined on a continuous scale, then C is uniquely specified and it can be written as :

$$C(u,v;\theta) = F(F_x^{-1}(u), F_y^{-1}(v)), \quad 0 \le u,v \le 1. \tag{5.10}$$

where the quantiles F_x^{-1} and F_y^{-1} are the inverses of F_x and F_y, respectively.

A copula contains all the information about the dependence between two or more random variables. The joint distribution F is thus specified by its marginal distributions and the copula function which binds them together. Compared to the traditional approach for constructing the joint distribution, the main advantage of the approach defined in (5.10) is that we can construct many bivariate distributions by linking any two univariate distributions, not necessarily of the same type, with any copula (Genest and Favre, 2007; Trivedi et al., 2005).

The choice of an appropriate copula depends on the suitability of the copula's dependence parameter for describing the dependence structure in the data. There are several copula families with different ways of modeling dependence in the literature, (Joe, 1997; Trivedi et al., 2005; Nelson, 2006). For the current work, our candidate models include the Frank, Clayton, Joe and Gumbel families, which belong to the class of Archimedean copulas. [4] The flexibility and analytical tractability of this class of copulas make them a suitable tool in many applications. In particular, Archimedean copulas possess a nice mathematical properties and are easy to construct. In general, a bivariate Archimedean copula can be defined as

$$C(u,v;\theta) = \phi^{-1}[\phi(u) + \phi(u)]$$

where ϕ denotes the generating function and ϕ^{-1} its inverse. ϕ is a continuous strictly decreasing function from $I = [0,1]$ to $[0,\infty]$ such that $\phi(1) = 0$. The function ϕ is called a generator of copula C.

Copula models have been mostly used for dependence between continuous random variables. However, there is a growing interest in the application of copula models for mixed (discrete and continuous) outcomes (de Leon and Wu, 2011; Dobra and Lenkoski, 2011; Song et al., 2009; Song, 2007).

[4] See Genest and Mackay ,1986, for a thorough discussion of bivariate Archimedean copulas.

In this paper, the bivariate data underlying our empirical analysis are mixed outcomes. We model the dependence structure of these outcomes using a bivariate Archimedean copula. In the present study, S denotes a prospective groom's level of education which is measured on a discrete scale, while W denotes his premarital income which is measured on a continuous scale. We denote by $f_s(., \theta_s)$ and $f_\omega(., \theta_\omega)$ the probability densities for S and W respectively, with θ_s and θ_ω as respective vectors of parameters. The likelihood for a single mixed observation (s, ω) is the joint density, which is obtained as in Zimmer (2013) by combining differentiating and differencing for the continuous and discrete outcomes, respectively, as in

$$f(s, \omega; \theta_s, \theta_\omega, \theta) = f_\omega(\omega, \theta_\omega)\Big\{C_v\big[F_s(s), F_\omega(\omega); \theta\big] - C_v\big[F_1(s-1), F_2(w); \theta\big]\Big\},$$

where F_s and F_ω denote the cumulative distribution functions (cdf) of S and W respectively ; and C_v is the derivative of the copula function in (5.10) over its second argument, $C_v(u, v, \theta) = \partial C(u, v, \theta)/\partial v$.

Given n independent bivariate mixed observations, the expression of the log-likelihood becomes

$$l(\Theta) = \sum_{i=1}^{n} \log\left\{f_\omega(w_i, \theta_\omega)\right\} + \sum_{i=1}^{n} \log\Big\{C_v\big[F_s(s_i), F_\omega(w_i); \theta\big] - C_v\big[F_s(s_i - 1), F_\omega(w_i); \theta\big]\Big\},$$

where $\Theta = (\theta_s, \theta_\omega, \theta)$. Parameters can be estimated using the maximum likelihood method (mle). Using this method, the estimator of θ_s, θ_ω, and θ are those values that maximize the log-likelihood function l. However, because the mle jointly estimates the margins and the dependence parameter of the copula, this approach could be computationally intensive in the case of high dimensional of model parameters Favre et al. (2004).

Inference functions for margins (IFM) can also be used. As outlined in Joe (1997), this approach is to perform the estimation in two steps, separately maximizing the univariate likelihoods $l(\theta_j) = \sum_{i=1}^{n} \log f_j(\theta_j)$, over θ_j , to get $\hat{\theta}_j$, $j = s, \omega$, and then maximizing $l(\hat{\theta}_s, \hat{\theta}_\omega, \theta)$ over θ to obtain $\hat{\theta}$. The IFM is more attractive in practice as it significantly reduces the computational cost, especially in higher-dimensional problems.

In the present study, we are only interested in the estimation of the dependence parameter θ, and this can be done without any prior knowledge of the margins Genest and Favre (2007). In this case, we use the pseudo log-likelihood function which can be written as :

$$l_n(\theta) = \sum_{i=1}^{n} \log\Big\{C_v\big[\hat{F}_s(s_i), \hat{F}_\omega(w_i); \theta\big] - C_v\big[\hat{F}_s(s_i - 1), \hat{F}_\omega(w_i); \theta\big]\Big\}, \tag{5.11}$$

where

$$\hat{F}_s(t) = \frac{1}{n+1}\sum_{i=1}^{n} I(S_i \leq t)$$

and

$$\hat{F}_\omega(w_i) = \frac{1}{n+1}\mathtt{Rank}(\omega)$$

are the empirical estimates of the discrete and continuous random variables, respectively.

The maximum pseudo likelihood estimator (mple) of θ is thus obtained by maximizing $l_n()$:

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta} l_n(\theta).$$

We can use the associated variance of $\hat{\theta}_n$ as the measure of its variability. However, this variance can lead to severe underestimation of $Var(\hat{\theta}_n)$ because it ignores the variability associated with $\hat{F}_s$ and $\hat{F}_\omega$. To overcome this drawback, we use the *Jackknife Resampling Method*. The jackknife approach consists of first computing the Jackknife estimator $\hat{\theta}_{n,i}$ from the jackknife sample
$\{(s_1,w_1),\ldots,(s_{i-1},w_{i-1}),(s_{i+1},w_{i+1}),\ldots,(s_n,w_n)\}$, for $i = 1,\ldots,n$, then computing the jackknife variance as follows :

$$Var_{jack}(\hat{\theta}_n) = \frac{n-1}{n}\sum_{i=1}^{n}\left[\hat{\theta}_{n,i} - \hat{\theta}_n(.)\right]^2,$$

where $\hat{\theta}_n(.) = \sum_{i=1}^{n} \hat{\theta}_{n,i}/n$ is the mean of Jackknife estimators.

To select the most suitable copula model, various information-theoretic criteria are available. We use the *Akaike Information Criterion* (AIC) which is the most popular and commonly used model selection criterion. The formula of the AIC is the following :

$$AIC = -2l_n(\hat{\theta}_n) + 2k;$$

where $\hat{\theta}_n$ is the estimate of θ and k is the number of estimated parameters, and hence $k = 1$. Models that have smaller values of AIC are better models.

5.6 Application to Niger's data

In this section, we build a bivariate data set representing men's socioeconomic outcomes, namely their individual level of education (S) and income (W). These data are extracted from the 2006 DHS for Niger using the women's questionnaire. The DHS ask surveyed women aged 15 - 49 to report among other things (i) their age at first marriage, (ii) their husband level of education, if married, and (iii) their marital household's wealth. The S denoting a man's number of years of schooling completing is therefore directly extracted from respondents answers, as is a woman's age at first marriage which allows us to divide married women's respondents in two groups, including those married off as children and those who married after reaching the legal age.

Constructing the series of observations on married men's pre-marital income is more complicated because the DHS does not contain this information. This means we have to construct a proxy for the variable W denoting a married man's pre-marital income. Since the DHS contain information on household standard of living, we use this information to build an appropriately scaled wealth index. To make this wealth index a good proxy of men's pre-marital income, we restrict our attention to the group of married women aged $20-29$ at the time of the survey. This choice is justified by the fact that in sub-Saharan Africa, women in that age-group are usually in the pick of their reproduction, which may lead them to devote more time in the childbearing activity and less in the economic activity. Men whose wives are in that age group are therefore most likely to be the main providers in the household, which makes the constructed wealth index, W, a good proxy of the married man income. Furthermore, although this income may be different from the husband's pre-marital income, our main point is that for households where the wife is aged $20-29$, such difference is not explained by marriage, in the sense that it is less likely to be influenced by a significant contribution from child labor in family farms or family-owned businesses. This is because most women in that age-group are likely to have fewer or no teenagers at all at that stage of their reproductive lives.

We organize our data in two samples. One sample consists of a series of bivariate observations $\{(S_1,W_1),....,(S_{n_0},W_{n_0})\}$ pertaining to men whose wives were underage at the time of the marriage, with sample size n_0 ; the other also consists of a series of bivariate observations $\{(S_1,W_1),....,(S_{n_1},W_{n_1})\}$ pertaining to married men whose wives had the legal age at the time of marriage, with sample size n_1.

We first summarize available information from the data for each series of independent observation, and for each sample. Table 3 displays, for each sample m the sample size (n_m), with $m = 0,1$, the average number of years of schooling completed ($\bar{s}_m$) and rescaled wealth index ($\bar{\omega}_m$) among married men, the median of the distributions of years of schooling (s_m^{med}) and wealth indexes (ω_m^{med}), respectively, in addition to the boundaries of each set of observations.

	n_m	$s_m^{\min}$	$s_m^{\max}$	$\omega_m^{\min}$	$\omega_m^{\max}$	$\bar{s}_m$	$\bar{\omega}_m$	s_m^{med}	ω_m^{med}
sample $m = 0$	2033	0	19	0	4.9141	1.0571	0.7216	0.3847	2.6863
sample $m = 1$	578	0	20	0	5.3910	3.1747	1.3299	0.6287	4.9639

TABLE 5.3 – Summary statistics

5.6.1 Measure and tests of dependence

The Kendall tau correlation coefficient is one of the most popular measures used to detect the presence of dependence among observations. The Kendall tau is a nonparametric measure of association which can be defined as based on the number of concordances and discordances in paired observations.

$$\tau_n = \frac{P_n - Q_n}{\binom{n}{2}},$$

where P_n and Q_n are the number of concordances and discordances, respectively in paired observations, and

$$\binom{n}{2} = \frac{n(n+1)}{2}$$

is the total number of pairs observations.

In the current work, we used Kendall τ_{bn} which is the correction of τ_n for ties. The Kendall tau-b is given by

$$\tau_{bn} = \frac{P_n - Q_n}{\sqrt{(P_n + Q_n + S_0)(P_n + Q_n + W_0)}},$$

where S_0 is the number of pairs tied only on the S variable and W_0 is the number of pairs tied only on the W variable.

The sample Kendall τ will be used to test for independence between the two random variables under the null hypothesis $H_0 : \tau = 0$. The test statistics is

$$Z = \sqrt{\frac{9n(n-1)}{2(2n+5)}}|\tau_{bn}|.$$

Table 5.4 below reports the results of the independence test.

Bride's age < 18		Bride's age ≥ 18	
τ_n	P-value	τ_n	P-value
0.22	0.00	0.46	0.00

TABLE 5.4 – The empirical Kendall tau-b and the p-value of the independence test.

Test results presented in Table 5.4 show that the dependence parameter is significantly different from zero ($P-value < 0.001$) suggesting the presence of dependence between the observations of the two series, for both samples.

5.6.2 Modeling dependence

Having formally detected the presence of dependence in both sets of data, the next step in this work is to formulate a probability model for the joint distribution of the level of education (S) and the pre-marital income (W). We adopt a copula-based approach, and restrict of our choice of candidate copula models to the class of Archimedean copulas. In particular, we contrast the *Gumbel*, *Frank*, *Joe* and *Clayton* copula models in terms of their respective performance in capturing the dependence in the bivariate data. These families have been chosen for their simplicity and their flexibility. We report in Table 5.5 below the estimates of the AIC values obtained under these models.

	Bride's age < 18				Bride's age ≥ 18			
	Clayton	Gumbel	Frank	Joe	Clayton	Gumbel	Frank	Joe
AIC	2668.96	3305.34	2385.89	3848.75	1153.20	1276.92	1078.19	1388.19

TABLE 5.5 – The AIC criterion for copula models

The best copula model is one that yields the smallest AIC. Therefore, on the basis of Table 5, we select the Frank copula for both sets of data, since it yields the smallest AIC values. Graphical tools confirm this verdict. Indeed Figure 2-(a) and 2-(b) below show how well the Frank copula (red dots) fits the data (blue dot).

Based on these results, we select the Frank copula for both sets of data, since it yields the smallest AIC values. The estimate of dependence parameter, its jackknife standard error and the estimate of its corresponding Kendall *tau* are presented in Table 5.6 below.

5.6.3 Hypothesis test

Recall that the main goal of this work is gauge the similarity between socioeconomic outcomes of men who choose to marry underage girls differ and those who choose to marry grown-up instead. The problem at hand can be described as follows.

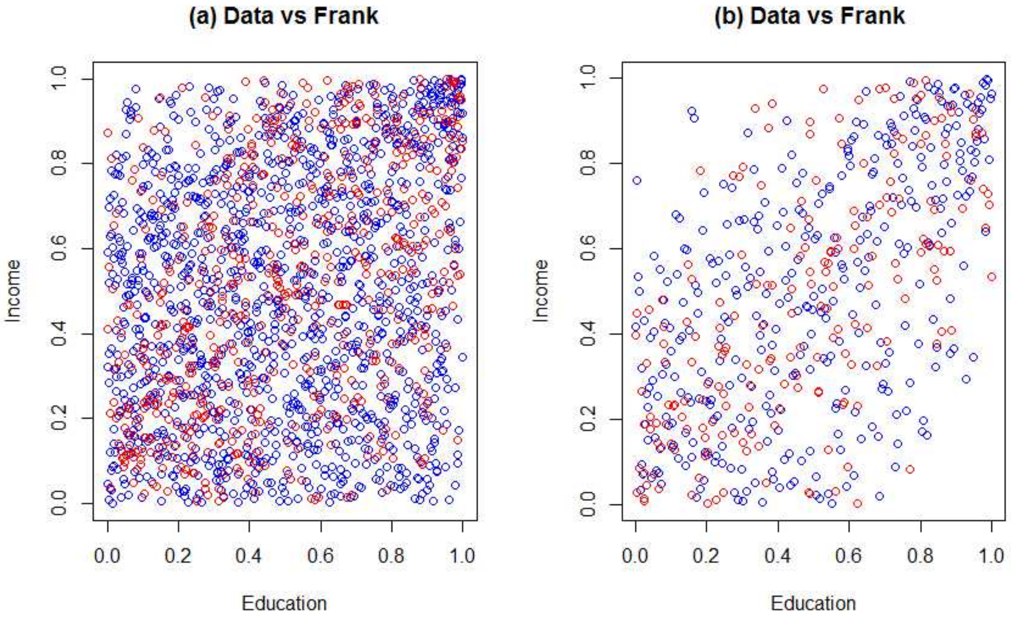


FIGURE 5.2 – Frank copula and the data : (a) Bride's age $<$ 18 and (b) Bride's age $>$ 18

	Bride's age $<$ 18	Bride's age $\geq$ 18
$\hat{\theta}$	2.68	5.36
$s.e.$	0.30	0.55
$\hat{\tau}$	0.28	0.48

TABLE 5.6 – The mple of dependence parameter, its standard error and its corresponding Kendall tau under selected copula model (Frank).

Suppose we face two independent bivariate samples. The first sample, $\{(S_1, W_1), ..., (S_{n_1}, W_{n_1})\}$, is taken from a joint distribution function F with a margins F_s (discrete) and F_ω (continuous) ; the second sample, $\{(S_1, W_1), ..., (S_{n_2}, W_{n_2})\}$, is taken from a joint distribution function G with a margins G_s (discrete) and G_ω (continuous). F and G are joints distributions of the socioeconomic outcomes of men who choose to marry underage girls and older women, respectively. The copulas associated with the first and second samples are determined, for any bivariate realization pair (s, ω), by

$$F(s, \omega) \quad : \quad = C[F_s(s), F_w(\omega); \theta_1]$$

$$G(s, \omega) \quad : \quad = C[G_s(s), G_w(\omega); \theta_2]$$

where θ_1 and θ_2 denote the dependence parameters, respectively for the first and second bivariate sample.

The aim of this paper can thus be construed as one of testing for equality between the two dependence structures $F(s, \omega)$ and $G(s, \omega)$. Because we are only interested in dependence, we leave the behavior of the margins out, so that our hypothesis test is not equivalent to testing for $F_s(s) = G_s(s)$ and $F_w(\omega) = G_w(\omega)$.

Instead, under the proposed model, this test is equivalent to testing the null hypothesis

$$H_0 : \theta_1 = \theta_2,$$

where θ_1 and θ_2 are the true dependence parameters associated with sample 1 and sample 2 respectively. Since we want to show that the dependence between the random variables S and W is weaker in sample 1 than in sample 2, we consider the lower-tailed alternative hypothesis

$$H_1 : \theta_1 < \theta_2.$$

The estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ of the true dependence parameters θ_1 and θ_2 respectively are obtained from (5.11) above for each $j = 1,2$.

5.6.4 Test statistic and results

To obtain a consistent test, we rely on the Wald test for H_0 whose test statistic is given by

$$Z_0 = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\hat{V}\left(\hat{\theta}_1\right) + \hat{V}\left(\hat{\theta}_2\right)}},$$

where $\hat{V}\left(\hat{\theta}_1\right)$ and $\hat{V}\left(\hat{\theta}_2\right)$ are the jackknife estimators of the variance of $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. Under the null hypothesis H_0, the test statistic Z_0 follows the standard Normal distribution $N(0;1)$. If the value z_0 of Z_0 is computed from the data, then the P-value of the test is given by $P - value = \Upsilon(z_0)$, where $\Upsilon(.)$ denotes the standard normal cumulative distribution function. At the usual $\alpha = 0.05$ level of significance, H_0 is rejected if $P - value < 0.05$. Table 5.7 below reports the computed value of the test statistic Z_0 and the associated P-value.

$H_0 : \theta_1 = \theta_2$		Niger
$H_1 : \theta_1 < \theta_2$	z_0	-4.26
	p-value	< 0.0001

TABLE 5.7 – The statistic and the P-value of the left one-sided Wald test of no difference between groups.

On basis of the test results reported in Table 5.7, the null hypothesis of equality between the two dependence structures is rejected, in favor of the alternative that the positive dependence between education and premarital income is strong for men who married legal-age women than for those who marry underage girls.

5.7 Conclusion

Public discussions of development highlight child marriage as a persistent barrier to social and economic progress in sub-Saharan Africa. This sub-continent of Africa is home to eight of the ten countries with the highest prevalence levels of child marriage in the world (UNFPA 2012). Public policies aimed at eradicating this harmful practice are therefore an imperative to enhance sustainable development in this region, but are yet to be well-understood, particularly on the demand-side.

In this paper, we developed a theory of men's decision on bride type highlighting the socioeconomic characteristics of men who prefer child brides (understood as brides under UNFPA's legal age of 18). The main ingredients of this theory include (i) the desired characteristics of a typical bride (i.e., reproductive fitness, submissiveness, and economic contribution to the household), and (ii) prospective grooms' level of education and pre-marital income as they interact with the desired characteristics of a typical bride to influence their choice of the type of bride (child bride or legal-age bride) to have. The model predicts that men who are less educated and/or have a low level of pre-marital income are those most likely to prefer child brides, whereas those with higher levels of education and pre-marital income are more likely to prefer legal-age bride.

The intuition behind this prediction is as follows. On the one hand, relative to a prospective groom with a low level of education, a prospective groom with high level of education may more clearly perceive the inferior reproductive fitness of a child bride. However, because reproductive fitness is not the only desired bride's characteristic, having a high level of education, though necessary, is not sufficient for a prospective groom to prefer a legal-age bride over a child bride. On the other hand, a prospective groom with a high level of income may be indifferent between a child bride and a legal-age bride from the viewpoint of both their capacity to contribute the economic prosperity of their future household and the extent to which they can be sexually controlled. However, he may still lean towards a child bride ; if he has too low a level of education to clearly perceive a child bride's inferior reproductive fitness. This explains why, in our model, prospective grooms with a high level of both education and pre-marital are those most likely to choose to have legal-age brides.

We formally tested this theory using Niger's 2006 DHS data. For this purpose, we divided the bivariate data extracted from this DHS into two samples of married men. One such sample consists of independent observations of education and pre-marital income among men whose wives were underage (i.e., aged less than 18) at the start of their marriage, and the other consists of similar bivariate data pertaining to men whose wives had the legal age (i.e., aged 18 or above). We restricted attention to men married to women aged 20 – 29 at the time of the survey, to ensure the data used provided the best possible proxy for each married man's pre-marital income. Descriptive statistics show that men whose wives were underage at the start of the marriage have in average a lower level of education and a lower level of pre-marital income than those whose wives were of legal age at the start of the marriage.

For each sample, we performed an independence test for the bivariate data based on the sample Kendall *tau* to detect the presence of dependence between the two socioeconomic outcomes of interest. The test results led to the rejection of the null hypothesis of independence at 0.05 significance level. Given the existence of positive dependence in the data, we adopted a copula-based approach to summarize the structure of this dependence by a single parameter known as the dependence parameter. We then estimated this parameter using a maximum pseudo-likelihood method (mple), combining it with a *Jackknife resampling method* to correct for the underestimation of the variance of the estimator resulting from the use of mple. The main goal of the estimation process was to provide a basis for constructing the test statistic needed to perform a parametric test of equality between two dependence structures, under the null hypothesis that the association between education and pre-marital income is identical for men who married underage girls and those who married legal-age women.

Our Wald-type hypothesis test leads to the rejection of the null hypothesis of equality, in favor of the alternative that the positive dependence between these two socioeconomic outcomes is twice as strong for men who married legal-age women than for those who married underage girls. This finding can be interpreted as follows. First, recall that men whose wives were underage at the start of the marriage have in average a lower level of education and a lower level of pre-marital income than those whose wives were of legal age at the start of the marriage. Second, this feature of the data, combined with the fact that the positive dependence between education and pre-marital income is twice as strong for men who married legal-age women than for those who married underage girls, suggests that men who married legal-age women are more likely to have both a higher level of education and a higher level of pre-marital income. Third, a man with either a low level of education or a low level of pre-marital income is therefore more likely to marry an underage girl, because dependence between these two variables is weaker for men married to underage girls. These findings are of practical interest for public policy aimed at curbing the demand for underage brides in developing countries. They suggest that intervention aimed at discouraging the demand for child brides should target young men who have a low level of education and/or a low level of income.

Conclusion

Dans cette thèse, nous avons developpé une nouvelle classe de modèles à effets aléatoires pour analyser des données de Bernoulli échangeables en forme de grappe. Les modèles conditionnels sachant l'effet aléatoire, et les modèles marginaux obtenus en intégrant par rapport à l'effet aléatoire sont disponibles. Nous avons montré que la version marginale de nos modèles est associée à la classe des modèles de copules Archimédiennes multidimensionnelles.

Nos modèles sont attractifs dans le sens où une variété de modèles paramétriques, avec une grande variété de structure de dépendance, est disponible pour la corrélation intra-grappe. Les grappes de tailles différentes, ainsi que des données avec covariables au niveau des grappes, de même qu'au niveau des unités sont facilement prises en compte. L'un des avantages potentiels de nos modèles est que la fonction de vraisemblance a une expression explicite, ce qui facilite la mise en œuvre de la méthode du maximum de vraisemblance pour l'estimation des paramètres.

Au chapitre 2, nous avons considéré le cas le plus simple où aucune covariable n'est disponible. Nous avons proposé la méthode du maximum de vraisemblance pour l'estimation du coefficient de corrélation intra-grappe (ICC) pour plusieurs spécifications de copules Archimédiennes. La sélection d'un modèle particulier est effectuée en utilisant le critère d'information d'Akaike (AIC). La procédure comprend l'estimation du maximum de vraisemblance et la méthode d'intervalle de confiance par vraisemblance profilée. Nous avons fait des études de simulation pour mesurer la performance de la méthode de vraisemblance profilée sous nos modèles en termes de taux de couverture et de longueur de l'intervalle de confinace, et la sensibilité de notre approche à la spécification d'un modèle de copule. La procédure que nous proposons a aussi été testée sur des données réelles. Nous comparons notre méthode à celle proposée sous le modèle Béta-Binomial, et la méthode d'intervalle Wald modifié proposée par Zou and Donner (2004). L'une des conclusions importantes de ces études est que l'intervalle de confiance par vraisemblance profilée obtenu sous nos modèles présente de belles propriétés pour le taux de couverture et de longueur d'intervalle de confiance, même dans le cas où le nombre de grappe est petit. La sélection de modèle est une étape importante.

Au chapitre 3, nous avons considéré une extension des modèles du chapitre 2 pour accommoder des covariables au niveau des grappes, afin de modéliser l'hétérogénéité dans les probabilités de capture lors d'une expérience de capture-recapture dans une population fermée. Dans ce contexte, nos modèles sont utilisés pour modéliser l'hétérogénéité résiduelle qui n'est pas prise en compte par les covariables mesurées sur les unités capturées. Plusieurs modèles sont disponibles pour l'hétérogénéité non observée et la probabilité de capture marginale est modélisée en utilisant les fonctions de liens logit et log-log complémentaire. Les paramètres sont estimés en utilisant la vraisemblance conditionnelle construite à partir des observations collectées sur les unités capturées au moins une fois. Ceci généralise le modèle de Huggins (1991) qui ne tient pas compte de l'hétérogénéité résiduelle. La sensibilité de l'inférence à la spécification d'un modèle a également été étudiée par des simulations. Un exemple numérique a aussi été présenté.

Le chapitre 4 a considéré les modèles des chapitres 2 et 3, et traite de la prédiction dans les petites régions. Nous avons proposé la méthode empirique pour estimer des proportions. Nous avons obtenu une expression explicite des prédicteurs empiriques de Bayes et de leurs variances a posteriori pour les vraies proportions régionales. Nous avons proposé la méthode du Jackknife proposée par Jiang and Lahiri (2002), afin d'obtenir des estimations de l'incertitude associée aux estimateurs empiriques. Nous avons également présenté des résultats empiriques obtenus à partir de données simulées et réelles.

Le chapitre 5 de cette thèse concerne l'analyse des caractéristiques socio-économiques des hommes qui ont une préférence à épouser des jeunes filles de moins de 18 ans. Pour se faire, nous avons considéré les données de l'EDS 2006 du Niger et avons utilisé les copules Archimédiennes bidimentionelles pour modéliser l'association entre le niveau d'éducation des hommes et leur revenu pré-marital. Nous avons construit la vraisemblance pour un échantillon issu de ce couple de variables aléatoires mixtes, et avons déduit une estimation du paramètre de dépendance via une procédure semi-paramétrique où les marges sont estimées par leurs équivalents empiriques. Nous avons utilisé la méthode du jackknife pour estimer l'erreur type associée à l'estimation du paramètre de dépendance. Nous avons proposé la méthode de Wald, pour tester l'égalité entre l'association des caractéristiques socio-économiques des hommes qui épousent des jeunes filles mineures et celle des hommes qui se marient avec des femmes âgées. Les résultats du test ont validé ainsi notre théorie selon laquelle les hommes qui développent une préférence à épouser des jeunes filles mineurs sont caractérisés par un faible niveau d'éducation et un faible revenu pré-marital, lorsqu'on les compare à ceux qui ne le font pas.

Les modèles que nous avons proposés dans cette thèse sont très flexibles pour traiter la corrélation intra-grappe avec une structure de dépendance échangeable dans une expérience de Bernoulli. Nous pensons que l'utilisation de ces modèles pour tenir compte de la corrélation inter-grappe pourrait potentiellement être un sujet qui mérite de profondes réflexions. De plus au lieu d'utiliser un effet aléatoire spécifique à chaque grappe pour tenir compte de la corrélation, il serait intéressant dans le futur, d'incorporer une variable aléatoire qui varie dans le temps afin d'explorer ces modèles pour des données longitudinales où la structure de dépendance est souvent non échangeable.

Bibliographie

Ahmed, M. and Shoukri, M. (2010). A bayesian estimator of the intracluster correlation coefficient from correlated binary responses. *J. Data Sci.*, 8 :127–137.

Aitkin, M., Liu, C., and Chadwick, T. (2009). Bayesian model comparison and model averaging for small-area estimation. *The Annals of Applied Statistics*, 3(1) :199–221.

Alanko, T. and Duffy, J., C. (1996). Compound binomial distribution for modeling consumption data. *The Statistician*, 45 :269–286.

Ananth, C., V. and Preisser, J., S. (1999). Bivariate logistic regression : Modeling the association of small for gestational age births in twin gestations. *Stat. Med.*, 18 :2011–2023.

Beilei, W. (2013). *Contributions to Copula Modeling of Mixed Discrete-Continuous Outcomes.* PhD Dissertation, Department of Mathematics and Statistics, University of Calgary, Calgariy, Alberta, Canada.

Burnham, K., P. (1972). Estimation of population size in multinomial capture-recapture studies when capture probabilities vary among animals. *Ph.D. thesis, Oregon State University, Corvallis.*

Burnham, K., P. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43 (4) :783–791.

Catalano, P., J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, 16 :883–900.

Catalano, P., J. and Ryan, L., M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87 :651–658.

Chakraborty, H., Moore, J., Carlo, W., A., Hartwell, T., D., and Wright, L., L. (2009). A simulation based technique to estimate intracluster correlation for a binary variable. *Contemporary Clinical Trials*, 30 :71–80.

Chambers, J., M., Mallows C., L., and Stuck, B., W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association.*, 71 :340–344.

Cockburn and al. (2015). A quantitative assessment of the added value men attach to having a child bride. *manuscript*.

Coull, B., A. and Agresti, A. (1999). The use of mixed logit to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55 :294–301.

Darroch, J.-P., Fienberg, S., E., Glonek, G., F.-V., and Junker, B., W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Amer. Statist. Assoc.*, 88 :1137–1148.

de Leon, A., R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters*, 75 :49–57.

de Leon, A., R. and Carrière, K., C. (2007). General mixed-data model : extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35 :533–548.

de Leon, A., R. and Carrière Chough, K. (2010). *Mixed-outcome data. In : Encyclopedia of Biopharmaceutical Statistics (S. C. Chow, Ed.). 3rd ed. pp. 817-822.* Chapman and Hall., New York.

de Leon, A., R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcomes. *Statistics in Medicine*, 30 :175–185.

Dempster, A. and Tomberlin, T. (1980). The analysis of census undercount from a postenumeration survey, proceedings of the conference on census undercount. *Arlington, Va.*, 1 :88–94.

Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5 :969–993.

Dorazio, R., M. and Royle, J., A. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59 :351–364.

Eldridge, S. and Kerry, S. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Wiley, New York.

Embrechts, P., McNeil, A., J., and Straumann, D. (1999). Correlation : Pitfalls and alternatives. *Risk*, 12 :69–71.

Faes, C. (2013). *Hierarchical modeling of endpoints of different types with generalized linear mixed models. In : Analysis of Mixed Data : Methods & Applications (A. R. de Leon and K. Carrière Chough, Eds.). pp. 125-138.* Chapman & Hall/CRC, .

Favre, A.-C., Adlouni, S., E., Perreault, L., N., T., and B, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40 :W01101.

Feng, Z. and Crizzle, J., E. (1992). Correlated binomial variates : Properties of estimators of intraclass correlation and its effect on sample size calculation. *Stat. Med.*, 11 :1600–1614.

Field, E. and Ambrus, A. (2008). Early marriage, age of menarche and female schooling attainment in bangladesh. *Journal of Political Economy*, 116 :881–930.

Fitzmaurice, G., M. and Laird, N., M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53 :110–122.

Fitzmaurice, G., M. and Laud, N., M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90 :845–852.

Fleiss, J., L. and Cuzick, J. (1979). The reliability of dichotomous judgments : Unequal numbers of judges per subject. *Applied Psychological Measurement*, 3 :537–542.

Frees, E., W. and Valdez, E., A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2 :1–25.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12 :347–368.

Genest, C. and MacKay, R. (1986). Copules Archimédiennes et familles bidimensionnelles dont les marges sont données. *The Canadian Journal of Statistics*, 14 :145–159.

Genest, C. and Nešlehová, J. (2007). A primer on copula for count data. *The Austin Bulletin*, 37 :475–515.

Genest, C. and Rivest, L., P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88 :1034–1043.

Gueorguieva, R. (2013). *Random effects models for joint analysis of repeatedly measured discrete and continuous outcomes. In : Analysis of Mixed Data : Methods & Applications (A. R. de Leon and K. Carrière Chough, Eds.). pp. 109-123.* Chapman & Hall/CRC, .

Gueorguieva, R., V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96 :1102–1112.

Hirakawa, A. (2012). Adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase i in oncology trials. *Statistics in Medicine*, 31 :516–532.

Huggins, R., M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76 :130–140.

Huggins, R., M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics*, 47 :725–732.

Huggins, R., M. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statistics and Probability Letters.*, 54 :147–152.

Hwang, W.-H. and R. Huggins, R., M. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92 :229–233.

Jensen, R. and Thornton, R., L. (2003). Early female marriage in the developing world. *Oxfam Journal of Gender and Development*, 11 :9–19.

Jiang, J. and Lahiri, P. S. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *The Annals of Statistics*, 30(6) :1782–1810.

Jiang, J. and Lahiri, P. S. (2006a). Estimation of finite population domain means : a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101 :301–311.

Jiang, J. and Lahiri, P. S. (2006b). Mixed model prediction and small area estimation. *Test*, 15 :1–96.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* Chapman & Hall/CRC, New York.

Joe, H. (2015). *Dependence modeling with copulas.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, New York.

Jordanger, L., A. and Tjostheim, D., B. (2014). Model selection of copulas : Aic versus a cross validation copula information criterion. *Statistics and Probability Letters*, 92 :249–255.

Kang, J. and Yang, Y. (2013). Joint modeling of mixed count and continuous longitudinal data. *In : Analysis of Mixed Data : Methods & Applications, Chapman & Hall /CRC*, pages 63–79.

Kjersti, A. (2004). Modelling the dependence structure of financial assets : A survey of four copulas. *Norwegian Computing Center Applied Research and Development.*

Kuk, A., Y. C. (2004). A generalized estimating equation approach to modelling foetal response in developmental toxicity studies when the number of implants is dose dependent. *Applied Statististics*, 52 :52–61.

Légaré, F., Labrecque, M., LeBlanc, A., Njoya, M., Laurier, C., Côté, L., Godin, G., Thivierge, R. L., O'Connor, A., and S., S.-J. (2011). Training family physicians in shared decision making for the use of antibiotics for acute respiratory infections : a pilot clustered randomized controlled trial. *Health Expect.*, 1 :96–110.

Liang, K., Y., Qaqish, B., and Zeger, S., L. (1992). Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc. Ser. B.*, 54 :3–40.

Liang, K., Y. and Zeger, S., L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73 :13–22.

Lin, L., Bandyopadhyay, D., Lipsitz, S., and Sinha, D. (2010). Association models for clustered data with binary and continuous responses. *Biometrics*, 66 :287–293.

Link, W., A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59 (4) :1123–1130.

Lipsitz, S., R., Laird, N. M., and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data : using the odds ratio as a measure of association. *Biometrika*, 78 :153–160.

Liu, B. (2009). Adaptive hierarchical bayes estimation of small area proportions. *Social Statistics Section JSM 2009.*

MacGibbon, B. and Tomberlin, T. (1997). Small area estimation of proportions via empirical bayes techniques. *Survey Methodology*, 15 :237–252.

Madsen, R., W. (1993). Generalized binomial distribution. *Comm. Statist. Theory Methods*, 22 :3065–3086.

Mai, J.-M. and Scherer, M. (2012). *Simulating Copulas ; Stochastic Models, Sampling Algorithms and Applications*. Series in Quantitative Finance : Volume 4. World Scientific Publishing Company London, New York.

Malec, D., Sedransk, J., Moriarity, C., and LeClere, F. (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association*, 92 :815–826.

Marshall, A., W. and Olkin, I. (1993). Families of multivariate distributions. *Journal of the American Statistical Association*, 83 :834–841.

McCulloch, C. (2003). Generalized linear mixed models,. *NSF-CBMS Regional Conference Series in Probability and Statistics7. Beachwood, OH : Institute of Mathematical Statistics.*

McCulloch, C. (2008). Joint modeling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, 17 :53–73.

McCulloch, C., E. and Searle, S., R. (2001). *Generalized Linear and Mixed Models.* Wiley, New York.

McCulloch, C., E., Searle, S., R., and H., N. (2008). *Generalized Linear and Mixed Models.* Wiley, New York.

Meester, S. and Mackay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics*, 50 :954–963.

Nelson, R., B. (2006). *An Introduction to Copulas.* Springer-Verlag, New York.

Nguyen, M., C. and Quentin, W., D. B. (2012). Measuring child marriage. *Economics Bulletin*, 31 :398–411.

Nikoloulopoulos, A., K. and Karlis, D. (2008). Multivariate logit copula model with an application to dental data. *Stat. Med.*, 27 :6393–6406.

Nikoloulopoulos, A., K. and Karlis, D. (2010). Modeling multivariate count data using copulas. *Communications in Statistics-Simulations & Computation*, 39 :172–187.

Nikolouplopoulos, A., K. (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference*, 139 :3878–3890.

Norris, J., L. and Pollock, K., H. (1996). Nonparametric mle under two closed-capture models with heterogeneity. *Biometrics*, 52 :639–649.

Ochi, Y. and Prentice, R., L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71 :531–542.

Olkin, I. and Tate, R., F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32 :448–465.

Otis, D., L., Burnham, K., P., White, G., C., and Anderson, D. R. (1978). Statistical inference from capture data on closed populations. *Wildlife Monographs*, 62 :33–37.

Pals, S., L., Beaty, B., L., Posner, S., F., and Bull, S., S. (2009). Estimates of intraclass correlation for variables related to behavioral hiv/std prevention in a predominantly african american and hispanic sample of young women. *Health Education & Behavior*, 36 :182–194.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28(1) :40–68.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56 :434–442.

Pollock, K. H., Hines, J. E., and Nichols, J. D. (1984). The use of auxiliary covariables in capture-recapture and removal experiments. *Biometrics*, 40 :329–340.

Prentice, R., L. (1986). Binary regression using extended beta-binomial distribution, with discution of correlation induced by covariate measurement errors. *Journal of the American Statistical Assiciation*, 81 :321–327.

Rao, J., N. K. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology, New York.

Ridout, M., S., Demétrio, C. G. B., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, 55 :137–148.

Rivest, L.-P. and Baillargeon, S. (2014). Capture-recapture methods for estimating the size of a population : Dealing with variable capture probabilities. *Statistics in Action : A Canadian Outlook*, 40 :289–304.

Saha, K., K. (2012). Profile likelihood-based confidence interval of the intraclass correlation for binary outcome data sampled from clusters. *Stat. Med.*, 31 :3982–4002.

Schweizer, B. and Wolff, E., F. (1981). On non parametric measures of dependence for random variables. *Annals of Statististic*, 8 :870–885.

Shoukri, M., M., Kumar, P., and Colak, D. (2011). Analyzing dependent proportions in cluster randommized trials : Modeling inter-cluster correlation via copula function. *Comput. Statist. Data Anal.*, 55 :1226–1235.

Sklar, A. (1959). Fonctions de repartition à n dimensions et leurs marges. *Publication of the Institute of Statistics of Paris 8*, pages 229–231.

Song, P., X.-K. (2007). *Correlated Data Analysis : Modeling, Analytics, and Applications.* Chapman & Hall /CRC.

Song, P., X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65 :60–68.

Stefanescu, C. and Turnbull, B., W. (2003). Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics*, 59 :18–24.

Stoklosa, J. and Huggins, R., M. (2012). A robust p-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics and Data Analysis*, 56 :408–417.

Teixeira-Pinto, A. and Normand, S.-L., T. (2009). Correlated bivariate continuous and binary outcomes : Issues and applications. *Statistics in Medicine*, 28 :1753–1773.

Tomberlin, T. J. (1988). Predicting accident frequencies for drivers classified by two factors,. *Journal of the American Statistical Association*, 83 :309–321.

Tounkara, F. and Rivest, L.-P. (2014). Some new random effect models for correlated binary responses. *Dependence Modeling*, 2 :73–87.

Tounkara, F. and Rivest, L.-P. (2015). Mixture regression models for closed population capture re-capture data. *Biometrics*, 71 :721–730.

Trégouët, D.-A., Ducimetière, P., Bocquet, V., Visvikis S., Soubrier, F., and L., T. (1999). A parametric copula model for analysis of familial binary data. *Am. J. Hum. Genet*, 64 :886–893.

Trivedi, P., K., David, M., Z., and Vogelstein, R. (2005). Copula modeling : An introduction for practitionersctives. *Foundations and Trends in Econometrics*.

Turner, R., M., Omar, R., Z., and Thompson, S., G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat. Med.*, 20 :453–472.

Turner, R., M., Omar, R., Z., and Thompson, S., G. (2006). Constructing intervals for the intracluster correlation coefficient using bayesian modelling, and application in cluster randomized trials. *Stat. Med.*, 25 :1443–1456.

Van Dyk, A. C. (2001). Traditional African beliefs and customs : Implications for aids education and prevention in africa. *South African Journal of Psychology*, 31 :60–66.

Vogelstein, R. (2013). Ending child marriage : How elevating the status of girls advances u.s. foreign policy objectives. *Council on Foreign Relations*, 11 :9–19.

Wahhaj, Z. (2014). A theory of child marriage. *Unpublished, the University of Kent available online at http ://www.kent.ac.uk/economics/staff/staffhost/zaki-wahhaj/Child*.

Williams, D., A. (1975). The analysis of binary response from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31 :949–952.

Williams, D., A. (1982). Extra-binomial variation in logistic linear models. *J. Appl. Statist.*, 31 :305–309.

Williams, D., A. (1988). Estimation bias using beta-binomial distribution in teratology. *Biometrics*, 44 :305–309.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80 :513–524.

Zimmer, D. and Trivedi, P. (2006). Using trivariate copulas to model sample selection and treatment effects : Application to family health care demand. *Journal of Business & Economic Statistics*, 24 :63–76.

Zimmer, David, M. (2013). Analysis of mixed outcomes in econometrics : Applications in health economics, chapter 11, in de leon, a.r and k.c. chough eds, analysis of mixed data : Methods and applications.

Zou, G. and Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics*, 60 :807–811.

Annexe A

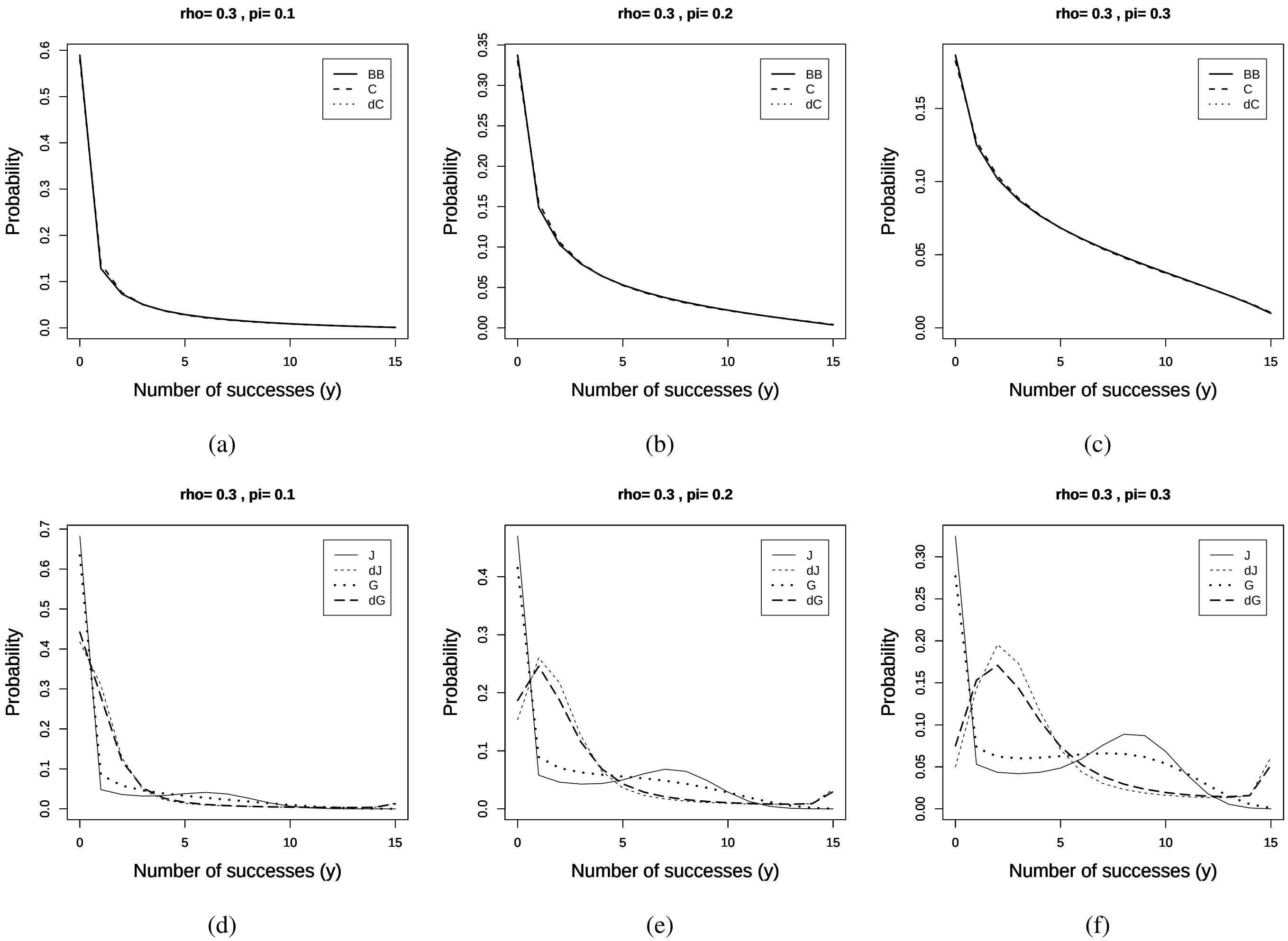
Supplementary material

A.1 Supplementary material of Chapter 2

Web Appendix A : Graphical comparison with others distributions

Small cluster comparisons

We can see from graphs (a), (b) and (c) of Web Figure 2.1 that the probability mass functions (pmf) of the beta-binomial's, the Clayton's and the dClayton's models are nearly identical. On the other hand, in (c), (d) and (e), the pmfs for dJoe and dGumbel distributions are very similar.



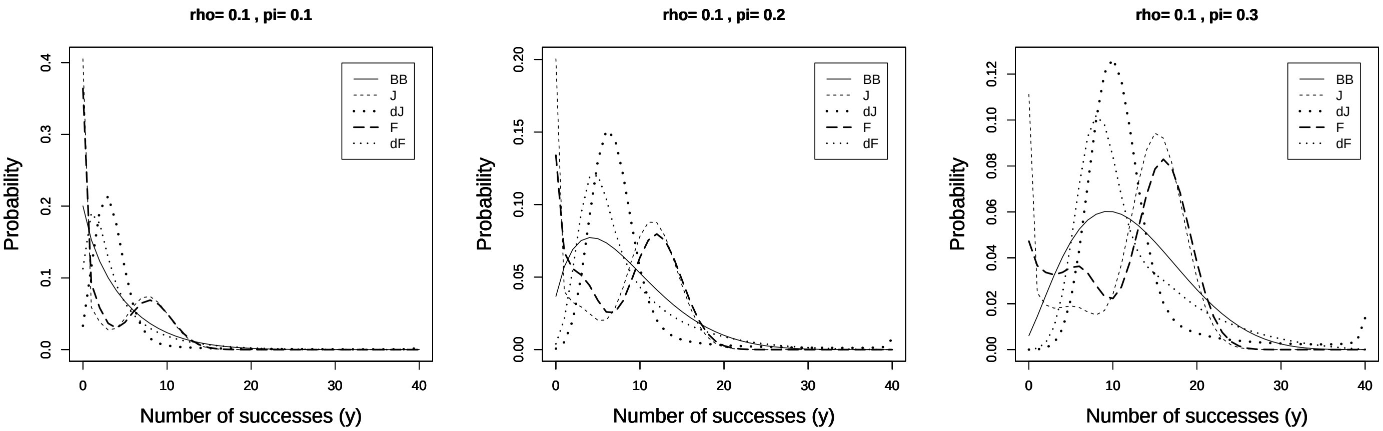
(a) (b) (c)

(d) (e) (f)

Web figure 2.1 : A comparison of the probability mass functions for cluster size 15 under some small clusters models, for $\rho = 0.3$ and $\pi = 0.1, 0.2, 0.3$

Larger cluster comparisons

The graphs in Web Figure 2.2 show that the pmfs of all distributions are different, except for dF's and dJ's models, whose shapes are similar.



(a) (b) (c)

Web Figure 2.2 : A comparison of the probability mass functions for cluster size 40 under the beta-binimial (BB), Joe (J), dJoe (dJ), Frank (F) and dFrank (dF), for $\rho = 0.1$ and π=0.1 (a), 0.2 (b) ,0.3 (b)

Web Appendix B : Simulation results

Results for small cluster's models

			$\pi = 0.1$				π =0.3			
			J's data		dJ's data		J's data		dJ's data	
ρ	k	Models	CIL	Cov	CIL	Cov	CIL	Cov	CIL	Cov
0.1	100	ZD	0.36	99.90	0.35	98.10	0.22	99.30	0.21	95.80
		BB	0.23	96.00	0.18	79.00	0.18	96.30	0.17	87.40
	250	ZD	0.24	100.00	0.24	98.10	0.14	98.30	0.14	94.80
		BB	0.15	97.50	0.12	76.70	0.12	93.80	0.11	87.90
0.3	100	ZD	0.45	100.00	0.44	98.50	0.28	98.80	0.28	97.10
		BB	0.34	97.40	0.31	79.56	0.24	96.30	0.23	90.20
	250	ZD	0.32	100.00	0.31	97.30	0.18	99.20	0.18	97.10
		BB	0.22	97.10	0.20	76.90	0.15	95.30	0.15	88.20

Web Table 2.1 :Small clusters simulations's results : Confidence interval length (CIL) and empirical coverage (COV) of two-sided confidence intervals for ρ with a nominal 95% confidence level for Zou and Donner's confidence interval and the beta-binomial PLCI, when the data comes from J's, dJ's and dIG's distributions

In web Table 2.1, We report results on Zou and Donner's confidence intervals and for the beta-binomial PLCI, when data J's and dJ's models. The Zou and Donner's confidence interval shows very conservative behavior as it is much wider than the PLCI. The beta-binomial PLCI shows in general a severe under-coverage when data are from the dJ distribution. However, when data are from J's model, this method provides a coverage slightly larger than the nominal value of 95%.

Results for large cluster simulations

In Web Table 2.2, we reported Confidence interval length (CIL) and empirical coverage (COV) of Zou and Donner's and the beta-binomial PLCI, when data are from J's, dJ's, F's and dF's distributions. We can see that ZD's method is very conservative. On the other hand, the beta-binomial PLCI is sensitive to a model misspecification, as it provides coverage below the nominal value of 95%, for J's and dJ's data.

				J's data		dJ's data		F's data		dF's data	
ρ	π	K	Models	CIL	Cov	CIL	Cov	CIL	Cov	CIL	Cov
0.05	0.1	10	ZD	0.66	100.00	0.63	90.57	0.66	100.00	0.63	99.84
			BB	0.34	89.95	0.17	81.89	0.30	94.52	0.16	96.42
		25	ZD	0.49	100.00	0.49	92.28	0.49	100.00	0.48	100.00
			BB	0.16	76.35	0.08	50.94	0.15	85.31	0.08	88.89
		50	ZD	0.37	100.00	0.37	96.62	0.37	100.00	0.37	100.00
			BB	0.11	56.97	0.05	37.69	0.10	68.37	0.05	81.64
	0.3	10	ZD	0.48	100.00	0.45	88.03	0.46	100.00	0.43	98.99
			BB	0.28	53.39	0.20	65.24	0.21	86.16	0.15	94.12
		25	ZD	0.29	100.00	0.27	95.07	0.29	100.00	0.27	99.89
			BB	0.14	51.33	0.09	58.95	0.12	75.00	0.08	85.09
		50	ZD	0.19	100.00	0.18	96.36	0.19	100.00	0.18	100.00
			BB	0.09	45.93	0.06	50.85	0.08	71.31	0.05	82.08
0.1	0.1	10	ZD	0.67	100.00	0.64	93.18	0.67	100.00	0.65	100.00
			BB	0.45	97.03	0.20	66.35	0.41	97.99	0.21	92.86
		25	ZD	0.52	100.00	0.51	97.34	0.52	100.00	0.52	100.00
			BB	0.27	79.34	0.10	37.06	0.24	88.68	0.12	83.75
		50	ZD	0.41	100.00	0.40	97.44	0.41	100.00	0.41	100.00
			BB	0.18	39.20	0.06	33.52	0.17	63.00	0.08	73.40
	0.3	10	ZD	0.49	100.00	0.47	96.73	0.48	100.00	0.46	99.88
			BB	0.33	75.19	0.23	60.41	0.27	93.02	0.19	90.22
		25	ZD	0.32	100.00	0.31	97.84	0.32	100.00	0.31	99.90
			BB	0.19	61.99	0.12	57.09	0.16	85.96	0.11	82.86
		50	ZD	0.23	100.00	0.22	96.92	0.23	100.00	0.22	100.00
			BB	0.13	42.93	0.08	51.18	0.11	77.57	0.08	79.72

Web Table 2.2 : Larger cluster simulation's results :Confidence interval length (CIL) and empirical coverage (COV) of two-sided confidence intervals for ρ with a nominal 95% level for Zou and Donner's confidence interval and the beta-binomial PLCI, when the data comes from J's, dJ's F's and dF's distributions

A.2 Supplementary material of Chapter 3

Web Appendix A : The proportion of correct classification

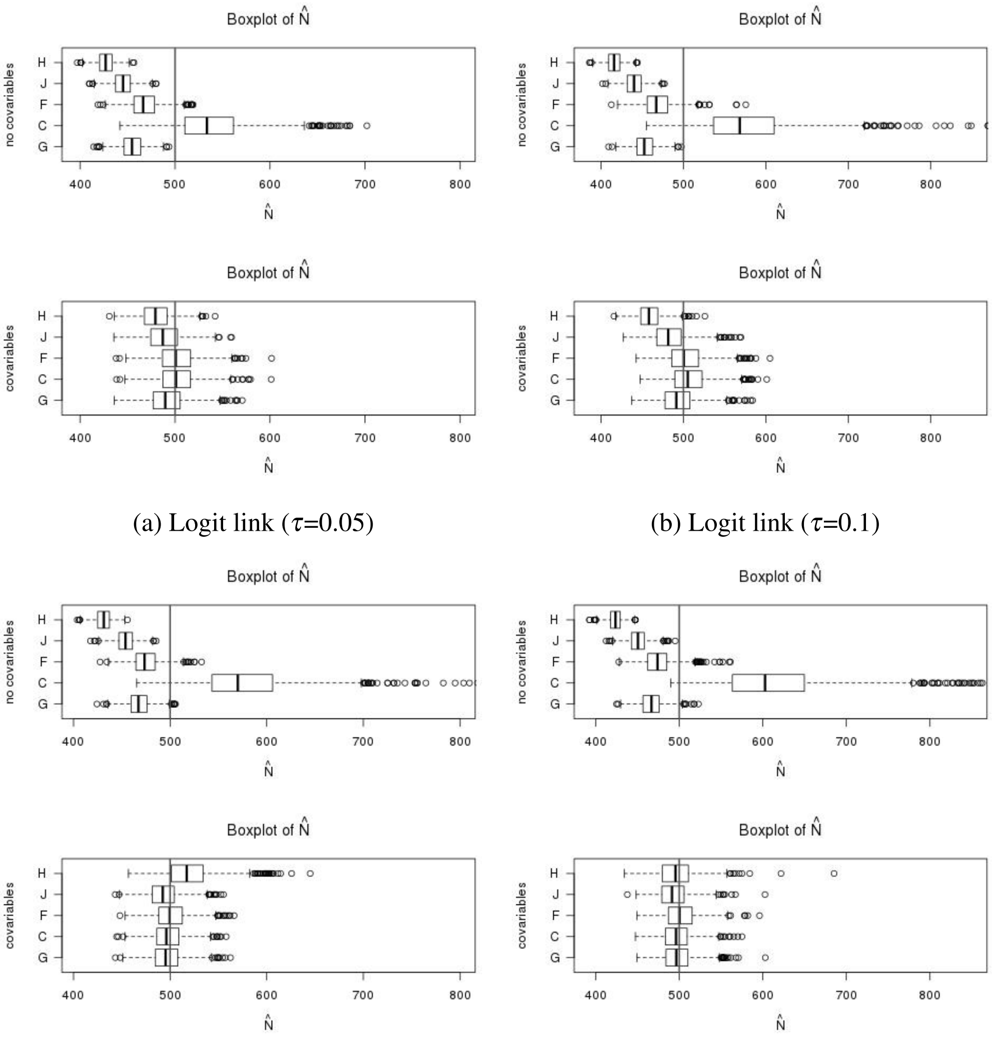
The accuracy of the AIC model selection criterion was measured using the proportions of times that a model was selected when the data was simulated from another model. The results are presented in Web Table 3.1 ; the columns are the models from which the data is simulated and the rows are the models that minimize the AIC. Web Table 3.1 highlights that the extreme models in Figure 1, Joe and Clayton, are more easily identifiable than Frank's and Gumbel's. The two link functions give similar results so we focus on the Logit case. The small population ($N = 100$) and small heterogeneity ($\tau = 0.05$) scenario gives the smallest proportion , 37 %, of correct classification which is defined as the mean diagonal percentage. It is slightly larger than 25%, the result that would be obtained if the models were indistinguishable. Correct classifications increase to 54% for $N = 100$ and $\tau = 0.2$ and to 82% for $N = 500$ and $\tau = 0.2$.

		$\tau = 0.05$				$\tau = 0.1$				τ =0.2			
		Data				Data				Data			
N	Models	G	C	F	J	G	C	F	J	G	C	F	J
100	G	20(24)	25(24)	19(22)	17(22)	22(19)	12(11)	11(12)	16 (17)	34 (31)	7 (5)	16(12)	22(17)
	C	17(14)	44(42)	28(23)	14(12)	14(13)	55(60)	28(25)	6 (6)	8 (7)	66(69)	27 (19)	2 (2)
	F	28(30)	24(25)	37(40)	21(24)	26(28)	27(25)	47(49)	18 (19)	21 (26)	25(24)	49(58)	10(13)
	J	36(32)	7 (9)	15(15)	49(43)	38(40)	5 (4)	14(14)	60 (58)	37 (36)	2 (2)	8 (10)	66(68)
250	G	24(24)	13(13)	10(13)	22(20)	35(35)	6 (5)	14 (13)	25(22)	54(55)	2(1)	12(10)	26 (20)
	C	12(13)	55(58)	28(23)	4 (6)	4(6)	69(76)	21(16)	0 (1)	1 (1)	78 (86)	19 (8)	0 (0)
	F	26(28)	25(26)	50(52)	19(22)	22(23)	23 (19)	59 (66)	7 (10)	15 (16)	20 (13)	68 (80)	3 (4)
	J	38(36)	6 (3)	12(12)	55(53)	39(37)	2 (1)	6 (5)	68 (67)	30(28)	0 (0)	1(2)	72(76)
500	G	34(34)	6(7)	10(12)	23(25)	54(51)	2(2)	12(10)	25(24)	71(70)	0(0)	5(7)	18(16)
	C	4(5)	65(66)	24(18)	2(1)	0(1)	78(85)	14(9)	0(0)	0(0)	91(95)	8 2	0(0)
	F	24(24)	26(25)	59(65)	11(13)	12(16)	20(14)	74 (79)	2(2)	6(10)	9(5)	86(91)	0(0)
	J	38(37)	3(2)	7(5)	65(61)	33(32)	0 (0)	1(2)	73 (73)	22(19)	0(0)	0(0)	82(83)
750	G	37(38)	2 (3)	8(13)	24 (26)	62 (62)	0(1)	8(7)	22(22)	80(79)	0(0)	2(3)	13(12)
	C	4 (3)	74 (74)	19 (15)	0 (1)	0 (0)	86 (91)	10 (5)	0 (0)	0 (0)	95 (98)	5 (1)	0 (0)
	F	20 (21)	22 (23)	69 (69)	6 (7)	8 (12)	14 (8)	82 (87)	1 (1)	3 (5)	5 (2)	93 (96)	0 (0)
	J	40 (38)	2 (1)	4 (3)	70 (66)	29 (26)	0 (0)	0(0)	77(76)	17(16)	0(0)	0(0)	86 (88)

Web Table 3.1 : The proportions of times, expressed in percentages, that one of the models in Table 1 minimizes the AIC criterion, when the data sets are generated from one of the models in Table 1. Results when the marginal capture probability is modeled using either a logit or a Log-Log (results in parentheses) link function are given.

Web Appendix B : Sampling distribution properties

In Web Figures 3.1-3.3, we present results for Frank's, Gumbel's and Joe's data, respectively. These results shows that if no covariate is introduced, the sampling distribution can be highly skewed and models tend to give slightly larger population size estimate particulary when τ increases. These founding are more pronounced under the model of Clayton, compared to others. On the other hand, if the covariate is used the skewness in sampling distributions is greatly reduced and estimators obtained under all models are unbiased, except that of H's model which is negatively biased compared to others. Results also appear to be little better with the complementary Log-Log link, in comparison with the logit link function.



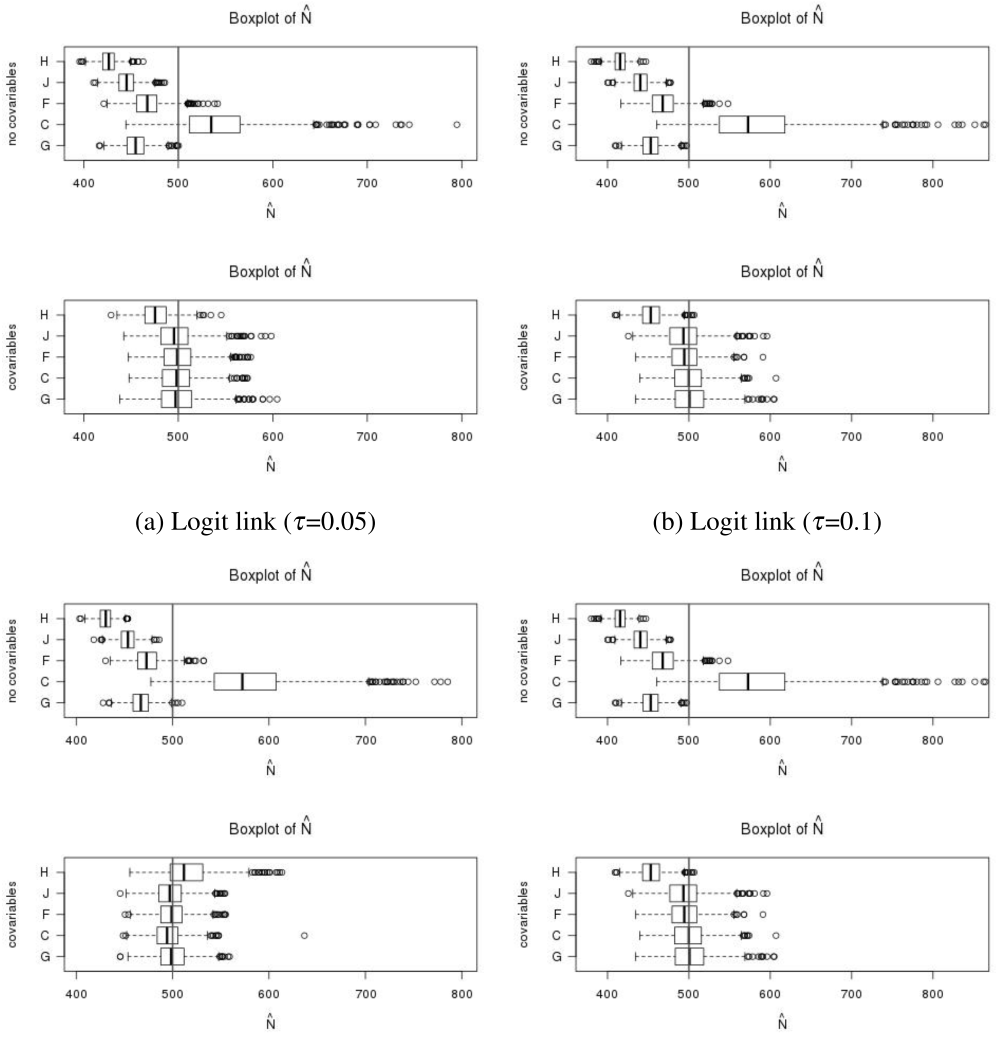
(a) Logit link (τ=0.05)

(b) Logit link (τ=0.1)

(c) c-Log-Log link (τ=0.05)

(d) c-Log-Log link (τ=0.1)

Web Figure 3.1 : Boxplots of population size estimators obtained under Gumbel (G), Clayton (C), Frank (F), Joe (J) and Zero-Truncated Binomial Regression (H) models without covariates (above) and models with single covariates (below), for **Frank data**, $N = 500$ and τ=0.05, 0.1 . The marginal probability of capture is modeled using Logit (graphs (a)-(b)) and complementary-Log-Log (c-Log-Log) (graphs (c)-(d)) link functions.



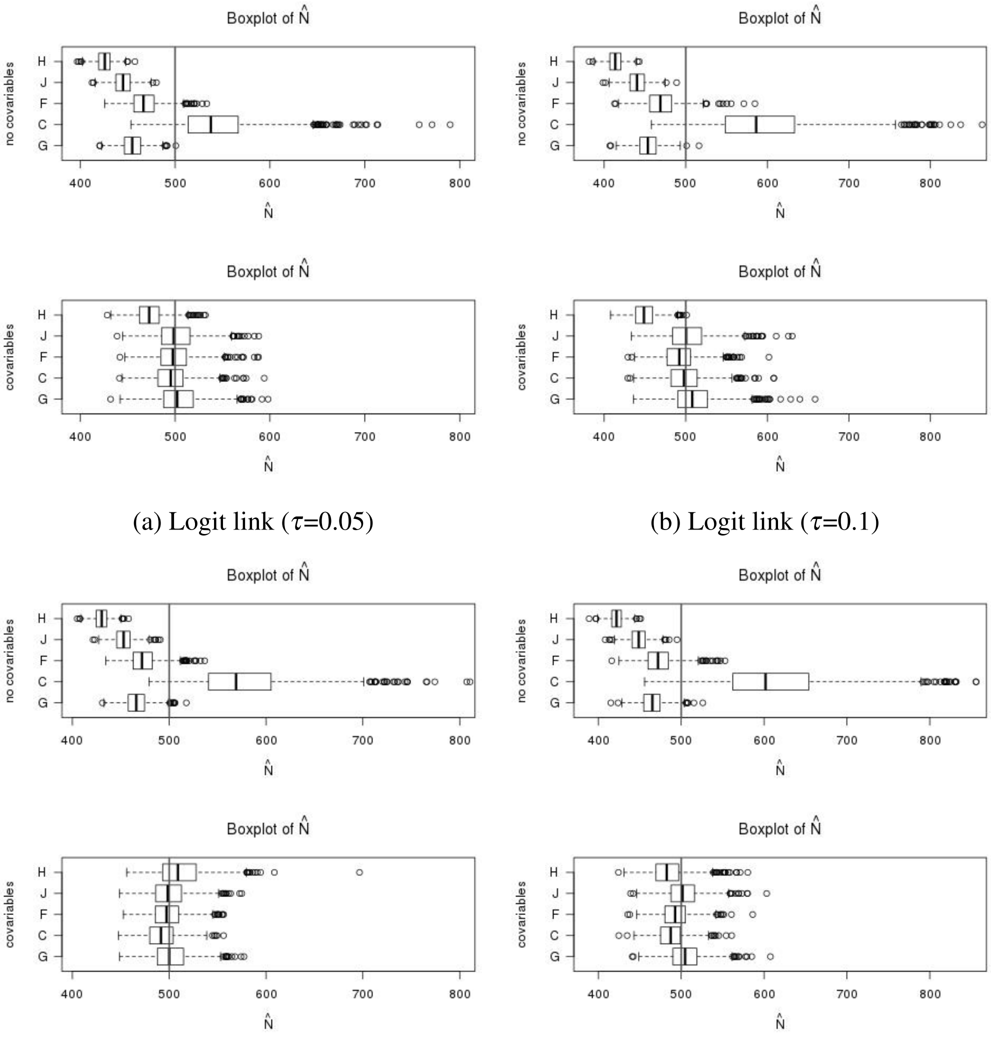
(a) Logit link (τ=0.05)

(b) Logit link (τ=0.1)

(c) c-Log-Log link (τ=0.05)

(d) c-Log-Log link (τ=0.1)

Web Figure 3.2 : Boxplots of population size estimators obtained under Gumbel (G), Clayton (C), Frank (F), Joe (J) and Zero-Truncated Binomial Regression (H) models without covariates (above) and models with single covariates (below), for **Gumbel data**, $N = 500$ and τ=0.05, 0.1. The marginal probability of capture is modeled using Logit (graphs (a)-(b)) and complementary-Log-Log (c-Log-Log) (graphs (c)-(d)) link functions.



(a) Logit link (τ=0.05)

(b) Logit link (τ=0.1)

(c) c-Log-Log link (τ=0.05)

(d) c-Log-Log link (τ=0.1)

Web Figure 3.3 : Boxplots of population size estimators obtained under Gumbel (G), Clayton (C), Frank (F), Joe (J) and Zero-Truncated Binomial Regression (H) models without covariates (above) and models with single covariates (below), for **Joe data**, $N = 500$ and τ=0.05, 0.1 . The marginal probability of capture is modeled using Logit (graphs (a)-(b)) and complementary-Log-Log (c-Log-Log) (graphs (c)-(d)) link functions.

Web Appendix C : Confidence intervals

Web Table 3.2 reports results on confidence interval (CI) that assume that the model is selected correctly. It evaluates Chao's CI for N under the models of Gumbel, Clayton, Frank, and Joe. Chao's CI has generally good statistical properties ; its coverage is close to the nominal 95% level and its average length is relatively small. These results also show that models are sensitives to a link function ; the CIL under the Log-Log link function is always shorter for all models, compared with those obtained with the Logit link function. This result was expected since n was on average 2% larger with the Log-Log link function. Note also that the expected confidence lengths for all methods increase with the heterogeneity level. Thus Chao's CI constructed using the four models investigated in this section are reliable statistical techniques.

				$\tau = 0.05$		τ =0.1		$\tau = 0.2$	
N	Link	Data	Models	CIL	Cov	CIL	Cov	CIL	Cov
100	Logit	G	G	49.01	93.80	60.81	92.70	149.76	94.20
		C	C	46.81	94.00	54.55	94.10	79.44	94.60
		F	F	47.61	95.30	58.86	93.60	93.71	93.90
		J	J	49.48	94.20	62.39	92.70	131.20	93.40
	Lg-Lg-c	G	G	38.56	93.60	44.85	92.99	68.00	94.79
		C	C	37.73	93.04	41.03	94.81	57.46	94.58
		F	F	37.92	95.36	43.93	94.33	68.30	94.07
		J	J	38.33	92.02	46.25	93.91	67.55	95.61
250	Logit	G	G	65.29	93.60	79.46	94.00	112.38	94.10
		C	C	65.72	93.70	73.17	94.50	92.84	94.50
		F	F	67.78	93.60	76.73	93.70	102.99	95.00
		J	J	67.22	94.40	82.26	93.50	133.99	94.10
	Lg-Lg-c	G	G	53.41	92.12	63.18	94.11	83.56	94.71
		C	C	52.27	91.45	58.91	95.12	76.75	94.72
		F	F	54.00	94.01	61.87	95.46	79.81	94.65
		J	J	53.43	94.43	62.87	94.46	86.23	94.81
500	Logit	G	G	87.45	91.60	104.59	94.90	147.39	94.90
		C	C	87.31	94.90	98.30	93.60	122.63	94.80
		F	F	91.00	95.20	103.25	94.80	136.75	95.10
		J	J	89.71	95.20	108.47	94.00	160.91	94.70
	Lg-Lg-c	G	G	72.37	94.51	83.59	93.46	108.47	95.82
		C	C	71.78	94.70	78.96	95.83	101.37	94.48
		F	F	73.58	94.53	82.66	94.60	105.24	95.08
		J	J	72.43	93.23	85.31	94.68	111.43	95.62
750	Logit	G	G	104.90	91.50	126.64	95.30	173.37	94.70
		C	C	107.16	95.40	118.18	94.10	147.47	95.10
		F	F	109.72	95.10	124.42	94.10	160.25	94.90
		J	J	109.50	94.90	130.36	96.50	186.88	95.00
	Lg-Lg-c	G	G	86.69	92.78	100.21	94.55	129.94	95.47
		C	C	85.97	93.88	94.70	94.99	121.46	94.28
		F	F	88.69	95.76	98.77	96.00	127.71	93.86
		J	J	89.73	92.99	101.15	94.94	132.49	95.15

Web Table 3.2 : Confidence interval length (CIL) and coverage (COV) of two-sided confidence intervals for N with a nominal 95 confidence level for data generated for 8 models

Web Table 3.3 report results of confidence intervals based on Zero-Truncated Binomial Regression (H) model. The Chao's confidence intervals based on this model shows in general serous under-coverage. Its coverage probabilities go to zero as N or τ increases.

				$\tau = 0.05$		$\tau = 0.1$		$\tau = 0.2$	
N	Link	Data	Models	CIL	Cov	CIL	Cov	CIL	Cov
100	Logit	G	H	32.41	90.70	27.12	74.00	18.78	33.10
		C	H	34.61	93.20	31.87	88.90	25.73	60.90
		F	H	32.86	93.30	29.36	82.30	20.72	39.80
		J	H	30.91	87.70	25.37	70.80	17.42	25.90
	c-Log-Log	G	H	27.78	92.99	23.35	82.11	17.62	46.68
		C	H	29.17	93.14	25.97	88.40	21.42	62.95
		F	H	27.66	93.45	24.34	85.01	19.04	50.10
		J	H	26.84	89.49	22.76	80.71	16.12	34.90
250	Logit	G	H	46.00	83.50	38.12	51.60	26.27	4.70
		C	H	49.64	91.20	44.39	71.00	35.33	25.80
		F	H	47.55	86.30	40.55	59.40	29.44	7.90
		J	H	44.66	81.30	35.71	42.70	23.80	2.70
	c-Log-Log	G	H	39.79	89.39	34.59	62.74	25.35	8.04
		C	H	41.28	90.24	37.42	74.80	30.43	25.79
		F	H	40.62	87.92	36.14	70.03	26.94	13.54
		J	H	38.86	87.55	33.33	57.86	23.51	4.99
500	Logit	G	H	61.78	71.10	51.24	22.10	35.65	0.00
		C	H	66.40	83.40	60.55	52.10	47.45	3.30
		F	H	64.41	78.40	54.95	32.00	39.74	0.20
		J	H	59.26	65.40	48.29	14.70	31.74	0.00
	c-Log-Log	G	H	54.01	78.54	46.83	34.87	34.38	0.31
		C	H	57.31	87.36	50.98	57.68	41.13	3.78
		F	H	55.26	83.18	48.93	48.07	36.84	1.02
		J	H	52.91	72.62	44.98	28.32	31.68	0.10
750	Logit	G	H	74.98	60.00	62.30	10.50	42.45	0.00
		C	H	81.29	79.90	72.14	32.10	57.49	0.30
		F	H	77.66	69.00	66.70	17.10	47.24	0.00
		J	H	72.79	51.90	58.76	4.60	37.50	0.00
	c-Log-Log	G	H	65.22	70.28	56.23	18.31	41.15	0.00
		C	H	68.27	80.94	61.18	35.75	49.67	0.42
		F	H	66.96	74.43	58.62	25.13	44.82	0.20
		J	H	64.87	66.91	53.99	10.43	38.01	0.00

Web Table 3.3 : Confidence interval length (CIL) and empirical coverage (COV) of two-sided PLCI for N with a nominal 95 confidence level calculated using Huggins (1991) estimator for the M_h model for data generated using the models of Table 1

Annexe B

Proofs

B.1 Proof of Chapter 3

For the proof of this theorem, we need the notion of a subadditive function :

Definition : We say that a function f defined on $[0,\infty)$ is subadditive if and only if

$$f(x_1+\cdots+x_t) \leq f(x_1)+\cdots+f(x_t) \tag{B.1}$$

for all $t \geq 2$ and $x_j \in [0,\infty]$, $j=1,\ldots,t$.

For all models in Table 3.1, if $\alpha_1 \leq \alpha_2$ then the function $\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}$ is subadditive.

For Gumbel's family, the function $\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}(x) = x^{(\alpha_1+1)/(\alpha_2+1)}$ is concave, hence from the Corollary 4.4.4 and Example 4.14 in Nelson (2006) the function $\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}$ is subadditive. For Clayton's family, $\psi_{\alpha_1}^{-1}(x)/\psi_{\alpha_2}^{-1}(x) = (\alpha_2/\alpha_1)(1-x)^{\alpha_1-\alpha_2}$ is nondecreasing function on $(0,1)$, hence from the corollary 4.4.5 in Nelson (2006) the function $\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}$ is subadditive. For Frank's model, the proof is given in (Nelson, 2006, Corollary 4.4.6), see also (Genest and Mackay, 1986, Example 5.3). For Joe's family, the inverse of the Laplace transform is $\psi_{\alpha}^{-1}(x) = \log\{1-(1-x)^{\alpha+1}\}$ and its derivative is $(\psi_{\alpha}^{-1})'(x) = (\alpha+1)(1-x)^{\alpha}/\{1-(1-x)^{\alpha+1}\}$. If $\alpha_1 \leq \alpha_2$ then the quotient $(\psi_{\alpha_1}^{-1})'(x)/(\psi_{\alpha_2}^{-1})'(x) = (\alpha_1+1)/(\alpha_2+1)(1-x)^{\alpha_1-\alpha_2}\big(1-(1-x)^{\alpha_2+1}\big)/\big(1-(1-x)^{\alpha_1+1}\big)$ is nondecreasing on $(0,1)$ since its derivative is positive, because it is nondecreasing on $(0,1)$ and null at 0. Thus, it follows from (Nelson, 2006, Corollary 4.4.6) that $\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}$ is subadditive.

So for these models, it follows from the definition (B.1) that

$$\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}(tx) \leq t\psi_{\alpha_1}^{-1} \cdot \psi_{\alpha_2}(x) \tag{B.2}$$

for all $x > 0$.

Let $1-\rho = \psi_{\alpha_2}(x)$ and $x = \psi_{\alpha_2}^{-1}(1-\rho)$. Applying the decreasing function ψ_{α_1} to both sides of (B.2) yields

$$\psi_{\alpha_2}\{t\psi_{\alpha_2}^{-1}(1-\rho)\} \geq \psi_{\alpha_1}\{t\psi_{\alpha_1}^{-1}(1-\rho)\}$$

and

$$1-\psi_{\alpha_1}\{t\psi_{\alpha_1}^{-1}(1-\rho)\}\geq 1-\psi_{\alpha_2}\{t\psi_{\alpha_2}^{-1}(1-\rho)\}$$

which completes the proof.