UNIVERSITÉ LAVAL

Régression avec copules pour des données hiérarchiques

Thèse

Talagbe Gabin Akpo

Doctorat en statistique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Résumé

Dans cette thèse, nous proposons un modèle multivarié pour la modélisation des données en grappes. Le modèle proposé, que nous nommons « d-copule échangeable », permet d'écrire la distribution jointe de nd variables aléatoires mesurées sur n unités de la grappe.

Le modèle de d-copule échangeable fait intervenir trois copules et d lois marginales. Il possède des propriétés de flexibilité et de maniabilité dues à sa forme explicite. Nous montrons que la d-copule échangeable est une généralisation du modèle linéaire mixte avec ordonnées à l'origine aléatoires. En effet, lorsque les copules sont toutes normales et les lois marginales sont normales, alors les deux modèles sont équivalents. Nous utilisons le modèle de d-copule échangeable pour faire de la prédiction. Ensuite, nous nous intéressons particulièrement au cas de d=2 variables pour étudier ses propriétés. Nous expliquons la procédure séquentielle pour sélectionner les cinq éléments entrant dans la construction du modèle de 2-copule échangeable. L'estimation des paramètres du modèle de 2-copule échangeable se fait en utilisant deux méthodes d'estimation : la méthode IFM généralisée ou la méthode du maximum de vraisemblance. Nous démontrons que les estimateurs associés aux paramètres du modèle de 2-copule échangeable sont convergents et asymptotiquement normaux que l'on utilise la méthode IFM généralisée ou celle par maximum de vraisemblance. Nous comparons ces deux méthodes d'estimation par le biais d'une étude Monte-Carlo.

Finalement, nous montrons la modélisation de données en utilisant un modèle de 2-copule échangeable. Les données proviennent d'une étude effectuée au centre de Londres, dans le cadre du «Junior School Project (JSP)». Nous construisons des courbes de prédiction en utilisant la méthode de 2-copule échangeable que nous comparons à celles obtenues avec le modèle linéaire mixte et le modèle de régression ordinaire avec une copule.

Abstract

In this thesis, we propose a multivariate model for modeling clustered data. The proposed model, which we name "d-copula", allows us to write the joint distribution of nd random variables measured on n units of the cluster.

The exchangeable d-copula model involves three copulas and d marginal laws. It has properties of flexibility and handiness due to its explicit form. We show that the exchangeable d-copula is a generalization of the linear mixed model with random intercepts. Indeed, when the copulas are all normal and the marginal laws are normal, then the two models are equivalent. We use the exchangeable d-copula model to make predictions. Then, we focus on the case of d=2 variables to study its properties. We explain the sequential procedure for selecting the five elements that go into the construction of the exchangeable 2-copula model. The estimation of the parameters of the exchangeable 2-copula model is done using two estimation methods : the generalized IFM method or the maximum likelihood method. We show that the estimators associated with the parameters of the exchangeable d-copula model are convergent and asymptotically normal whether using the generalized IFM or the maximum likelihood method. We compare these two estimation methods by means of a Monte-Carlo study.

Finally, we show the construction of an exchangeable 2-copula model from observed data. The data come from a study in central London, as part of the «Junior School Project (JSP)». We construct prediction curves using the exchangeable 2-copula method and compare them to those obtained with the mixed linear model and the ordinary regression model with one copula.

Table des matières

Liste des tableaux

Liste des figures

Aux trois femmes
d'inspiration : ma mère
Abla Alice Esseh, ma
bien-aimée épouse Fatima
Zougouri Akpo et Étchèlissa
Zeyna Maureen Akpo

Remerciements

J'aimerais formuler mes chaleureux remerciements et ma profonde gratitude à l'endroit de mon directeur de thèse, Professeur Louis-Paul Rivest, qui n'a cessé de m'encourager malgré les nombreuses péripéties et les embûches sur le chemin doctoral. Son enthousiasme pour le projet, son soutien et sa patience ont contribué grandement à la réussite de ce projet doctoral. Il m'a permis de tracer les sillons de l'excellence.

Mes remerciements au Dr. Jules Brice Tchatchueng, qui a cru en moi et m'a donné l'opportunité de gravir les échelons. Recevez les bénédictions de Dieu le Tout-Puissant.

J'adresse mes remerciements au Dr. Raymond Hounnonkpe, au Dr. Ali Ben Charif, à Sophie Baillargeon avec qui j'ai eu des discussions fructueuses dans la rédaction de ma thèse. Je vous suis reconnaissant du temps que vous m'avez consacré.

Je tenais à dire merci à mes amis Mamane Samri, Érika Lupien-Angel, Dr. Mamadou Yauck, Pr. Bernard Lamond, Dr. Judicael Alladatin et Christine Bérubé. Les moments de solidarité et de discussions m'ont permis de faire émerger de nouvelles idées.

Je tiens également à dire merci à tou-te-s les professeur-e-s du Département de Mathématiques et de Statistique avec qui j'ai discuté tout au long de notre projet. Vous avez été des exemples, des sources d'inspiration dignes de vos fonctions respectives et représentent pour moi une réussite dans le transfert de connaissances. Un sincère merci à Catherine Levesque pour sa bonne humeur et l'accueil toujours chaleureux.

Je remercie enfin celles et ceux qui me sont très proches et que j'ai quelque peu oubliés ces dernières années pour achever cette thèse en particulier ma soeur Pélagie Akpo. Leurs attentions et encouragements m'ont accompagnée tout au long de la rédaction. Je suis très reconnaissant à mes feus parents, Kodjo et Abla Alice Akpo. Que la terre leur soit légère. Enfin, je ne manquerai pas de dire mon attachement et ma constante motivation à véhiculer les valeurs sacerdotales que m'ont inculquées M. Édouard Adje et M. Paulin Étienne Édah. Encore merci pour ces caractères que vous m'avez aidé à

forger.

Notations et symboles utiles

Dans cette thèse, plusieurs notations et symboles ont été utilisés. Certaines notations portant sur les copules s'inspirent de celles utilisées dans Joe (2014, page 107), Czado (2019, page 77) ou encore de Chang (2019).

Notations et symboles utilisés dans le chapitre 1

d	:	un nombre positif égal au nombre de variables mesurées sur un individu.
A^T ou u^T	:	transposée de la matrice A ou du vecteur u.
$\mathbf{X} = (X^1, \ldots, X^{d-1})^T$	:	un vecteur aléatoire de dimension $(d-1)$.
$\mathbf{X}$ et $\mathbf{x}$	:	la notation $\mathbf{X}$ majuscule représente un vecteur aléatoire et $\mathbf{x}$ est la valeur observée du vecteur aléatoire.
$\mathbf{Z} = (\mathbf{X}^T, Y)^T$	:	un vecteur aléatoire de dimension d.
F	:	fonction de répartition unidimensionnelle dont la densité est f.
$F_{d,n}$	:	la fonction de répartition d'un ensemble de n vecteurs dont chacun est de longueur d.
$f_{d,n}$	:	la fonction de densité associée à $F_{d,n}$.
C, c	:	copule, la densité de la copule.
$\bar{C}$,	:	copule de survie de C.
C_{ij}	:	la copule bivariée associée au vecteur (U_i, U_j).
$C_{ij;S}$	:	la copule conditionnelle bivariée associée au vecteur (U_i, U_j) sachant un ensemble de variables S.
$C_{i\|S}$	:	la fonction de distribution conditionnelle de U_i sachant un ensemble de variables S.
$\mathbb{E}$	:	espérance d'une variable aléatoire.
$\mathbb{P}$	:	probabilité d'un évènement.
$\mathbb{V}$	:	variance d'une variable aléatoire.
$\mathbb{R}$	:	ensemble des nombres réels.
$\sim$	:	distribuer selon la loi.
$\Gamma(\alpha)$	:	la fonction gamma évaluée à α.
$\Sigma(\rho, n)$	:	matrice carrée d'ordre n dont tous les éléments hors diagonale sont égaux à ρ et les éléments de la diagonale sont égaux à 1.

ϕ	:	la fonction de densité de la loi normale centrée réduite.
$B_{p,q}$	:	la fonction de répartition d'une loi bêta de paramètres p et q.
cor	:	corrélation de Pearson entre deux variables.
cov	:	covariance entre deux variables.
ρ_N	:	la corrélation de Pearson calculée avec des scores normaux. Le score normal de a est $\Phi^{-1}(a)$.
ρ_S	:	le rho de Spearmann.
τ	:	le tau de Kendall.

Notations et symboles additionnels utilisés dans le chapitre 2

$\boldsymbol{U}$	:	un vecteur aléatoire de dimension $(d-1)$.
$\boldsymbol{u}$	:	observation de dimension $(d-1)$.
$C_{d,n}$	:	la copule ou une fonction de répartition multivariée d'un ensemble de n vecteurs aléatoires dont chacun est de dimension d.
$C_{d,1}$	:	la copule ou une fonction de répartition multivariée d'un vecteur aléatoire de dimension d.
$C_{1,n}$	:	la famille de fonctions de répartition multivariées d'un ensemble de n vecteurs aléatoires dont chacun est de dimension 1 notée aussi $\left(C_{1,n}\right)$. Elle s'obtient de la notation $C_{d,n}$ pour $d=1$.
$\equiv$	:	équivalent à.
$\approx$	:	approximativement égal à.
$\overset{\mathcal{L}}{=}$	:	même distribution.
$\xrightarrow{\mathcal{L}}$	:	convergence en distribution.
$\xrightarrow{\mathcal{P}}$	:	convergence en probabilité.
$\|\|.\|\|_d$	:	la norme euclidienne sur $\mathbb{R}^d$ appliqué à un vecteur ou à une matrice.
$\otimes$	:	produit tensoriel.
$\mathbf{1}_n$	:	$\underbrace{(1,\ldots,1)^T}_{\text{taille } n}$, vecteur de taille n dont chaque composante est 1.
$\mathbf{I}_n$	:	matrice identité d'ordre n.
$\mathbf{0}_n$	:	matrice carrée nulle.
$\mathbf{J}_n$	:	matrice $n \times n$ dont tous les termes sont 1.
$\mathcal{L}$	:	le logarithme de la vraisemblance d'un modèle.
∇_θ	:	le gradient par rapport au vecteur θ.

Notations et symboles additionnels utilisés dans les chapitres 3 et 4

AIC	:	Akaike Information Criterion.
MV	:	Maximum de Vraisemblance.
IFM	:	Inference Function for Margins.
m	:	le nombre de grappes.
X	:	la variable aléatoire indépendante.
Y	:	la variable aléatoire dépendante.
x_{ji}	:	observation de la variable explicative de l'individu $i = 1, \ldots, n_j$, de la grappe $j = 1, \ldots, m$.
y_{ji}	:	observation de la variable dépendante de l'individu i, de la grappe $j = 1, \ldots, m$.
$\mathbf{x}_j$ et $\mathbf{X}_j$	:	$\mathbf{x}_j = (x_{j1}, \ldots, x_{jn_j})^T$, un vecteur de taille n_j de la grappe j et $\mathbf{X}_j = (X_{j1}, \ldots, X_{jn_j})^T$, le vecteur aléatoire correspondant.
$\mathbf{y}_j$ et $\mathbf{Y}_j$	:	$\mathbf{y}_j = (y_{j1}, \ldots, y_{jn_j})^T$, un vecteur de taille n_j de la grappe j et $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jn_j})^T$, le vecteur aléatoire correspondant.
$F(.;\alpha)$	:	la fonction de répartition d'une loi paramétrique de paramètre α.
$B(\alpha, \beta)$	:	Fonction beta qui se calcule par $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$.
$\mathcal{B}(.;\alpha, \beta)$	:	loi bêta à deux paramètres α et β positifs.
$F_{\mathcal{B}}$	:	la fonction de répartition de la loi bêta à deux paramètres.
$\mathcal{GB}3(.;\alpha, \beta, \lambda)$	:	loi bêta à trois paramètres α, β et λ.
$F_{\mathcal{GB}3}$	:	la fonction de répartition de la loi bêta à trois paramètres.
$\mathcal{G}(\theta)$	:	la matrice d'information de Godambe.
$\mathcal{I}(\theta)$	:	la matrice d'information de Fisher.
$\sup_\theta$	:	le supremum sur θ d'une fonction.

Introduction générale

La modélisation multidimensionnelle pour données en grappes est un problème statistique complexe et diverses approches sont proposées dans la littérature. Les modèles linéaires à effets aléatoires sont fréquemment utilisés pour l'analyse de données hiérarchiques, voir par exemple Verbeke et Molenberghs (2000, page 23), Albert (2012) et Brown et Prescott (2014, page 384). Le modèle de Battese *et al.* (1988) est une façon d'explorer ces données. Des applications de ce modèle sont faites dans divers domaines comme en agriculture, Brown et Prescott (2014, page 384), en santé, Brown et Prescott (2014, page 392) et en médecine. Le modèle linéaire mixte se base sur plusieurs hypothèses. La première hypothèse est la normalité des résidus avec homogénéité de la variance et la normalité des effets aléatoires. La deuxième hypothèse est que le modèle prédictif construit est linéaire. Cependant, l'hypothèse de normalité s'avère être très restrictive et le problème de modélisation de loi en plusieurs dimensions est récurrent. Les domaines des assurances et de la finance constituent des exemples où la distribution des variables tend à être asymétrique (loi beta, gamma, etc), voir Cherubini *et al.* (2014).

La modélisation avec les copules est une approche intéressante. En effet, elles permettent de modéliser la dépendance entre deux ou plusieurs variables aléatoires indépendamment des lois marginales, voir Sklar (1959). Les copules constituent un puissant outil pour faire de la modélisation multivariée, voir Joe (2014, page 7). Elles peuvent aussi être utilisées pour faire de la prédiction, voir Kumar (2007) ou Crane et Van der Hoek (2008). Nous présentons au chapitre 1, section 1.6 de la thèse, la régression avec les copules pour expliquer cette méthode. Cette méthodologie a été reprise par Noh *et al.* (2013) dans un cas multidimensionnel. Ces auteurs étudient les propriétés asymptotiques des estimateurs des paramètres. Plus particulièrement, dans le cas bivariée, ceux-ci construisent des courbes de prédiction pour illustrer la pertinence de la méthode. L'usage des copules pour faire de la prédiction devient de plus en plus fréquent dans plusieurs domaines. Par exemple, Kraemer *et al.* (2013) s'appuient sur les copules

pour construire des modèles "robustes" de prédiction dans l'étude des sinistres en assurance. De même, Genest *et al.* (2013) fait de la prédiction d'une variable dépendante binaire avec les copules en environnement et Li *et al.* (2016), en ingénierie électrique, modélise la dépendance en matière de coût. Récemment, pour la tarification des franchises en assurance, Shi et Lee (2022) ont utilisés les copules pour construire un modèle de prédiction. Par ailleurs, la régression sur des données longitudinales en utilisant les copules est aussi proposée dans la littérature, voir entre autres Shi et Zhang (2011), Wang *et al.* (2019) et Wang et Shan (2020). Shi *et al.* (2011) analysent des données hiérarchiques à l'aide d'un modèle utilisant des copules. Pour des données avec beaucoup de zéros, Zhang *et al.* (2020) proposent un modèle utilisant les copules et qui prend en compte cette spécificité. En particulier, dans le cas de données en grappes, des auteurs proposent l'utilisation des copules archimédiennes pour faire la prédiction, voir Su *et al.* (2019). En dehors de la prédiction avec les modèles utilisant les copules, d'autres auteurs s'intéressent aux erreurs quadratiques des prédicteurs de nouvelles observations, voir par exemple Acar *et al.* (2019). En effet, dans l'approche de modélisation des données en grappes, Acar *et al.* (2019) reprennent la méthodologie de prédiction en utilisant les copules mais en calculant les erreurs quadratiques moyennes pour plusieurs familles de copules. Ceci permet de faire une évaluation comparative des modèles prédictifs utilisant les copules.

L'utilisation de modèles construits à partir des copules pour faire la prédiction nécessite la connaissance de la distribution multivariée des variables aléatoires. En utilisant l'équation (1.2), la distribution multivariée s'obtient à partir des lois marginales et d'une copule multivariée en faisant usage du théorème du Sklar, voir Sklar (1959). En effet, la première méthode de construction est l'utilisation directe des copules multivariées pour obtenir la distribution multivariée. La deuxième méthode, plus souple et flexible, utilise les vignes à la place de la copule multivariée, voir Joe et Kurowicka (2011). Les modèles en vigne permettent d'établir une liaison entre des variables aléatoires, voir Kurowicka et Cooke (2006, page 81). En effet, elles permettent de décomposer une distribution multivariée en des produits de copules bivariées et de marginales, voir Joe (2014, page 107), Czado et Nagler (2022). Elles se construisent aussi sur les données discrètes, voir Panagiotelis *et al.* (2012). La décomposition en vigne est très flexible puisqu'il existe une infinité de copules bivariées. Dans la décomposition en vigne de d variables aléatoires, nous notons plusieurs formes de décomposition en fonction de la stratégie utilisée. En effet, il existe la vigne R, la vigne C, la vigne D, etc., toutes dépendantes de la forme du premier "arbre" de la décomposition. Nous considérons

ici la vigne D en guise d'illustration pour écrire la fonction de densité de la copule multivariée.

Soit $Z_1, \ldots, Z_d$, d variables aléatoires de fonctions de répartition $F_1, \ldots, F_d$ respectivement. La fonction de densité de la copule associée au vecteur aléatoire $(Z_1, \ldots, Z_d)$, en utilisant la décomposition en D-vigne pour $u_i \in [0,1], i = 1, \ldots, d$ s'écrit

$$\prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{ii+j;i+1:i+j-1} \left\{ C_{i|i+1:i+j-1}(u_i|u_{i+1}, .., u_{i+j-1}), C_{i+j|i+1:i+j-1}(u_{i+j}|u_{i+1}, .., u_{i+j-1}) \right\}.$$

La fonction $c_{ii+j;i+1:i+j-1}$ est la fonction de densité de la copule bivariée conditionnelle associée au vecteur aléatoire (Z_i, Z_{i+j}) sachant le vecteur $(Z_{i+1}, \ldots, Z_{i+j-1})$. La fonction $C_{i|i+1:i+j-1}$ est la distribution conditionnelle de la variable aléatoire $F_i(Z_i)$ sachant le vecteur $\{F_{i+1}(Z_{i+1}), \ldots, F_{i+j-1}(Z_{i+j-1})\}$. Ce résultat est inspiré des travaux de Stöber *et al.* (2013) et de Joe (2014, page 108). Nous présentons succinctement la décomposition en vigne à la section 1.7 pour $d = 4$ variables aléatoires. Il est possible de faire une réduction des décompositions en vigne en utilisant les hypothèses dites simplificatrices que nous expliquons dans le chapitre 1, section 1.7 de cette thèse. Plus précisément, une de ces hypothèses simplificatrices consiste à tronquer certains arbres de la décomposition en vigne. La seconde hypothèse stipule que la copule conditionnelle ne dépend pas de la variable sur laquelle nous conditionnons, voir exemple 1.9. Cependant, l'utilisation très fréquente d'hypothèses sur la vigne pour réduire la dimension de l'espace des paramètres peut avoir des limites, voir Han *et al.* (2017). En effet, une hypothèse d'indépendance pour une relation entre deux variables aléatoires peut s'avérer fausse. D'autre part, la décomposition en vigne fait appel à un grand nombre de copules bivariées et le travail devient fastidieux. Pour ce faire, les algorithmes se positionnent comme une solution sérieuse pour faire le travail dans le choix des arbres, voir Joe (2014, page 259) et Hobaek Haff (2013). Par ailleurs, des auteurs comme Spanhel et Kurz (2019) proposent des méthodes d'approximation des copules conditionnelles des arbres supérieurs de la décomposition. L'idée commune à toutes les constructions en vigne est de tester plusieurs copules ayant une forme particulière pour sélectionner la "meilleure" pour chaque relation. Cette méthode, bien que cela soit une solution acceptable, pose aussi deux problématiques. La première est la forme des copules proposées pour approximer les copules bivariées et la deuxième est la complexité de calculs due à l'existence d'un grand nombre de copules bivariées à ajuster.

À partir de la distribution multivariée construite, nous prédisons facilement en utilisant la loi conditionnelle, voir section 1.6. De plus, Cooke *et al.* (2019), montre spécifiquement

comment nous pouvons partir d'une distribution jointe de variables aléatoires dont une variable est dépendante et le reste, des covariables, en utilisant les vignes, pour construire un modèle de prédiction, voir aussi Atique et Attoh-Okine (2016) et Smith et Klein (2021). Nous traitons des données en grappes qui ont des particularités liées à la dépendance à l'intérieur des grappes. La plupart des modèles existants, Battese *et al.* (1988), Graf *et al.* (2018) Rivest *et al.* (2016) ou Acar *et al.* (2019), utilisé pour modéliser les données en grappes sont des modèles conditionnels. En effet, pour un vecteur de covariables $\mathbf{X}$ et une variable dépendante Y, les équations du modèle de régression linéaire mixte et du modèle de régression avec copules donnent la loi Y sachant $\mathbf{X}$. Autrement dit, pour construire le prédicteur, nous supposons que le vecteur de covariables $\mathbf{X}$ est connu.

Dans notre thèse, la stratégie innovante développée consiste à modéliser la distribution jointe de l'ensemble des vecteurs $(\mathbf{X}_i, Y_i)_{i=1}^n$ dans une grappe à n individus, voir la définition 2.4. Elle propose une toute nouvelle vision qui permet de considérer $\mathbf{X}$ comme un vecteur aléatoire et Y comme une variable aléatoire puis de donner la loi jointe de $(\mathbf{X}^T, Y)^T$. Ceci nous a conduits à la formulation d'un modèle nommé d-modèle échangeable, voir la définition 2.4. En partant du modèle proposé et en conditionnant sur les covariables $\mathbf{X}$, nous aboutissons aux modèles existants d'une part. D'autre part, ce modèle proposé a une expression analytique explicite et possède les propriétés de flexibilité. Pour ce faire, nous faisons usage à la fois des copules, de leurs propriétés et de la décomposition en vigne pour déboucher sur un modèle à la fois flexible, souple et facilement maniable. Plus spécifiquement, nous définissons un nouveau modèle adapté aux données en grappes en partant de la décomposition en vigne et qui repose sur une méthodologie différente des modèles existants. En conséquence, nous définissons quatre objectifs spécifiques.

Objectif spécifique 1 Proposer un modèle multivarié pour les données en grappes faisant intervenir des copules et qui intègre le modèle linéaire mixte comme cas particulier, puis étudier ses propriétés.

Objectif spécifique 2 Suggérer des méthodes d'estimation de ses paramètres, puis étudier les propriétés asymptotiques des estimateurs des paramètres du modèle.

Objectif spécifique 3 Construire des prédictions pour des observations futures dans une grappe connaissant les données des individus observés dans cette grappe.

Objectif spécifique 4 Utiliser le modèle proposé pour analyser un jeu de données et faire des prédictions. Comparer les prédictions à celles obtenues avec un modèle linéaire mixte standard.

Nous avons organisé la thèse en quatre chapitres. Dans le chapitre 1, nous recensons et énonçons les propriétés des copules et les éléments importants de la régression copule. Nous faisons aussi un bref survol des mesures de corrélation, de la décomposition en vigne et de la régression avec les modèles de copules sans oublier la présentation du modèle Battese *et al.* (1988). Le chapitre 2 aborde la définition du concept de d-copule en se basant sur trois éléments : l'échangeabilité multidimensionnelle, l'indépendance conditionnelle et la compatibilité qui sont définies clairement. Nous nous intéressons par la suite aux propriétés de ce nouveau modèle et ainsi qu'à son utilisation pour construire des modèles prédictifs. Le modèle proposé généralise le modèle de Battese *et al.* (1988) et l'alternative proposée par Rivest *et al.* (2016). Dans le chapitre 3, nous traitons de la méthode de construction et de l'estimation de la d-copule proposée au chapitre 2. Nous étudions les propriétés de convergence et de normalité asymptotique des estimateurs issus des méthodes d'estimation du modèle. En effet, les estimateurs des paramètres proposés s'obtiennent par maximisation de la vraisemblance globale ou des vraisemblances marginales en utilisant la méthode IFM généralisée, expliquée dans Joe et Xu (1996). Enfin, dans le chapitre 4, nous utilisons le modèle de 2-copule échangeable pour construire des courbes de prédictions par école, des notes en mathématiques.

Les annexes A, B et C sont les démonstrations des résultats ou des explications techniques des chapitres 1, 2 et 3 respectivement. La partie informatique pour obtenir les figures, les estimations et les évaluations de prédicteur dans toute cette thèse se trouve aussi dans les annexes et une partie à l'annexe D.

Chapitre 1

Théorie générale sur les copules et la régression

Dans ce chapitre, $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathbb{P})$, où $\mathcal{B}$ est une tribu est l'espace probabilisé de dimension d.

1.1 Introduction

Les copules ont connu un développement important au cours des dernières années. En effet, dès leur première définition formelle grâce au théorème de Sklar en 1959 jusqu'à des applications récentes en assurance et en finance pour modéliser les risques, voir Brahim *et al.* (2018) et Cherubini *et al.* (2014), plusieurs objectifs sont poursuivis quant à leur utilisation et leur mise en œuvre. Premièrement, elles sont utilisées pour faire de la modélisation à cause de leur flexibilité pour construire des distributions jointes. Deuxièmement, elles permettent de construire des modèles probabilistes pour des phénomènes complexes. Les copules sont aussi appliquées pour l'étude des données temporelles, voir (Vaz de Melo Mendes et Aiube, 2011), et spatiales, voir (Quessy *et al.*, 2015). Il y a un très grand nombre de copules voire une infinité si on ne se restreint pas aux copules paramétriques. Ceci crée une multitude de choix lors de leur utilisation. Nous présentons dans les sections qui suivent les notions essentielles sur les copules tout en définissant certaines familles spécifiques, leurs propriétés et d'autres outils comme des mesures de corrélations qui sont calculés à partir des copules. Nous terminons ce chapitre en faisant un bref survol de la méthode de construction des distributions multivariées à partir des copules en vigne, voir Joe et Kurowicka (2011). Nous présentons aussi le modèle de régression mixte de Battese *et al.* (1988). Finalement, la régression

avec copules est brièvement abordée à la fin du chapitre.

1.2 Notions générales sur les copules

Le but de cette section est de donner des notions de base sur les copules.

Définition 1.1. ***Fonction de répartition***

Soit $(X_1, \ldots, X_d)$*, un vecteur aléatoire de dimension* d*. La fonction de répartition* F *de* $(X_1, \ldots, X_d)$ *est*

$$F(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d), \quad x_1, \ldots, x_d \in \mathbb{R}\cdot \tag{1.1}$$

Définition 1.2. ***Copule***

Une fonction C *de* $d \geq 2$ *variables est une copule si elle est une fonction de répartition définie sur* $[0,1]^d$ *dont les lois marginales sont des lois uniformes sur* $[0,1]$*.*

La copule permet de créer une relation entre les lois marginales de variables aléatoires. Le théorème de Sklar constitue une base de la théorie des copules.

Théorème 1.1. ***Théorème de Sklar***

Si F *est la fonction de répartition jointe de* d *variables aléatoires* $X_1, \ldots, X_d$*, de marginales* $F_1, \ldots, F_d$*, alors il existe une copule* C *définie de* $[0,1]^d$ *sur* $[0,1]$ *telle que*

$$F(x_1, x_2, \ldots, x_d) = C\left\{F_1(x_1), \ldots, F_d(x_d)\right\}, \quad x_1, \ldots, x_d \in \mathbb{R}. \tag{1.2}$$

Inversement, pour toute copule C et pour toutes fonctions de distributions, chacune univariée, $F_1, \ldots, F_d$, la fonction F définie par l'équation (1.2), est une fonction de répartition multivariée, voir Joe (1997, page 7). De plus, si les marges sont continues, alors la copule C, de l'équation (1.2) est unique, Sklar (1959).

Le théorème de Sklar permet donc de construire des distributions multivariées à partir d'une copule et des lois marginales. De ce théorème, on déduit l'expression de la copule à partir de la distribution multivariée et des marginales. Elle s'écrit

$$C(u_1, \ldots, u_d) = F\left\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right\}, \quad u_1, \ldots, u_d \in [0,1],$$

où pour tout u_j, $F_j^{-1}(u_j) = \inf\left\{z : F_j(z) \geq u_j\right\}, j = 1, \ldots, d$ est l'inverse de F_j.

Lorsqu'elle existe, la densité de la copule se calcule de la façon suivante :

$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \partial u_2 \ldots \partial u_d} = \frac{f\left\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right\}}{f_1\left\{F_1^{-1}(u_1)\right\} \times \ldots \times f_d\left\{F_d^{-1}(u_d)\right\}}, \tag{1.3}$$

où f est la fonction de densité du vecteur $(X_1, \ldots, X_d)$ et f_j est la densité de la marge $X_j, j = 1, 2, \ldots, d$.

Le théorème de Sklar est fondamental dans la détermination de la distribution multivariée à partir des copules. Il a forgé l'idée des copules pour une utilisation pratique et constitue une référence de la théorie sur les copules. Pour caractériser une copule, on utilise le théorème suivant que nous énonçons et en s'inspirant de Mai et Scherer (2012).

Théorème 1.2. *Une fonction $C : [0,1]^d \to [0,1]$ est une copule si et seulement si les conditions suivantes sont vérifiées :*

i) $C(u_1, \ldots, u_i, \ldots, u_d) = 0$ *si au moins un* u_i *est nul,* $i \in \{1, \ldots, d\}$ *;*

ii) $C(u) = u_j$ *pour* $u = (u_1, \ldots, u_d) \in [0,1]^d$ *et tous les éléments de* u *sont égaux à 1 sauf la* j*è composante qui est* u_j *;*

iii) C *est d-croissante c'est-à-dire pour* $u = (u_1, \ldots, u_d)$ *et* $v = (v_1, \ldots, v_d)$ *tel que* $u_i \leq v_i, i = 1, \ldots, d$*, on a :* $V_C([u, v]) \geq 0$*, où*

$$[u, v] = [u_1, v_1] \times [u_2, v_2] \times \ldots \times [u_d, v_d],$$

et

$$V_C([u, v]) = \Delta_{u_d}^{v_d} \Delta_{u_{d-1}}^{v_{d-1}} \ldots \Delta_{u_1}^{v_1} C(x_1, \ldots, x_d),$$

avec

$$\begin{aligned} \Delta_{u_k}^{v_k} C(x_1, \ldots, x_d) &= C(x_1, \ldots, x_{k-1}, v_k, x_{k+1}, \ldots, x_d) - \\ &\quad C(x_1, \ldots, x_{k-1}, u_k, x_{k+1}, \ldots, x_d). \end{aligned}$$

Nous présentons maintenant quelques exemples de copules classiques utilisées dans la quantification de la dépendance. Elles répondent à la caractérisation du théorème 1.2.

Exemple 1.1. *La copule d'indépendance notée souvent* C_π^d *est*

$$C_\pi^d(u_1, \ldots, u_d) = \prod_{i=1}^{d} u_i,\ u_1, \ldots, u_d \in [0,1]. \tag{1.4}$$

Cette copule C_π^d *permet de représenter le concept de l'indépendance entre les variables aléatoires.*

Exemple 1.2. *La copule comonotone, encore appelée copule de dépendance maximale, notée souvent M_d est*

$$M_d(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d), \quad u_1, \ldots, u_d \in [0, 1]. \tag{1.5}$$

Cette copule est la fonction de répartition du vecteur aléatoire $(U_1, \ldots, U_d) = (U, \ldots, U)$, où U est une variable aléatoire distribuée uniformément sur $[0, 1]$. Elle s'appelle aussi la borne supérieure de Fréchet.

Exemple 1.3. ***La copule de survie***

On note C une copule de dimension d associée au vecteur $(U_1, \ldots, U_d)$ où les variables aléatoires sont uniformément distribuées sur $[0, 1]$. La copule de survie provenant de la copule C, est la copule associée au vecteur aléatoire $(1 - U_1, \ldots, 1 - U_d)$, voir Nelsen (2006, page 33). En particulier, pour une copule bivariée C, la copule de survie $\bar{C}$ est

$$\bar{C}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2), \quad u_1, u_2 \in [0, 1]. \tag{1.6}$$

Les développements les plus conséquents sont faits sur les copules en dimension 2. Nous donnons quelques exemples de copules bivariées sur $[0, 1]^2$ avec leurs densités, obtenues à partir de l'équation (1.3), définies pour $u_1, u_2 \in [0, 1]$. Pour réaliser le tableau, nous utilisons la référence Joe (2014, page 159)

TABLEAU 1.1 – Exemples de copules bivariées classiques et leurs densités

Copule	Expression	Densité
Gumbel $\alpha \geq 1$	$\exp\left[-\left\{(-\log u_1)^\alpha + (-\log u_2)^\alpha\right\}^{1/\alpha}\right]$	c_α
FGM $-1 \leq \delta \leq 1$	$u_1 u_2 \left\{1 + \delta(1 - u_1)(1 - u_2)\right\}$	$1 + \delta(1 - 2u_1)(1 - 2u_2)$
Clayton $\theta > 0$	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$\frac{(1+\theta)}{(u_1 u_2)^{\theta+1}} \left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-\frac{1}{\theta}-2}$
Joe $\delta \geq 1$	C_δ	c_δ
BB1 $\theta > 0, \delta \geq 1$	$\left[1 + \left\{(u_1^{-\theta} - 1)^\delta + (u_2^{-\theta} - 1)^\delta\right\}^{1/\theta}\right]^{-1/\theta}$	$c_{\theta,\delta}$

La densité de la copule de Gumbel, notée c_α pour $u, v \in]0,1[$ est

$$c_\alpha = C_\alpha(u_1,u_2)\left[\psi_\alpha(u_1)+\psi_\alpha(u_2)\right]^{\frac{1}{\alpha}-2}\left[\alpha-1+\left\{\psi_\alpha(u_1)+\psi_\alpha(u_2)\right\}^{\frac{1}{\alpha}}\right]\frac{\psi_{\alpha-1}(u_1)\psi_{\alpha-1}(u_2)}{u_1u_2},$$

où nous avons

$$C_\alpha(u_1,u_2)=\exp\left[-\left\{(-\log u_1)^\alpha+(-\log u_2)^\alpha\right\}^{1/\alpha}\right],\quad \psi_\alpha(t)=(-\log t)^\alpha,\quad t>0.$$

La copule de Joe est

$$C_\delta(u,v)=1-\left\{(1-u_1)^\delta+(1-u_2)^\delta-(1-u_1)^\delta(1-u_2)^\delta\right\}^{1/\delta}, \tag{1.7}$$

et la densité de la copule de Joe, notée c_δ pour $u_1, u_2 \in]0,1[$ est

$$c_\delta = \left\{\bar{u}_1^\delta+\bar{u}_2^\delta-\bar{u}_1^\delta\bar{u}_2^\delta\right\}^{1/\theta-2}\bar{u}_1^{\delta-1}\bar{u}_2^{\delta-1}\left[\delta-1+\bar{u}_1^\delta+\bar{u}_2^\delta-\bar{u}_1^\delta\bar{u}_2^\delta\right]\cdot \tag{1.8}$$

où $\bar{u}_1 = 1-u_1$ et $\bar{u}_2 = 1-u_2$, voir Joe (2014, page 170).

La densité de la copule BB1, notée $c_{\theta,\delta}$ pour $u_1, u_2 \in]0,1[$ est

$$\begin{aligned} c_{\theta,\delta} &= \left\{1+(x+y)^{1/\delta}\right\}^{-1/\theta-2}(x+y)^{1/\delta-2}\left[\theta(\delta-1)+(\theta\delta+1)(x+y)^{1/\delta}\right]\cdot \\ &\quad (xy)^{1-1/\delta}(u_1u_2)^{-\theta-1}, \end{aligned} \tag{1.9}$$

où $x=(u_1^{-\theta}-1)^\delta$ et $y=(u_2^{-\theta}-1)^\delta$, voir Joe (2014, page 190).

Remarque 1.1. ***Somme convexe de copules***
Soit $\kappa \in [0,1]$ et on définit une fonction C_κ par

$$C_\kappa(u_1,\ldots,u_d)=\kappa C_\pi^d(u_1,\ldots,u_d)+(1-\kappa)M_d(u_1,\ldots,u_d),\ u_1,\ldots,u_d\in[0,1], \tag{1.10}$$

où C_π^d et M_d sont respectivement données par les équations (1.4) et (1.5). La fonction C_κ est une copule. Elle se nomme, la copule de Fréchet-Mardia (Nelsen, 2006, page 15).

Proposition 1.1. *Si $(X_1,\ldots,X_d)$ est un vecteur aléatoire dont la copule associée est C alors C est la copule associée au vecteur $(\varphi_1(X_1),\ldots,\varphi_d(X_d))$ si $\varphi_1,\ldots,\varphi_d$ sont des fonctions continues et strictement croissantes de $\mathbb{R}$ vers $\mathbb{R}$, voir Joe (2014, page 8).*

La proposition 1.1 permet de savoir que la copule associée à un vecteur aléatoire ne change pas lorsqu'on applique un vecteur de transformation croissante. Le théorème suivant donne le résultat que toutes les copules sont comprises entre deux bornes.

Théorème 1.3. ***Bornes de Fréchet-Hoeffding***

Pour toute copule C, on a :

$$W_d(u) = \max\left\{\sum_{i=1}^{d} u_i - d + 1, 0\right\} \leq C(u) \leq M_d(u), \;\; u = (u_1, \ldots, u_d) \in [0,1]^d,$$

où M_d est la copule définie à l'équation (1.5). Une démonstration de ce résultat se trouve dans Nelsen (2006, page 47). Dans le cas bivarié ($d = 2$), W_2 est la copule associée au vecteur aléatoire $(U, 1-U)$, où U suit une loi uniforme sur $[0,1]$. W_2 s'appelle la copule borne inférieure de Fréchet.

1.3 Quelques familles de copules paramétriques

Il existe plusieurs familles de copules qui permettent de modéliser la dépendance. Dans les sous-sections qui suivent, nous parcourons les copules archimédiennes, les copules elliptiques et une copule asymétrique.

1.3.1 Copules archimédiennes

La copule archimédienne est définie à partir de fonctions dites complètement monotones.

Définition 1.3. ***Fonction complètement monotone***

Une fonction ψ est dite complètement monotone si elle est infiniment différentiable sur $[0, +\infty[$ et

$$(-1)^j \psi^{(j)}(t) \geq 0, \quad j \in \mathbb{N}^*,$$

où $\psi^{(j)}$ est la dérivée d'ordre j de ψ.

Exemple 1.4. *La transformée de Laplace-Stieltjes, $\psi(t) = \mathbb{E}(e^{-Zt})$, d'une variable aléatoire strictement positive Z, est une fonction complètement monotone, voir Feller (1972, page 449).*

Remarque 1.2. *Si ψ_θ est une fonction complètement monotone alors la fonction C définie par*

$$C(u_1, \ldots, u_d; \theta) = \psi_\theta\left\{\psi_\theta^{-1}(u_1) + \ldots + \psi_\theta^{-1}(u_d)\right\}, \;\; u_1, \ldots, u_d \in [0,1], \tag{1.11}$$

est une copule archimédienne de générateur ψ_θ^{-1} et de paramètre θ pour tout entier $d > 1$.

La densité de la copule archimédienne de l'équation (1.11) donne l'expression

$$c(u_1,\ldots,u_d;\theta)=\frac{\psi_\theta^{(d)}\left\{\psi_\theta^{-1}(u_1)+\ldots+\psi_\theta^{-1}(u_d)\right\}}{\prod_{i=1}^{d}\psi_\theta'\left\{\psi_\theta^{-1}(u_i)\right\}}, \tag{1.12}$$

voir McNeil et Nêslehovà (2009), page 3075.

Les copules archimédiennes vérifient la définition d'échangeabilité unidimensionnelle c'est-à-dire pour toute permutation $\{\pi(1),\ldots,\pi(d)\}$ de $\{1,\ldots,d\}$, nous avons

$$C(u_1,\ldots,u_d;\theta)=C\left(u_{\pi(1)},\ldots,u_{\pi(d)};\theta\right),$$

voir (Mai et Scherer, 2012, page 39). La notion d'échangeabilité multidimensionnelle est formellement définie au chapitre 2, section 2.1. Vu l'existence d'une multitude de familles de copules archimédiennes, l'utilisation dépend de la situation à laquelle nous sommes confrontés. Nous donnons un exemple de copule archimédienne, la copule de Frank.

Exemple 1.5. ***Copule de Frank***

La copule de Frank de paramètre θ en dimension d est

$$C_\theta(u_1,\ldots,u_d;\theta)=-\frac{1}{\theta}\log\left\{1+\frac{\prod_{i=1}^{d}(e^{-\theta u_i}-1)}{(e^{-\theta}-1)^{d-1}}\right\},\ \ \theta>0\cdot \tag{1.13}$$

C'est une copule archimédienne de générateur $\psi_\theta(t)=-\log\left\{1+e^{-t}(e^{-\theta}-1)\right\}/\theta$, $t\in\mathbb{R}_+$. En effet, ψ_θ est la transformée de Laplace d'une variable aléatoire logarithmique de paramètre $(1-e^{-\theta})$. On rappelle qu'une variable aléatoire Z suit une loi logarithmique de paramètre $(1-e^{-\theta})$ si sa fonction de probabilité est

$$P(Z=k)=\frac{\left(1-e^{-\theta}\right)^k}{k\theta},\ \ k=1,2,\ldots. \tag{1.14}$$

À partir de la copule de Frank, on retrouve la copule d'indépendance lorsque $\theta\longrightarrow 0$, voir (1.4). Lorsque $\theta\longrightarrow\infty$, la copule de Frank donne la copule comonotone, voir (1.5).

1.3.2 Copules elliptiques

Les copules elliptiques sont définies à partir de lois elliptiques. La majorité des copules elliptiques n'ont pas une forme explicite à cause de l'écriture sous forme intégrale de la

distribution multidimensionnelle dont elles proviennent. Nous donnons une caractérisation de la loi elliptique. Nous supposons dans cette sous-section que $\mathbf{Z}$ est un vecteur aléatoire de dimension d.

Définition 1.4. ***Caractérisation d'une loi elliptique***

On dit que le vecteur aléatoire $\boldsymbol{Z}$ suit une distribution elliptique centrée à μ et de matrice de dispersion Σ, qu'on note $\boldsymbol{Z} \sim \mathcal{E}_d(\mu, \Sigma)$, si et seulement si $\boldsymbol{Z}$ peut s'écrire en termes de distribution en loi sous la forme

$$\boldsymbol{Z} \overset{\mathcal{L}}{=} \mu + dRAU, \tag{1.15}$$

où

- *la matrice A provient de la décomposition de Cholesky de Σ ($\Sigma = AA^T$, A est une matrice triangulaire inférieure) ;*
- *R est une variable aléatoire positive et*
- *U est un vecteur aléatoire uniformément distribué sur l'hypersphère unité de $\mathbb{R}^d$ et indépendant de R.*

Sous l'hypothèse que Σ est de rang plein, la densité de R est

$$f(r) = \frac{(2\pi)^{-d/2}}{\Gamma(d/2)} r^{d-1} g(r^2), \quad r \in \mathbb{R}_+,$$

et la densité de $\boldsymbol{Z}$ est

$$h_g(\boldsymbol{z}) = |\Sigma|^{-1/2} g\left\{(\boldsymbol{z}-\mu)^T \Sigma^{-1} (\boldsymbol{z}-\mu)\right\}, \quad \boldsymbol{z} \in \mathbb{R}^d,$$

où la fonction Γ est définie par $\Gamma(t) = \int_0^\infty y^{t-1} \exp(-y) dy$ et $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ est appelée la fonction générateur.

Cette définition provient des travaux de Frahm *et al.* (2003). Une loi elliptique multivariée a ses marginales elliptiques de même loi. La copule associée à une loi elliptique est une copule elliptique, voir Mai et Scherer (2012, page 171). Nous donnons deux exemples de copule usuelle de loi elliptique en dimension d.

Exemple 1.6. ***Copule normale***

Si $\boldsymbol{Z}$ est un vecteur aléatoire gaussien de dimension d, centré et de matrice de corrélation P, alors la copule associée est

$$C_P(u_1, \ldots, u_d) = \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-d/2} |P|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{z}^T P^{-1} \boldsymbol{z}\right) dz,$$

où Φ est la fonction de répartition de la loi normale standard. La densité de la copule associée en dimension d est

$$c_(u_1, \ldots, u_d; P) = |P|^{-1/2} \exp\left\{-\frac{1}{2} w^T (P^{-1} - I_d) w\right\},$$

où $w = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d))^T$, voir Xue-Kun Song (2000) et Φ^{-1} l'inverse de la fonction de répartition de la loi normale centrée réduite .

La copule normale est une copule symétrique (elle est égale à la copule de survie associée). Elle est égale à la copule d'indépendance si la matrice de corrélation est celle identité (toutes les corrélations sont nulles).

Exemple 1.7. ***Copule de Student***

La copule de Student s'écrit

$$C_(u_1, \ldots, u_d; P) = \int_{-\infty}^{F_{\nu,P}^{-1}(u_1)} \ldots \int_{-\infty}^{F_{\nu,P}^{-1}(u_d)} f(\boldsymbol{z}; \nu, P) d\boldsymbol{z},$$

où $u_1, \ldots, u_d \in [0,1]$ et $F_{\nu,P}$ est la fonction de répartition d'une loi de Student à ν degrés de liberté, de matrice de corrélation P et $F_{\nu,P}^{-1}$ est la fonction inverse. La densité d'une loi de Student multivariée en dimension d, à $\nu > 2$ degrés de liberté, de matrice de corrélation P, inversible, est

$$f(\boldsymbol{z}; \nu, P) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{d/2}|P|^{1/2}} \left(1 + \frac{\boldsymbol{z}^T P^{-1} \boldsymbol{z}}{\nu}\right)^{-(\nu+d)/2}, \quad \boldsymbol{z} \in \mathbb{R}^d.$$

Cette copule permet de capter de la dépendance aux extrêmes, positive ou négative, voir la section 1.4.4.

1.3.3 La copule bêta généralisée

Nous donnons quelques définitions concernant les variables aléatoires suivant des lois gamma, bêta et MGB2 dans un premier temps avant de définir la copule bêta dans un deuxième temps.

Définition 1.5. ***Distribution gamma***

Une variable aléatoire Y suit une loi gamma de paramètres a et λ, positifs, que l'on note $Y \sim \mathcal{G}(a, \lambda)$ si sa densité de probabilité f s'exprime

$$f(y; a, \lambda) = \frac{\lambda^a}{\Gamma(a)} e^{-\lambda y} y^{a-1}, \quad y \in \mathbb{R}_+.$$

Si $\lambda = 1$, on parle de loi gamma de paramètre a noté $\mathcal{G}(a)$ et si $a = 1$, on parle de loi exponentielle de paramètre λ.

Définition 1.6. ***Distribution bêta***

Une variable aléatoire Y suit une loi bêta de paramètres α et β, positifs et notée $Y \sim \mathcal{B}(\alpha, \beta)$ si sa densité de probabilité g est

$$g(y; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1\cdot \tag{1.16}$$

Une propriété connue sur la construction la somme des lois gamma.

Propriété 1.1. *Si Y_1 est une variable aléatoire suivant une loi gamma de paramètre α, Y_2 une autre variable aléatoire suivant loi gamma de paramètre β, indépendante de Y_1 alors la variable aléatoire $Y_1/(Y_1 + Y_2)$ suit une loi bêta de paramètres α et β. Les variables aléatoires $Y_1/(Y_1 + Y_2)$ et $Y_1 + Y_2$ sont indépendantes et $Y_1 + Y_2$ suit une loi gamma de paramètre $\alpha + \beta$.*

Définition 1.7. ***Distribution MGB2***

On considère $a = (a_1, \ldots, a_d)$, $b = (b_1, \ldots, b_d)$ et $p = (p_1, \ldots, p_d)$ où b_i et p_i sont positifs pour tout i. Un vecteur aléatoire $\boldsymbol{Z}$ de dimension d suit une loi bêta généralisée multivariée de deuxième espèce, de paramètres a, b, p et q noté $MGB2(a, b, p, q)$ si et seulement si sa fonction de densité h est

$$h(\boldsymbol{z}; a, b, p, q) = \frac{\Gamma(\tilde{p} + q) \prod_{i=1}^{d} |a_i| \boldsymbol{z}_i^{a_i p_i - 1}}{\Gamma(p_1) \ldots \Gamma(p_d)\Gamma(q) \prod_{i=1}^{d} b_i^{a_i p_i} \left\{ 1 + \sum_{i=1}^{d} \left(\frac{z_i}{b_i} \right)^{a_i} \right\}^{\tilde{p}+q}}, \tag{1.17}$$

où $\boldsymbol{z} = (z_1, \ldots, z_d)^T \in \mathbb{R}_+^d$ et q, positif et $\tilde{p} = p_1 + \ldots + p_d$, voir Cockriel et McDonald (2018) pour plus de détail sur cette loi multivariée.

Lorsque $d = 1$ et $a = b = (1, \ldots, 1)$, on parle de loi GB2 de paramètres p et q.

Remarque 1.3. *Si Y_1 et Y_2 sont deux variables aléatoires indépendantes suivant respectivement la loi gamma de paramètres p et q, positifs, alors la variable aléatoire Y_1/Y_2 suit une loi GB2 unidimensionnelle de paramètres p et q.*

Nous définissons par la suite la copule bêta à partir de la distribution MGB2.

Définition 1.8. ***Copule bêta***

La densité d'une copule bêta en dimension d, de paramètres $p_1, \ldots, p_d, q$, positifs, s'écrit

$$c(u_1, \ldots, u_d; p_1, \ldots, p_d, q) = \frac{\Gamma(q)^{d-1}\Gamma\left(\sum_{i=1}^{d} p_i + q\right)}{\prod_{i=1}^{d}\Gamma\left(p_i + q\right)} \cdot \frac{\prod_{i=1}^{d}(1+x_i)^{p_i+q}}{\left(1+\sum_{i=1}^{d} x_i\right)^{\sum_{i=1}^{d} p_i + q}}, \tag{1.18}$$

où x_i est défini par

$$x_i = \frac{B_{p_i,q}^{-1}(u_i)}{1 - B_{p_i,q}^{-1}(u_i)},$$

où $B_{p_i,q}$ est la fonction de répartition d'une loi bêta de paramètres p_i et q, voir Yang et al. (2011) pour des informations complémentaires sur la copule bêta.

Cette copule est obtenue à partir de la densité MGB2($\mathbf{1}$,$\mathbf{1}$,p,q) où $p = (p_1, \ldots, p_d)$ de l'équation (1.17). Cette copule permet de modéliser une dépendance asymétrique. Si $p_1 = p_2 = \ldots = p_d = 1$, alors la copule bêta est la copule de survie associée à une copule de Clayton en dimension d de paramètre q, voir Yang *et al.* (2011).

La propriété suivante permet de rendre compte de la généralisation de la copule bêta par rapport à la copule normale ordinaire. Les résultats particuliers sur la copule bêta sont obtenus lorsque la dimensionnalité est 2.

Propriété 1.2. *Soit p_1, p_2 et q des paramètres réels positifs. Considérons la copule bêta bivariée de paramètres p_1, p_2 et q.*

1) Pour p_1 et p_2, fixés, nous avons le résultat suivant

$$\lim_{q \longrightarrow \infty} C(u_1, u_2; p_1, p_2, q) = u_1 u_2. \tag{1.19}$$

2) Si p_1, p_2 et q sont très grands de telles sortes que $p_1/q \longrightarrow c_1$ et $p_2/q \longrightarrow c_2$ lorsque p tend vers l'infini alors

$$\lim_{q \longrightarrow \infty} C(u_1, u_2; p_1, p_2, q) = \Phi_\rho\left\{\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right\}, \tag{1.20}$$

où le paramètre est $\rho = \sqrt{\frac{c_1 c_2}{(1+c_1)(1+c_2)}}$.

Une démonstration de cette propriété se trouve en annexe de l'article de Yang *et al.* (2011).

1.3.4 Copule de Khoudraji bivariée

Définition 1.9. *Une fonction C est une copule de `Khoudraji` bivariée si elle peut s'écrire sous la forme*

$$C(u_1, u_2; \kappa_1, \kappa_2) = C_1\left(u_1^{1-\kappa_1}, u_2^{1-\kappa_2}\right) C_2\left(u_1^{\kappa_1}, u_2^{\kappa_2}\right), \quad \kappa_1, \kappa_2 \in [0,1], \tag{1.21}$$

pour $u_1, u_2 \in [0,1]$ et les fonctions C_1 et C_2 sont des copules bivariées, voir Mukherjee et al. (2018).

- Si $\kappa_1 \neq \kappa_2$, alors la copule C est asymétrique.

- Si C_1 est une copule d'indépendance et C_2 est une copule symétrique de paramètre κ alors la copule C est de paramètres $(\kappa, \kappa_1, \kappa_2)$.

Nous avons présenté les notions générales sur les copules. Nous traitons maintenant des outils qui permettent de quantifier la structure de dépendance entre deux ou plusieurs variables aléatoires. La corrélation linéaire par exemple permet de donner une idée de la dépendance linéaire entre deux variables aléatoires. D'autres types de corrélations bivariées existent et peuvent s'exprimer facilement en fonction des copules. Dans la suite de ce chapitre, nous allons présenter quelques mesures de corrélations.

1.4 Relations de dépendance entre variables

Les corrélations constituent des mesures pour déterminer l'importance des liens entre variables. On dispose de plusieurs mesures de corrélation dont la mesure de Pearson, de Kendall et Spearman. Nous présentons les deux dernières et une autre mesure de dépendance appelée dépendance caudale.

1.4.1 Corrélation de Spearman

Le coefficient de corrélation de Spearman noté ρ_S est une mesure qui permet de quantifier le degré d'association entre deux variables aléatoires. Pour deux variables aléatoires X_1 et X_2, de loi bivariée continue et de marges F_1 et F_2, la corrélation de Spearman est définie par :

$$\rho_S(X_1, X_2) = \frac{\mathbb{E}\left\{F_1(X_1)F_2(X_2)\right\} - \mathbb{E}\left\{F_1(X_1)\right\}\mathbb{E}\left\{F_2(X_2)\right\}}{\sqrt{\mathbb{V}\left\{F_1(X_1)\right\}\mathbb{V}\left\{F_2(X_2)\right\}}}.$$

La corrélation de Spearman peut être estimée à partir des données, voir Joe (2014, page 56). Il existe une relation entre la corrélation de Spearman d'un vecteur aléatoire bivarié et la copule qui lui est associée.

Théorème 1.4. *Si C est la copule associée au vecteur aléatoire (X_1, X_2), alors*

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3\cdot$$

Voir Cherubini et al. *(2014, page 97) pour plus de détail.*

1.4.2 Le tau de Kendall

Le tau de Kendall se base sur la définition 1.10 portant sur les notions de concordance ou de discordance.

Définition 1.10. *Soit (x_1, y_1) et (x_2, y_2), deux points de $\mathbb{R}^2$.*

- *On dit que les couples (x_1, y_1) et (x_2, y_2) sont concordants si $(x_1 - x_2)(y_1 - y_2) > 0$.*
- *On dit que les couples (x_1, y_1) et (x_2, y_2) sont discordants si $(x_1 - x_2)(y_1 - y_2) < 0$.*

La Figure 1.1 donne une illustration de la définition de points concordants ou discordants.

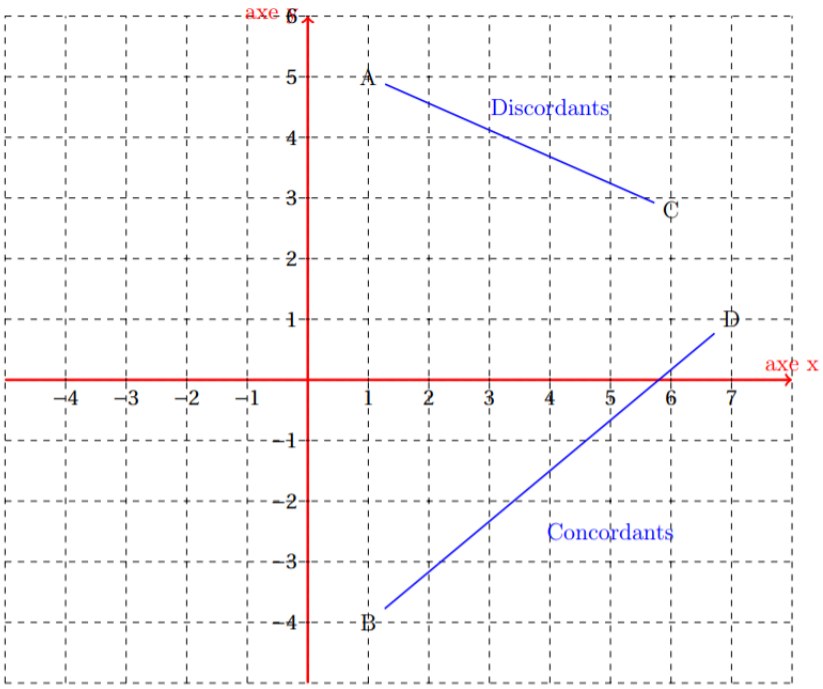


FIGURE 1.1 – Illustration de la définition 1.10 : $A(1,5)$ et $C(6,2.8)$ sont discordants ; $B(1,-4)$ et $D(7,1)$ sont concordants.

Définition 1.11. *On définit le tau de Kendall, noté τ, par la différence entre la probabilité d'observations concordantes et la probabilité d'observations discordantes. Soit (X_1, Y_1) et (X_2, Y_2), deux vecteurs aléatoires continus, indépendants et distribués comme le vecteur aléatoire (X, Y). Le tau de Kendall du vecteur aléatoire (X, Y) s'exprime par*

$$\tau = \mathbb{P}\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - \mathbb{P}\{(X_1 - X_2)(Y_1 - Y_2) < 0\}.$$

En utilisant les données sur n points, l'estimation du tau de Kendall τ est

$$\hat{\tau} = \frac{\text{(nombre de paires concordantes)-(nombre de paires disconcordantes)}}{1/2 \cdot n \cdot (n-1)}. \quad (1.22)$$

Si C est la copule associée au vecteur aléatoire (X, Y), alors le tau de Kendall se calcule

$$\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 = 4\mathbb{E}\{C(U, V)\} - 1,$$

où (U, V) est un vecteur aléatoire de loi C. Voir Cherubini *et al.* (2014, page 97-98) pour de plus amples explications. Le tau de Kendall échangeable est défini dans le travail de Romdhani *et al.* (2014a). C'est une version modifiée du tau de Kendall pour spécifiquement les données échangeables en grappes. Le cas particulier des copules archimédiennes se donne dans le théorème 1.5.

Théorème 1.5. *Soit C, une copule archimédienne de générateur ψ, alors le tau de Kendall associé à cette copule est*

$$\tau = 1 + 4\int_0^1 \frac{\psi^{-1}(t)}{(\psi^{-1})'(t)}dt, \tag{1.23}$$

voir Genest et Rivest (1993) pour une démonstration du résultat.

Nous présentons, au tableau 1.2, quelques copules avec leurs tau de Kendall et de Spearman. Ces résultats proviennent des références comme Nelsen (2006, page 163), Shemyakin et Kniazev (2017, page 244).

TABLEAU 1.2 – Corrélation de Spearman et Kendall pour trois copules avec les paramètres entre parenthèses.

Copule	Domaine du paramètre	Tau de Spearman	Tau de Kendall
Clayton (θ)	$[0, +\infty[$	-	$\theta/(\theta+2)$
Gumbel (α)	$[1, +\infty[$	-	$1 - 1/\alpha$
Normale (ρ)	$[-1, 1]$	$(6/\pi)\arcsin(\rho/2)$	$(2/\pi)\arcsin(\rho)$

1.4.3 Corrélation des scores normaux

Elle se définit à partir de la corrélation de Pearson en appliquant la loi normale standard aux marges de deux variables aléatoires. Nous donnons une définition de la corrélation de score normal.

Définition 1.12. *Pour deux variables aléatoires X_1 et X_2 dont les fonctions de répartitions sont F_1 et F_2 respectivement, on définit le score normal ρ_N entre X_1 et X_2 par*

$$\rho_N = cor\left[\Phi^{-1}\left\{F_1(X_1)\right\}, \Phi^{-1}\left\{F_2(X_2)\right\}\right],$$

où cor est la corrélation de Pearson. Cette formulation de la corrélation de score normal est due à Joe (2014, page 58).

En notant c, la fonction de densité de la copule associée au vecteur aléatoire (X_1, X_2), nous avons

$$\rho_N = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t_1 t_2 \phi(t_1)\phi(t_1) c\left\{\Phi(t_1), \Phi(t_2)\right\} dt_1 dt_2.$$

Une autre forme de dépendance existe, relativement aux queues supérieure ou inférieure des distributions. Nous la présentons dans la section suivante.

1.4.4 Dépendance caudale d'un vecteur bivarié

Nous donnons une définition de la dépendance caudale de manière formelle.

Définition 1.13. *La dépendance caudale est une notion locale qui apporte une description de la dépendance au niveau des queues de la distribution d'un vecteur aléatoire bivarié. Cette dépendance est quantifiée dans les queues inférieures et supérieures. Le texte de Joe (2014, page 62) explique clairement la notion de dépendance caudale et y donne des propriétés.*

Nous notons λ_U et λ_L, les dépendances supérieure et inférieure respectivement d'un vecteur aléatoire (X_1, X_2). En considérant C, la copule associée au vecteur (X_1, X_2), nous définissons ces coefficients par

$$\lambda_U = \lim_{u \longrightarrow 1^-} \left\{ \frac{1 - 2u + C(u,u)}{1-u} \right\}, \quad \lambda_L = \lim_{u \longrightarrow 0^+} \left\{ \frac{C(u,u)}{u} \right\}. \tag{1.24}$$

Ils sont utiles pour étudier la dépendance de valeurs extrêmes (proche de 0 ou de 1) des distributions.

1.5 Ajustement d'un modèle de copule paramétrique

Pour ajuster un modèle de copule paramétrique C_θ, de paramètre θ, associée à un vecteur aléatoire de dimension d, $\mathbf{Z} = (X^1, \ldots, X^{d-1}, Y)^T$, on fait généralement une hypothèse sur les lois marginales. Nous supposons que le vecteur $\mathbf{Z}$ est observé sur un échantillon de n unités et noté $\mathbf{z}_i = (x_i^1, \ldots, x_i^{d-1}, y_i)^T, i = 1, \ldots, n$. Dans le cas où les fonctions de répartitions sont respectivement des lois G pour Y, $F_1, \ldots, F_{d-1}$ pour $\mathbf{X}$, de paramètres $\beta, \alpha_1, \ldots, \alpha_{d-1}$, alors l'ajustement est paramétrique. Dans le cas contraire, les marginales sont estimées de manière par la fonction de répartition empirique et le modèle considéré est semi-paramétrique, voir Joe (2014, page 17). Deux méthodes d'estimation des paramètres existent : la méthode du maximum de vraisemblance globale et la méthode IFM (*Inference Function for Margins*) proposée par Joe et Xu (1996). Nous notons $(\beta, \alpha_1, \ldots, \alpha_{d-1}, \theta)$, les paramètres du modèle.

La méthode du maximum de vraisemblance globale : Elle consiste à maximiser la vraisemblance globale dans le cas paramétrique ou semi-paramétrique.

Dans le cas d'un modèle paramétrique, la log-vraisemblance globale est

$$\begin{aligned}\mathcal{L}(\beta, \alpha_1, \ldots, \alpha_{d-1}, \theta) &= \sum_{i=1}^{n}\sum_{j=1}^{d-1} \log\left\{f_j(x_i^j; \alpha_j)\right\} + \sum_{i=1}^{n} \log\left\{g(y_i; \beta)\right\} \\ &+ \sum_{i=1}^{n} \log\left[c_\theta\left\{G(y_i;\beta), F_1(x_i^1;\alpha_1), \ldots, F_{d-1}(x_i^{d-1};\alpha_{d-1}); \theta\right\}\right]. \end{aligned} \tag{1.25}$$

Les estimateurs obtenus en maximisant (1.25) sont convergents, asymptotiquement efficaces et normaux.

Dans le cas d'un modèle semi-paramétrique, les marginales G et $F_j, j = 1, \ldots, d-1$ de l'équation (1.25) sont estimées de manière non paramétrique par

$$\hat{F}_j(x) = n^{-1}\sum_{i=1}^{n} I(x_i^j < x),\ \hat{G}(y) = n^{-1}\sum_{i=1}^{n} I(y_i < y),$$

alors la pseudo log-vraisemblance à maximiser est

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log\left[c_\theta\left\{\hat{G}(y_i), \hat{F}_1(x_i^1), \ldots, \hat{F}_{d-1}(x_i^{d-1}); \theta\right\}\right]. \tag{1.26}$$

L'ajustement d'une copule en utilisant la méthode semi-paramétrique est expliqué aussi dans Joe (2014, page 247). Les estimateurs obtenus en maximisant (1.26) sont convergents et normaux, voir Genest et Rivest (1993) et Chen et Fan (2006).

Une autre méthode d'ajustement des copules existe et s'appelle *Inference Function for Margins* (IFM).

La méthode IFM : la méthode IFM proposée par Joe et Xu (1996) se déroule en deux étapes.

- La première étape consiste à estimer les paramètres β et α_j des lois G et F_j respectivement par $\tilde{\beta}$ et $\tilde{\alpha}_j$, $j = 1, \ldots, d-1$ en maximisant les log-vraisemblance marginales. Les estimations sont

$$\tilde{\beta} = \underset{\beta}{\text{argmax}}\left[\sum_{i=1}^{n} \log\left\{g(y_i;\beta)\right\}\right],\ \tilde{\alpha}_j = \underset{\alpha_j}{\text{argmax}}\left[\sum_{i=1}^{n} \log\left\{f_j(x_i^j;\alpha_j)\right\}\right].$$

- La deuxième étape consiste à estimer le paramètre θ de la copule par $\tilde{\theta}$ en maximisant la pseudo log-vraisemblance et

$$\tilde{\theta} = \underset{\theta}{\text{argmax}}\left[\sum_{i=1}^{n} \log\left[c_\theta\left\{g(y_i;\tilde{\beta}), F_1(x_i^1;\tilde{\alpha}_1), \ldots, F_{d-1}(x_i^{d-1};\tilde{\alpha}_{d-1}); \theta\right\}\right]\right].$$

Les estimateurs obtenus par cette méthode IFM, sous les conditions de régularité, sont convergents et asymptotiquement normaux, Joe (1997) et Joe (2005).

Notre thèse traite des données présentant une dépendance entre les observations dans une grappe et une dépendance entre les variables d'un même individu. Les méthodes expliquées précédemment sont étudiées pour s'adapter à cette spécificité. Ainsi, dans le chapitre 3, après avoir expliqué en détail les différentes étapes d'ajustement des copules du modèle postulé au chapitre 2 et les méthodes d'estimation des paramètres, nous étudions les propriétés asymptotiques des estimateurs dans le cas où $d = 2$. Pour la suite du chapitre 1, nous présentons la régression avec les modèles de copules.

1.6 Régression avec les modèles de copules

La régression avec les modèles de copules est discutée dans Crane et Van der Hoek (2008) et mise en œuvre aussi par Das (2015) pour faire de la prédiction. Nous nous mettons dans le cas où nous avons une variable dépendante Y continue sur $\mathbb{R}$ et des covariables $\mathbf{X} = (X^1, X^2, \ldots, X^{d-1})^T$, continues. Dans cette section, nous voulons évaluer $\mathbb{E}(Y|\mathbf{X})$. La méthode de régression avec les copules cherche à prédire Y sachant que $\mathbf{X}$ à l'aide de l'espérance conditionnelle $\mathbb{E}(Y|\mathbf{X})$ en utilisant les copules. Elle s'écrit à l'aide de la copule et des lois marginales déterminant la distribution conjointe de $(\mathbf{X}, Y)$. Ceci demande théoriquement, la connaissance de la loi conditionnelle de Y sachant $\mathbf{X}$. Cette loi conditionnelle s'écrit en fonction de la copule pour $(\mathbf{X}, Y)$ et des lois marginales pour $\mathbf{X}$ et Y.

1.6.1 Expression théorique du prédicteur de Y sachant $\mathbf{X}$ avec la copule

Nous supposons que $C(.;\theta)$ est la copule de paramètres θ associée au vecteur aléatoire $(\mathbf{X}, Y)$, de dimension d dont les marges sont G pour Y et $F = (F_1, \ldots, F_{d-1})^T$ pour $\mathbf{X}$, où $F_j, j = 1, \ldots, d-1$ est la fonction de répartition associée à la variable aléatoire X_j. La densité de probabilité $f_{Y|\mathbf{x}}$, associée à la variable aléatoire Y conditionnellement à $\mathbf{x} = (x^1, \ldots, x^{d-1})^T$ est

$$f_{Y|\mathbf{x}}(y) = g(y)\frac{c\left\{G(y), F_1(x^1), \ldots, F_{d-1}(x^{d-1}); \theta\right\}}{c_{\mathbf{X}}\left\{F_1(x^1), \ldots, F_{d-1}(x^{d-1})\right\}},$$

où $c_{\mathbf{X}}$ est la densité de la copule associée à $\mathbf{X}$. Les densités c_θ et $c_{\mathbf{X}}$ sont

$$c(u_0, \ldots, u_{d-1}; \theta) = \frac{\partial^d C(u_0, u_1, \ldots, u_{d-1}; \theta)}{\partial u_0 \ldots \partial u_{d-1}}, \quad u_0, \ldots, u_{d-1} \in [0, 1],$$

et

$$c_{\mathbf{X}}(u_1,\ldots,u_{d-1}) = \frac{\partial^{d-1}C(1,u_1,\ldots,u_{d-1};\theta)}{\partial u_1\ldots\partial u_{d-1}}.$$

L'espérance conditionnelle de Y sachant $\mathbf{x}$ donne

$$\mathbb{E}(Y|\mathbf{X}=\mathbf{x}) = \frac{\int_{\mathbb{R}} yg(y)c\left\{G(y),F_1(x^1),\ldots,F_{d-1}(x^{d-1});\theta\right\}dy}{c_{\mathbf{X}}\left\{F_1(x^1),\ldots,F_{d-1}(x^{d-1})\right\}}, \tag{1.27}$$

En faisant un changement de variable, l'équation (1.27) peut s'écrire d'une autre façon par :

$$\mathbb{E}(Y|\mathbf{X}=\mathbf{x}) = \frac{\int_0^1 G^{-1}(u_0)c\left\{u_0,F_1(x^1),\ldots,F_{d-1}(x^{d-1});\theta\right\}du_0}{c_{\mathbf{X}}\left\{F_1(x^1),\ldots,F_{d-1}(x^{d-1})\right\}}, \tag{1.28}$$

Voir Noh *et al.* (2013), Leong et Valdez (2005) et Schucany *et al.* (1978) pour une formule similaire. Une évaluation de tel prédicteur est aussi présentée dans l'article de Acar *et al.* (2019) où les auteurs commentent les différences entre la prédiction avec la régression linéaire. De même, cette évaluation de prédicteur a fait l'objet de l'article de Bernard et Czado (2015) où ils présentent la régression quantile avec les copules conditionnelles. Cependant, la méthode de Acar *et al.* (2019) présentent des limites dû au fait que les lois marginales sont estimées de manière non paramétrique et elle se limite simplement aux variables continues. Nous donnons deux cas de régression par copule en dimension 2 dans l'exemple qui suit.

Exemple 1.8. ***En dimension*** $d=2$***, on détermine l'expression théorique du prédicteur dans deux cas spécifiques***

- ***La copule de Clayton de paramètre*** θ_1 ***est associée à*** (X_1,Y) ***et*** $Y\sim\mathcal{N}(\mu,\sigma^2)$

  L'espérance conditionnelle de Y *sachant* X_1 *est donc*

  $$\mathbb{E}(Y|X_1) = \mu+\sigma\mathbb{E}\left\{\Phi^{-1}(T^{-1/\theta_1})\right\}, \tag{1.29}$$

  où la variable aléatoire T *est de densité* f_T *donnée par*

  $$f_T(t) = \frac{(1/\theta_1+1)\left\{F_1(X_1)\right\}^{-(\theta_1+1)}}{(t+F_1(X_1)^{-\theta_1}-1)^{1/\theta_1+2}},\ \ t>1\cdot$$

  Nous rappelons que f_T *est la densité de* $C_{1|2}(v|u)$ *dans le tableau 1.3 pour la copule de Clayton.*

- ***La copule FGM de paramètre*** θ_2 ***est associée à*** (X_1,Y) ***et*** $Y\sim\mathcal{N}(\mu,\sigma^2)$

  L'espérance conditionnelle de Y *sachant* X_1 *est*

  $$\mathbb{E}(Y|X_1) = \mu+\sigma\frac{\theta_2}{\sqrt{\pi}}\left\{2F_1(X_1)-1\right\}\cdot \tag{1.30}$$

Le résultat de l'équation (1.30) *est obtenu aussi dans Hoang* et al. *(2019). Les développements pour obtenir les résultats des équations* (1.29) *et* (1.30) *se trouvent en annexe A, section A.2.*

L'exemple 1.8 donne une idée des expressions théoriques qu'on obtient avec deux copules usuelles dans le cas bivarié en utilisant la méthode de régression avec une copule. L'identification de la copule associée au vecteur aléatoire $(\mathbf{X}, Y)$ constitue une étape importante pour avoir le meilleur prédicteur en utilisant la méthode de régression avec les copules. La section 1.5 donne un aperçu de l'ajustement d'un modèle de copules aux données.

Nous avons, dans les sections antérieures, présenté les copules et quelques types de dépendance et vu la difficulté à trouver une distribution multivariée associée à un vecteur aléatoire, d'autres alternatives s'offrent à nous. Nous présentons donc une construction de copules multivariées à partir des vignes.

1.7 Construction des copules en vignes

Les vignes sont utilisées pour construire des distributions multivariées à l'aide d'une série de copules bivariées. Joe (2014) présente la méthode, Bedford et Cooke (2002) ont été les premiers à proposer une théorie de construction qui est approfondie par Aas *et al.* (2009). La vigne est composée de plusieurs arbres et le type de décomposition dépend de la forme du premier arbre. Il existe plusieurs types de vignes parmi lesquelles nous avons la vigne D, la vigne R et la vigne C.

Pour faire de la modélisation par les vignes, il y a trois étapes : premièrement, il faut sélectionner le type de vigne en fonction du premier arbre. Deuxièmement, il faut sélectionner les copules bivariées impliquées et terminer par la troisième étape, en faisant une estimation des paramètres. Par exemple, Aas *et al.* (2009), précise que, pour une distribution multivariée en dimension d, il existe $d!$ manières différentes de la décomposer en R-vigne. Du fait de leur grande flexibilité dans la modélisation multidimensionnelle, bon nombre d'outils comme les algorithmes d'exploration des arbres sont utilisés pour en sélectionner la décomposition la plus "adaptée" pour une situation. Nous donnons un exemple de construction de copule en D-vigne pour quatre variables aléatoires. Nous sommes inspirés de la méthode de construction des vignes proposée dans le livre de Mai et Scherer (2012, page 185).

1.7.1 Construction d'une D-vigne en dimension 4

Nous considérons 4 variables aléatoires notées 1, 2, 3 et 4 et on souhaite construire une D-vigne associée à ces variables. Spécifiquement, nous cherchons la densité de la copule associée aux 4 variables aléatoires. La décomposition fait intervenir 3 arbres dont le premier arbre est la base de la décomposition.

Première étape de la décomposition (arbre I) Elle consiste à construire le premier arbre en ordonnant les variables aléatoires puis en joignant les plus proches voisines les unes aux autres. Les traits pleins représentent les copules et les ronds, les variables aléatoires. La figure 1.2 donne un aperçu du premier arbre.

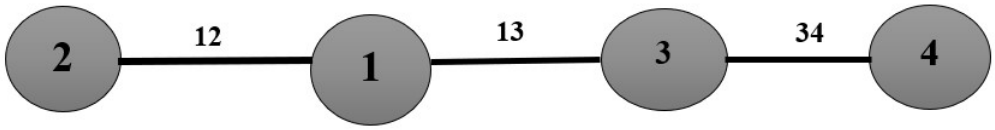


FIGURE 1.2 – Illustration graphique de la première ligne (arbre I) de la D-vigne de 4 variables.

Sur la figure 1.2, les copules bivariées intervenantes sont respectivement notées C_{12}, C_{13} et C_{34} associées aux vecteurs $(1,2)$, $(1,3)$ et $(3,4)$. De cette première étape, la contribution de l'arbre I à la densité de la copule multivariée est

$$c_{12}(u_1,u_2) \times c_{13}(u_1,u_3) \times c_{34}(u_3,u_4)\cdot$$

Deuxième étape de la décomposition (arbre II) Après la première étape, les binômes de variables aléatoires de la figure 1.2 sont les composantes du deuxième l'arbre. Elle consiste à joindre par la même occasion les plus proches voisines pour obtenir des copules conditionnelles se présentant sur la figure 1.3.

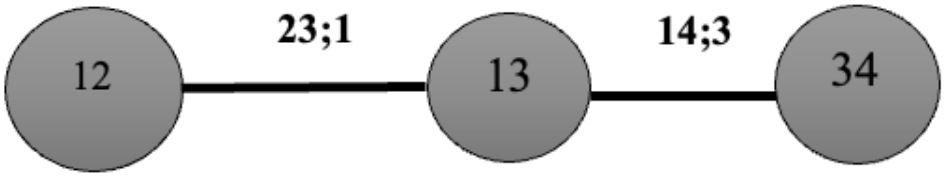


FIGURE 1.3 – Illustration graphique de la deuxième ligne (arbre II) de la D-vigne.

La copule de liaison entre les vecteurs binômes $(1,2)$ et $(1,3)$ est la copule conditionnelle $C_{23;1}$. Pour celle associée aux vecteurs $(3,4)$ et $(1,3)$ est $C_{14;3}$. La contribution de cet arbre à la densité multivariée est

$$c_{23;1}\left\{C_{2|1}(u_2|u_1), C_{3|1}(u_3|u_1)\right\} \times c_{14;3}\left\{C_{1|3}(u_1|u_3), C_{4|3}(u_4|u_3)\right\},$$

où les fonctions de distribution conditionnelles s'écrivent

$$C_{2|1}(u_2|u_1) = \int_0^{u_2} c_{12}(u_1, w)dw, \quad C_{3|1}(u_3|u_1) = \int_0^{u_3} c_{13}(u_1, w)dw, \tag{1.31}$$

et

$$C_{1|3}(u_1|u_3) = \int_0^{u_1} c_{13}(u_3, w)dw, \quad C_{4|3}(u_4|u_3) = \int_0^{u_4} c_{34}(u_3, w)dw. \tag{1.32}$$

Troisième étape de la décomposition (arbre III) La dernière étape de cette décomposition consiste à considérer les liens conditionnels de la deuxième étape comme les variables pour construire l'arbre 3.

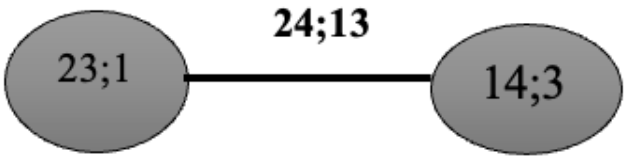


FIGURE 1.4 – Illustration graphique de l'arbre III de la vigne D.

La contribution en termes de densité de l'arbre III à la densité multivariée globale est

$$c_{24;13}\left\{C_{2|13}(u_2|u_1, u_3), C_{4|13}(u_4|u_1, u_3)\right\},$$

où

$$C_{2|13}(u_2|u_1, u_3) = \int_0^{C_{2|1}(u_2|u_1)} c_{23;1}\left\{w, C_{3|1}(u_3|u_1)\right\} dw, \tag{1.33}$$

et

$$C_{4|13}(u_4|u_1, u_3) = \int_0^{C_{4|3}(u_4|u_3)} c_{14;3}\left\{w, C_{1|3}(u_1|u_3)\right\} dw\cdot \tag{1.34}$$

Les distributions conditionnelles de chaque étape sont obtenues en utilisant un résultat de Joe (1996). Il provient aussi de la référence (Czado, 2019, chapitre 4, page 92).

À la fin de ces trois étapes, nous aboutissons à la décomposition globale en D-vigne que nous résumons sur un seul graphique.

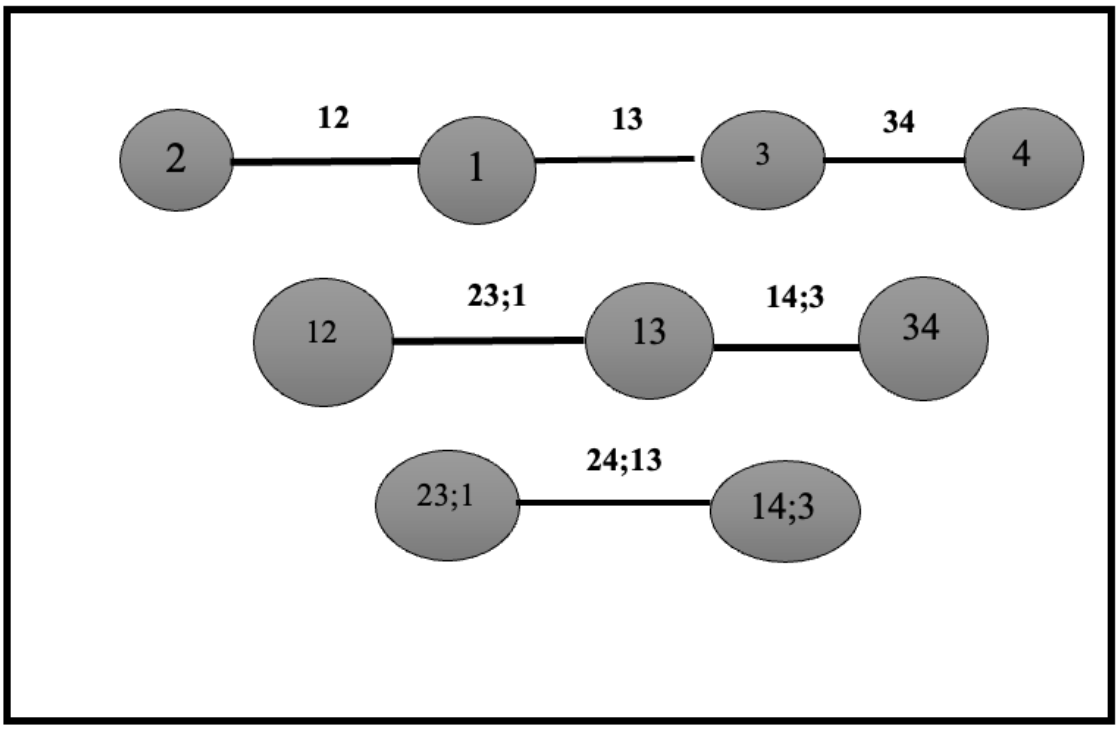


FIGURE 1.5 – Représentation graphique globale de la décomposition en D-vigne de 4 variables aléatoires.

Finalement, la densité de la copule associée à cette décomposition en D-vigne s'écrit

$$\begin{aligned} c(u_1,u_2,u_3,u_4) &= c_{12}(u_1,u_2)\times c_{13}(u_1,u_3)\times c_{34}(u_3,u_4)\times \\ &\quad c_{23;1}\left\{C_{2|1}(u_2|u_1),C_{3|1}(u_3|u_1)\right\}\times c_{14;3}\left\{C_{1|3}(u_1|u_3),C_{4|3}(u_4|u_3)\right\}\times \\ &\quad c_{24;13}\left\{C_{2|13}(u_2|u_1,u_3),C_{4|13}(u_4|u_1,u_3)\right\}, \end{aligned} \tag{1.35}$$

où $u_1,\ldots,u_4 \in [0,1]$. L'équation (1.35) fait intervenir 6 copules bivariées, dont les copules conditionnelles $C_{14;3}$, $C_{23;1}$ et $c_{24;13}$.

Dans la sous-section suivante, nous donnons, dans des cas particuliers, quelques résultats théoriques des fonctions de répartitions conditionnelles obtenues à partir des copules.

1.7.2 Quelques fonctions de répartitions conditionnelles univariées particulières obtenues à partir des copules

Dans les équations (1.31) et (1.32), les fonctions conditionnelles se calculent grâce à Joe (1996) et (Czado, 2019, chapitre 4, page 92). Nous présentons dans le tableau 1.3

la fonction de distribution conditionnelle $C_{2|1}(u_2|u_1)$ et son inverse $C_{2|1}^{-1}(w|u_1)$, $u_1, u_2 \in [0,1]$ et $w \in [0,1]$ pour quelques copules particulières dans le cas univarié.

TABLEAU 1.3 – Quelques fonctions de répartition conditionnelles $C_{2|1}$ et leurs inverses $C_{2|1}^{-1}$ obtenues à partir des formules analytiques des copules concernées.

Copule	Expression de la distribution conditionnelle et inverse
Normale $\rho \in (-1,1)$	$C_{2\|1}(u_2\|u_1) = \Phi\left\{\frac{\Phi^{-1}(u_2)-\rho\Phi^{-1}(u_1)}{\sqrt{1-\rho^2}}\right\}$ $C_{2\|1}^{-1}(w\|u_1) = \Phi\left\{\sqrt{1-\rho^2}\Phi^{-1}(w)+\rho\Phi^{-1}(u_1)\right\}$
FGM $\delta \in [-1,1]$	$C_{2\|1}(u_2\|u_1) = u_2\{1+\delta(1-u_2)(1-2u_1)\}$ $C_{2\|1}^{-1}(w\|u_1) = \frac{1+\delta(1-2u_1)-\sqrt{\delta(1-2u_1)\{\delta(1-2u_1-4w)+2\}+1}}{2\delta(1-2u_1)}$
Clayton $\theta > 0$	$C_{2\|1}(u_2\|u_1) = u_1^{-\theta-1}(u_1^{-\theta}+u_2^{-\theta}-1)^{-1/\theta-1}$ $C_{2\|1}^{-1}(w\|u_1) = \left[(w^{-\theta/(\theta+1)}-1)u_1^{-\theta}+1\right]^{-1/\theta}$
Frank $\delta \in \mathbb{R}\backslash\{0\}$	$C_{2\|1}(u_2\|u_1) = \frac{e^{-\delta u_1}(1-e^{-\delta u_2})}{-e^{-\delta}-e^{-\delta(u_1+u_2)}+e^{-\delta u_1}+e^{-\delta u_2}}$ $C_{2\|1}^{-1}(w\|u_1) = -1/\delta \exp\left\{1-\frac{w(1-e^{-\delta})}{e^{-\delta u_1}+w(1-e^{-\delta u_1})}\right\}$
Bêta $p_1, p_2, q > 0$	$C_{2\|1}(u_2\|u_1) = B_{p_2,p_1+q}\left\{\frac{B_{p_2,q}^{-1}(u_2)(1-B_{p_1,q}^{-1}(u_1))}{1-B_{p_1,q}^{-1}(u_1)B_{p_2,q}^{-1}(u_2)}\right\}$ $C_{2\|1}^{-1}(w\|u_1) = B_{p_2,q}\left\{\frac{B_{p_2,p_1+q}^{-1}(w)}{1-B_{p_1,q}^{-1}(u_1)+B_{p_1,q}^{-1}(u_1)B_{p_2,p_1+q}^{-1}(w)}\right\}$

Certains résultats du tableau 1.3 peuvent être retrouvés en utilisant la copule associée mais aussi ils proviennent des références comme Bernard et Czado (2015).

Sous le logiciel **R**, les formules utilisées pour calculer la distribution conditionnelle et son inverse à partir de certaines copules se trouvent dans la documentation du package `VineCopula`, voir Schepsmeier *et al.* (2018). Pour la copule bêta, la dérivation de la densité conditionnelle et de son inverse se trouve en annexe A, section A.1. Pour les fonctions de distributions conditionnelles des équations (1.33) et (1.34), elles peuvent s'écrire en fonction des fonctions de répartitions conditionnelles présentées au tableau 1.3.

1.7.3 Deux stratégies de construction des copules en vigne

Pour les décompositions en vigne, nous utilisons des stratégies de construction pour simplifier la décomposition. L'explication naturelle est que l'on souhaite que la dépendance soit relativement faible dans les branches de niveaux supérieurs (exemple arbre III de la figure 1.5) d'une part. D'autre part, dans le cadre d'une régression linéaire simple par exemple, la prédiction d'une variable sur un individu ne dépend pas de la valeur des autres individus. Nous explicitons ici les deux hypothèses standard que nous faisons sur les décompositions en vigne, voir Stöber *et al.* (2013).

• **Hypothèse simplificatrice** Elle est utilisée souvent en pratique et s'énonce par : "la copule conditionnelle ne dépend pas des variables sur lesquelles on conditionne".

• **Hypothèse d'indépendance** Lorsque nous tronquons une vigne (couper un arbre) par exemple en enlevant précisément le troisième arbre, cela revient à supposer que toutes les copules associées à cet arbre soient la copule d'indépendance.

Exemple 1.9. ***Application des hypothèses dans le cas de la décomposition en vigne de la sous-section 1.7.1***
L'hypothèse simplificatrice appliquée aux densités impliquées à la vigne de l'équation (1.35), s'écrit de la façon suivante

- *la copule conditionnelle $C_{23;1}$ ne varie pas en fonction de la variable 1 ;*
- *la copule conditionnelle $C_{14;3}$ ne varie pas en fonction de la variable 3 ;*
- *la copule conditionnelle $C_{24;13}$ ne varie pas en fonction des variables 1 et 3.*

Pour la décomposition en D-vigne de la sous-section 1.7.1 et présentée sur la figure 1.5, nous avions deux possibilités :

- *vigne tronquée au troisième arbre signifie que la copule $C_{24;13}$ est la copule d'indépendance ;*
- *vigne tronquée au deuxième arbre signifie que les copules $C_{23;1}$ et $C_{14;3}$ sont des copules d'indépendance.*

1.8 Modèle de Battese, Harter et Fuller (1988)

La régression ordinaire, en dimension d, étudie la liaison statistique entre une variable continue Y, unidimensionnelle, et des variables indépendantes $\mathbf{x}$. Ici, les unités sont regroupées en des grappes et la grappe peut contribuer à la prédiction de Y. Les modèles

mixtes ou modèles à effets aléatoires de Battese *et al.* (1988), développés aussi par Diggle *et al.* (1994) et McCulloch et Searle (2001) permettent de prendre en considération l'effet grappe dans la prédiction de Y. Dans cette section, nous définissons ce modèle clairement.

Pour définir le modèle linéaire mixte, nous considérons que la variable dépendante pour un individu i de la grappe j est noté Y_{ji} et appelons $\mathbf{x}_{ji}$, la variable indépendante associée. Nous considérons s_j, un échantillon pour la grappe j et de taille n_j. L'équation du modèle de régression s'écrit

$$Y_{ji} = \mathbf{x}_{ji}^T\beta + \nu_j + e_{ji}, \quad i = 1, \ldots, n_j, \ j = 1, \ldots, m, \tag{1.36}$$

où les hypothèses suivantes sont faites sur le modèle

$$\nu_j \sim \mathcal{N}(0, \sigma_\nu^2), \qquad e_{ji} \sim \mathcal{N}(0, \sigma_e^2),$$

ν_j est l'effet aléatoire pour la grappe j, e_{ji} est l'erreur expérimentale et ν_j est indépendant de e_{ji}. Le prédicteur de Y sur un nouvel individu de la grappe j ayant pour variable explicative $\mathbf{x}_{jN}$ est alors donné par

$$\hat{Y}_{jN} = \mathbf{x}_{jN}^T\beta + \frac{\sigma_\nu^2}{\sigma_e^2/n_j + \sigma_\nu^2} \sum_{i \in s_j} \left(\frac{Y_{ji} - \mathbf{x}_{ji}^T\beta}{n_j} \right),$$

sous l'hypothèse que les paramètres β, σ_ν et σ_e sont connus. Ce modèle permet de faire de la prédiction au niveau des individus dans une grappe. Pour l'analyse des données en grappes, la méthode de Battese, Harter et Fuller est la méthode standard utilisée, voir Battese *et al.* (1988).

Dans ce chapitre descriptif et de présentation, nous avons fait un bref survol des différents outils et éléments généraux existant sur les copules. Ces outils trouvent leur utilisation dans cette thèse où la construction de distribution jointe nécessite les copules d'une part et d'autre part des distributions marginales. Nous donnons quelques utilisations de copules pour modéliser la structure de dépendance et pour faire de la modélisation avec des méthodes préexistantes comme la régression avec les copules présentée à la section 1.6. Dans le chapitre 2, nous proposons un modèle échangeable pour les données en grappes en spécifiant les propriétés qui le rendent souple et généralisant des méthodes standards d'analyse des données en grappes. En bref, nous partons des hypothèses vérifiables pour les données en grappes pour définir de nouveaux concepts qui sont utilisés dans le modèle construit.

Chapitre 2

Une nouvelle classe de modèle pour des données hiérarchiques

Ce chapitre porte sur la modélisation de données hiérarchiques dans le cas multidimensionnel à l'aide de modèles probabilistes. Il traite du cas où d variables sont mesurées sur chaque unité d'observation et ces unités sont dans des grappes. Nous sommes donc en présence de deux types de relations de dépendance. La première caractérise les liens entre les variables mesurées sur la même unité et la deuxième porte sur les relations entre les unités d'une même grappe. Ce type de données se modélise souvent avec un modèle normal à effets mixtes (McCulloch et Searle, 2001). Nous nous intéressons plus particulièrement au cas où les unités dans une grappe sont échangeables et nous introduisons d'abord une définition d'échangeabilité multidimensionnelle.

Pour définir le modèle à proposer, l'indice i, $i = 1, \ldots, n$, dénote une des n unités de la grappe. Nous considérons $\mathbf{Z}_i$, le vecteur $d \times 1$ pour la $i$ème unité. L'objectif de ce chapitre est de suggérer des fonctions de répartition $F_{d,n}$, définies sur $\mathbb{R}^{nd}$, pour la loi conjointe du vecteur aléatoire $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ observé sur les unités d'une même grappe. Le vecteur aléatoire $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ est un ensemble de n vecteurs aléatoires dont chacun est de dimension d. Nous nous intéressons plus particulièrement à des familles de loi où $F_{d,n}$ se décompose en une copule en dimension nd et un ensemble de d distributions marginales.

Ce chapitre comporte deux grandes parties. La première suggère une notion d'échangeabilité multidimensionnelle, étudie son impact et donne quelques exemples de lois échangeables. La deuxième partie s'intéresse à un contexte de régression : la première composante du vecteur $\mathbf{Z}$ est le vecteur des variables indépendantes et la dernière com-

posante est la variable dépendante. Nous suggérons une notion d'indépendance conditionnelle partielle pour restreindre et rendre flexible la classe des modèles à considérer dans ce contexte. Une méthode générale de construction de familles de copules échangeables satisfaisant la condition d'indépendance conditionnelle est ensuite proposée. De nouvelles familles sont enfin construites en guise d'illustration sans oublier de proposer des méthodes de simulation et l'utilisation du modèle échangeable pour faire de la prédiction des données en grappes.

2.1 Notion d'échangeabilité multidimensionnelle

L'échangeabilité est une notion classique qui se rencontre dans des situations où l'ordre de mesure des variables sur des individus de la population étudiée peut changer sans impacter la distribution des variables mesurées. Mai et Scherer (2012, page 39) définissent l'échangeabilité unidimensionnelle et donnent des exemples de modèle vérifiant cette condition. Dans le cas de copules bivariées, Ghiselli (2013) a présenté un générateur de famille de copules échangeables. Il faut aussi noter que l'échangeabilité est la symétrie naturelle dans le cas bivarié. La définition 2.1 d'échangeabilité multidimensionnelle que nous proposons est une extension de l'échangeabilité unidimensionnelle définie par Aldous (1985, page 5), Hill (2003) et Mai et Scherer (2012) pour un nombre fini et infini de variables aléatoires. Nous donnons une définition de cette notion en dimension plus grande que deux, de manière formelle.

Définition 2.1. ***Échangeabilité en dimension*** d ***ou*** d***-échangeabilité.***

Une famille de fonctions de répartition $\mathcal{F}_d = \left\{F_{d,n}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) : \boldsymbol{z}_i \in \mathbb{R}^d, n = 2, 3 \ldots\right\}$ *est échangeable en dimension* d *si, pour tout* $n \geq 2$, $F_{d,n} \in \mathcal{F}_d$ *satisfait les conditions suivantes*

i) ***Invariance de loi :*** *Pour toute permutation* $\{\pi(1), \ldots, \pi(n)\}$ *de* $\{1, 2, \ldots, n\}$,

$$F_{d,n}\left(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\right) = F_{d,n}\left\{\boldsymbol{z}_{\pi(1)}, \ldots, \boldsymbol{z}_{\pi(n)}\right\}. \tag{2.1}$$

ii) ***Fermeture sur les marges :*** *Pour* $1 \leq p \leq n$,

$$F_{d,p}\left(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\right) = F_{d,n}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p, \infty, \ldots, \infty). \tag{2.2}$$

Remarque 2.1. *L'échangeabilité définie par Mai et Scherer (2012) est celle où le nombre de variables est* $d = 1$. *Particulièrement, lorsque* $d = 1$, *la définition 2.1 se réduit à celle de Mai et Scherer (2012, page 39). Notre contribution consiste à définir cette notion étendue de l'échangeabilité pour la modélisation.*

L'échangeabilité multidimensionnelle permet d'écrire la distribution de probabilité d'un vecteur aléatoire d'une manière assez réduite et compacte. Elle est une notion qui trouve des applications intéressantes en médecine, en hydrologie (Durante et Okhrin, 2015), en statistique multidimensionnelle, etc. Elle s'utilise aussi pour faire de l'estimation dans les petits domaines, voir Rivest *et al.* (2016). Il existe de nombreux modèles échangeables et nous en énumérons deux.

Exemple 2.1. *Pour $d = 1$ et si $Z_1, \ldots, Z_n$, sont n variables aléatoires indépendantes alors le vecteur $(Z_1, \ldots, Z_n)$ est échangeable en dimension 1.*

Exemple 2.2. ***Copule archimédienne hiérarchique***

La fonction C définie par :

$$C(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = C_{\psi_0}\left\{C_{\psi_1}\left(\boldsymbol{u}_1\right), \ldots, C_{\psi_1}\left(\boldsymbol{u}_n\right)\right\},$$

où $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id}) \in [0,1]^d$, $i = 1, \ldots, n$, $\psi_0^{-1} \circ \psi_1$ est une fonction complètement monotone. Les fonctions C_{ψ_0} et C_{ψ_1} sont des copules archimédiennes de générateurs ψ_0 et ψ_1 respectivement en dimension n et d, et d étant le nombre de variables. La fonction C est une copule échangeable en dimension d.

La copule C considérée dans l'exemple 2.2 est un cas particulier d'échangeabilité. En effet, la fonction C est un exemple de copule archimédienne hiérarchique en dimension d, Joe (1996).

L'échangeabilité d'un ensemble de vecteurs aléatoires impacte la loi de ce dernier. Elle peut servir de pont pour faciliter une meilleure construction et une spécification des paramètres d'un modèle dans le but d'une inférence. Ainsi, la proposition 2.1, sous l'hypothèse d'échangeabilité multidimensionnelle, nous énonçons un résultat sur la forme de la matrice de corrélation d'un vecteur aléatoire multidimensionnel. En effet, la définition d'échangeabilité multidimensionnelle étant nouvelle, nous donnons la forme de la matrice de corrélation associée comme acquis que nous proposons.

Proposition 2.1. *Soit $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n\}$ est un ensemble de n vecteurs aléatoires dont chacun est de dimension d, vérifiant la définition d'échangeabilité en dimension d (voir définition 2.1). Alors, les matrices de corrélation $nd \times nd$, de Pearson, de Spearman,*

des scores normaux ou de Kendall entre ces n vecteurs sont de la forme :

$$I_n \otimes \Sigma_w + J_n \otimes \Sigma_b = \begin{pmatrix} \Sigma_w + \Sigma_b & \Sigma_b & \dots & \Sigma_b \\ \Sigma_b & \Sigma_w + \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \dots & \Sigma_b & \Sigma_w + \Sigma_b \end{pmatrix}, \tag{2.3}$$

où I_n est la matrice identité d'ordre n, J_n est la matrice carrée d'ordre n où tous les éléments sont 1 et les matrices Σ_w et Σ_b sont des matrices symétriques d'ordre d. Le produit tensoriel se note $\otimes$. De plus, pour les corrélations de Pearson, de Spearman et des scores normaux, les matrices Σ_w et Σ_b sont semi-définies positives.

Démonstration. Soit $\{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n\}$, un ensemble de n vecteurs aléatoires dont chacun est de dimension d et vérifiant la condition de d-échangeabilité. Alors la variance de chaque sous-vecteur $\mathbf{Z}_i, i = 1, \dots, n$ est la même, et est elle-même une matrice de variance-covariance notée ici A. De plus, la covariance entre deux vecteurs $\mathbf{Z}_i$ et $\mathbf{Z}_{i'}$, $i, i^{'} = 1, \dots, n$, avec $i \neq i^{'}$ de dimension d ne dépend pas de i et $i^{'}$ et nous la notons Σ_b.

$$\begin{pmatrix} A & \Sigma_b & \dots & \Sigma_b \\ \Sigma_b & A & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \dots & \Sigma_b & A \end{pmatrix} = \begin{pmatrix} \Sigma_w + \Sigma_b & \Sigma_b & \dots & \Sigma_b \\ \Sigma_b & \Sigma_w + \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \dots & \Sigma_b & \Sigma_w + \Sigma_b \end{pmatrix}.$$

Ainsi donc, la matrice de variance-covariance du vecteur $\{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_n\}$ se met sous la forme $(A = \Sigma_w + \Sigma_b)$. Pour tout vecteur $u \in \mathbb{R}^d$, $\left\{u^T\mathbf{Z}_1, u^T\mathbf{Z}_2, \dots, u^T\mathbf{Z}_n\right\}$ est échangeable univariée. Considérons $\mathbf{Z}_i$ et $\mathbf{Z}_{i'}$ pour deux unités quelconques i et $i^{'}$ et la matrice de corrélation entre $u^T\mathbf{Z}_i$ et $u^T\mathbf{Z}_{i'}$ est

$$\begin{pmatrix} u^T\left(\Sigma_w + \Sigma_b\right)u & u^T\Sigma_b u \\ u^T\Sigma_b u & u^T\left(\Sigma_w + \Sigma_b\right)u \end{pmatrix}. \tag{2.4}$$

D'après Mai et Scherer (2012, page 41), cette matrice est définie non-négative et la corrélation entre les deux variables aléatoires est positive ou nulle aboutissant donc $u^T\Sigma_w u \geq 0$ et $u^T\Sigma_b u \geq 0$. □

Nous donnons, dans deux exemples, la matrice de corrélation de modèles échangeables.

Exemple 2.3. ***Copule elliptique multivariée***

Soit $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n)$, un ensemble de n vecteurs aléatoires dont chacun est de dimension d et qui suit une distribution elliptique de matrice de variance-covariance Σ de la forme (2.3). La copule associée à ce vecteur aléatoire est une copule elliptique échangeable en dimension d. Cet exemple constitue la réciproque de la proposition 2.1 dans le cas de la loi elliptique multidimensionnelle.

L'exemple qui suit constitue la suite de l'exemple 2.2.

Exemple 2.4. ***Copule archimédienne hiérarchique (suite)***

La matrice de corrélation du vecteur aléatoire du modèle hiérarchique considéré dans l'exemple 2.2 peut se réduire à la forme

$$\begin{pmatrix} \Sigma_{\rho_1} & \rho_0 J_d & \ldots & \rho_0 J_d \\ \rho_0 J_d & \Sigma_{\rho_1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_0 J_d \\ \rho_0 J_d & \ldots & \rho_0 J_d & \Sigma_{\rho_1} \end{pmatrix}, \quad \Sigma_{\rho_1} = \begin{pmatrix} 1 & \rho_1 & \ldots & \rho_1 \\ \rho_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_1 \\ \rho_1 & \ldots & \rho_1 & 1 \end{pmatrix},$$

où J_d est la matrice carrée d'ordre d où tous les éléments sont 1 et ρ_0, ρ_1 sont les corrélations obtenues respectivement à partir des copules C_{ψ_0} et C_{ψ_1}.

Exemple 2.5. ***Modèle d'analyse de la variance à un facteur multivarié***
Considérons le modèle d'analyse de la variance multivariée à un facteur aléatoire. L'erreur aléatoire de ce modèle s'écrit

$$\boldsymbol{Z}_i = \boldsymbol{a} + \boldsymbol{\nu}_i,\ i = 1, \ldots, n, \tag{2.5}$$

où on note Σ_b, la matrice de dispersion de $\boldsymbol{a}$ et $\boldsymbol{\nu}_i = (\boldsymbol{\nu}_{i1}, \ldots, \boldsymbol{\nu}_{id})^T$ a pour matrice de dispersion Σ_w. De plus, $\boldsymbol{\nu}_i$ et $\boldsymbol{a}$ sont indépendant.

La matrice de dispersion associée au modèle de l'équation (2.5) est exactement de la forme de l'équation (2.3). En conséquence, le modèle linéaire mixte est un cas particulier de modèles échangeables.

La présentation brève de la notion d'échangeabilité nous permet d'avoir une idée sur les possibles utilisations et les propriétés particulières intrinsèques comme la forme de la matrice de corrélation. Ainsi, les copules échangeables peuvent être utilisées pour faire de la modélisation de données dans les cas particuliers où les individus sont interchangeables. Ceci nous motive à postuler des modèles ayant des propriétés d'échangeabilité pour des données hiérarchiques et multiniveaux.

Dans les sections qui suivent, nous proposons un cadre et de nouveaux concepts dans le but de construire un modèle échangeable à base de copules et qui inclut les modèles de base comme le modèle linéaire mixte de Battese *et al.* (1988) et le modèle de l'équation (2.5), voir le livre de Bray et Maxwell (1985) pour plus de précision sur le modèle. Par ailleurs, la construction des modèles multivariés tend à donner des pistes de réflexion impliquant les copules, voir Huang (2020) et Smith *et al.* (2020). Plus spécifiquement, nous présentons, les notions comme l'indépendance conditionnelle et la compatibilité, utiles pour construire le modèle échangeable.

2.2 Notion d'indépendance conditionnelle

Une notion importante que nous utilisons pour la souplesse des modèles multidimensionnels est l'indépendance conditionnelle dont nous donnons une définition dans cette section. Elle existe dans la littérature sous forme d'indépendance conditionnelle des évènements de probabilités Menard et Raoult (1978), de variables aléatoires Su (2004, page 2). Cependant, ces définitions faites étaient dans le cadre univarié, avec un ensemble de variables aléatoires. En conséquence, nous proposons la définition de l'indépendance dans le cas multivarié. L'indépendance conditionnelle a un impact sur un vecteur aléatoire échangeable et sur sa matrice de corrélation.

Nous considérons un vecteur aléatoire $\mathbf{Z} = (\mathbf{X}^T, Y)^T$ continu de dimension d, constitué d'une variable dépendante notée Y et de $(d-1)$ variables explicatives notées $\mathbf{X}$. Spécifiquement, pour faire un parallèle avec des observations en grappes, dans une grappe à n individus, nous considérons que le vecteur aléatoire sur un individu i est noté $\mathbf{Z}_i = (\mathbf{X}_i^T, Y_i)^T$.

Définition 2.2. ***Indépendance conditionnelle***

On parle d'indépendance conditionnelle lorsque conditionnellement à certaines variables, deux ensembles de variables aléatoires sont indépendants. Nous nous intéressons à deux cas de figure d'indépendance conditionnelle.

Indépendance conditionnelle partielle : *on parle d'indépendance conditionnelle partielle si et seulement si les lois conditionnelles vérifient la relation :*

$$\textit{Pour tout } \ i, \quad (Y_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_n) \overset{\mathcal{L}}{=} (Y_i|\boldsymbol{X}_i)\,. \tag{2.6}$$

Indépendance conditionnelle totale : *elle s'exprime en termes de distribution conditionnelle par :*

$$\textit{Pour tout } i, \quad (Y_i|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n,Y_1,\ldots,Y_{i-1},Y_{i+1},\ldots,Y_n) \overset{\mathcal{L}}{=} (Y_i|\boldsymbol{X}_i)\cdot \tag{2.7}$$

Le regroupement dans une grappe à n individus crée une corrélation intra-variables (entre Y_i et $\mathbf{X}_i$) d'un même individu et inter-individus (entre $\mathbf{X}_1,\ldots,\mathbf{X}_n$) au sein d'une grappe. Pour un modèle prédictif, l'indépendance conditionnelle partielle dans cette grappe signifie qu'étant donné des covariables $\mathbf{X}_1$ pour un individu, les covariables $\mathbf{X}_2,\ldots,\mathbf{X}_n$ pour les $(n-1)$ autres unités du grappe n'influencent pas la prédiction de Y_1.

Conséquence : il faut noter que si un modèle vérifie la définition d'indépendance conditionnelle totale traduite par l'équation (2.7) alors elle est également en parfaite adéquation avec celle partielle, (2.6) c'est-à-dire (2.7) $\Longrightarrow$ (2.6).

Nous construisons un modèle échangeable avec une hypothèse d'indépendance conditionnelle à partir des vignes. En particulier, à partir de la D-vigne du chapitre 1, section 1.7, nous montrons les contraintes à imposer pour que la décomposition obtenue soit échangeable. À partir de l'équation (1.35), on cherche les conditions ou les hypothèses à formuler sur la décomposition pour que le modèle considéré vérifie l'hypothèse d'indépendance conditionnelle partielle. Ceci est l'objet de la section 2.3.

2.3 Modèle 2-échangeable et vérifiant la condition d'indépendance conditionnelle partielle

Dans toute cette section, nous considérons que $d = 2$ et que X est une variable aléatoire. Nous construisons dans cette section une copule en dimension 4 vérifiant les propriétés d'indépendance conditionnelle partielle et d'échangeabilité bivariée (2-échangeabilité).

Pour contextualiser, nous supposons que nous avons deux grappes dont chacune contient une unité. Considérons Z_1 et Z_2 deux vecteurs aléatoires de dimension $d = 2$ associés aux deux grappes. Posons $Z_1 = (X_1, Y)^T$ et $Z_2 = (X_2, Y_2)^T$. Les variables X_1, Y_1, X_2 et Y_2 sont notés respectivement 1, 2, 3 et 4 pour faciliter la présentation sur la figure. Notons qu'il y a la corrélation entre X_1 et Y_1 d'une part et entre X_1 et X_2 d'autre part. Pour ces quatre variables aléatoires, nous proposons exactement la décomposition en D-vigne de la figure 1.5. Cette décomposition nécessite six copules bivariées, voir l'équation

(1.35). Notre objectif ici est de chercher les contraintes à imposer à ces six copules pour que les propriétés d'indépendance conditionnelle partielle et d'échangeabilité bivariée soient vérifiées.

Condition d'échangeabilité : L'échangeabilité en dimension 2 signifie que la copule associée à $(\mathbf{Z}_1, \mathbf{Z}_2)$ est la même que celle associé à $(\mathbf{Z}_2, \mathbf{Z}_1)$. Sous la condition que la copule de l'équation (1.35) soit échangeable, nous concluons

- les copules de l'arbre II sont des copules identiques ;
- la symétrie de l'arbre I : les copules C_{12} et C_{34} de la figure 1.5 sont identiques.

La décomposition en D-vigne de la figure 1.5 prend donc une forme plus réduite qui se présente sur la figure 2.1.

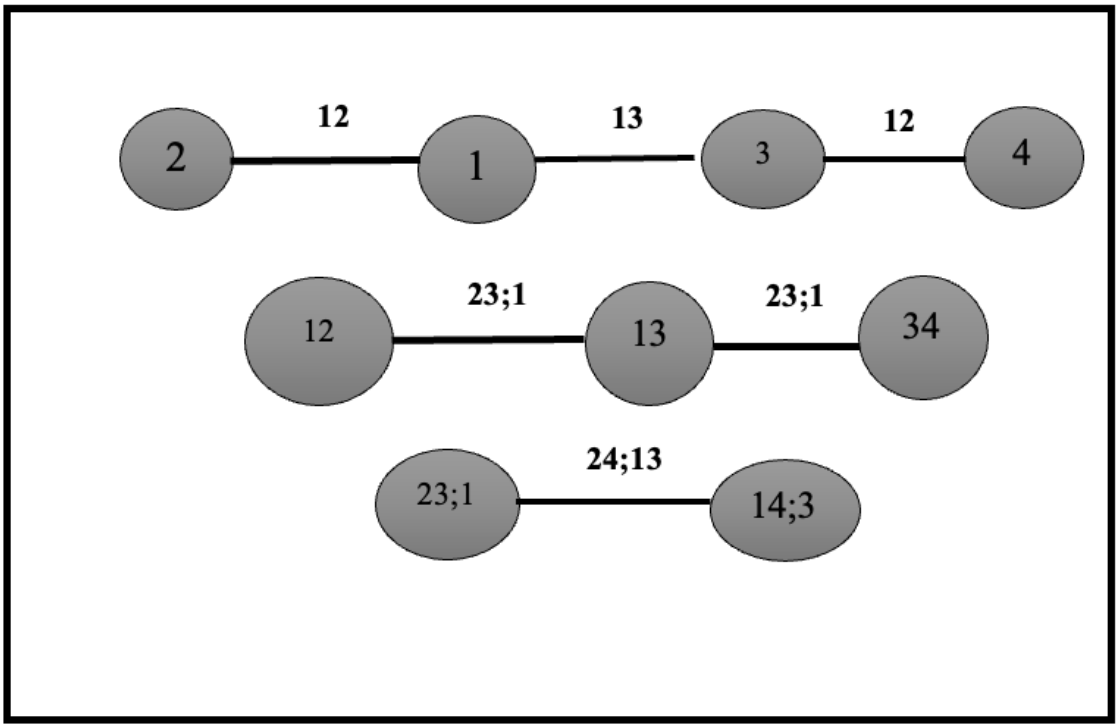


FIGURE 2.1 – Représentation graphique de la D-vigne de 4 variables avec symétrie des arbres I et II.

La densité globale de l'équation (1.35) revient à

$$\begin{aligned} c\{(u_1,u_2),(u_3,u_4)\} &= c_{13}(u_1,u_3)c_{12}(u_1,u_2)c_{12}(u_3,u_4) \\ &\times c_{23;1}\left\{C_{2|1}(u_2|u_1), C_{3|1}(u_3|u_1)\right\} c_{23;1}\left\{C_{1|3}(u_1|u_3), C_{4|3}(u_4|u_3)\right\} \\ &\times c_{24;13}\left\{C_{2|13}(v_1|u_1,u_2), C_{4|13}(v_2|u_1,u_2)\right\}. \end{aligned} \tag{2.8}$$

Condition d'indépendance conditionnelle partielle : Si la densité de copule de l'équation (2.8) vérifie la propriété d'indépendance partielle, alors nous avons les résultats suivants

- les fonctions de distribution conditionnelle s'expriment analytiquement par

$$C_{2|13}(u_2|u_1,u_3) = C_{2|1}(u_2|u_1), \;\; C_{4|13}(u_4|u_1,u_3) = C_{4|3}(u_4|u_3)\cdot \tag{2.9}$$

- La famille de copule conditionnelle $C_{23;1}$ est la copule d'indépendance qui signifie

$$C_{23;1}(u_2,u_3) = u_2 \cdot u_3, \quad c_{23;1}(u_2,u_3) = 1\cdot$$

En appliquant les conclusions de l'hypothèse d'indépendance conditionnelle partielle, par remplacement des équations (2.9) et (2.3) dans (2.8), nous avons

$$\begin{aligned} c\{(u_1,u_2),(u_3,u_4)\} &= c_{13}(u_1,u_3) \times \{c_{12}(u_1,u_2)c_{12}(u_3,u_4)\} \\ &\times \; c_{24;13}\left\{C_{2|1}(u_2|u_1), C_{2|1}(u_4|u_3)\right\}, \end{aligned} \tag{2.10}$$

où $u_1,u_2,u_3,u_4 \in [0,1]$. La copule obtenue est une copule 2-échangeable en dimension 4 et vérifiant la propriété d'indépendance conditionnelle partielle. Elle fait intervenir deux types de copules bivariées : deux copules symétriques C_{13} et $C_{24;13}$ et une copule bivariée quelconque, C_{12}.

Exemple 2.6. ***Modèle beta 2-échangeable ayant la propriété d'indépendance conditionnelle partielle***

Les calculs pour la justification de certains résultats se trouvent en annexe B, section B.1.

L'objectif ici est de construire un modèle échangeable en dimension 4 vérifiant les conditions d'échangeabilité et d'indépendance conditionnelle partielle de l'équation (2.10) en utilisant des lois bêta. Le modèle construit s'inspire de Graf et al. (2018). Le modèle 2-échangeable proposé ici cherche la loi jointe des variables aléatoires alors que dans l'article, les auteurs s'intéressent à la loi conditionnelle des variables sachant les variables explicatives.

Nous nous mettons dans le cas de $d = 2$ et nous considérons quatre variables aléatoires X_1, Y_1, X_2 et Y_2 telles que

$$X_1 = \frac{B_1}{B_1+A}, \quad X_2 = \frac{B_2}{B_2+A}, \quad Y_1 = \frac{1}{1-X_1}\cdot\frac{C_1}{D}, \quad Y_2 = \frac{1}{1-X_2}\cdot\frac{C_2}{D}, \tag{2.11}$$

et les hypothèses suivantes sont faites :

— *la variable aléatoire A suit la loi gamma de paramètre q.*

— *les variables aléatoires B_1 et B_2 suivent chacune la loi gamma de paramètre p.*

— *les variables aléatoires C_1 et C_2 suivent chacune la loi gamma de paramètre r.*

— *la variable aléatoire D suit la loi gamma de paramètre $p+q$.*

— *les variables aléatoires A, B_1, B_2, C_1 et C_2 sont indépendantes.*

Nous obtenons de ces hypothèses et en utilisant aussi la propriété 1.1 portant sur la loi gamma et bêta, les conclusions suivantes.

- *Le vecteur aléatoire (X_1, Y_1, X_2, Y_2) vérifie la définition d'échangeabilité unidimensionnelle de la définition 2.1, c'est-à-dire, $(X_1, Y_1, X_2, Y_2) \overset{\mathcal{L}}{=} (X_2, Y_2, X_1, Y_1)$;*
- *La loi conditionnelle de Y_2 sachant X_2 et X_1 dépend uniquement de X_2, c'est-à-dire $(Y_2|X_2, X_1) \overset{\mathcal{L}}{=} (Y_2|X_2)$.*

Nous déterminons les copules associées respectivement aux vecteurs aléatoires (X_1, X_2) et (X_1, Y_1).

La copule associée au vecteur (X_1, X_2) est la copule bêta de paramètres p, p et q. Cette copule s'adapte à partir de l'équation (1.18) pour $p_1 = p_2 = p$ et $q = q$.

Par ailleurs, en posant $D = B_1 + A$, la copule associée au vecteur aléatoire (Y_1, X_1) est la copule bêta de paramètres p, r et q.

La forme de la copule associée au vecteur aléatoire (X_1, Y_1, X_2, Y_2), en s'inspirant de l'équation (2.10), pour $u_1, v_1, u_2, v_2 \in [0,1]$ est

$$c(u_1, v_1, u_2, v_2) = c_{13}(u_1, u_2)c_{12}(u_1, v_1)c_{12}(u_2, v_2)c_{24;13}\left\{C_{2|1}(v_1|u_1), C_{2|1}(v_2|u_2)\right\} \cdot \quad (2.12)$$

La distribution conditionnelle $C_{2|1}$ vient du tableau 1.3 lorsque C_{12} est une copule beta. Les copules impliquées sont

- *C_{13} est la copule bêta de paramètres p, p et q ;*
- *C_{12} est la copule bêta de paramètres p, r et q ;*
- *$C_{24;13}$ est la copule bêta de paramètres r, r et $p+q$.*

Les équations (2.10) et (2.12) correspondent point pour point à la forme du modèle échangeable dans le cas particulier où les copules impliquées sont des copules bêta.

Le modèle 2-échangeable considéré constitue donc une extension et une forme de construction du modèle de Graf *et al.* (2018). Précisément, les auteurs s'intéressent à la loi conditionnelle de (Y_1, Y_2) sachant (X_1, X_2) alors que nous nous intéressons à la loi du vecteur

(X_1, X_2, Y_1, Y_2). En effet, en supposant connu le vecteur (X_1, X_2), la loi de (Y_1, Y_2) est une loi $\mathcal{MGB}2(r, r, p)$, voir (1.7) et nous l'obtenons aussi avec notre modèle de copules. Ce que décrit précisément Graf *et al.* (2018) dans leur article.

En partant de l'équation (2.10) dans laquelle les composantes sont normales (copules normales et lois marginales normales), la propriété d'échangeabilité et d'indépendance conditionnelle de la définition (2.2) impacte conséquemment sur la forme particulière de la matrice de corrélation d'une loi échangeable. Ainsi, la forme de la matrice de corrélation dans ce cas se donne dans l'exemple 2.1.

Corollaire 2.1. ***Modèle normal 2-échangeable ayant la propriété d'indépendance conditionnelle partielle***

Soit (X_1, Y_1, X_2, Y_2), un vecteur aléatoire de dimension 4 dont la copule provient de l'équation (2.10). Nous faisons les hypothèses suivantes

- *la copule C_{13} est la copule normale de corrélation ρ_1 ;*
- *la copule C_{12} est la copule normale de corrélation ρ_2 ;*
- *la copule $C_{24;13}$ est la copule normale de corrélation ρ_3.*

Le vecteur aléatoire (X_1, X_2, Y_1, Y_2) suit une loi normale multivariée dont la matrice de corrélation est

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_1\rho_2 \\ \rho_1 & 1 & \rho_1\rho_2 & \rho_2 \\ \rho_2 & \rho_1\rho_2 & 1 & \rho_1\rho_2^2 + \rho_3(1-\rho_2^2) \\ \rho_1\rho_2 & \rho_2 & \rho_1\rho_2^2 + \rho_3(1-\rho_2^2) & 1 \end{pmatrix}. \tag{2.13}$$

Démonstration Soit (X_1, Y_1, X_2, Y_2), un vecteur de dimension 4 suivant une loi multivariée dont la densité de copule vient de l'équation (2.10) sous l'hypothèse que les copules impliquées soient normales. Alors la matrice de corrélation de (X_1, X_2, Y_1, Y_2) est de la forme

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_0 \\ \rho_1 & 1 & \rho_0 & \rho_2 \\ \rho_2 & \rho_0 & 1 & \rho_{00} \\ \rho_0 & \rho_2 & \rho_{00} & 1 \end{pmatrix},$$

car la copule C_{13} donne la corrélation ρ_1 entre X_1 et X_2 d'une part. D'autre part, la copule C_{12} donne la corrélation ρ_2 entre X_1 et Y_1. Nous posons aussi $\rho_0 = \text{cor}(X_1, Y_2)$

et $\rho_{00} = \text{cor}(Y_1, Y_2)$. Pour la corrélation ρ_0, nous avons

$$\rho_0 = \mathbb{E}(X_1 Y_2) = \mathbb{E}\left\{\mathbb{E}(X_1 Y_2 | X_2)\right\} = \rho_1 \rho_2.$$

Par ailleurs, la loi de (Y_1, Y_2) sachant (X_1, X_2) est de matrice de variance-covariance

$$\begin{pmatrix} 1-\rho_2^2 & \rho_{00}-\rho_1\rho_2^2 \\ \rho_{00}-\rho_1\rho_2^2 & 1-\rho_2^2 \end{pmatrix}.$$

Or $\rho_3 = \text{cor}(Y_1, Y_2 | X_1, X_2)$ donc en posant $\rho_3 = (\rho_{00} - \rho_1\rho_2^2)/(1-\rho_2^2)$, nous obtenons le résultat attendu.

Dans la suite de cette section, nous établissons la relation entre le résultat obtenu à l'équation (2.13) et le modèle linéaire mixte.

Relation entre le modèle échangeable (2.10) et le modèle de Battese *et al.* (1988)

Nous considérons quatre variables aléatoires X_1, X_2, Y_1 et Y_2 telle que X_1 et X_2 suivent chacune une loi normale, centrée et réduite. Considérons les variables aléatoires A et $E_i, i = 1, 2$ suivant une loi normale centrée réduite et indépendante chacune de X_i. L'objectif ici est de construire un modèle prédictif de Y_i sachant X_i avec le modèle 2-échangeable de l'équation (2.10). Par la suite, nous le comparons au modèle de Battese *et al.* (1988), voir (1.36).

L'équation du modèle de prédiction est

$$Y_i = \rho_2 X_i + \sqrt{\rho_3}(1-\rho_2^2)^{1/2} A + (1-\rho_2^2)^{1/2}(1-\rho_3)^{1/2} E_i \; i = 1, 2. \tag{2.14}$$

Nous formulons les conclusions suivantes sur le modèle proposé.

- les marginales de Y_1 et Y_2 sont des lois normales, centrées et réduites ;
- la copule associée à (X_1, Y_1, X_2, Y_2) est la copule normale échangeable et sa matrice de corrélation est (2.13) ;
- le modèle considéré vérifie la propriété d'indépendance conditionnelle partielle puisque sachant X_1 et X_2, Y_1 ne dépend que de X_1 ;
- la décomposition de l'équation (2.14) est exactement le modèle de Battese *et al.* (1988) de l'équation (1.36) avec les correspondances suivantes

$$\beta X_i \equiv \rho_2 X_i, \quad \nu_j \equiv \sqrt{\rho_3}(1-\rho_2^2)^{1/2} A, \quad e_i \equiv (1-\rho_2^2)^{1/2}(1-\rho_3)^{1/2} E_i. \tag{2.15}$$

La propriété d'indépendance conditionnelle partielle appliquée à un vecteur aléatoire échangeable multivarié permet d'aboutir à des résultats déjà connus comme le modèle de Battese *et al.* (1988). Ceci souligne la généralisation des modèles avec les propriétés d'échangeabilité malgré l'hypothèse faible d'indépendance partielle. Pour un vecteur de variable aléatoire de loi normale multivariée en dimension 4 dont la matrice de corrélation est sous la forme (2.3), la condition particulière $\rho_0 = \rho_1\rho_2$ permet d'avoir la propriété d'échangeabilité avec indépendance conditionnelle partielle en dimension $d = 2$. Cette forme particulière de la matrice de corrélation nous conduit à d'autres classes de modèle multidimensionnel. Nous définissons dans la section suivante une nouvelle classe de modèle multidimensionnel intégrant les copules multivariées échangeables et ses propriétés.

2.4 Formulation mathématique du modèle de d-copule échangeable

Dans cette section, nous définissons un nouveau modèle en utilisant des notions comme l'échangeabilité, l'indépendance conditionnelle et la compatibilité. Pour cela, nous définissons le concept de compatibilité entre deux copules.

2.4.1 Notion de compatibilité de deux copules en dimension d

La notion de compatibilité définit une relation particulière entre deux copules dont une est échangeable. Nous donnons une définition formelle.

Définition 2.3. *Nous considérons d, un entier plus grand ou égal à 2. Nous notons $\left(C^{(1)}_{d-1,n}\right)$, la copule strictement positive appartenant à une famille de copules échangeables de n vecteurs dont chacun est de dimension $d-1$ de la définition 2.4. Soit $C^{(2)}_{d,1}$, la copule de dimension d telle que la copule $C^{(1)}_{d-1,1}$ se lie à $C^{(2)}_{d,1}$ par la relation définie*

$$c^{(1)}_{d-1,1}(\boldsymbol{u}) = \int_0^1 c^{(2)}_{d,1}(\boldsymbol{u}, t)dt, \ \ \boldsymbol{u} \in [0,1]^{d-1}. \tag{2.16}$$

On dit que les copules $C^{(2)}_{d,1}$ et $C^{(1)}_{d-1,1}$ sont d-compatible.

Si $d = 2$, l'équation (2.16) est toujours vraie pour toute copule $C^{(2)}_{d,1}$.

La notion de compatibilité est une notion naturelle qui existe entre une fonction densité de probabilité et les densités marginales par exemple. Cependant, nous l'introduisons

sur les copules et donnons ici le nom de compatibilité pour avoir un formalisme. En dimension d, nous donnons un exemple de copules qui vérifient la notion de compatibilité.

Exemple 2.7. ***Copule normale***

Considérons une copule $C^{(2)}_{d,1}$ associée à un vecteur aléatoire de dimension d suivant une loi normale échangeable et la famille $C^{(1)}_{d-1,1}$ est celle de la marginale en dimension $(d-1)$. Les copules $C^{(1)}_{d-1,1}$ et $C^{(2)}_{d,1}$ sont d-compatible et vérifient l'équation (2.16).

La notion de compatibilité est nécessaire pour construire le d-modèle. Notre objectif dans la construction du modèle probabiliste consiste à écrire la distribution jointe associée à des variables aléatoires structurées en grappes. Nous donnons la formulation mathématique du modèle proposé dans la section qui suit.

2.4.2 Définition mathématique de la d-copule échangeable

Le modèle proposé s'applique pour la construction de loi pour des observations hiérarchiques. Nous définissons la forme d'une d-copule échangeable.

Définition 2.4. ***Un d-modèle en dimension n***

Nous considérons des copules $\left(C^{(1)}_{d-1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$ appartenant à des familles de copules échangeables respectivement en dimension $(d-1)$ et 1. Soit $C^{(2)}_{d,1}$, la copule d-compatible avec la copule $C^{(1)}_{d-1,1}$, de dimension $(d-1)$, au sens de l'équation (2.16). On appelle d-modèle en dimension n, une famille de fonctions $c_{d,n}$ définie par

$$\begin{aligned} c_{d,n}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_n, v_n)\right\} &= c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \times \prod_{i=1}^{n} \left\{ \frac{c^{(2)}_{d,1}(\boldsymbol{u}_i, v_i)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_i)} \right\} \\ &\quad \times c^{(3)}_{1,n}\left\{C_{d|1:d-1}(v_1|\boldsymbol{u}_1), \ldots, C_{d|1:d-1}(v_n|\boldsymbol{u}_n)\right\}, \quad (2.17) \end{aligned}$$

avec $\boldsymbol{u}_i \in [0,1]^{d-1}$, $v_i \in [0,1]$, $i = 1, \ldots, n$. La distribution conditionnelle $C_{d|1:d-1}$ est la fonction de répartition de la dième variable sachant les $(d-1)$ premières variables, représentant la première composante. Elle se déduit de la copule $C^{(2)}_{d,1}$ pour $\boldsymbol{u} \in [0,1]^{d-1}$ par

$$C_{d|1:d-1}(v|\boldsymbol{u}) = \frac{1}{c^{(1)}_{d-1,1}(\boldsymbol{u})} \cdot \int_0^v c^{(2)}_{d,1}(\boldsymbol{u}, t)dt, \qquad v \in [0,1]\cdot \quad (2.18)$$

Le d-modèle en dimension n, donné par l'équation (2.17) possède certaines propriétés que nous pouvons énoncer dans la propriété 2.1.

Propriété 2.1. *Le d-modèle en dimension n défini par l'équation (2.17) vérifie les propriétés suivantes :*

(i) $c_{d,n}$ *est une fonction positive ;*

(ii) $c_{d,n}$ *est une fonction de densité d'une copule ;*

(iii) le d-modèle est fermé sur les marges ;

(iv) le d-modèle satisfait la condition d'échangeabilité multidimensionnelle en dimension d donnée à la définition 2.1 ;

(v) le d-modèle vérifie la condition d'indépendance conditionnelle partielle de l'équation (2.6).

Démonstration. Nous considérons le d-modèle en dimension n de l'équation (2.17). On note $(u_1, \ldots, \boldsymbol{u}_n)$, un vecteur de dimension $n(d-1)$, $(v_1, \ldots, v_n)$, un vecteur de dimension n et

$$d\boldsymbol{u}_i = du_{i1} \times \ldots \times du_{i(d-1)}, \;\; i = 1, \ldots, n\cdot$$

(i) Montrons que la fonction $c_{d,n}$ est positive.
Dans la définition de la fonction $c_{d,n}$, les copules $\left(C^{(1)}_{d-1,n}\right)$, $C^{(2)}_{d,1}$, $C^{(1)}_{d-1,1}$ et $\left(C^{(3)}_{1,n}\right)$ sont des fonctions positives compte tenu du fait qu'elles sont des copules. Donc nous concluons que $c_{d,n}$ est une fonction positive.

(ii) Récursivement, nous faisons une intégration du d-modèle en dimension n jusqu'à aboutir à la copule marginale en dimension 2 donnée par

$$c_{d,2}(\boldsymbol{u}_1, v_1) = c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1)\cdot$$

En conséquence, nous avons

$$\begin{aligned}\int_{[0,1]^{nd}} c_{d,n}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_n, v_n)\right\} d\boldsymbol{u}_1 \ldots d\boldsymbol{u}_n dv_1 \ldots dv_n &= \int_{[0,1]^d} c(\boldsymbol{u}_1, v_1) d\boldsymbol{u}_1 dv_1 \\ &= \int_{[0,1]^d} c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1) d\boldsymbol{u}_1 dv_1\end{aligned}$$

d'où

$$\int_{[0,1]^{nd}} c_{d,n}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_n, v_n)\right\} d\boldsymbol{u}_1 \ldots d\boldsymbol{u}_n dv_1 \ldots dv_n = 1\cdot$$

Ceci conclut le fait que $c_{d,n}$ est une fonction densité d'une copule.

(iii) Nous recherchons la densité de la copule en dimension $(n-1)$ à partir du d-modèle. Posons $A = c\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_{n-1}, v_{n-1})\right\}$ pour simplifier la présentation

des équations. Nous trouvons :

$$\begin{aligned}
A &= \int_{[0,1]^d} c_{d,n}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_n, v_n)\right\} d\boldsymbol{u}_n dv_n \\
&= \prod_{i=1}^{n-1} \left\{ \frac{c_{d,1}^{(2)}(\boldsymbol{u}_i, v_i)}{c_{d-1,1}^{(1)}(\boldsymbol{u}_i)} \right\} \int_{[0,1]^d} c_{d-1,n}^{(1)}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \times \\
&\quad \frac{c_{d,1}^{(2)}(\boldsymbol{u}_n, v_n)}{c_{d-1,1}^{(1)}(\boldsymbol{u}_n)} c_{1,n}^{(3)}\left\{(C_{d|1:d-1}(v_1|\boldsymbol{u}_1), \ldots, C_{d|1:d-1}(v_n|\boldsymbol{u}_n)\right\} d\boldsymbol{u}_n dv_n \\
&= \prod_{i=1}^{n-1} \left\{ \frac{c_{d,1}^{(2)}(\boldsymbol{u}_i, v_i)}{c_{d-1,1}^{(1)}(\boldsymbol{u}_i)} \right\} \int_{[0,1]^{d-1}} c_{d-1,n}^{(1)}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \\
&\quad \times \int_0^1 c_{1,n}^{(3)}\left\{C_{d|1:d-1}(v_1|\boldsymbol{u}_1), \ldots, C_{d|1:d-1}(v_{n-1}|\boldsymbol{u}_{n-1}), w_n\right\} dw_n d\boldsymbol{u}_n,
\end{aligned}$$

où nous posons $t_n = C_{d|1:d-1}(v_n|\boldsymbol{u}_n)$. Ceci conduit à

$$\begin{aligned}
c\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_{n-1}, v_{n-1})\right\} &= c_{d-1,n-1}^{(1)}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{n-1}) \times \prod_{i=1}^{n-1} \left\{ \frac{c_{d,1}^{(2)}(\boldsymbol{u}_i, v_i)}{c_{d-1,1}^{(1)}(\boldsymbol{u}_i)} \right\} \times \\
&\quad c_{1,n-1}^{(3)}\left\{C_{d|1:d-1}(v_1|\boldsymbol{u}_1), .., C_{d|1:d-1}(v_{n-1}|\boldsymbol{u}_{n-1})\right\},
\end{aligned}$$

car les copules échangeables $\left(C_{d-1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont fermées sur les marges. Ainsi, nous avons

$$c\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_{n-1}, v_{n-1})\right\} = c_{d,n-1}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_{n-1}, v_{n-1})\right\}.$$

La forme du d-modèle en dimension n est la même que le d-modèle en dimension $(n-1)$. Nous concluons donc que le d-modèle en dimension n est fermée sur les marges.

(iv) Pour toute permutation $\{\pi(1), \ldots, \pi(n)\}$ de $\{1, \ldots, n\}$, nous avons

$$c_{d,n}\left\{(\boldsymbol{u}_1, v_1), \ldots, (\boldsymbol{u}_n, v_n)\right\} = c_{d,n}\left\{(\boldsymbol{u}_{\pi(1)}, v_{\pi(1)}), \ldots, (\boldsymbol{u}_{\pi(n)}, v_{\pi(n)})\right\},$$

donc le d-modèle en dimension n est échangeable.

(v) Sans perdre de généralité, l'objectif est de montrer que nous avons le résultat $F(v_j|\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n) = C_{d|1:d-1}(v_j|\boldsymbol{u}_j)$ pour tout $j = 1, \ldots, n$.

La densité de la distribution conditionnelle de la variable aléatoire V_j sachant le vecteur aléatoire $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_j, \ldots, \boldsymbol{U}_n)$ est

$$f\left(v_j|\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n\right) = \frac{c_2\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n, v_j\right)}{c_1\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n\right)}, \tag{2.19}$$

où c_1 et c_2 sont des densités des copules respectivement associées aux vecteurs aléatoires $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_j, \ldots, \boldsymbol{U}_n)$ et $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_j, \ldots, \boldsymbol{U}_n, V_j)$ or nous avons

$$c_1(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n) = c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n) \cdot \tag{2.20}$$

En effet, de l'équation (2.17), nous avons

$$c_1(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n) = \int_{[0,1]^n} c_{d,n}\{(\boldsymbol{u}_1, v_1) \ldots, (\boldsymbol{u}_n, v_n)\}\, dv_1 \ldots dv_n \cdot$$

Ainsi, nous avons

$$\begin{aligned} c_1(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) &= c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \int_{[0,1]^n} \prod_{i=2}^{n} \left\{ \frac{c^{(2)}_{d,1}(\boldsymbol{u}_i, v_i)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_i)} \right\} \\ &\quad \times c^{(3)}_{1,n}\{F(v_1|\boldsymbol{u}_1), \ldots, F(v_n|\boldsymbol{u}_n)\}\, dv_1 \ldots dv_n \\ &= c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \int_{[0,1]^n} c^{(3)}_{1,n}(w_1, \ldots, w_n)\, dw_1 \ldots dw_n, \end{aligned}$$

où nous posons $w_i = C_{d|1:d-1}(v_i|\boldsymbol{u}_i)$, $i = 1, \ldots, n$ d'où le résultat de l'équation (2.20). Par ailleurs, la densité c_2, dans (2.19) s'obtient de la manière suivante

$$\begin{aligned} c_2(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, v_1) &= c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \frac{c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_1)} \int_{[0,1]^{n-1}} \prod_{i=2}^{n} \left\{ \frac{c^{(2)}_{d,1}(\boldsymbol{u}_i, v_i)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_i)} \right\} \\ &\quad \times c^{(3)}_{1,n}\left\{C_{d|1:d-1}(v_1|\boldsymbol{u}_1), \ldots, C_{d|1:d-1}(v_n|\boldsymbol{u}_n)\right\} dv_2 \ldots dv_n \\ &= c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \frac{c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_1)} C_{d|1:d-1}(v_1|\boldsymbol{u}_1) \cdot \end{aligned}$$

L'équation (2.19) revient donc à

$$\begin{aligned} \frac{c_2(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, v_1)}{c_1(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)} &= \frac{c^{(2)}_{d,1}(\boldsymbol{u}_j, v_j)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_j)} \times \\ &\quad \int_{[0,1]^{n-1}} c^{(3)}_{1,n}\left\{C_{d|1:d-1}(v_1|\boldsymbol{u}_1), w_2, \ldots, w_n\right\} dw_2 \ldots dw_n \\ &= C_{d|1:d-1}(v_1|\boldsymbol{u}_1) \frac{c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_1)} \cdot \end{aligned}$$

Finalement, la densité conditionnelle est

$$f(v_j|\boldsymbol{u}_1, \ldots, \boldsymbol{u}_j, \ldots, \boldsymbol{u}_n) = \frac{c^{(2)}_{d,1}(\boldsymbol{u}_j, v_j)}{c^{(1)}_{d-1,1}(\boldsymbol{u}_j)} \int_0^{v_j} c^{(2)}_{d,1}(\boldsymbol{u}_j, t) dt \cdot$$

Ce qui confirme la propriété d'indépendance conditionnelle du d-modèle en dimension n.

□

Nous montrons précédemment que le d-modèle en dimension n est à la fois une densité de copule échangeable et vérifie la propriété d'indépendance conditionnelle. Nous l'appelons donc une ***d-copule échangeable en dimension*** n.

Le modèle de d-copule échangeable fait intervenir des copules : deux familles de copules échangeables $\left(C^{(1)}_{d-1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$ et une copule quelconque $C^{(2)}_{d,1}$, compatible avec $\left(C^{(1)}_{d-1,n}\right)$. Nous enonçons une propriété d'indépendance liée au modèle de d-copule échangeable en dimension n.

Théorème 2.1. *Soit $(\boldsymbol{U}_1, V_1), \ldots, (\boldsymbol{U}_n, V_n)$, un ensemble de n vecteurs aléatoires dont chacun est de dimension d et dont la densité provient de l'équation (2.17). Si nous posons $W_i = C_{d|1:d-1}(V_i|\,\boldsymbol{U}_i)$, $i = 1, \ldots, n$, où la fonction de répartition $C_{d|1:d-1}$ provient de l'équation (2.18), alors les vecteurs aléatoires $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ et $(W_1, \ldots, W_n)$ sont indépendants. De plus, les vecteurs $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ et $(W_1, \ldots, W_n)$ suivent les lois de famille de copules respectives $\left(C^{(1)}_{d-1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$ avec des marginales normales.*

Démonstration. Soient $(\boldsymbol{U}_1, V_1), \ldots, (\boldsymbol{U}_n, V_n)$, un ensemble de n vecteurs aléatoires où chacun est de dimension d et a une densité de copule qui vérifie la définition de la d-copule échangeable en dimension n. Nous déterminons la loi du vecteur aléatoire $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n, W_1, \ldots, W_n)$ où $W_i = C_{d|1:d-1}(V_i|\,\boldsymbol{U}_i)$ vient de l'équation (2.18).

En utilisant la formule du changement de variables multivariées, le déterminant de la matrice jacobienne est

$$|J| = \frac{c^{(2)}_{d,1}(\boldsymbol{u}_1, v_1) \times \ldots \times c^{(2)}_{d,1}(\boldsymbol{u}_n, v_n)}{c^{(1)}_{d-1}(\boldsymbol{u}_1) \times \ldots \times c^{(1)}_{d-1}(\boldsymbol{u}_n)}.$$

La densité du vecteur aléatoire $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n, W_1, \ldots, W_n)$ est

$$c(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, w_1, \ldots, w_n) = c^{(1)}_{d-1,n}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \times c^{(3)}_{1,n}(w_1, \ldots, w_n)\,.$$

Ceci conclut la preuve du théorème sur l'indépendance entre le vecteur aléatoire de dimension $n(d-1)$, $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ et le vecteur aléatoire de dimension n, $(W_1, \ldots, W_n)$ et donne la loi de chacun des vecteurs aléatoires. □

Le théorème précédent, sert de base pour suggérer une méthode de simulation de données suivant un modèle de d-copule échangeable en dimension n.

2.5 Simulation d'un modèle de d-copule échangeable

Nous présentons dans cette section, une procédure de simulation du vecteur (U, V) dans le cas particulier où il provient de la définition de la d-copule échangeable (voir définition 2.4). Nous présentons l'algorithme pour une grappe de taille n. Nous rappelons que le modèle multivarié échangeable utilisé est de l'équation (2.17) et nous nous mettons dans le cas où le nombre de variables mesurées par individu est d, avec $(d-1)$ variables explicatives. La procédure de simulation se déroule en trois étapes.

Algorithme de simulation de la d-copule échangeable en dimension n

Étape 1 : Simuler n vecteurs dont chacun a une dimension $(d-1)$, $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ selon la famille de copule échangeable $\left(C^{(1)}_{d-1,n}\right)$.

Étape 2 : Simuler un ensemble $(W_1, \ldots, W_n)$ de n vecteurs dont chacun est de dimension d selon la famille de copule échangeable $\left(C^{(3)}_{1,n}\right)$.

Étape 3 : Résoudre l'équation en V_i définie par

$$W_i = C_{d|1:d-1}(V_i|\,\boldsymbol{U}_i), \quad i = 1, \ldots, n,$$

où $C_{d|1:d-1}$ vient de l'équation (2.18).

Remarque 2.2. *Pour la simulation des familles de copules échangeables, nous donnons deux références usuelles en fonction du type de famille de copules échangeables. Pour la famille de copules archimédiennes échangeables, voir Wu* et al. *(2007). Pour une famille de copules elliptiques échangeables, voir Mai et Scherer (2012, page 179).*

Exemple 2.8. ***Algorithme présenté en dimension*** $d = 2$

Dans cet exemple, les trois copules utilisées pour la simulation sont :

- *la famille de copule $\left(C^{(1)}_{d-1,n}\right)$ est une copule de Frank multivariée échangeable de paramètre $\theta > 0$ en dimension n ;*
- *la copule $\left(C^{(3)}_{1,n}\right)$ est une copule normale multivariée échangeable de paramètre ρ en dimension n ;*
- *la copule $C^{(2)}_{d,1}$ est une copule bêta bivariée de paramètres p_1, p_2 et q, tous positifs.*

Nous souhaitons simuler $(\boldsymbol{U}_1, V_1), \ldots, (\boldsymbol{U}_n, V_n)$.

Étape 1 : Simulation de $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ ***suivant la famille de copule échangeable*** $\left(C^{(1)}_{d-1,n}\right)$***, copule de Frank à partir de Wu et al. (2007)***

1. *Simuler* $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ *suivant une loi exponentielle de paramètre 1, mutuellement indépendantes.*
2. *Simuler une variable aléatoire positive* Y *suivant une loi logarithmique avec comme support les entiers positifs dont la transformée de Laplace est* ψ *définie par*

$$\psi(t) = -\frac{1}{\theta}\log\left\{1 + e^{-t}(e^{-\theta} - 1)\right\}, \quad t \in \mathbb{R}_+. \tag{2.21}$$

   La variable aléatoire Y *suit une loi logarithmique de paramètre* θ*, voir* (1.14) *avec support les entiers positifs.*
3. *Poser* $U_i = \psi^{-1}\left(\frac{\boldsymbol{Z}_i}{Y}\right),\ i = 1, \ldots, n.$

Ces trois points nous permettent d'obtenir le vecteur $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$.

Sous le logiciel R les fonctions `BiCopSim` *et* `rCopula` *du package* `VineCopula` *permettent facilement de faire cette simulation.*

Étape 2 : Simulation de $W_1, \ldots, W_n$ ***suivant la famille de copule échangeable*** $\left(C^{(3)}_{1,n}\right)$

Simuler $(W_1, \ldots, W_n)$ *suivant la copule normale échangeable en dimension* n *pour la corrélation fixée* ρ*, voir Mai et Scherer (2012, page 179).*

Étape 3 : Obtention de $V_1, \ldots, V_n$

Nous obtenons V_i *grâce à la formule*

$$V_i = B_{p_2,q}\left\{\frac{B^{-1}_{p_2,p_1+q}(W_i)}{1 - B^{-1}_{p_1,q}(U_i) + B^{-1}_{p_1,q}(U_i)B^{-1}_{p_2,p_1+q}(W_i)}\right\}, \quad i = 1, \ldots, n. \tag{2.22}$$

Sous le logiciel R par exemple, les fonctions `uniroot`*,* `nlme` *ou* `optim` *permettent d'obtenir la solution exacte de cette équation.*

Nous expliquons la procédure d'obtention de l'équation (2.22) *en détail dans l'annexe A, section A.1. À la fin des étapes 1, 2 et 3, nous obtenons alors* $(\boldsymbol{U}_1, V_1), \ldots (\boldsymbol{U}_n, V_n)$.

Dans la suite de ce chapitre, nous utilisons le modèle de d-copule échangeable en dimension n pour construire un modèle prédictif pour les données en grappes. Cette prédiction se base sur la connaissance des observations au sein de la grappe.

2.6 Prédiction à partir d'un modèle d-copule échangeable

Nous considérons $\mathbf{X}$, un vecteur aléatoire de dimension $(d-1)$ et Y, une variable aléatoire. Supposons que nous sommes dans une grappe à n individus et nous connaissons le vecteur $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$, observations de $\mathbf{Z} = (\mathbf{X}^T, Y)^T$. En considérant que nous connaissons le vecteur indépendant $\mathbf{x}_0$, de dimension $(d-1)$ sur un $(n+1)$ième individu de cette grappe. Nous souhaitons prédire Y_0, la valeur de la variable dépendante associée à $\mathbf{x}_0$. Nous notons G la fonction de répartition univariée associée à Y et F, celle associée au vecteur aléatoire $\mathbf{X}$. L'objectif est de déterminer l'expression théorique de la prédiction de Y sachant $\mathbf{X}$ par l'espérance conditionnelle de Y sachant $\mathbf{X}$.

La densité de la distribution $f_{d,n+1}$ de l'ensemble des variables aléatoires noté par $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n), (\mathbf{X}_0, Y_0)\}$ dans une grappe à n individus s'écrit

$$
\begin{aligned}
&f_{d,n+1}\left\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), (\mathbf{x}_0, y_0)\right\} = c^{(1)}_{d-1,n+1}\left\{F(\mathbf{x}_1), \ldots, F(\mathbf{x}_n), F(\mathbf{x}_0)\right\} \times \\
&f(\mathbf{x}_0) g(y_0) \frac{c^{(2)}_{d,1}\left\{F(\mathbf{x}_0), G(y_0)\right\}}{c^{(1)}_{d-1,1}\left\{(F(\mathbf{x}_0)\right\}} \prod_{i=1}^{n} \left[f(\mathbf{x}_i) g(y_i) \frac{c^{(2)}_{d,1}\left\{F(\mathbf{x}_i), G(y_i)\right\}}{c^{(1)}_{d-1,1}\left\{F(\mathbf{x}_i)\right\}} \right] \times \\
&c^{(3)}_{1,n+1}\left[C_{d|1:d-1}\left\{G(y_1)|F(\mathbf{x}_1)\right\}, \ldots, C_{d|1:d-1}\left\{G(y_n)|F(\mathbf{x}_n)\right\}, C_{d|1:d-1}\left\{G(y_0)|F(\mathbf{x}_0)\right\} \right].
\end{aligned}
\tag{2.23}
$$

En partant de l'équation (2.23) et en notant $h_{\mathbf{x}_0}$, la densité conditionnelle de la variable Y_0 sachant $\mathbf{x}_0$, $\mathbf{z}_1, \ldots, \mathbf{z}_n$, nous avons

$$
\mathbb{E}\left(Y_0 | \mathbf{x}_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \int_{\mathbb{R}} y h_{\mathbf{x}_0}(y) dy, \tag{2.24}
$$

où $h_{\mathbf{x}_0}$ s'écrit

$$
h_{\mathbf{x}_0}(y) = g(y) \frac{c_{d,n+1}\left[\left\{F(\mathbf{x}_1), G(y_1)\right\}, \ldots, \left\{F(\mathbf{x}_n), G(y_n)\right\}, \left\{F(\mathbf{x}_0), G(y)\right\}\right]}{c_X\left[\left\{F(\mathbf{x}_1), G(y_1)\right\}, \ldots, \left\{F(\mathbf{x}_n), G(y_n)\right\}, F(\mathbf{x}_0)\right]}, \ y \in \mathbb{R}, \tag{2.25}
$$

et $c_{d,n+1}$ est la densité de la copule provenant de la d-copule échangeable, définition 2.4 par l'équation (2.17). En partant du modèle postulé pour une grappe à $(n+1)$ individus

et en remplaçant dans l'équation (2.24), l'expression du prédicteur est

$$\begin{aligned}\mathbb{E}\left(Y_0|\mathbf{x}_0,\mathbf{z}_1,\ldots,\mathbf{z}_n\right) &= \int_{\mathbb{R}} yg(y)\frac{c^{(2)}_{d,1}\left\{F(\mathbf{x}_0),G(y)\right\}}{c^{(1)}_{d-1,1}\left\{F(\mathbf{x}_0)\right\}}\times \\ &\frac{c^{(3)}_{1,n+1}\left[C_{d|1:d-1}\left\{G(y_1)|F(\mathbf{x}_1)\right\},\ldots,C_{d|1:d-1}\left\{G(y)|F(\mathbf{x}_0)\right\}\right]}{c^{(3)}_{1,n}\left[C_{d|1:d-1}\left\{G(y_1)|F(\mathbf{x}_1)\right\},\ldots,C_{d|1:d-1}\left\{G(y_n)|F(\mathbf{x}_n)\right\}\right]}dy.\end{aligned} \tag{2.26}$$

Ce prédicteur dépend en partie de la densité de la variable aléatoire Y notée g réajustée avec la copule $C^{(2)}_{d,1}$ et la copule $\left(C^{(3)}_{1,n}\right)$. Ainsi, en supposant que la famille de copule $\left(C^{(3)}_{1,n}\right)$ est une copule d'indépendance alors le prédicteur de l'équation (2.26) est exactement le prédicteur de l'équation (1.28) du chapitre 1. Le prédicteur de l'équation (2.26) est donc une généralisation de la régression copule de la section 1.6 en tenant compte de l'effet des observations dans la grappe. En faisant un changement de variable, le prédicteur de l'équation (2.26) est de la forme suivante

$$\begin{aligned}\mathbb{E}(Y_0|\mathbf{x}_0,\mathbf{z}_1,\ldots,\mathbf{z}_n) &= \int_0^1 G^{-1}\left[F^{-1}\left\{z|F(\mathbf{x})\right\}\right]\times \\ &\frac{c^{(3)}_{1,n+1}\left[C_{d|1:d-1}\left\{G(y_1)|F(\mathbf{x}_1)\right\},\ldots,C_{d|1:d-1}\left\{G(y_n)|F(\mathbf{x}_n)\right\},z\right]}{c^{(3)}_{1,n}\left[C_{d|1:d-1}\left\{G(y_1)|F(\mathbf{x}_1)\right\},\ldots,C_{d|1:d-1}\left\{G(y_n)|F(\mathbf{x}_n)\right\}\right]}dz.\end{aligned} \tag{2.27}$$

Nous étudions dans la section qui suit la prédiction les conclusions suivantes par rapport au prédicteur dans des cas spécifiques.

2.7 Étude du prédicteur avec la 2-copule échangeable

Dans cette section, nous considérons que nous avons une variable explicative X et une variable dépendante Y. Dans toute la section, pour les besoins d'utilisation, la densité de probabilité d'un ensemble de vecteurs $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ à partir d'une 2-copule échangeable s'écrit

$$f_{2,n}\left\{(x_1,y_1),\ldots,(x_n,y_n)\right\} = c_{2,n}\left[\left\{F(x_1),G(y_1)\right\},\ldots,\left\{F(x_n),G(y_n)\right\}\right]\prod_{i=1}^{n}f(x_i)g(y_i), \tag{2.28}$$

avec $c_{2,n}$ définie pour $(u_i,v_i)\in[0,1]^2$ par

$$\begin{aligned}c_{2,n}\left\{(u_1,v_1),\ldots,(u_n,v_n)\right\} &= c^{(1)}_{1,n}(u_1,\ldots,u_n)\times\prod_{i=1}^{n}\left\{c^{(2)}(u_i,v_i)\right\} \\ &\times c^{(3)}_{1,n}\left\{C_{2|1}(v_1|u_1),\ldots,C_{2|1}(v_n|u_n)\right\}.\end{aligned}$$

2.7.1 Quelques résultats de l'expression du prédicteur dans des cas particuliers

Nous considérons que X_0 est une variable aléatoire de densité de probabilité f. Nous rappelons que $\mathbf{z}_i = (x_i, y_i)^T,\ i = 1, \ldots, n$.

- Si les copules $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules d'indépendance de l'équation (1.4), qui signifie que l'effet grappe est absent et impliquant que les deux vecteurs aléatoires X_0 et Y_0 ne sont pas corrélées alors à partir de (2.27), nous avons

  $$\mathbb{E}\left(Y_0|X_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \mathbb{E}(Y_0)\cdot$$

  Le prédicteur est donc une constante égale à l'espérance de Y quelque soit la grappe dans laquelle se trouve X_0.

- Si la copule $C^{(2)}$ converge vers la copule de la borne supérieure de Fréchet voir l'équation (1.5), le théorème 2.2, de la sous-section 2.7.4 donne un résultat dans le cas bivarié.

- Nous avons aussi le résultat sur l'espérance du prédicteur

  $$\mathbb{E}\left\{\mathbb{E}\left(Y_0|X_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)|X_0\right\} = \mathbb{E}(Y_0|X_0)\cdot$$

  En effet, $\mathbb{E}\left(Y_0|X_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ dépend de X_0 et donc en prenant l'espérance et en utilisant l'équation (2.25), nous avons

  $$\begin{aligned}\mathbb{E}\left\{\mathbb{E}\left(Y_0|X_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)|X_0\right\} &= \int_{\mathbb{R}^{d-1}} \left\{\int_{\mathbb{R}} y h_{X_0}(y)dy\right\} f(x)dx \\ &= \int_{\mathbb{R}} y g(y) \int_{\mathbb{R}^{d-1}} f(x) \times \\ &\quad \frac{c\left[\{F(x_1), G(y_1)\}, \ldots, \{F(x), G(y)\}\right]}{c_X\left[\{F(x_1), G(y_1)\}, \ldots, \{F(x_n), G(y_n)\}, F(x)\right]} dxdy,\end{aligned}$$

  Or

  $$\int_{\mathbb{R}^{d-1}} f(x) \frac{c\left[\{F(\mathbf{x}_1), G(y_1)\}, \ldots, \{F(x_n), G(y_n)\}, \{F(x), G(y)\}\right]}{c_X\left[\{F(x_1), G(y_1)\}, \ldots, \{F(x_n), G(y_n)\}, F(x)\right]} dx = 1\cdot$$

  Donc nous concluons que $\mathbb{E}\left\{\mathbb{E}\left(Y_0|X_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)|X_0\right\} = \mathbb{E}(Y_0|X_0)$.

Dans la suite de cette partie, nous présentons les résultats de l'évaluation du prédicteur dans le cas où la famille de copule $\left(C_{1,n}^{(3)}\right)$ est la copule normale.

2.7.2 Évaluation du prédicteur de Y_0 pour une copule $C_{1,n}^{(3)}$ normale échangeable

Dans cette section, nous présentons une évaluation du prédicteur donné à l'équation (2.27) dans le cas spécifique $d = 2$ où la famille de copule $\left(C_{1,n}^{(3)}\right)$ est une copule normale de paramètre

ρ_3 et les marginales F et G sont des lois normales, centrées et réduites. Les deux objectifs poursuivis sont :

— Évaluer théoriquement, l'espérance conditionnelle de Y_0 sachant x_0 et $\mathbf{z}_1, \ldots, \mathbf{z}_n$ notée $\mathbb{E}_{\delta_2}(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n)$. $\mathbb{E}_{\delta_2}$ est l'espérance indexée par le paramètre δ_2 de la copule $C^{(2)}$;

— Évaluer numériquement le prédicteur $\mathbb{E}_{\delta_2}(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n)$ pour quelques familles de copule $C^{(2)}$ puis construire les courbes de prédiction associées.

En partant de l'équation (2.27) dans le cas de la dimension $d = 2$ et en posant $w_i = \Phi^{-1}\left[C_{2|1}\left\{G(y_i)|F(x_i)\right\}\right]$, $i = 1, \ldots, n$, et le changement de variable nous donne

$$\begin{aligned}\mathbb{E}_{\delta_2}(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n) &= \int_{\mathbb{R}} G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(w)|F(x_0)\right\}\right]\phi(w) \\ &\times \frac{c_{1,n+1}^{(3)}\left\{\Phi(w_1), \ldots, \Phi(w_n), \Phi(w); \rho_3\right\}}{c_{1,n}^{(3)}\left\{\Phi(w_1), \ldots, \Phi(w_n); \rho_3\right\}} dw,\end{aligned}$$

où $C_{2|1}^{-1}$ est l'inverse de la fonction, obtenue à partir de (2.18) pour $d = 2$. La fonction de répartition conditionnelle $C_{2|1}$ varie en fonction du choix de la copule $C^{(2)}$ et ses expressions pour quelques cas particuliers de copule est référencée par le tableau 1.3. Nous rappelons le résultat Joe (2014, page 9) qui s'énonce : si $(W_1, \ldots, W_n, W)$ est de copule associée normale multivariée échangeable $\left(C_{1,n}^{(3)}\right)$ alors le vecteur aléatoire $\left\{\Phi^{-1}(W_1), \ldots, \Phi^{-1}(W_n), \Phi^{-1}(W)\right\}$ suit une loi normale multivariée de matrice de corrélation $\Sigma(n+1, \rho_3)$ échangeable de paramètre ρ_3. Nous rappelons que la matrice $\Sigma(n+1, \rho_3)$ est la matrice carrée d'ordre $n+1$ dont les termes de la diagonale sont 1 et les termes hors de la diagonale sont ρ_3. De plus, la variable $\Phi^{-1}(W)$ suit une loi normale centrée réduite et $\Phi^{-1}(W)$ sachant $\left\{\Phi^{-1}(w_1), \ldots, \Phi^{-1}(w_n)\right\}$ suit aussi une loi normale, voir Rivest *et al.* (2016), de moyenne μ_0 et de variance σ_0 définies par

$$\mu_0 = \frac{n\rho_3\overline{w}_n}{1 + (n-1)\rho_3}, \qquad \sigma_0^2 = \frac{(1-\rho_3)(1+n\rho_3)}{1+(n-1)\rho_3}, \tag{2.29}$$

et $\overline{w}_n$ est la moyenne de quantiles provenant de la normale centrée réduite définie par

$$\overline{w}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi^{-1}\left[C_{2|1}\left\{G(y_i)|F(x_i)\right\}\right]. \tag{2.30}$$

En utilisant ce résultat, nous avons donc

$$\phi(w)\frac{c_{1,n+1}^{(3)}\left\{\Phi(w_1), \ldots, \Phi(w_n), \Phi(w); \rho_3\right\}}{c_{1,n}^{(3)}\left\{\Phi(w_1), \ldots, \Phi(w_n); \rho_3\right\}} = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left\{-\frac{1}{2\sigma_0^2}(w-\mu_0)^2\right\}, \ w \in \mathbb{R}.$$

Le vecteur aléatoire $\left\{C_{2|1}\left\{G(Y_1)|F(X_1)\right\}, \ldots, C_{2|1}\left\{G(Y_n)|F(X_n)\right\}\right\}$ suit la loi multivariée de copule $\left(C_{1,n}^{(3)}\right)$. De plus, dans le cas où la taille n est grande, la variable aléatoire donnée par

$\sum_{i=1}^{n} \Phi^{-1}\left[C_{2|1}\left\{G(Y_i)|F(X_i)\right\}\right]/n$ converge vers la loi normale centrée de variance ρ_3.
La densité conditionnelle h_{x_0} de la variable Y_0 sachant x_0, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ est

$$h_{x_0}(y) = \frac{g(y)c^{(2)}\left\{F(x_0), G(y)\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\Phi^{-1}\left[C_{2|1}\left\{G(y)|F(x_0)\right\}\right] - \mu_0)^2\right\}}{\sqrt{2\pi}\sigma_0 \phi\left(\Phi^{-1}\left[C_{2|1}\left\{G(y)|F(x_0)\right\}\right]\right)}, \tag{2.31}$$

pour $y \in \mathbb{R}$ où μ_0 et σ_0 sont donnés par (2.29). En faisant un changement de variable $t = \left(\Phi^{-1}\left[C_{2|1}\left\{G(y)|F(x_0)\right\}\right] - \mu_0\right)/\sigma_0$, nous obtenons une forme réduite du prédicteur qui est

$$\mathbb{E}_{\delta_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \int_{\mathbb{R}} G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\mu_0 + \sigma_0 t)|F(x_0)\right\}\right]\phi(t)dt\cdot \tag{2.32}$$

Le prédicteur dépend de la taille n de la grappe et de $c^{(2)}$ via le paramètre μ_0. Dans les courbes de prédictions présentées plus bas, nous tenons compte de cette variation associée à μ_0 en faisant le choix du quantile de la loi normale centrée et de variance ρ_3 associée à la valeur de $\overline{w}_n$. La variance de l'erreur de prédiction

$$\begin{aligned}\mathbb{V}_{\delta_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) &= \int_{\mathbb{R}} \left\{G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\mu_0 + \sigma_0 t)|F(x_0)\right\}\right]\right\}^2 \phi(t)dt \\ &- \left\{\mathbb{E}_{\delta_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)\right\}^2\cdot\end{aligned} \tag{2.33}$$

Lorsque les marginales F et G sont des lois normales, centrées et réduites, le prédicteur s'écrit

$$\mathbb{E}_{\delta_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \int_{\mathbb{R}} \Phi^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\mu_0 + \sigma_0 t)|\Phi(x_0)\right\}\right]\phi(t)dt\cdot \tag{2.34}$$

C'est l'espérance conditionnelle de Y_0 sachant x_0, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ lorsque la famille de copules $\left(C_{1,n}^{(3)}\right)$ est une copule normale et les lois marginales de X et Y sont la loi normale standard.

Le programme sous le logiciel R, pour faire une évaluation numérique de l'intégrale de l'équation (2.34) en utilisant la méthode de Gauss-Hermite se trouve à l'annexe B.4.

Supposons que $C^{(2)}$ est une copule normale bivariée de paramètre ρ_2. Finalement, nous avons le résultat

$$\mathbb{E}_{\rho_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right) = \rho_2 x_0 + \frac{n\rho_3}{1+(n-1)\rho_3}(\overline{y}_n - \rho_2\overline{x}_n), \quad \mathbb{V}_{\rho_2}\left(Y_0|x_0, \mathbf{z}_1, \mathbf{z}_n\right) = (1-\rho_2^2)(1-\rho_3^2),$$

en remplaçant δ_2 par ρ_2. La partie $\frac{n\rho_3}{1+(n-1)\rho_3}(\overline{y}_n - \rho_2\overline{x}_n)$ étant considérée comme un résidu. Ce prédicteur est une droite et donc notre modèle généralise le cas particulier de modèle linéaire mixte.

2.7.3 Construction des courbes de prédiction dans quelques cas particuliers de la copule bivariée $C^{(2)}$

Nous utilisons le prédicteur obtenu à l'équation (2.34). Les résultats issus de l'évaluation sont présentés sur les graphiques pour observer l'effet de la grappe et de la dépendance entre les deux variables sur le modèle de prédiction. Nous rappelons que les lois marginales F et G sont la loi normale standard.

Les graphiques obtenus se basent sur les choix que nous énumérons.

- nous considérons des grappes ayant le même nombre $n = 20$ unités.
- nous choisissons la valeur de la corrélation régie par la copule $\left(C^{(3)}_{1,n}\right)$ de paramètre ρ_3 comme étant $\tau_3 = 0.1$ ($\rho_3 = 0.16$), pour marquer la faible importance de la corrélation résiduelle dans le modèle considéré.
- En notant $F_{\overline{W}_n}$, la fonction de répartition de $\overline{W}_n$, nous avons $F_{\overline{W}_n}(\overline{w}_n) = \Phi(\overline{w}_n/\sqrt{\rho_3})$. Les valeurs prises par $\overline{w}_n$ de l'équation (2.30) sont respectivement les quantiles $\overline{w}_n = -0.84$ (résidus faibles avec $F_{\overline{W}_n}(\overline{w}_n) = 0.018$), $\overline{w}_n = 0$ (résidus moyens avec $F_{\overline{W}_n}(\overline{w}_n) = 0.5$) et $\overline{w}_n = 0.84$ (résidus forts avec $F_{\overline{W}_n}(\overline{w}_n) = 0.982$).
- nous considérons trois tau de Kendall $\tau_2 = 0.3$ (faible corrélation), $\tau_2 = 0.6$ (moyenne corrélation) et $\tau_2 = 0.8$ (corrélation élevée) qui sont associés aux choix des copules $C^{(2)}$.

Pour construire lesdits graphiques à partir du tau de Kendall, nous calculons premièrement le paramètre de la copule spécifiée en utilisant la relation entre le paramètre de la copule et le tau de Kendall (méthode d'inversion du tau de Kendall avec la fonction *BiCopTau2Par* du package *VineCopula*). Cette valeur du paramètre de la copule s'utilise dans la formule (2.34) pour calculer la valeur prédite.

Nous présentons dans cette section, pour deux types de copules $C^{(2)}$ (Clayton et Frank), les courbes de prédiction. Les autres graphiques des courbes de prédiction se trouvent à l'annexe B.4 pour plusieurs choix de copules $C^{(2)}$ avec le programme R d'obtention des résultats.

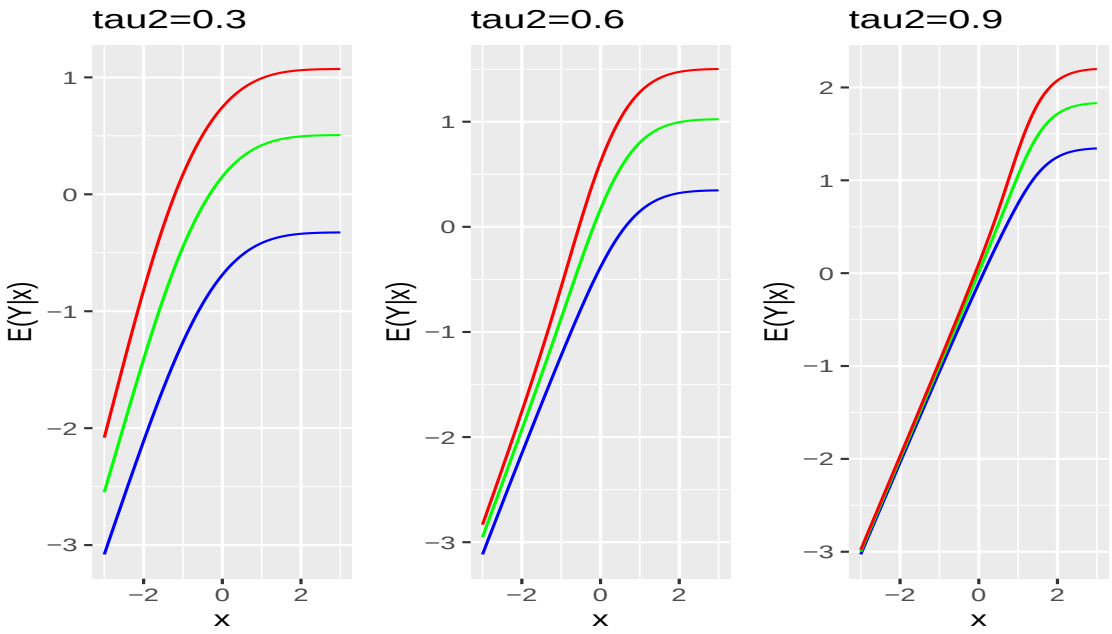


FIGURE 2.2 – Courbes de prédiction avec une copule de Clayton pour $C^{(2)}$ en fonction de l'importance des résidus $\overline{w}_n = -0.84$ (courbe bleue), $\overline{w}_n = 0$ (courbe verte) et $\overline{w}_n = 0.84$ (courbe rouge) en fixant $n = 20$.

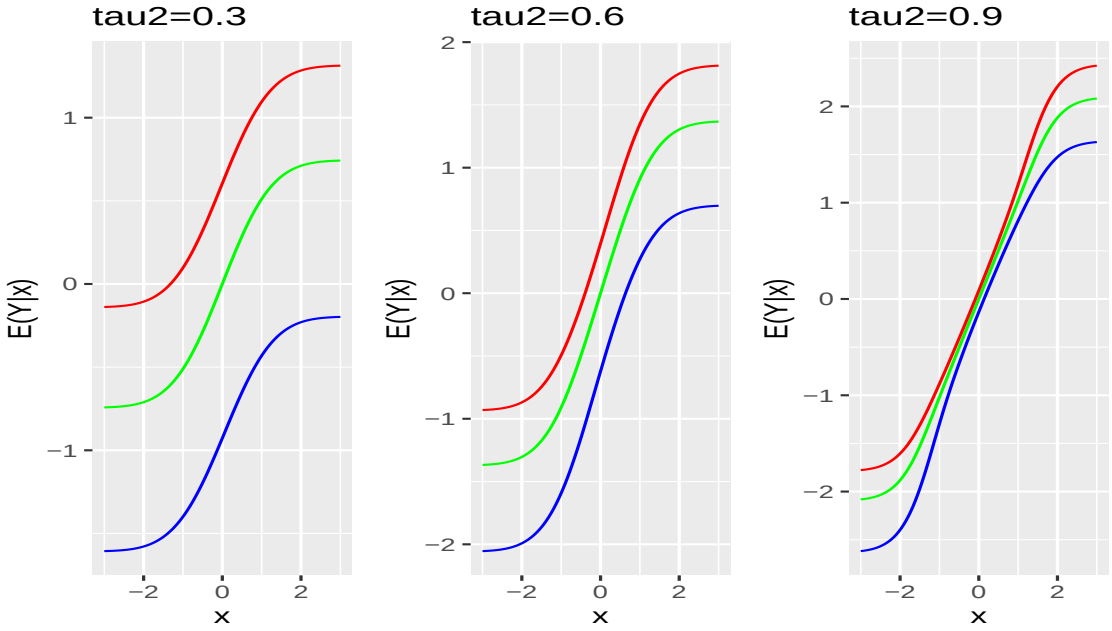


FIGURE 2.3 – Courbes de prédiction avec une copule de Frank en fonction de l'importance des résidus $\overline{w}_n = -0.84$ (courbe bleue), $\overline{w}_n = 0$ (courbe verte) et $\overline{w}_n = 0.84$ (courbe rouge) en fixant $n = 20$.

Globalement, ces graphiques montrent deux aspects des courbes de prédiction. Premièrement, le modèle échangeable modifie le parallélisme noté dans les courbes de prédiction du modèle de Battese *et al.* (1988). Plus précisément, les courbes de prédictions ne sont plus des droites parallèles mais des courbes curvilignes dépendantes de la copule $C^{(2)}$ qui tentent de capturer la non-linéarité. Deuxièmement, si la dépendance issue de la copule $C^{(2)}$ est forte, la courbe de prédiction de nouvelles observations prend la forme de la dépendance associée à celle-ci. En particulier, si la corrélation de la copule $C^{(2)}$ est forte et la copule considérée est une copule de Clayton alors la courbe de prédiction à une dépendance dans les petites valeurs pour n'importe quelles valeurs de dépendance associée à la famille de copules $\left(C_{1,n}^{(3)}\right)$.

2.7.4 Étude du cas limite du prédicteur du modèle de 2-copule échangeable

Le prédicteur de l'équation (2.32), obtenu à partir du modèle de 2-copule échangeable s'illustre aux figures 2.2 et 2.3 dans le cas où la copule $C^{(2)}$ est une copule de Clayton ou une copule de Frank. D'autres illustrations se trouvent à l'annexe B, section B.4. De ces différents graphiques, nous constatons que les courbes de prédictions tendent à se confondre et à se rapprocher de la droite d'équation $y = x$ au fur et à mesure que la corrélation τ_2 augmente et s'approche de 1. Nous rappelons que lorsque la dépendance convergence vers 1, la copule $C^{(2)}$ est la copule borne supérieure de Fréchet. Nous discutons théoriquement dans cette section de la convergence du prédicteur $\mathbb{E}_{\delta_2}\left(Y_0|x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n\right)$ lorsque la copule $C^{(2)}$ tend vers la copule borne supérieure de Fréchet (comonotonie). Pour ce faire, nous énonçons le théorème 2.2 qui résume le résultat.

Théorème 2.2. *Considérons les hypothèses **H1**, **H2** et **H3** définies par*

- ***H1 :*** *Les points $x_1, \ldots, x_n$, de $\mathbb{R}$ sont connus et fixés dans une grappe.*
- ***H2 :*** *Les lois marginales F et G sont absolument continues ;*
- ***H3 :*** *Pour tout x_i, il existe une suite de points $y_i^{(\delta_2)}$, dépendant de δ_2 associée à la copule $C^{(2)}$ par*

$$\lim_{\delta_2 \longrightarrow +\infty} \left\{y_i^{(\delta_2)}\right\} = G^{-1}\left\{F(x_i)\right\}, \ i = 1, \ldots, n,$$

et il existe $\varepsilon \in]0, 1[$ telle que

$$\varepsilon < \lim_{\delta_2 \longrightarrow +\infty} \left[\int_0^{G\left\{y_i^{(\delta_2)}\right\}} c^{(2)}\left\{F(x_i), t; \delta_2\right\} dt\right] < 1 - \varepsilon.$$

*Considérons la 2-copule échangeable en dimension n. Si la copule $C^{(2)}$ converge vers la borne supérieure de Fréchet alors sous les hypothèses **H1**, **H2** et **H3**, le prédicteur défini par l'équa-*

tion (2.32) *converge et nous avons*

$$\lim_{\delta_2 \longrightarrow +\infty} \mathbb{E}_{\delta_2}\left(Y_0 | x_0, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\right) = G^{-1}\left\{F(x_0)\right\}\cdot \tag{2.35}$$

Démonstration.
Nous voulons démontrer que sous les hypothèses ***H1***, ***H2*** et ***H3***, nous obtenons le résultat de l'équation (2.35).

Soit H_{x_0}, la fonction de répartition de la variable aléatoire Y_0 sachant $(x_0, \mathbf{z}_1, \ldots, \mathbf{z}_n)$, calculée à partir de la densité de l'équation (2.31). En utilisant le fait que si $C^{(2)}$ converge vers la borne supérieure de Fréchet (dépendance parfaite), donnée par l'équation (1.5), le tau τ_2 de Kendall associé à la copule converge vers 1 et nous avons

$$\lim_{\delta_2 \longrightarrow +\infty} C_{2|1}\left\{w | F(x_0)\right\} = 1, \quad F(x_0) < w \in [0,1] \backslash \left\{F(x_0)\right\}\cdot$$

De plus, pour la fonction inverse, nous avons

$$\lim_{\delta_2 \longrightarrow +\infty} \left[C_{2|1}^{-1}\left\{z | F(x_0)\right\}\right] = F(x_0), \quad z \in (0,1]\cdot$$

Si la copule $C^{(2)}$ converge vers la borne supérieure de Fréchet, la fonction de répartition H_{x_0} converge vers la fonction de répartition dégénérée H_0 donnée par

$$H_0(y) = \begin{cases} 1 & \text{si } y \geq G^{-1}\left\{F(x_0)\right\} \\ 0 & \text{sinon}\cdot \end{cases}$$

En utilisant l'hypothèse ***H3***, nous avons donc pour n fixé,

$$|\overline{w}_n| < \Phi^{-1}(1-\varepsilon) \implies |\mu_0| < \frac{n\rho_3 \Phi^{-1}(1-\varepsilon)}{1+(n-1)\rho_3}, \text{ pour } \delta_2 \longrightarrow \infty,$$

où μ_0 provient de l'équation (2.29). En partant de l'hypothèse ***H2***, la fonction définie par

$$w \longmapsto G^{-1}\left[F^{-1}\left\{\Phi(\mu_0 + \sigma_0 w) | F(x_0)\right\}\right]\phi(w),$$

est intégrable sur $[0,1]$ et σ_0 provient de l'équation (2.29). En utilisant la propriété d'interchangeabilité entre la limite et l'intégrale, l'équation (2.32) converge vers μ donnée par

$$\mu(x_0) \quad = \quad \int_{\mathbb{R}} \lim_{\delta_2 \longrightarrow \infty} \left(G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\mu_0 + \sigma_0 w) | F(x_0)\right\}\right]\right)\phi(w)dw = G^{-1}\left\{F(x_0)\right\}\cdot$$

En conséquence, la fonction μ est donnée par

$$\mu(x_0) = G^{-1}\left\{F(x_0)\right\}\cdot$$

Ceci conclut la démonstration du résultat. □

Ce résultat signifie que dans le cas d'une dépendance comonotone entre X et Y, les courbes de prédiction de Y sachant X, données par l'équation (2.27) sont identiques pour toute les grappes et la fonction de prédiction est exactement l'équation (2.35).

2.7.5 Calcul du prédicteur à partir du modèle de 2-copule échangeable pour $d = 2$ dans un autre cas particulier de copules $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$

Dans cette section, nous considérons le cas où $n = 2$ et les copules bivariées $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules FGM de paramètres respectifs θ_2 et θ_3, voir le tableau 1.1 pour la copule FGM. Les lois marginales F et G sont normales. Nous considérons que nous avons des données groupées sur lesquels deux variables sont mesurées $d = 2$. Notons (X_1, Y_1) les valeurs observées pour (X, Y) d'un individu. Nous supposons connu X_2, la valeur observée pour X d'un nouvel individu et on utilise le modèle se basant sur la 2-copule échangeable en dimension 2 pour prédire Y_2 pour cet individu. Nous notons F et G, les fonctions de répartitions de X et Y respectivement qui sont normales, centrées et réduites.

L'expression du prédicteur de Y_2 sachant X_1, Y_1 et X_2, à partir de l'équation (2.27) est

$$\begin{aligned}\mathbb{E}(Y_2|X_1, Y_1, X_2) &= \frac{\theta_2 \{2\Phi(X_2) - 1\}}{\sqrt{\pi}} \\ &\quad + \theta_3 \left[1 - 2\Phi(Y_1) \left\{1 + C_{2|1} \{\Phi(Y_1)|\Phi(X_1)\}\right\}\right] k(X_2) \qquad (2.36)\end{aligned}$$

où la fonction k est

$$k(X_2) = \left[-\frac{1}{\sqrt{\pi}} + \theta_2 \{2\Phi(X_2) - 1\} \left\{\frac{1}{\sqrt{\pi}} + 0.2\theta_2(1 - 2\Phi(X_2))\right\}\right] \cdot$$

Voir l'annexe B.2, pour plus de détail dans les calculs. Cette expression du prédicteur dépend de X_2, des valeurs des variables sur les autres individus et des résidus représentés par $C_{2|1} \{\Phi(Y_1)|\Phi(X_1)\} = \Phi(Y_1) [1 + \theta_2 \{1 - \Phi(Y_1)\} \{1 - 2\Phi(X_1)\}]$. Nous étudions, graphiquement, l'effet des résidus sur la valeur prédite. En considérant les cas particuliers des valeurs de $C_{2|1} \{\Phi(Y_1)|\Phi(X_1)\}$, nous construisons trois courbes pour situer leur effet, du petit au plus grand résidu sur la valeur prédite de Y_2.

L'équation (2.36), se décompose en deux parties, une qui donne l'effet sur la prédiction de X_2 par $\theta_2 \{2\Phi(X_2) - 1\} / \sqrt{\pi}$. L'autre partie donne l'effet de la grappe qui est la fonction croissante k. En faisant une petite comparaison avec la régression copule du chapitre 1, équation (1.30), nous avons les équivalences suivantes

$$\theta_3 \left[1 - 2\Phi(Y_1) \left\{1 + C_{2|1} \{\Phi(Y_1)|\Phi(X_1)\}\right\}\right] k(X_2) \equiv \mu \cdot \qquad (2.37)$$

et

$$\sigma \frac{\theta_2}{\sqrt{\pi}} \{2F_1(X_1) - 1\} \equiv \frac{\theta_2 \{2\Phi(X_2) - 1\}}{\sqrt{\pi}}, \qquad (2.38)$$

où $\mu = 0$, $\sigma = 1$ et F_1 est la fonction de probabilité de la loi normale centrée réduite.

Nous présentons le graphique avec les trois courbes de prédiction.

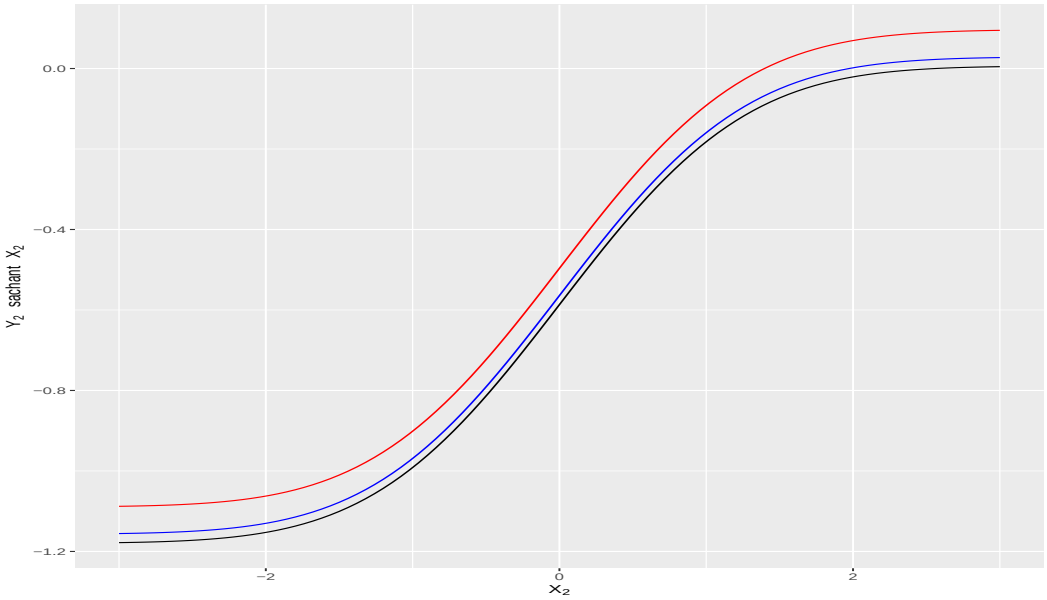


FIGURE 2.4 – Graphiques du prédicteur pour $\theta_2 = 0.7$, pour $\theta_3 = 0.2$ en fixant les valeurs de $C_{2|1}\{\Phi(Y_1)|\Phi(X_1)\} = 0.4$ (Courbe en noire) et $C_{2|1}\{\Phi(Y_1)|\Phi(X_1)\} = 0.8$ (Courbe en rouge) puis de l'indépendance des résidus ($\theta_3 = 0$, courbe en bleue).

Ceci donne un aperçu que plus les résidus sont élevés, plus le prédicteur change de comportement (forme croissante). Nous concluons à travers ce graphique que plusieurs types de copules impliquées peuvent être considérés théoriquement, faisant varier le modèle de dépendance que nous cherchons à modéliser dans les données.

2.8 Conclusion

Dans ce chapitre 2, nous proposons la construction d'un modèle échangeable nommé «d-copule échangeable» pour des données hiérarchiques et intégrant trois propriétés : (1) l'échangeabilité (2) l'indépendance conditionnelle et (3) la compatibilité entre deux copules. Nous avons aussi étudié les propriétés de ce modèle. Par la suite, nous montrons que ce modèle généralise le modèle de Battese, Harter et Fuller, voir Battese *et al.* (1988). Nous construisons un modèle de prédiction avec le cas particulier de la 2-copule échangeable tout en évaluant numériquement et graphiquement l'équation de prédiction.

Dans la suite de cette thèse, nous étudions dans un premier temps la procédure d'ajustement de la 2-copule échangeable, l'estimation des paramètres et les théories asymptotiques pour un modèle paramétrique.

Chapitre 3

Inférence paramétrique pour un modèle de 2-copule échangeable

Dans ce chapitre, les contributions majeures que nous présentons sont :

— La procédure séquentielle du choix des composantes d'un modèle de 2-copule échangeable.

— La méthode d'estimation IFM généralisée des paramètres du modèle à partir de celle existante.

3.1 Présentation des données en grappes et formulation du modèle de 2-copule échangeable

Nous considérons un échantillon de m grappes où nous avons n_j unités dans la grappe j, $j = 1, \ldots, m$. Pour chaque unité de l'échantillon, on a observé deux variables aléatoires continues notées X et Y. Nous supposons que X est la variable explicative et Y est la variable d'intérêt. On adopte les notations suivantes pour les observations en grappes.

- Le vecteur d'observations de l'individu i, de la grappe j, $i = 1, \ldots, n_j$ et $j = 1, \ldots, m$ est noté $(x_{ji}, y_{ji})^T$.
- $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn_j})^T$, et $\mathbf{y}_j = (y_{j1}, \ldots, y_{jn_j})^T$ sont deux vecteurs d'observations de la grappe j.
- L'ensemble des observations des m grappes est donc $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)$.

Les fonctions de répartitions marginales sont notées F et G pour X et Y respectivement avec f et g, les densités associées à ces variables aléatoires.

Nous supposons que la loi jointe des variables aléatoires dans une grappe est spécifiée à partir d'une 2-copule échangeable (voir la définition 2.4) impliquant des distributions marginales F et G. Plus spécifiquement, la fonction de densité h conjointe des $2 \times n_j$ variables dans la grappe j s'écrit

$$\begin{aligned} h(\mathbf{x}_j, \mathbf{y}_j) \quad = \quad & c^{(1)}_{1,n_j} \left\{ F(x_{j1};\alpha), \ldots, F(x_{jn_j};\alpha); \delta_1 \right\} \times \\ & \prod_{i=1}^{n_j} f(x_{ji};\alpha) g(y_{ji};\beta) c^{(2)} \left\{ F(x_{ji};\alpha), G(y_{ji};\beta); \delta_2 \right\} \times \\ & c^{(3)}_{1,n_j} \left[C_{2|1} \left\{ G(y_{j1};\beta) | F(x_{j1};\alpha) \right\}, \ldots, C_{2|1} \left\{ G(y_{jn_j};\beta) | F(x_{jn_j};\alpha) \right\}; \delta_3 \right], \end{aligned} \tag{3.1}$$

avec $C_{2|1}$, la distribution conditionnelle donnée par

$$C_{2|1} \left\{ v|u \right\} = \partial C^{(2)}(u, v; \delta_2) / \partial u \cdot$$

Nous expliquons les composantes du modèle de 2-copule échangeable.

- Les fonctions de répartition marginales F et G appartiennent à des familles paramétriques de paramètres respectifs α et β de dimensions supérieures ou égales à 1.
- La copule bivariée $C^{(2)}$ appartient à la classe de copules bivariées paramétriques de paramètres δ_2.
- Les familles $\left(C^{(1)}_{1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$, pour n, un entier non nul, appartiennent à des familles de copules échangeables paramétriques, de dimension n de paramètres δ_1 et δ_3 respectivement.

Nous considérons que le modèle de 2-copule échangeable est complètement paramétrique. En somme, le vecteur des paramètres impliqués dans le modèle se note $\theta = (\alpha, \beta, \delta_1, \delta_2, \delta_3)$ appartenant à un ensemble de dimension p.

Dans ce chapitre, nous tentons de répondre à trois objectifs :

(1) expliquer la procédure du choix des composantes des deux lois marginales F, G et les trois copules $\left(C^{(1)}_{1,n}\right)$, $C^{(2)}$ et $\left(C^{(3)}_{1,n}\right)$ de l'équation (3.1).

(2) présenter deux méthodes d'estimation des paramètres θ du modèle de 2-copule échangeable : la méthode de maximisation de la vraisemblance globale et la méthode IFM généralisée s'inspirant de la méthode IFM proposée par Joe et Xu (1996, page 229) et

(3) étudier les propriétés asymptotiques et à échantillons finis des estimateurs pour des paramètres du modèle de 2-copule échangeable, obtenus par les deux méthodes d'estimation.

3.2 Procédure de choix des différentes composantes du modèle de 2-copule échangeable

Le modèle de 2-copule échangeable de l'équation (3.1) nécessite le choix des familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$, de la copule bivariée $C^{(2)}$ et des fonctions marginales F et G. Nous présentons ici la procédure du choix de chacune de ces composantes. Globalement, la procédure pour faire le choix de chaque composante se décline en deux points. Le premier point est une étape graphique dont l'objectif est de partir des représentations graphiques sur des valeurs observées pour proposer des composantes candidates pour l'élément recherché. Le deuxième point est une étape analytique qui consiste à sélectionner la composante ayant la plus petite valeur de l'AIC parmi les candidates proposées au premier point.

Nous rappelons que pour l'ajustement d'un modèle $\mathcal{M}$ ayant un nombre p de paramètres et dont la fonction de la pseudo-vraisemblance est $\mathcal{L}$, le critère AIC est

$$\mathrm{AIC}(\mathcal{M}, p) = -2\log(\mathcal{L}) + 2p, \tag{3.2}$$

voir Taniguchi et Hirukawa (2012) et Joe (2014, page 16) qui donnent des informations sur la calcul l'AIC pour l'ajustement d'un modèle de copules.

Pour un modèle associé à une famille de copules échangeables notée $\left(C_{1,n_j}\right)_{j=1}^{m}$ de paramètre θ. L'estimateur $\hat{\theta}$ de θ s'obtient grâce à une méthode d'estimation. Le critère AIC se définit par Czado (2019, Chapitre 8) sur des données $u = (u_{ji}), j = 1, \ldots, m, i = 1, \ldots, n_j$ ou des pseudo-observations $\tilde{u} = (\tilde{u}_{ji})$ par

$$\mathrm{AIC}(C_\theta, j = 1, \ldots, m, \hat{\theta}, u) = -2\sum_{j=1}^{m} \log\left\{c_{1,n_j}\left(\tilde{u}_{j1}, \ldots, \tilde{u}_{jn_j}; \hat{\theta}\right)\right\} + 2\mathrm{dimension}(\hat{\theta})\cdot \tag{3.3}$$

Cette formulation de l'AIC est applicable pour les fonctions de répartition marginales et pour la copule bivariée du modèle de 2-copule échangeable de l'équation (3.1).

Pour la suite de la section, nous présentons dans un premier temps, l'identification des lois marginales, la première étape de la construction du modèle.

3.2.1 Identification des fonctions de répartitions marginales

Les lois marginales F et G sont associées respectivement aux données (x_{ji}) et (y_{ji}), $i = 1, \ldots, n_j$ et $j = 1, \ldots, m$. Nous faisons l'hypothèse dans la construction du modèle de 2-copule échangeable que les lois marginales F et G appartiennent des familles de loi paramétrique. Plusieurs stratégies existent pour sélectionner une loi paramétrique univariée d'une variable aléatoire, Cousineau *et al.* (2004). Un choix de lois paramétriques s'implémente par exemple

dans le package `fitdistr`, Delignette-Muller et Dutang (2015). Nous présentons ici une des stratégies qui s'exécute en plusieurs points pour la marginale F.

Pour déterminer la loi F associée à la variable X, nous suggérons les étapes suivantes.

Étape 1 construire le graphique de la fonction de répartition empirique ou le graphe de la densité des données (x_{ji}). Une présentation graphique des données est généralement l'une des premières étapes d'une analyse exploratoire des données.

Étape 2 construire le graphique des lois paramétriques théoriques imaginées puis comparer le graphique fait à l'étape 1 sur les données observées à celui fait pour chaque loi théorique pour proposer des lois candidates. À cette étape, on peut tester, en premier lieu, la loi normale. Si les données sont à support borné, il faut en tenir compte dans la comparaison avec le graphique observé et proposer des lois tronquées sur le domaine au besoin. La comparaison est basée sur la forme du graphique fait à l'étape 1 et le graphique théorique des distributions considérées. Pour construire le graphique théorique, il faut estimer d'abord le vecteur de paramètres α de la loi paramétrique F en maximisant la vraisemblance. Cette vraisemblance se calcule en ne tenant pas compte de la dépendance à l'intérieur des grappes. Pour le vecteur de paramètres α, il est estimé par

$$\tilde{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left[\mathcal{L}_\alpha = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log \left\{ f(x_{ji}; \alpha) \right\} \right]. \tag{3.4}$$

Étape 3 calculer le critère AIC des lois candidates proposées à l'étape 2 pour sélectionner la loi ayant la plus petite valeur.

Les étapes 1, 2 et 3 nous permettent d'identifier la loi marginale F, utilisée dans la suite de notre procédure.

Pour la loi marginale G, on applique la même procédure sur les données (y_{ji}) et le vecteur de paramètres β de la loi G s'estime par

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\mathcal{L}_\beta = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log \left\{ g(y_{ji}; \beta) \right\} \right]. \tag{3.5}$$

Le livre de Thomopoulos (2017) et sa version complétée Thomopoulos (2018) sont pertinents pour sélectionner une loi de probabilité paramétrique qui répond à ce type de problème. En effet, dans ce livre, les auteurs parcourent les caractérisations de plusieurs distributions paramétriques : des lois à support infini et des lois à support borné. Une autre référence en matière de distribution à support borné est le livre de Karian et Dudewicz (2000, chapitres 2 et 3). Les auteurs présentent la généralisation des distributions paramétrique et une extension des distributions bêta. Nous ajustons une fonction de répartition en utilisant un programme

du logiciel ***R***, R Development Core Team (2008) provenant du package *fitdistrplus*, Delignette-Muller et Dutang (2015).

Si les observations sont indépendantes et identiquement distribuées, les équations (3.4) et (3.5) définissent des estimateurs du maximum de vraisemblance qui sont convergents, Casella et Berger (2002). Les équations (3.4) et (3.5) définissent dans ce cas des vraisemblances complètes du fait que la dépendance intra grappe est ignorée. Dans notre cas, on note une dépendance entre les observations au sein d'une grappe. Nous étudions donc dans la section 3.4, les propriétés asymptotiques des estimateurs issus de cette maximisation dite par étape, expliquées par la suite.

À la fin de l'identification des lois marginales F et G, nous construisons des pseudo-observations associées à toutes les observations des grappes et définies par

$$\tilde{u}_{ji} = F(x_{ji};\tilde{\alpha}), \qquad \tilde{v}_{ji} = G(y_{ji};\tilde{\beta}),\ i = 1,\ldots,n_j,\ j = 1,\ldots,m\cdot \tag{3.6}$$

Ces pseudo-observations $(\tilde{u}_{ji},\tilde{v}_{ji})$ s'utilisent pour caractériser la dépendance entre X et Y représentée par la copule $C^{(2)}$. Dans la suite, nous présentons la procédure pour faire le choix de la copule bivariée $C^{(2)}$ et des familles de copules échangeables $\left(C^{(1)}_{1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$.

3.2.2 Choix des deux familles de copules échangeables et de la copule bivariée du modèle de 2-copule échangeable

Nous expliquons comment faire le choix de la copule bivariée $C^{(2)}$ associée à (X,Y) et des familles de copules échangeables $\left(C^{(1)}_{1,n}\right)$ et $\left(C^{(3)}_{1,n}\right)$. La famille de copules échangeables $\left(C^{(1)}_{1,n}\right)$ est associée à n variables aléatoires. La dimensionnalité de cette famille change en fonction de la taille n, nombre d'unité dans la grappe. La famille de copules échangeables $\left(C^{(3)}_{1,n}\right)$ est associée aux pseudo-observations résiduelles que nous définissons. Nous détaillons pour chacune d'elle, la procédure de sélection en nous basant sur une approche à la fois descriptive et analytique.

• Choix de la copule bivariée $C^{(2)}$ associée au vecteur (X,Y)

La copule bivariée $C^{(2)}$ se construit à partir des pseudo-observations $(\tilde{u}_{ji},\tilde{v}_{ji})$, $i = 1,\ldots,n_j$ et $j = 1,\ldots,m$. Pour identifier la copule bivariée $C^{(2)}$, nous proposons la stratégie suivante en quelques points.

(a) Faire une représentation graphique du nuage de points $(\tilde{u}_{ji},\tilde{v}_{ji})$ sur le carré unité.

(b) Restreindre le choix des copules bivariées à des copules candidates en comparant le graphique construit à la représentation graphique des copules bivariées théoriques. Autrement dit,

proposer des lois candidates à partir de la représentation graphique. Pour la représentation de la copule bivariée théorique candidate, il faut estimer le vecteur de paramètres δ_2 associé, en maximisant la pseudo log-vraisemblance et l'estimateur est donné par

$$\tilde{\delta}_2 \quad = \quad \underset{\delta_2}{\operatorname{argmax}} \left[\mathcal{L}_2 = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log \left\{ c^{(2)} \left(\tilde{u}_{ji}, \tilde{v}_{ji}; \delta_2 \right) \right\} \right] . \tag{3.7}$$

L'estimateur de δ_2, obtenue via cette maximisation dépend, des paramètres $\tilde{\alpha}$ et $\tilde{\beta}$. Pour cette étape de proposition de copules bivariées possibles, Joe (2014, chapitre 5, page 223) recommande aussi d'analyser la dépendance caudale (asymétrique, symétrie réflexive) en calculant les corrélations suivantes et en les comparant deux à deux. Nous considérons deux variables aléatoires U et V suivant des lois uniformes.

$$\operatorname{cor}(U, V | U \geq 1/2, V \geq 1/2), \;\; \operatorname{cor}(U, V | U \leq 1/2, V \leq 1/2),$$

et

$$\operatorname{cor}(U, V | U \leq 1/2, V \geq 1/2), \;\; \operatorname{cor}(U, V | U \geq 1/2, V \leq 1/2) \cdot$$

Ces corrélations permettent de comprendre si les copules ayant de la dépendance asymétrique ou radiale sont à inclure dans l'ensemble des copules candidates, voir Chang et Joe (2020).

(c) Faire un choix unique parmi les copules de la classe de copules bivariées candidates en minimisant le critère AIC, voir (3.2).

La copule bivariée sélectionnée est celle ayant le petit AIC. Une référence aux méthodes généralement utilisées pour ajuster une copule bivariée comme $C^{(2)}$ est Joe (2014, chapitre 5, page 223).

Dans la section 3.4, nous précisons le comportement asymptotique de l'estimateur du vecteur de paramètres associé à la copule bivariée $C^{(2)}$. Les prochaines composantes sont les deux familles de copules échangeables $\left(C_n^{(1)}\right)$ et $\left(C_n^{(3)}\right)$ dont nous expliquons les choix.

• Choix des copules échangeables associées respectivement aux pseudo-observations dans la même grappe et aux pseudo-observations résiduelles

Il s'agit d'identifier les familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$. Pour faire le choix de $\left(C_{1,n}^{(1)}\right)$, on utilise les pseudo-observations $(\tilde{u}_{j1}, \ldots, \tilde{u}_{jn_j})$ avec $j = 1, \ldots, m$. Nous présentons la procédure du choix de $\left(C_{1,n}^{(1)}\right)$ dans un premier temps, en utilisant l'approche expliquée par Rivest *et al.* (2016) et se déroulant en deux étapes.

Méthode graphique d'identification de familles de copule échangeable : La première chose à faire est de classer les m grappes en fonction de la médiane des pseudo-observations $\tilde{u}_{ji}$ calculés sur les grappes. Si on note Me, la médiane d'une grappe, nous avons le classement $\text{Me}_{(1)} < \ldots < \text{Me}_{(m)}$, où $\text{Me}_{(i)}$ représente la $i$ème statistique d'ordre de l'ensemble des médianes des grappes. Nous calculons par la suite les scores normaux des pseudo-observations $(\tilde{u}_{ji})$ qui est $\Phi^{-1}(\tilde{u}_{ji})$. Nous construisons sur un même graphique les boîtes à moustache des scores normaux des grappes classées en ordre croissant des médianes comme expliquées précédemment. Ce graphique, nous l'appelons *graphique observé* que nous comparons à des *graphiques simulés* pour certaines familles échangeables. Cette comparaison permet de proposer des copules candidates. Pour les données observées, on calcule aussi le tau de Kendall échangeable $\hat{\tau}_E$, voir Romdhani *et al.* (2014a).

Pour faire la simulation de données selon une copule échangeable $\left(C_{1,n}^{(1)}\right)$, de dimension n, nous calculons d'abord le vecteur de paramètres δ_1 de la copule candidate $\left(C_{1,n}^{(1)}\right)$. Cette simulation utilise des fonctions intégrées du logiciel **R**, R Development Core Team (2008) avec le package *copula*, Wu *et al.* (2007). Nous avons deux éventualités de calcul de δ_1 :

- Si le vecteur de paramètres δ_1 est de dimension 1, on inverse simplement le tau de Kendall échangeable $\hat{\tau}_E$ en fonction de la copule échangeable spécifiée en utilisant le tableau 1.2 pour quelques cas de copules. Ceci s'explique par le fait que le tau de Kendall de l'équation (1.23) et le tau de Kendall échangeable sont équivalents, voir Romdhani *et al.* (2014a) et Romdhani *et al.* (2014b). Nous utilisons la fonction `itau` sous le logiciel **R** par exemple ;

- En général, si le vecteur de paramètres δ_1 est de dimension supérieure ou égale à 1, alors nous l'estimons à partir des données des m grappes par

$$\tilde{\delta}_1 \quad = \quad \underset{\delta_1}{\text{argmax}} \left[\mathcal{L}_1 = \sum_{j=1}^{m} \log \left\{ c_{1,n_j}^{(1)} \left(\tilde{u}_{j1}, \ldots, \tilde{u}_{jn_j}; \delta_1 \right) \right\} \right]. \tag{3.8}$$

La log-vraisemblance $\mathcal{L}_1$ dépend de $\tilde{u}_{ji}$, dépendant à son tour de $\tilde{\alpha}$, estimé à l'étape d'ajustement de la fonction de répartition F. Le vecteur de paramètres estimé $\tilde{\delta}_1$ permet de simuler des données de taille n pour chaque grappe et dont la dépendance se modélise à l'aide de la famille de copules échangeables $\left(C_{1,n}^{(1)}\right)$.

En bref, pour proposer des familles de copules échangeables comme candidates pour la famille $\left(C_{1,n}^{(1)}\right)$, il faut tester chacune des copules échangeables connues et plausibles si la boîte à moustache par grappe obtenue théoriquement par simulation est proche du *graphique observé*. Pour faire cela, nous

(★) simulons des observations par grappe (en fonction de sa taille) selon une copule échangeable existante et fixée et ayant pour paramètres $\tilde{\delta}_1$;

(★★) construisons le même graphique des données simulées (boîte à moustache des scores normaux, classée par ordre des médianes des grappes) que nous appelons *graphique simulé* puis

(★★★) comparons les *graphiques observés* et *simulés* en termes de ressemblance basée sur la forme de la représentation.

Les trois points consécutifs (★), (★★) et (★★★) nous permettent de tirer la conclusion si cette copule échangeable simulée est une copule candidate.

Cette première étape constitue la méthode graphique permettant d'identifier un nombre fini k_0 de familles de copules échangeables candidates.

Méthode analytique du choix d'une famille de copules échangeables : Nous utilisons le critère AIC pour sélectionner une copule échangeable parmi un ensemble de copules candidates. Pour chaque copule candidate identifiée, nous calculons le critère AIC, voir (3.3). On obtient donc les valeurs $\text{GAIC}_1, \ldots, \text{GAIC}_{k_0}$ où GAIC_i est la valeur de l'AIC du modèle $i \in \{1, \ldots, k_0\}$. Nous comparons les valeurs de AIC, calculées pour identifier la copule échangeable associée à la petite valeur de GAIC et correspondant exactement à $\min(\text{GAIC}_1, \ldots, \text{GAIC}_{k_0})$.

La sélection de la famille de copules échangeables $\left(C^{(3)}_{1,n}\right)$ vient après l'identification de la copule $C^{(2)}$. En effet, elle est associée aux pseudo-observations résiduelles $(\tilde{w}_{ji})\,,\ i = 1, \ldots, n_j$, $j = 1, \ldots, m$ définis à partir de $C^{(2)}$ par

$$\tilde{w}_{ji} = C_{2|1}(\tilde{v}_{ji}|\tilde{u}_{ji}) = \int_0^{\tilde{v}_{ji}} c^{(2)}(\tilde{u}_{ji}, z; \tilde{\delta}_2)dz\cdot \tag{3.9}$$

Après avoir calculé les pseudo-observations résiduelles $\tilde{w}_{ji}, i = 1, \ldots, n_j,\ j = 1, \ldots, m$, on applique la même méthode de recherche de familles de copules échangeables, expliquées précédemment pour trouver la famille de copules échangeables $\left(C^{(3)}_{1,n}\right)$ où $(\tilde{u}_{ji})$ est remplacé par $(\tilde{w}_{ji})$ de l'équation (3.9). Le vecteur de paramètres δ_3 de la famille de copules $\left(C^{(3)}_{1,n}\right)$ est estimé en maximisant la pseudo log-vraisemblance et nous avons

$$\tilde{\delta}_3 \quad = \quad \underset{\delta_3}{\text{argmax}} \left[\mathcal{L}_3 = \sum_{j=1}^{m} \log\left\{c^{(3)}_{1,n_j}(\tilde{w}_{j1}, \ldots, \tilde{w}_{jn_j}; \delta_3)\right\}\right]. \tag{3.10}$$

Le choix de la famille de copules échangeables se base sur la méthode Rivest *et al.* (2016). Elle consiste à répéter l'étape d'identification graphique de chaque famille de copules échangeables.

Pour ajuster et estimer le vecteur de paramètres d'une famille de copules échangeables sur des données en grappes, Grover *et al.* (2020) ont proposé de maximiser la pseudo log-vraisemblance. Ils démontrent les propriétés asymptotiques de normalité des estimateurs, voir aussi Prenen *et al.* (2017) et Su et Lin (2019).

D'une part, le choix des familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont proposés dans la littérature comme Rivest *et al.* (2016) et Grover *et al.* (2020). D'autre part, le choix de la copule bivariée référencée par Nelsen (2017, chapitre 8) et Joe (2014, chapitre 5). La procédure de construction, de choix et d'estimation des paramètres du modèle de la 2-copule échangeable suit la procédure décrite à l'algorithme 3.1.

TABLEAU 3.1 – Procédure d'ajustement étape par étape d'un modèle de 2-copule échangeable.

Procédure chronologique de construction du modèle de 2-copule échangeable

- Identifier les lois marginales F et G et estimer les vecteurs de paramètres α et β respectivement associés, à partir des équations (3.4) et (3.5).
- Calculer les pseudo-observations de l'équation (3.6) à partir des lois marginales.
- Choisir la famille de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et estimer le vecteur de paramètres δ_1 à partir de l'équation (3.8).
- Choisir la copule $C^{(2)}$ et estimer le vecteur de paramètres δ_2 à partir de l'équation (3.7).
- Calculer les pseudo-observations résiduelles de l'équation (3.9) à partir de la copule bivariée $C^{(2)}$.
- Choisir la famille de copules échangeables $\left(C_{1,n}^{(3)}\right)$ puis estimer le vecteur de paramètres δ_3 à partir de l'équation (3.10).

La procédure d'ajustement étant présentée, nous nous intéressons aux méthodes d'estimation des paramètres du modèle de 2-copule échangeable.

3.3 Estimation des paramètres du modèle de 2-copule échangeable

Dans cette section, nous présentons en résumé deux méthodes d'estimation du modèle de 2-copule échangeable.

3.3.1 Estimation des paramètres par la méthode IFM généralisée

L'estimation $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3)$ appartenant à un ensemble Θ, de dimension p par la méthode IFM généralisée s'obtient en résolvant l'équation d'estimation

$$\left(\partial \mathcal{L}_\alpha / \partial \alpha^T, \partial \mathcal{L}_\beta / \partial \beta^T, \partial \mathcal{L}_1 / \partial \delta_1^T, \partial \mathcal{L}_2 / \partial \delta_2^T, \partial \mathcal{L}_3 / \partial \delta_3^T\right) = \mathbf{0}^T. \tag{3.11}$$

Les différentes étapes de la procédure proposée pour l'ajustement du modèle de 2-copule échangeable est une version généralisée de la méthode IFM classique présentée par Joe (1997). Elles sont similaires à un résultat de Joe (2014, page 229) pour la méthode d'estimation par étape pour l'ajustement d'un modèle de copule. C'est ce qui explique que nous la nommons "méthode IFM généralisée". En effet, si les familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules d'indépendance alors la méthode IFM généralisée est la méthode IFM classique. À partir de l'équation (3.11), la solution $\tilde{\theta}$ proposée se ramène à

$$\sum_{j=1}^{m} \psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \tilde{\theta}) = \mathbf{0}^T, \tag{3.12}$$

où pour $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$ et $\psi(\mathbf{x}, \mathbf{y}, n; \theta) = \left(\psi_1^{(n)}, \psi_2^{(n)}, \psi_3^{(n)}, \psi_4^{(n)}, \psi_5^{(n)}\right)^T$. La fonction score à cinq composantes qui se définissent par :

$$\psi_1^{(n)} = \sum_{i=1}^{n} \frac{\partial \log\{f(x_i;\alpha)\}}{\partial \alpha^T}, \qquad \psi_2^{(n)} = \sum_{i=1}^{n} \frac{\partial \log\{g(y_i;\beta)\}}{\partial \beta^T}, \tag{3.13}$$

et

$$\psi_3^{(n)} = \frac{\partial \log\left[c_{1,n}^{(1)}\{F(x_1;\alpha), \ldots, F(x_n;\alpha); \delta_1\}\right]}{\partial \delta_1^T}, \tag{3.14}$$

$$\psi_4^{(n)} = \sum_{i=1}^{n} \frac{\partial \log\left[c^{(2)}\{F(x_i;\alpha), G(y_i;\beta); \delta_2\}\right]}{\partial \delta_2^T}, \tag{3.15}$$

et

$$\psi_5^{(n)} = \frac{\partial \log\left[c_{1,n}^{(3)}\left\{C_{2|1}\{G(y_1;\beta)|F(x_1;\alpha)\}, \ldots, C_{2|1}\{G(y_n;\beta)|F(x_n;\alpha)\}; \delta_3\right\}\right]}{\partial \delta_3^T}. \tag{3.16}$$

L'équation (3.12) définie un m-estimateur, Stefanski et Boos (2002). Lorsque le nombre de grappes tend vers l'infini, la section 3.4 donne les résultats asymptotiques. Par rapport aux composantes de la fonction score associée à cette maximisation, nous avons le résultat suivant

Propriété 3.1. *Si* $\left(\psi_1^{(n)}, \psi_2^{(n)}, \psi_3^{(n)}, \psi_4^{(n)}, \psi_5^{(n)}\right)$ *est issu de la fonction score de l'équation* (3.12) *et provenant de la méthode IFM généralisée du modèle de 2-copule échangeable alors nous avons*

$$cov(\psi_1^{(n)}, \psi_3^{(n)}) = 0, \quad cov(\psi_1^{(n)}, \psi_5^{(n)}) = 0, \tag{3.17}$$

$$cov(\psi_2^{(n)}, \psi_3^{(n)}) = 0, \quad cov(\psi_2^{(n)}, \psi_5^{(n)}) = 0 \tag{3.18}$$

et

$$cov(\psi_3^{(n)}, \psi_4^{(n)}) = 0, \quad cov(\psi_3^{(n)}, \psi_5^{(n)}) = 0, \quad cov(\psi_4^{(n)}, \psi_5^{(n)}) = 0\cdot \tag{3.19}$$

Démonstration. Nous la donnons à l'annexe C, section C.1. □

Pour la suite, nous proposons une autre méthode d'estimation des paramètres.

3.3.2 Estimation des paramètres par la méthode du maximum de vraisemblance

Une autre méthode d'estimation du vecteur de paramètres θ est la méthode du maximum de vraisemblance globale. Elle consiste à estimer tous les paramètres simultanément en maximisant la log-vraisemblance globale pour le modèle de 2-copule échangeable à partir de la densité jointe dans une grappe. La log-vraisemblance pour les m grappes à partir de l'équation (3.1) s'écrit

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{j=1}^{m} \log\left[c_{1,n_j}^{(3)}\left\{C_{2|1}\left\{G(y_{j1};\beta)|F(x_{j1};\alpha)\right\}, \ldots, C_{2|1}\left\{G(y_{jn_j};\beta)|F(x_{jn_j};\alpha)\right\};\delta_3\right\}\right] \\
&+ \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log\left\{f(x_{ji};\alpha)f(y_{ji};\beta)\right\} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log\left[c^{(2)}\left\{F(x_{ji};\alpha), G(y_{ji};\beta);\delta_2\right\}\right] \\
&+ \sum_{j=1}^{m} \log\left[c_{1,n_j}^{(1)}\left\{F(x_{j1};\alpha), \ldots, F(x_{jn_j};\alpha);\delta_1\right\}\right]\cdot
\end{aligned} \tag{3.20}$$

L'estimation $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$ des paramètres s'obtient par maximisation complète de l'équation (3.20) et on note

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \left\{\mathcal{L}\left(\theta = (\alpha, \beta, \delta_1, \delta_2, \delta_3)\right)\right\}\cdot \tag{3.21}$$

L'estimation $\hat{\theta}$ de l'équation (3.21) est aussi la solution du système d'équations

$$\left(\partial\mathcal{L}/\partial\alpha^T, \partial\mathcal{L}/\partial\beta^T, \partial\mathcal{L}/\partial\delta_1^T, \partial\mathcal{L}/\partial\delta_2^T, \partial\mathcal{L}/\partial\delta_3^T\right) = \mathbf{0}^T\cdot \tag{3.22}$$

Nous écrivons aussi globalement la méthode utilisée pour estimer les paramètres sous forme d'équations d'estimation. L'équation (3.22) peut s'écrire sous la forme

$$\sum_{j=1}^{m}\Psi(\mathbf{x}_j,\mathbf{y}_j,n_j;\hat{\theta})=0, \tag{3.23}$$

où $\Psi(x,y,n;\theta)=\left(\psi_1^{(n)}+a_1^{(n)},\psi_2^{(n)}+a_2^{(n)},\psi_3^{(n)},\psi_4^{(n)}+a_3^{(n)},\psi_5^{(n)}\right)^T$ et le vecteur $(\psi_1^{(n)},\psi_2^{(n)},\psi_3^{(n)},\psi_4^{(n)},\psi_5^{(n)})$ est défini à l'équation (3.12). Les fonctions $a_1^{(n)}$, $a_2^{(n)}$ et $a_3^{(n)}$ s'écrivent

$$\begin{aligned} a_1^{(n)} &= \sum_{i=1}^{n}\frac{\partial\log\left[c^{(2)}\left\{F(x_i;\alpha),G(y_i;\beta);\delta_2\right\}\right]}{\partial\alpha^T}+\frac{\partial\log\left[c_{1,n}^{(1)}\left\{F(x_1;\alpha),\ldots,F(x_n;\alpha);\delta_1\right\}\right]}{\partial\alpha^T} \\ &+ \frac{\partial\log\left[c_{1,n}^{(3)}\left\{C_{2|1}\left\{G(y_1;\beta)|F(x_1;\alpha)\right\},\ldots,C_{2|1}\left\{G(y_n;\beta)|F(x_n;\alpha)\right\};\delta_3\right\}\right]}{\partial\alpha^T}, \end{aligned} \tag{3.24}$$

et

$$\begin{aligned} a_2^{(n)} &= \frac{\partial\log\left[c_{1,n}^{(3)}\left\{C_{2|1}\left\{G(y_1;\beta)|F(x_1;\alpha)\right\},\ldots,C_{2|1}\left\{G(y_n;\beta)|F(x_n;\alpha)\right\};\delta_3\right\}\right]}{\partial\beta^T} \\ &+ \sum_{i=1}^{n}\frac{\partial\log\left[c^{(2)}\left\{F(x_i;\alpha),G(y_i;\beta);\delta_2\right\}\right]}{\partial\beta^T}, \end{aligned} \tag{3.25}$$

et

$$a_3^{(n)}=\frac{\partial\log\left[c_{1,n}^{(3)}\left\{C_{2|1}\left\{G(y_1;\beta)|F(x_1;\alpha)\right\},\ldots,C_{2|1}\left\{G(y_n;\beta)|F(x_n;\alpha)\right\};\delta_3\right\}\right]}{\partial\delta_2^T}. \tag{3.26}$$

Les deux méthodes d'estimation étant présentées, nous étudions les propriétés asymptotiques des estimateurs des paramètres du modèle de 2-copule échangeable dans la suite de ce chapitre.

3.4 Étude asymptotique des estimateurs des paramètres du modèle de 2-copule échangeable

Dans cette partie, nous étudions la convergence en probabilité et la normalité asymptotique, lorsque m tend vers l'infini, des estimateurs du vecteur de paramètres θ, obtenus à partir de la méthode IFM généralisée et de la méthode du maximum de vraisemblance. Par la suite, nous comparons les deux méthodes dans un cas particulier.

3.4.1 Conditions de convergence et de normalité asymptotique

Nous formulons les hypothèses suivantes, utilisées pour étudier la convergence et la normalité asymptotique. Nous notons θ_0, la vraie valeur du vecteur de paramètres θ. Les hypothèses

formulées ici s'appuient sur celles émises par Tsiatis (2006, page 30) pour étudier la convergence et la normalité asymptotique d'un estimateur.

C0 : **Tailles bornées :** La taille de toutes les grappes est bornée, c'est-à-dire il existe $n_{\max}$, entier tel que $\max(n_1, \ldots, n_m) \leq n_{\max}$, pour tout m positif.

C1 : **Unicité :** θ_0 est l'unique solution des équations $\lim_{m \longrightarrow \infty} \left[\frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\left\{ \Psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta) \right\} \right] = 0$ ou $\lim_{m \longrightarrow \infty} \left[\frac{1}{m} \sum_{j=1}^{m} \mathbb{E}\left\{ \psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta) \right\} \right] = 0$ pour $\mathbf{X}_j$ et $\mathbf{Y}_j$ deux vecteurs aléatoires de dimension n_j et $(\mathbf{X}_j, \mathbf{Y}_j)$ est de densité de copule associée définie par (3.1).

C2 : **Identifiabilité :** Soient $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de dimension $n = 2, \ldots, n_{\max}$, appartenant à l'ensemble des entiers positifs. Nous avons respectivement

$$\text{Si} \quad \Psi(\mathbf{x}, \mathbf{y}, n; \theta_1) = \Psi(\mathbf{y}, \mathbf{x}, n; \theta_2) \Longrightarrow \theta_1 = \theta_2, \text{ pour tout } \mathbf{x}, \mathbf{y},$$

et

$$\text{Si} \quad \psi(\mathbf{x}, \mathbf{y}, n; \theta_1) = \psi(\mathbf{y}, \mathbf{x}, n; \theta_2) \Longrightarrow \theta_1 = \theta_2, \text{ pour tout } \mathbf{x}, \mathbf{y}.$$

C3 : **Compacité :** L'espace Θ des paramètres est un ensemble fermé et borné de $\mathbb{R}^p$.

C4 : **Dominance :** Il existe deux fonctions intégrables φ_{MV} et φ_{IFM}, indépendantes de θ et que pour tout $\mathbf{X}$ et $\mathbf{Y}$ deux vecteurs de dimension $n = 2, \ldots, n_{\max}$,

$$\sup_{\theta} \|\nabla_\theta \Psi(\mathbf{X}, \mathbf{Y}, n; \theta)\|_p \leq \varphi_{\text{MV}}(\mathbf{X}, \mathbf{Y}, n), \quad \text{telle que } \mathbb{E}\left\{\varphi_{\text{MV}}(\mathbf{X}, \mathbf{Y}, n)\right\} < \infty,$$

et

$$\sup_{\theta} \|\nabla_\theta \psi(\mathbf{Y}, \mathbf{X}, n; \theta)\|_p \leq \varphi_{\text{IFM}}(\mathbf{X}, \mathbf{Y}, n), \quad \text{telle que } \mathbb{E}\left\{\varphi_{\text{IFM}}(\mathbf{X}, \mathbf{Y}, n)\right\} < \infty,$$

où $\nabla_\theta u$ est l'opérateur dérivée partielle du vecteur u par rapport à chaque composante du vecteur θ.

C5 : **Non singularité :** Si $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires de dimension n alors les matrices $\mathbb{E}\left\{\nabla_\theta \Psi(\mathbf{Y}, \mathbf{X}, n; \theta_0)\right\}$ et $\mathbb{E}\left\{\nabla_\theta \psi(\mathbf{Y}, \mathbf{X}, n; \theta_0)\right\}$ sont non singulières pour $n = 2, \ldots, n_{\max}$ et

$$\frac{1}{m} \sum_{j=1}^{m} \nabla_\theta \Psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta) \xrightarrow{\mathcal{P}} \mathbb{E}\left\{\nabla_\theta \Psi(\mathbf{X}, \mathbf{Y}, n; \theta_0)\right\},$$

et

$$\frac{1}{m} \sum_{j=1}^{m} \nabla_\theta \psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta) \xrightarrow{\mathcal{P}} \mathbb{E}\left\{\nabla_\theta \psi(\mathbf{Y}, \mathbf{X}, n; \theta_0)\right\}.$$

Nous nous intéressons aux estimateurs de θ obtenus par la méthode du maximum de vraisemblance et par la méthode IFM généralisée et nous donnons le théorème suivant sur leurs propriétés asymptotiques.

Théorème 3.1. *Nous considérons un modèle de 2-copule échangeable dont le vecteur de paramètres θ s'estime par la méthode IFM généralisée et la méthode du maximum de vraisemblance globale lorsque m tend vers l'infini.*

*Sous les hypothèses **C0**, **C1**, **C2**, **C3** et **C4**, nous avons les résultats suivants*

i) ***Convergence en probabilité des estimateurs $\tilde{\theta}$ et $\hat{\theta}$.***

Les estimateurs de θ par la méthode IFM généralisée notée $\tilde{\theta}$ et par la méthode du maximum de vraisemblance globale notés $\hat{\theta}$ sont convergents, c'est-à-dire

$$\tilde{\theta} \xrightarrow{\mathcal{P}} \theta_0, \qquad \hat{\theta} \xrightarrow{\mathcal{P}} \theta_0, \quad \textit{pour} \;\; m \longrightarrow \infty.$$

ii) ***Normalité asymptotique de $\hat{\theta}$***

$$\sqrt{m}\left(\hat{\theta} - \theta_0\right) \overset{\mathcal{L}}{\sim} \mathcal{N}\left\{0, \mathcal{I}^{-1}(\theta_0)\right\}, \;\; m \longrightarrow \infty, \tag{3.27}$$

où la matrice d'information est

$$\mathcal{I}(\theta_0) = -\lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\nabla_\theta \Psi(\boldsymbol{X}_j, \boldsymbol{Y}_j, n_j; \theta_0)\right\}\right], \tag{3.28}$$

et le vecteur aléatoire $(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ est de loi définie par l'équation (3.1).

iii) ***Normalité asymptotique de $\tilde{\theta}$.***

$$\sqrt{m}\left(\tilde{\theta} - \theta_0\right) \overset{\mathcal{L}}{\sim} \mathcal{N}\left\{0, \mathcal{G}(\theta_0)\right\}, \;\; m \longrightarrow \infty, \tag{3.29}$$

où $\mathcal{G}(\theta_0)$ est la matrice d'information de Godambe, Cherubini et al. (2014). Elle est définie par

$$\mathcal{G}(\theta_0) = D^{-1}V_0\left(D^{-1}\right)^T, \;\; D = -\lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\nabla_\theta \psi(\boldsymbol{X}_j, \boldsymbol{Y}_j, n_j; \theta_0\right\}\right], \tag{3.30}$$

et la matrice V_0 est

$$V_0 = \lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\psi(\boldsymbol{X}_j, \boldsymbol{Y}_j, n_j; \theta_0)\psi(\boldsymbol{X}_j, \boldsymbol{Y}_j, n_j; \theta_0)^T\right\}\right]. \tag{3.31}$$

Démonstration.

Soient $\mathbf{X}_j$ et $\mathbf{Y}_j$ deux vecteurs aléatoires de dimension n_j chacun tel que $(\mathbf{X}_j, \mathbf{Y}_j)$ est de densité de copulem l'équation (3.1).

i) Sous les hypothèses **C0**, **C1**, **C2**, **C3** et **C4** et en utilisant la section 3.2 de Tsiatis (2006), $\tilde{\theta}$ et $\hat{\theta}$ convergent en probabilité vers θ_0 pour m tendant vers l'infini.

ii) Le vecteur de paramètres $\hat{\theta}$ s'obtient par maximisation de la vraisemblance globale donc il est convergent et asymptotiquement normal, Joe (2014, chapitre 1, page 15) et Joe (2014, chapitre 5, page 228) de matrice d'information $\mathcal{I}(\theta_0)$ comme un résultat connu. La matrice de la moyenne des matrices d'information provenant des grappes est $\{\mathcal{I}_1(\theta_0) + \ldots + \mathcal{I}_m(\theta_0)\}/m$ où $\mathcal{I}_j(\theta_0) = -\mathbb{E}\left\{\nabla_\theta \Psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)\right\}$ est la contribution de la grappe j à la matrice d'information totale. La matrice ∇_θ est la matrice des dérivées partielles d'un vecteur par rapport au vecteur de paramètres θ. Nous avons

$$\begin{aligned}\{\mathcal{I}_1(\theta_0) + \ldots + \mathcal{I}_m(\theta_0)\}/m &= -\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\nabla_\theta \Psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta_0)\right\} \\ &\xrightarrow{\mathcal{P}} -\lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\nabla_\theta \Psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta_0)\right\}\right].\end{aligned}$$

En conséquence, nous avons le résultat espéré sur la convergence de l'estimateur du maximum de vraisemblance globale.

iii) Le vecteur de paramètres θ est estimé à partir de l'équation (3.12). Nous avons donc

$$\sqrt{m}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left\{0, \mathcal{G}(\theta_0)\right\},\ m \longrightarrow \infty,$$

et la matrice de Godambe provient de l'équation (3.30). Pour justifier ce résultat, on considère l'expansion de Taylor de l'équation (3.12) au voisinage de θ_0 par

$$0 = \sum_{j=1}^{m}\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0) + \sum_{j=1}^{m}\nabla_\theta\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)(\tilde{\theta} - \theta_0) + o_p(\sqrt{m})\cdot$$

Nous avons par la suite

$$\sqrt{m}\sum_{j=1}^{m}\nabla_\theta\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)(\tilde{\theta} - \theta_0) = -\sqrt{m}\sum_{j=1}^{m}\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0) + o_p(\sqrt{m})\cdot$$

Soit on obtient alors

$$\sqrt{m}(\tilde{\theta} - \theta_0) = -\left[(m)^{-1}\sum_{j=1}^{m}\nabla_\theta\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)\right]^{-1}\left\{\sqrt{m}^{-1}\sum_{j=1}^{m}\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)\right\} + o_p(\sqrt{m}),$$

car la matrice est non singulière par ***C2***. Donc nous obtenons

$$\sqrt{m}(\tilde{\theta} - \theta_0) = -\left\{1/m\sum_{j=1}^{m}\nabla_\theta\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)\right\}^{-1}\left\{\sqrt{m}^{-1}\sum_{j=1}^{m}\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0)\right\} + o_p(\sqrt{m})\cdot$$

Or nous avons le résultat suivant, à partir de ***C5*** pour $m \longrightarrow \infty$ et de la loi faible des grands nombres, on a

$$(m)^{-1}\sum_{j=1}^{m}\nabla_\theta\psi(\mathbf{x}_j, \mathbf{y}_j, n_j; \theta_0) \xrightarrow{\mathcal{P}} \lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\nabla_\theta\psi(\mathbf{X}_j, \mathbf{Y}_j, n_j; \theta_0\right\}\right].$$

Le théorème central limite appliqué à l'estimateur $\tilde{\theta}$, nous avons le résultat que $\tilde{\theta}$ obéit une loi normale de moyenne θ_0 et de variance provenant de la matrice de Godambe. De plus, la variance de $\sum_{j=1}^{m}\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)/\sqrt{m}$ est

$$\begin{aligned}\mathbb{V}\left\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)\right\} &= \frac{1}{m}\sum_{j=1}^{m}\mathbb{V}\left\{\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)\right\}\\ &= \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)^T\right\},\end{aligned}$$

et donc pour $m\longrightarrow\infty$, nous avons

$$\mathbb{V}\left\{\frac{1}{\sqrt{m}}\sum_{j=1}^{m}\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)\right\}\xrightarrow{\mathcal{P}}\lim_{m\longrightarrow\infty}\left[\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left\{\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)\psi(\mathbf{X}_j,\mathbf{Y}_j,n_j;\theta_0)^T\right\}\right].$$

En conséquence, la matrice sandwich de l'estimateur $\tilde{\theta}$ donne le résultat espéré. □

Nous savons que la méthode du maximum de vraisemblance est plus précise que la méthode IFM voir Joe (2005). Dans la suite de ce chapitre, nous présentons la comparaison des deux méthodes d'estimation du modèle de 2-copule échangeable dans un cas particulier et par une étude Monte Carlo.

3.4.2 Comparaison des méthodes IFM généralisée et maximum de vraisemblance utilisées pour l'estimation des paramètres de la 2-copule échangeable

Pour avoir une idée de la marge entre les matrices de variance-covariance des deux estimateurs associés au modèle de 2-copule échangeable, nous comparons les variances asymptotiques de la méthode IFM généralisée et de la méthode du maximum de vraisemblance dans des cas particuliers. Dans les cas particuliers discutés ici, toutes les grappes sont de même taille n.

Cas particulier 1 : Nous considérons que les marginales F et G sont la loi normale centrée réduite et que les familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules d'indépendance. Le seul paramètre à estimer dans le modèle considéré est δ_2 et donc logiquement $\mathcal{G}(\delta_2)=\mathcal{I}^{-1}(\delta_2)$.

Cas particulier 2 : Nous considérons que les familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules d'indépendance. Le vecteur de paramètres dans ce cas est (α,β,δ_2).

Le modèle de 2-copule échangeable revient à

$$f(\mathbf{x}_j, \mathbf{y}_j; \alpha, \beta, \delta_2) = \prod_{i=1}^{n_j} F(x_{ji}; \alpha) G(y_{ji}; \beta) c^{(2)} \left\{ F(x_{ji}; \alpha), G(y_{ji}; \beta); \delta_2 \right\} \cdot \tag{3.32}$$

L'équation (3.32) correspond exactement à un cas particulier de l'ajustement d'un modèle de copule bivariée $C^{(2)}$ avec des lois marginales paramétriques F et G, voir section 1.5. Pour ce modèle, en faisant référence à Joe (2005) et Joe (2014, page 226), nous concluons à une convergence asymptotique des méthodes IFM généralisée et maximum de vraisemblance et que cette dernière est plus précise que la méthode IFM généralisée.

Nous élargissons la comparaison avec un autre exemple pour le modèle de 2-copule échangeable.

Cas particulier 3 : Les marginales F et G correspondent à la fonction de répartition de la loi normale centrée réduite. Les famille de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont de paramètres univariés δ_1 et δ_3 respectivement, $C^{(2)}$ est la copule bivariée de paramètre δ_2 univarié.

La procédure d'estimation comporte deux parties indépendantes : une partie liée à l'estimation du paramètre δ_1 et l'autre partie à l'estimation combinée des paramètres δ_2 et δ_3. L'inférence par rapport au paramètre δ_1 est similaire pour les deux méthodes c'est-à-dire les variances asymptotiques par IFM et maximum de vraisemblance de δ_1 sont égales. La comparaison des matrices de variance-covariance des méthodes IFM et maximum de vraisemblance porte donc sur (δ_2, δ_3).

On pose $A = \mathbb{E}\left\{\frac{\partial \psi_5^{(n)}}{\partial \delta_2}\right\}$ et $B = \mathbb{E}\left\{\frac{\partial^2 \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2^2}\right\}$. En rappelant que $a_3^{(n)} = \frac{\partial \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2}$, nous montrons que A et B s'écrivent par

$$A = -\mathbb{E}\left\{\psi_5^{(n)} \cdot a_3^{(n)}\right\} = -\text{cov}\left\{\psi_5^{(n)}, a_3^{(n)}\right\}, \quad B = -\mathbb{E}\left\{a_3^{(n)2}\right\} \cdot \tag{3.33}$$

Nous calculons les matrices de variance-covariance asymptotiques des deux méthodes d'estimation.

• Matrice de variance-covariance asymptotique de l'estimateur de (δ_2, δ_3) obtenu à partir de la méthode IFM

Le vecteur score par cette méthode pour (δ_2, δ_3) est $(\psi_4^{(n)}, \psi_5^{(n)})$ provenant des équations (3.14), (3.15) et (3.16).

La matrice de variance asymptotique de $(\tilde{\delta}_2, \tilde{\delta}_3)$, en utilisant la méthode IFM généralisée est

$$\mathcal{G}(\delta_2, \delta_3) = \begin{pmatrix} \frac{1}{\mathbb{V}\left\{\psi_4^{(n)}\right\}} & \frac{-\mathbb{E}\left\{\psi_5^{(n)} \cdot a_3^{(n)}\right\}}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\}} \\ \frac{-\mathbb{E}\left\{\psi_5^{(n)} \cdot a_3^{(n)}\right\}}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\}} & \frac{1}{\mathbb{V}\left\{\psi_5^{(n)}\right\}} + \frac{\mathbb{E}\left\{\psi_5^{(n)} \cdot a_3^{(n)}\right\}^2}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\}^2} \end{pmatrix} \cdot \tag{3.34}$$

Les justificatifs pour le résultat de l'équation (3.34) se trouvent à l'annexe C, section C.2.

• Matrice de variance-covariance asymptotique de l'estimateur de (δ_2, δ_3) obtenu à partir de la méthode de maximisation de la vraisemblance globale

Le vecteur score de cette méthode pour (δ_2, δ_3) est $(\Psi_4^{(n)}, \Psi_5^{(n)}) = (\psi_4^{(n)} + a_3^{(n)}, \psi_5^{(n)})$, où $a_3^{(n)}$ provient de l'équation (3.26).

La matrice de variance-covariance asymptotique de $(\hat{\delta}_2, \hat{\delta}_3)$, en utilisant la méthode du maximum de vraisemblance est

$$\mathcal{I}^{-1}(\delta_2, \delta_3) = \begin{pmatrix} \frac{\mathbb{V}\left\{\psi_5^{(n)}\right\}}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\} - B\mathbb{V}\left\{\psi_5^{(n)}\right\} - A^2} & \frac{A}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\} - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2} \\ \frac{A}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\} - B\mathbb{V}\left\{\psi_5^{(n)}\right\} - A^2} & \frac{\mathbb{V}\left\{\psi_4^{(n)}\right\} - B}{\mathbb{V}\left\{\psi_4^{(n)}\right\}\mathbb{V}\left\{\psi_5^{(n)}\right\} - B\mathbb{V}\left\{\psi_5^{(n)}\right\} - A^2} \end{pmatrix} \quad (3.35)$$

Les justificatifs pour le résultat de l'équation (3.35) se trouvent à l'annexe C, section C.2.

La matrice $\mathcal{G}(\delta_2, \delta_3) - \mathcal{I}^{-1}(\delta_2, \delta_3)$ est semi-définie positive et nous concluons donc que la variance asymptotique de la méthode du maximum de vraisemblance est plus petite que celle de la méthode IFM généralisée, voir annexe C.2.

Le résultat de la théorie asymptotique se confirme pour le modèle de 2-copule échangeable à travers le cas de la section 3.4.2, que la méthode du maximum de vraisemblance est plus précise que la méthode IFM généralisée. Cependant, les variances asymptotiques sont "proches" et la sous-section qui suit nous permet d'évaluer concrètement dans un cas simple, la différence entre les deux variances asymptotiques.

3.4.3 Évaluation des matrices de variance-covariance asymptotiques dans le cas où les copules du modèle de 2-copule échangeable sont toutes normales

Dans cette sous-section, nous évaluons les matrices des équations (3.34) et (3.35) où les familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont des copules normales échangeables dont les paramètres sont $\delta_1 = \rho_1$ et $\delta_3 = \rho_3$ respectivement d'une part et d'autre part la copule $C^{(2)}$ est une copule normale bivariée de paramètre $\delta_2 = \rho_2$.

Nous posons $r_0 = (1-\rho_2^2)\rho_1\rho_3 + 2\rho_2^2\rho_3^2$ et nous utilisons les notations suivantes pour simplifier l'écriture des matrices de variance-covariance asymptotiques

$$(\diamond) = 1 + (n-1)\rho_3, \ (\oplus) = n(1+\rho_2^2) + n(n-1)r_0, \ (\sim) = 1 + (n-1)\rho_3^2, \quad (3.36)$$

$$[**] = 2n(n-1)\rho_3^2\left[\rho_2^2(1-\rho_3)(\diamond) + (1-\rho_2^2)\left\{1 + (n-2)\rho_1 - (n-1)\rho_1\rho_3\right\}\right]. \quad (3.37)$$

La matrice de variance-covariance asymptotique de la méthode IFM généralisée, de l'équation (3.34) du vecteur $(\tilde{\rho}_2, \tilde{\rho}_3)$ revient à

$$\mathcal{G}(\rho_2,\rho_3) = \begin{pmatrix} \frac{(1-\rho_2^2)^2}{(\oplus)} & \frac{-2\rho_2\rho_3(1-\rho_3)(1-\rho_2^2)(\diamond)}{(\oplus)(\sim)} \\ \frac{-2\rho_2\rho_3(1-\rho_3)(1-\rho_2^2)(\diamond)}{(\oplus)(\sim)} & \frac{2(1-\rho_3)^2(\diamond)^2}{n(n-1)(\sim)}\left\{1+\frac{2n(n-1)\rho_2^2\rho_3^2}{(\oplus)(\sim)}\right\} \end{pmatrix}. \tag{3.38}$$

La matrice de variance-covariance asymptotique de la méthode du maximum de vraisemblance globale, de l'équation (3.35) du vecteur $(\hat{\rho}_2, \hat{\rho}_3)$ donne

$$\mathcal{I}^{-1}(\rho_2,\rho_3) = \begin{pmatrix} \frac{(1-\rho_2^2)^2(\sim)}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2} & \frac{-2\rho_2\rho_3(1-\rho_3)(1-\rho_2^2)(\diamond)}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2} \\ \frac{-2\rho_2\rho_3(1-\rho_3)(1-\rho_2^2)(\diamond)}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2} & \frac{2(1-\rho_3)(\diamond)}{n(n-1)}\left[\frac{(1-\rho_3)(\diamond)(\oplus)+[**]}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2}\right] \end{pmatrix}. \tag{3.39}$$

Les matrices des équations (3.38) et (3.39) s'obtiennent en remplaçant les équations (C.6), (C.7), (C.8) et (C.9) dans les résultats (3.34) et (3.35) respectivement. Les calculs théoriques effectués pour l'obtention des matrices de variance-covariance se trouvent à l'annexe C.3.

Par la suite, nous posons les rapports de variance R et $R^{'}$. Le rapport R est la variance asymptotique de la méthode du maximum de vraisemblance divisée par celle de la méthode IFM généralisée pour ρ_2 et s'écrit

$$R(\rho_1,\rho_2,\rho_3) = \frac{(\sim)(\oplus)}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2}. \tag{3.40}$$

Le rapport $R^{'}$ est la variance asymptotique de la méthode du maximum de vraisemblance divisée par celle de la méthode IFM pour ρ_3 et s'écrit

$$R^{'}(\rho_1,\rho_2,\rho_3) = \frac{(\oplus)(\sim)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2}\cdot\frac{(\oplus)(\sim)}{(\oplus)(\sim)+2n(n-1)\rho_2^2\rho_3^2}. \tag{3.41}$$

Nous formulons les conclusions suivantes :

- Les deux matrices de variance-covariance asymptotiques ne dépendent de ρ_1 que via $[**]$ et $(\oplus)$. Pour ρ_2 et ρ_3 fixés, R et $R^{'}$ sont des fonctions décroissantes de ρ_1 et donc la méthode IFM est toujours moins précise que la méthode du maximum de vraisemblance.
- Si le paramètre ρ_3 tend vers 0 (indépendance des résidus), alors $R = R^{'} = 1$ et donc les méthodes IFM généralisée et maximum de vraisemblance conduisent à la même matrice de variance-covariance asymptotiquement.
- Pour $n = 1$, autrement dit, un individu par grappe considéré alors $R = 1$ et la variance asymptotique de la méthode IFM généralisée et de la méthode du maximum de vraisemblance sont égales pour le paramètre ρ_2. Ceci correspond à la modélisation copule ordinaire utilisant la copule bivariée $C^{(2)}_{\rho_2}$ et l'effet grappe est inexistant. Ce résultat s'exprime dans le cas limite de n pour le paramètre ρ_3 par

$$\lim_{n\longrightarrow 1} R^{'}(\rho_1,\rho_2,\rho_3)=1\cdot$$

— Lorsque la taille des grappes tend vers l'infini, nous avons

$$\lim_{n\to+\infty} R^{'}(\rho_1,\rho_2,\rho_3)=1, \tag{3.42}$$

car

$$\lim_{n\longrightarrow\infty}\left\{\frac{(\oplus)(\sim)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}}{(\sim)(\oplus)+\frac{(\sim)[**]}{(1-\rho_3)(\diamond)}-2n(n-1)\rho_2^2\rho_3^2}\right\}=1,$$

et

$$\lim_{n\longrightarrow\infty}\left\{\frac{(\oplus)(\sim)}{(\oplus)(\sim)+2n(n-1)\rho_2^2\rho_3^2}\right\}=1,$$

Par ailleurs nous traçons les courbes du rapport des variances asymptotiques des estimateurs de ρ_2 et ρ_3. Nous caractérisons les rapports R et $R^{'}$ de deux façons en fixant la valeur de $\rho_1=0.2$.

- La première façon est de tracer des graphiques des rapports R et $R^{'}$ en fonction de la taille n des grappes pour ρ_1 fixé. Pour le graphique de R, nous fixons $\rho_3=0.1$ et nous faisons varier $\rho_2\in\{0.4,0.6,0.8\}$. Pour le graphique de $R^{'}$, nous fixons $\rho_2=0.8$ et nous faisons varier $\rho_3\in\{0.1,0.3,0.6\}$. Les valeurs de ρ_3 sont "faibles" car matérialisant la dépendance des résidus.
- La deuxième façon est de tracer les courbes de niveau de R et $R^{'}$.

Nous présentons 4 figures pour analyser la première façon de présenter les rapports R et $R^{'}$.

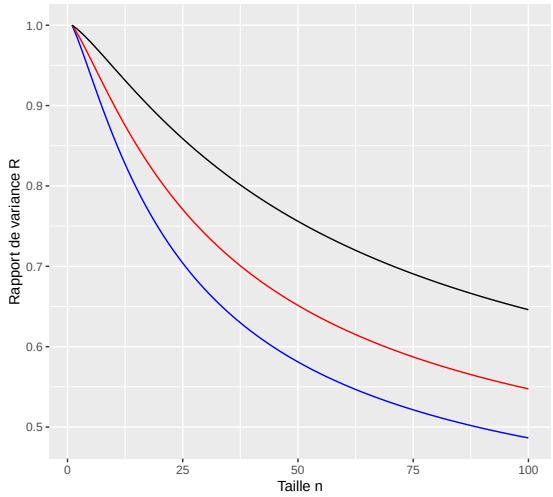


FIGURE 3.1 – Courbes de R en fonction de la taille n, $\rho_3=0.1$ pour $\rho_2=0.4$ (courbe bleue), $\rho_2=0.6$ (courbe rouge) et $\rho_2=0.8$ (courbe noire).

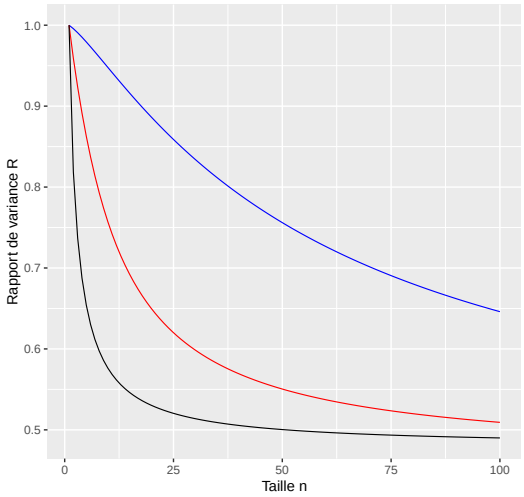


FIGURE 3.2 – Courbes de R sachant la taille n, ρ_2=0.8 pour $\rho_3=0.1$ (courbe bleue), $\rho_3=0.3$ (courbe rouge) et $\rho_3=0.6$ (courbe noire).

Commentaires sur les graphiques des figures 3.1 et 3.2

- Les courbes du rapport R décroit au fur et à mesure que la taille des grappes croit. La décroissance est rapide pour les valeurs faibles de ρ_2 (courbes en rouge et bleue). Ceci s'explique aisément, car l'écart-type dépend de la valeur de ρ_2.

- lorsque la corrélation ρ_3 est "forte" ($\rho_3 = 0.4$ ou $\rho_3 = 0.6$), l'on note une décroissance rapide du rapport R mais ceux-ci gardent une valeur constante lorsque la taille augmente. Les courbes en noire et rouge sont parallèles à partir d'une taille n avec une marge de différence faible.

Nous construisons par la suite des figures relatives à la variance asymptotique de l'estimateur du paramètre ρ_3.

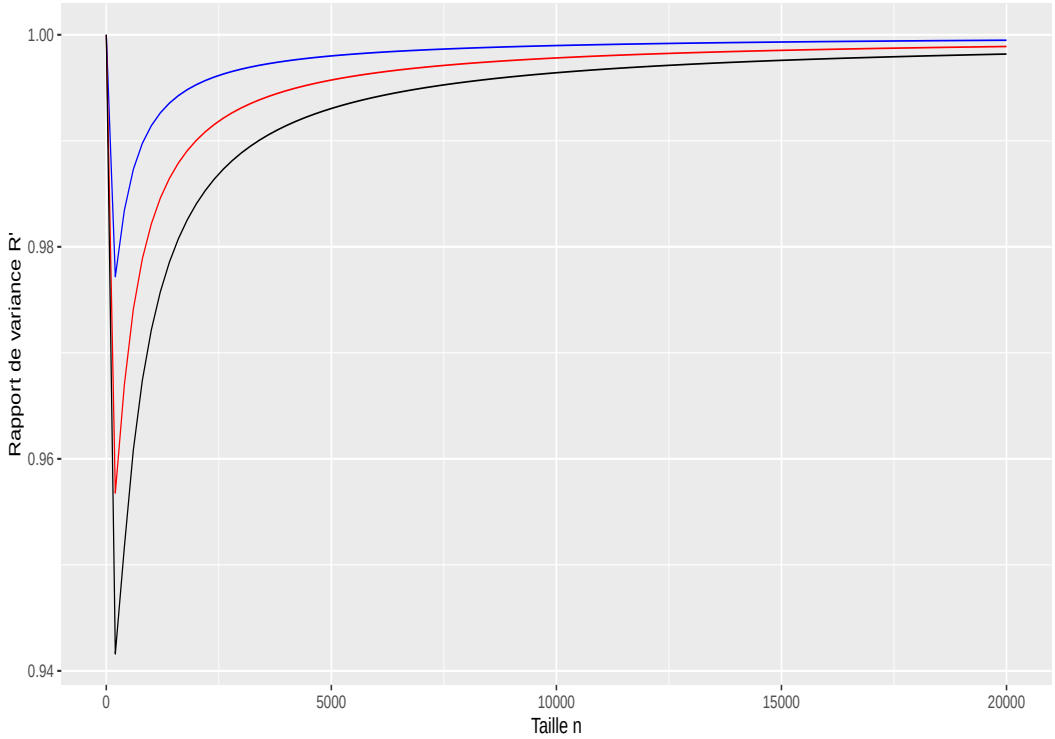


FIGURE 3.3 – Courbes de $R^{'}$ en fonction de la taille n, $\rho_3 = 0.1$ pour $\rho_2 = 0.4$ (courbe bleue), $\rho_2 = 0.6$ (courbe rouge) et $\rho_2 = 0.8$ (courbe noire).

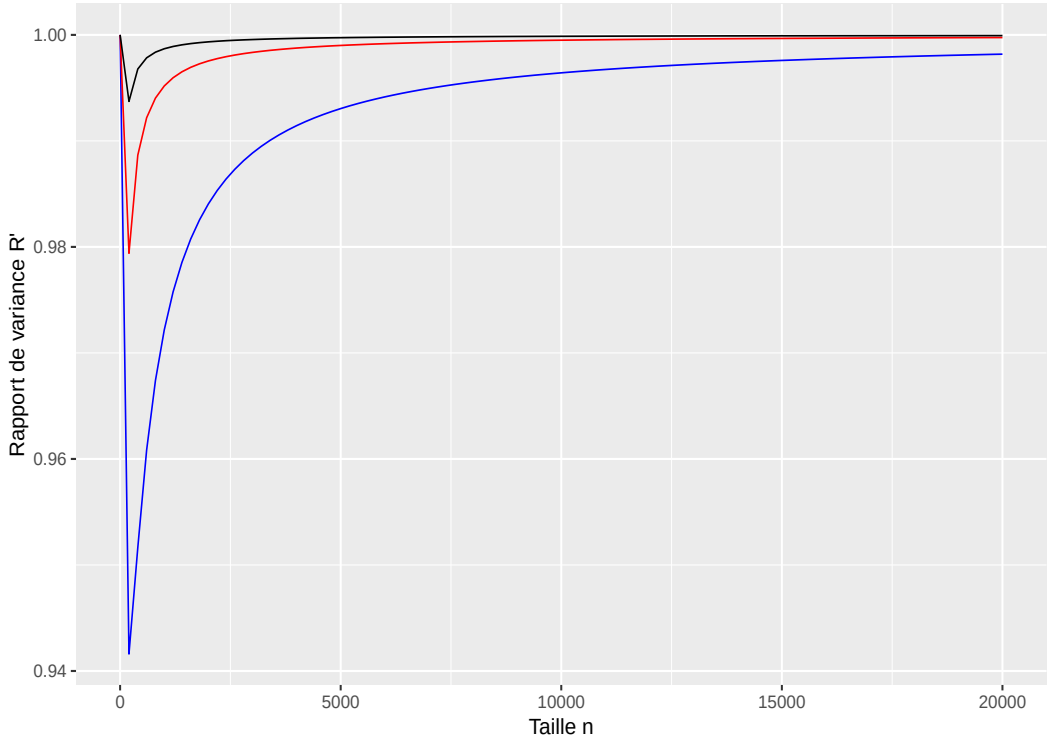


FIGURE 3.4 – Courbes de $R^{'}$ en fonction de la taille n, $\rho_2 = 0.8$ pour $\rho_3 = 0.1$ (courbe bleue), $\rho_3 = 0.3$ (courbe rouge) et $\rho_3 = 0.6$ (courbe noire).

Commentaire sur les figures 3.3 et 3.4

- la fonction $R^{'}$ est monotone et décroit rapidement quand ρ_2 est "forte" (courbe en noire). Un minimum existe pour les courbes et plus grand que 0.8. Les deux méthodes sont donc très proches. De plus, lorsque n tend vers l'infini, quelques soient les valeurs prisent par ρ_2, $R^{'}$ tendent vers 1 pour $\rho_1 = 0.2$ et donc les deux méthodes tendent à coïncider.

- la fonction $R^{'}$ est monotone et décroit rapidement quand ρ_2 est "forte". La conclusion de la figure 3.3 s'applique également et confirme l'équation (3.42).

Nous fixons par la suite la valeur de la corrélation $\rho_1 = 0.2$ et que la taille n de la grappe est $n = 30$ puis nous construisons les courbes de niveau.

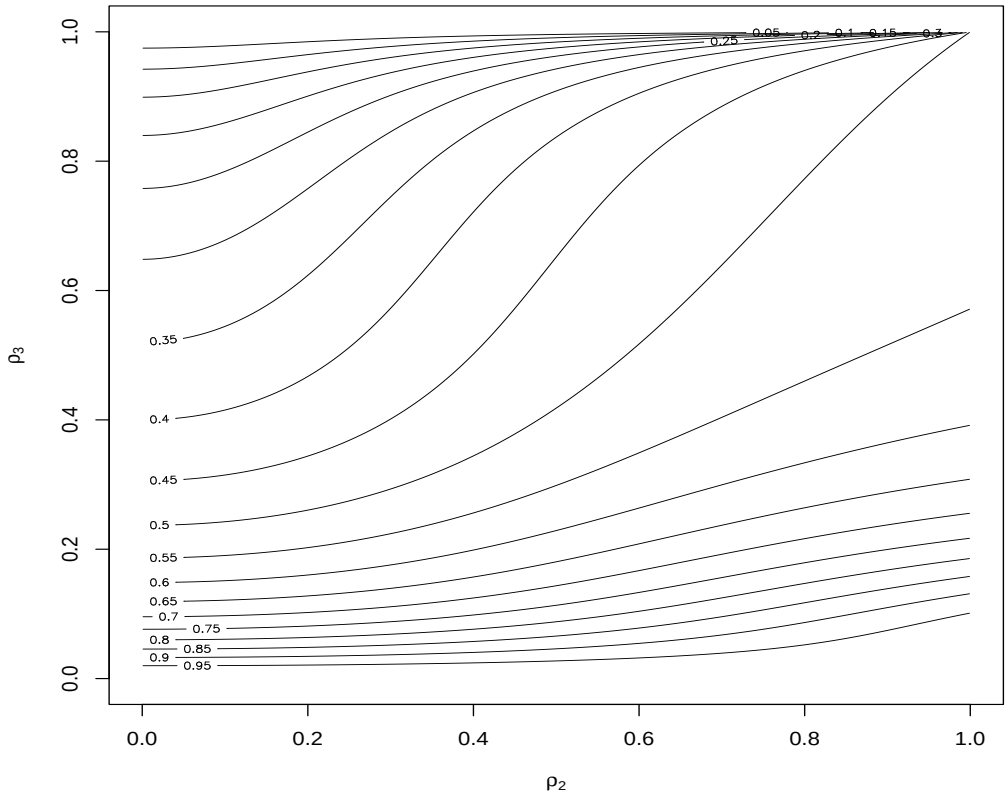


FIGURE 3.5 – Courbes de niveau pour le rapport R en fonction de (ρ_2, ρ_3) avec $n = 30$.

Discussion des courbes de niveau (ρ_2, ρ_3, R). Pour ce graphique, lorsque ρ_2 croit et ρ_3 est "petit", le rapport R des courbes de niveau varie entre 0.55 et 0.95. Ceci correspond au rapprochement entre les variances asymptotiques des méthodes IFM et maximum de vraisemblance. Tandis que lorsque ρ_3 est plus grand que 0.6 et que ρ_2 croit, le rapport R varie entre 0.05 et 0.3.

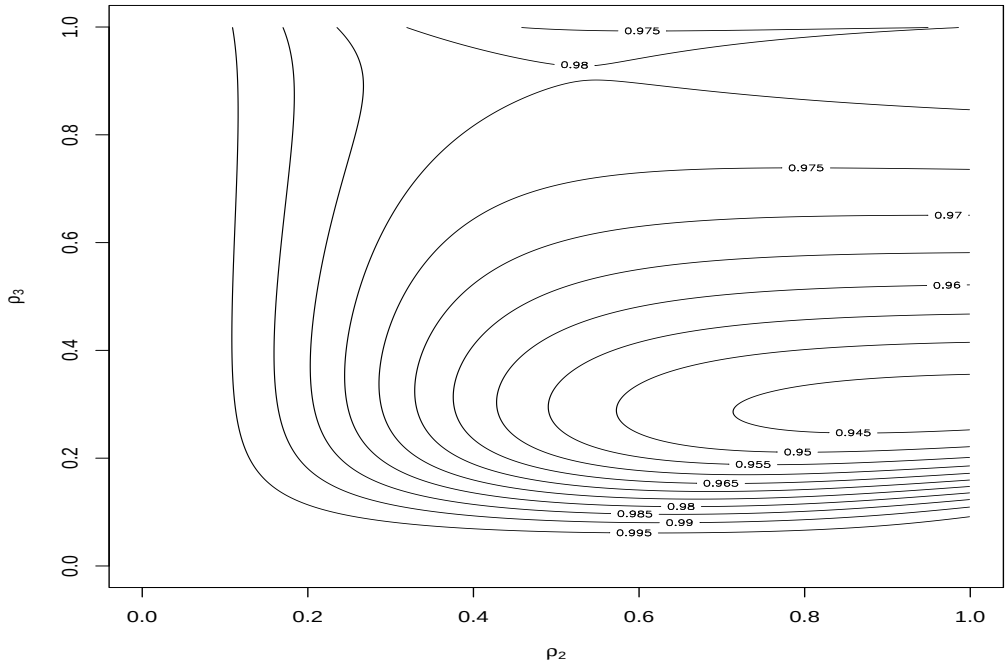


FIGURE 3.6 – Courbes de niveau pour le rapport $R^{'}$ en fonction de (ρ_2, ρ_3) avec $n = 30$.

Discussion des graphiques combinés (ρ_2, ρ_3) pour $R^{'}$. Pour les courbes de niveau relatives au rapport $R^{'}$, si les deux corrélations ρ_2 et ρ_3 croient, le rapport $R^{'}$ décroit donc la méthode du maximum reste plus précise que la méthode IFM pour ρ_2 ou ρ_3 fixé et la valeur la plus petite des lignes de niveau est 0.885.

Nous rappelons que dans la sous-section 3.4.3, nous effectuons un calcul explicite des matrices de variances-covariances asymptotiques issues des deux méthodes d'estimation. Dans ce cas, toutes les copules sont normales et les marginales sont connues. Cependant, lorsque les copules changent ou prennent la forme d'autres types de forme (*gumbel*, *Khoudraji*, etc), le calcul explicite s'avère difficile, voire impossible à obtenir. Nous avons comme objectif global, d'estimer les deux matrices de variance-covariance pour un nombre m de grappes fini, issues de la méthode IFM généralisée et de la méthode du maximum de vraisemblance en utilisant les données simulées. Ainsi, dans la section suivante, nous effectuons cette étude échantillonnale par la méthode Monte-Carlo.

3.5 Étude des propriétés échantillonnales des estimateurs des paramètres du modèle 2-copule échangeable par Monte-Carlo

Pour un modèle de 2-copule échangeable de paramètres θ, nous notons $\hat{\theta}$ et $\tilde{\theta}$, les estimateurs construits respectivement à partir de la méthode du maximum de vraisemblance et de la méthode IFM généralisée. Pour ces deux estimateurs, nous connaissons, des équations (3.29) et (3.27), les matrices de variance-covariance asymptotiques (lorsque le nombre m de grappes tend vers $+\infty$). Cependant, les calculs explicites ne sont pas toujours possibles sauf dans quelques cas particuliers. Nous souhaitons calculer l'espérance et la variance des estimateurs $\hat{\theta}$ et $\tilde{\theta}$ pour m fini. Pour ce fait, nous utilisons la méthode de Monte-Carlo qui nous permet d'approximer l'espérance et la variance des deux estimateurs. Ainsi, nous donnons une approximation de l'espérance et de la variance de l'estimateur $\hat{\theta}$ qui se définit par

$$\mathbb{E}_B(\hat{\theta}) \approx \frac{1}{B}\sum_{i=1}^{B}\theta_i, \quad \mathbb{V}_B(\hat{\theta}) \approx \frac{1}{B-1}\sum_{i=1}^{B}\left\{\theta_i - \mathbb{E}_B(\hat{\theta})\right\}\left\{\theta_i - \mathbb{E}_B(\hat{\theta})\right\}^T, \tag{3.43}$$

en fixant B, le nombre de simulations Monte-Carlo. La même approximation de l'équation (3.43) s'applique aussi à l'estimateur $\tilde{\theta}$. La racine carrée des éléments de la diagonale de la variance $\mathbb{V}_B$ donne l'écart-type, que nous calculons pour chacun des estimateurs dans les tableaux présentés.

Succinctement, l'objectif de la section est d'étudier, par simulation, les propriétés échantillonnales (le biais et la variance) des estimateurs $\hat{\theta}$ et $\tilde{\theta}$. Dans toute cette étude Monte-Carlo du modèle de 2-copule échangeable de l'équation (3.1), nous supposons que F est une loi normale de moyenne $\mu_1 = 0$ et de variance $\sigma_1^2 = 1$ et G est une loi normale de moyenne $\mu_2 = 0$ et de variance $\sigma_2^2 = 1$. Les copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ appartiennent à la famille normale de paramètres δ_1 et δ_3 respectivement. Nous fixons le tau de Kendall associé à la famille de copules échangeables $\left(C_{1,n}^{(1)}\right)$ à $\tau_1 = 0.2$. Cette corrélation de 0.2 comme étant la corrélation à l'intérieur des grappes et souvent faible et se situant entre 0.2 et 0.5. Nous fixons le tau de Kendall associé à la famille de copules $\left(C_{1,n}^{(3)}\right)$ à $\tau_3 = 0.1$. Cette corrélation est fixée à 0.1 parce que l'objectif est de minimiser toujours l'importance de cette copule. En effet, la famille de copules $\left(C_{1,n}^{(3)}\right)$ modélise une sorte de corrélation de résidus associés au modèle de régression. Les deux facteurs variables dans la simulation sont le tau de Kendall τ_2 associé à la copule $C^{(2)}$ et la taille m d'échantillon.

Tau de Kendall de la copule $C^{(2)}$

La copule $C^{(2)}$ est une copule bivariée quelconque (il en existe plusieurs) de paramètre δ_2. Nous fixons le tau de Kendall $\tau_2 \in \{0.4, 0.6\}$. Ces valeurs traduisent la corrélation entre les deux

variables dans le cas pratique. À partir du tau de Kendall et du type de copule spécifiée pour $C^{(2)}$, nous calculons le paramètre associé à partir du tableau 1.2 pour quelques types de copules bivariées avec les fonctions *iTau* ou *BiCopTau2Par* du package *copula*, du logiciel R. Dans le cas normal, pour $(\tau_1, \tau_3) = (0.2, 0.1)$, nous avons $(\delta_1, \delta_3) = (0.31, 0.16)$. Ce calcul s'utilise dans le cas de copules archimédiennes (*Clayton*, *Gumbel*, etc) ou de la copule elliptique. Dans le cas de la copule de *Khoudraji*, nous choisissons celle à trois paramètres. Pour cela, nous fixons un paramètre et nous recherchons, par essai-erreur les deux autres paramètres en utilisant le tau de Kendall τ_2 comme référence.

Taille d'échantillon des m grappes

Nous considérons trois cas de figure sur le nombre de grappes et deux scénarios concernant le type de grappes pour chaque cas de figure. Le nombre de grappes est $m = 10$, $m = 20$ et $m = 50$. Pour $m = 10$, $m = 20$, le nombre total d'individus est 300 et pour $m = 50$, le nombre total d'individus est 900.

- Pour $m = 10$, le scénario *Non Équilibré* est celui où nous avons 4 grappes de taille $n = 15$ chacune et 6 grappes de taille $n = 40$ chacune ;
- Pour $m = 10$, le scénario *Équilibré* est celui où toutes les grappes de tailles $n = 30$;
- Pour $m = 20$, le scénario *Non Équilibré* est celui où nous avons 10 grappes de taille $n = 10$ chacune et 10 grappes de taille $n = 20$ chacune ;
- Pour $m = 20$, le scénario *Équilibré* est celui où toutes les grappes de tailles $n = 15$;
- Pour $m = 50$, le scénario *Non Équilibré* est celui où nous avons 30 grappes de tailles $n = 10$ chacune et 20 grappes de taille $n = 30$ chacune ;
- Pour $m = 50$, le scénario *Équilibré* est celui où toutes les grappes de tailles $n = 18$.

Avant d'aborder l'étape de simulation des données et de présentation du résumé des résultats, nous rappelons explicitement les objectifs spécifiques de l'étude échantillonnale du modèle de 2-copule échangeable.

- Comparer, en termes de précision, la méthode IFM généralisée et la méthode du maximum de vraisemblance. En effet, la méthode du maximum de vraisemblance maximise la log-vraisemblance $\mathcal{L}$ de (3.44) pour estimer les paramètres. Les erreurs types sont les racines carrées des éléments de la diagonale de l'inverse de la matrice d'information de Fisher observée. La méthode IFM généralisée quant à elle, maximise des pseudo-vraisemblances $\mathcal{L}_\alpha$, $\mathcal{L}_\beta$, $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ des équations (3.4), (3.5), (3.7), (3.8) et (3.10) respectivement. Nous inversons les matrices d'information de Fisher observée de chaque

pseudo-vraisemblance pour calculer à nouveau des pseudo-erreurs types de chaque paramètre. Donc intuitivement, nous notons une perte d'information en utilisant des pseudo-vraisemblances. Nous nous posons la question de savoir, à quelle proportion la méthode IFM généralisée sous-estime ou sur-estime la variance des estimateurs du modèle de 2-copule échangeable.

- Comparer la précision d'estimation dans le cas où les grappes sont *Équilibré* et *Non Équilibré*. Plus explicitement, il s'agit de comprendre si le fait d'avoir un grand nombre de grappes avec de grandes tailles et un petit nombre de grappes avec de petites tailles, a un impact sur la précision des estimations.
- Analyser la précision d'estimation lorsque le choix de la copule $C^{(2)}$ change et surtout l'impact de ce changement sur la précision de la variance de l'estimateur du paramètre δ_1 ou δ_3. Il s'agit ici d'établir une conclusion sur l'impact d'un changement de copule $C^{(2)}$ passant d'une copule à un paramètre (*normale*, *Clayton*) à une copule plus complexe (copule de *Khoudraji* à trois paramètres) sur la précision des estimations surtout sur le paramètre δ_3 compte tenu du fait que les copules $C^{(2)}$ et $\left(C^{(3)}_{1,n}\right)$ sont liées par les pseudo-observations résiduelles de l'équation (3.9).

Toutes les vraisemblances considérées s'obtiennent précisément en utilisant les hypothèses sur F, G, $\left(C^{(1)}_{1,n}\right)$, $C^{(2)}$ et $\left(C^{(3)}_{1,n}\right)$. La log-vraisemblance globale à maximiser pour l'estimation des paramètres, par la méthode du maximum de vraisemblance est

$$\mathcal{L}(\theta) = \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log\left\{f(x_{ji};\mu_1,\sigma_1^2)\right\} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log\left\{G(y_{ji};\mu_2,\sigma_2^2)\right\} + \mathcal{L}_3(\theta), \quad (3.44)$$

où la fonction $\mathcal{L}_3$ s'écrit

$$\begin{aligned}\mathcal{L}_3(\theta) &= \sum_{j=1}^{m} \log\left\{c^{(1)}_{1,n_j}\left(u_{j1},\ldots,u_{jn_j};\delta_1\right)\right\} + \sum_{j=1}^{m}\sum_{i=1}^{n_j} \log\left\{c^{(2)}(u_{ji},v_{ji};\delta_2)\right\} \\ &+ \sum_{j=1}^{m} \log\left[c^{(3)}_{1,n_j}\left\{C_{2|1}(v_{j1}|u_{j1}),\ldots,C_{2|1}(v_{jn_j}|u_{jn_j});\delta_3\right\}\right],\end{aligned}$$

et $u_{ji} = F(x_{ji};\mu_1,\sigma_1^2)$, $v_{ji} = G(y_{ji};\mu_2,\sigma_2^2)$ et $C_{2|1}(v|u) = \partial C^{(2)}(u,v;\delta_2)/\partial u$.

Pour des questions de convergence de l'algorithme d'estimation, nous considérons une réparamétrisation η d'un paramètre δ dans $]0,1[$ par $\eta = \log\{\delta/(1-\delta)\}$. La transformation s'applique à δ_1 et δ_3 par

$$\eta_1 = \log\left(\frac{\delta_1}{1-\delta_1}\right), \quad \eta_3 = \log\left(\frac{\delta_3}{1-\delta_3}\right). \quad (3.45)$$

En résumé, les paramètres à estimer sont $\theta = (\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\eta_1,\delta_2,\eta_3)$ et par la suite nous utilisons la transformation inverse $\delta = h(\eta) = \exp(\eta)/\{1+\exp(\eta)\}$ de (3.45) pour obtenir le vecteur $(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\delta_1,\delta_2,\delta_3)$.

Si $\hat{\eta}$ et $\hat{\delta}$ sont des estimateurs de η et δ respectivement. L'estimateur de la variance de $\hat{\delta}$ s'écrit en fonction de l'estimateur de la variance de $\hat{\eta}$ par

$$\widehat{\mathbb{V}}(\hat{\delta}) = \left\{h^{'}(\hat{\eta})\right\}^2 \widehat{\mathbb{V}}(\hat{\eta}) = \frac{\exp(2\hat{\eta})}{\{1+\exp(\hat{\eta})\}^4} \cdot \widehat{\mathbb{V}}(\hat{\eta}), \tag{3.46}$$

car $h^{'}(\eta) = \exp(\eta)/\{1+\exp(\eta)\}^2$. Nous calculons les biais relatifs des estimateurs de la variance de $\hat{\theta}$ par

$$\textit{biais relatif} = \left\{\mathbb{V}_B(\hat{\theta})\right\}^{-1} \left[\mathbb{E}_B\left\{\widehat{\mathbb{V}}(\hat{\theta})\right\} - \mathbb{V}_B(\hat{\theta})\right], \tag{3.47}$$

où nous avons

$$\mathbb{E}_B\left\{\widehat{\mathbb{V}}(\hat{\theta})\right\} \approx \frac{1}{B}\sum_{i=1}^{B} \widehat{\mathbb{V}}(\hat{\theta})_i.$$

Les formules des équations (3.46) et (3.47) s'appliquent pour l'estimateur $\tilde{\theta}$ issu de la méthode IFM généralisée.

Nous présentons les différentes étapes de l'étude Monte-Carlo en fixant $B = 1000$. Nous présentons en annexe C.4, le programme R, utilisé pour l'obtention des résultats des simulations dans le cas où la copule $C^{(2)}$ est archimédienne ou *normale*. Le programme **R** se modifie facilement pour s'adapter lorsque $C^{(2)}$ est une copule de *Khoudraji*.

Nous fixons le vecteur de paramètres $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2, \delta_1, \delta_2, \delta_3)$ du modèle de 2-copule échangeable.

Procédure de simulation et d'estimation de la matrice de variance-covariance du modèle de 2-copule échangeable par Monte-Carlo

Étape 1 : Construire une matrice de données (X, Y) suivant la 2-copule échangeable

1. Nous simulons m vecteurs $u_j = \left(u_{j1}, \ldots, u_{jn_j}\right)$ de taille n_j en utilisant la copule échangeable $\left(C^{(1)}_{1,n_j}\right), j = 1, \ldots, m$ de paramètre δ_1 ;
2. Nous simulons m vecteurs $w_j = \left(w_{j1}, \ldots, w_{jn_j}\right)$ de taille n_j en utilisant la famille de copules échangeables $\left(C^{(3)}_{1,n_j}\right)$ de paramètre δ_3 ;
3. Nous calculons m vecteurs $v_j = \left(v_{j1}, \ldots, v_{jn_j}\right)$ en résolvant l'équation $w_{ji} = C_{2|1}(v_{ji}|u_{ji})$, où $C_{2|1}$ provient de l'équation (3.9) ;
4. Nous posons $x_{ji} = \mu_1 + \sigma_1 \Phi^{-1}(u_{ji})$ et $y_{ji} = \mu_2 + \sigma_2 \Phi^{-1}(v_{ji})$.

Étape 2 : Utiliser les données obtenues à l'étape 1 pour estimer les paramètres du modèle

Nous estimons le paramètre θ puis les erreurs types associés du modèle de 2-copule échangeable à partir des données simulées à l'étape 1 en maximisant
- la vraisemblance globale de l'équation (3.44) pour la méthode du maximum de vraisemblance et noté $(\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$ et les erreurs types associées en inversant la matrice d'information de Fisher observée de la vraisemblance globale, voir (3.20).
- les pseudo-vraisemblances issues des équations (3.4), (3.5), (3.7), (3.8) et (3.10) pour la méthode IFM, noté $(\tilde{\mu}_1, \tilde{\sigma}_1, \tilde{\mu}_2, \tilde{\sigma}_2, \tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3)$ et les pseudo-erreurs types associées à l'inverse de la matrice d'information de Fisher observée des pseudo-vraisemblances $\mathcal{L}_\alpha$, $\mathcal{L}_\beta$, $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ respectivement.
L'estimation de la variance est égale au carré de l'erreur type.

Étape 3 : Répéter les étapes 1 et 2

1. Répéter les étapes 1 et 2, B fois pour avoir B estimations de θ et B erreurs associées aux estimations.
2. Nous calculons par la suite la moyenne et l'écart-type associées à chaque paramètre pour chaque méthode grâce au B estimations.

3.5.1 Cas où la copule $C^{(2)}$ est *normale*

Nous présentons les résultats issus de la simulation pour ce cas spécifique du 2-copule échangeable.

TABLEAU 3.2 – Espérances des estimateurs et leurs écarts-types Monte Carlo entre parenthèses, pour le modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule *normale*.

$\tau_2(\delta_2)$	m	*Types*	*Méthode*	$\mu_1 = 0$	$\sigma_1 = 1$	$\mu_1 = 0$	$\sigma_2 = 1$	$\delta_1 = 0.31$	δ_2	$\delta_3 = 0.16$
0.4(0.59)		*Non*	MV	0.00(0.18)	0.99(0.08)	0.01(0.16)	0.99(0.05)	0.28(0.10)	0.58(0.05)	0.14(0.07)
		Équilibré	IFM	0.00(0.19)	0.98(0.08)	0.01(0.17)	0.99(0.06)	0.27(0.10)	0.58(0.06)	0.13(0.07)
	10	*Équilibré*	MV	0.01(0.18)	0.99(0.08)	0.01(0.15)	0.99(0.05)	0.29(0.10)	0.59**(0.05)**	0.14**(0.07)**
			IFM	0.00(0.19)	0.98(0.08)	0.00(0.15)	0.99(0.06)	0.27(0.11)	0.58**(0.06)**	0.13**(0.07)**
		Non	MV	0.00(0.13)	0.99(0.06)	0.00(0.11)	0.99(0.05)	0.29(0.08)	0.59(0.04)	0.15(0.06)
		Équilibré	IFM	0.00(0.14)	0.99(0.06)	0.00(0.12)	0.99(0.05)	0.28(0.08)	0.59(0.05)	0.14(0.06)
	20	*Équilibré*	MV	0.00(0.13)	0.99(0.06)	0.01(0.11)	1.00(0.05)	0.30(0.08)	0.59**(0.04)**	0.15**(0.06)**
			IFM	0.00(0.13)	0.99(0.06)	0.00(0.12)	0.99(0.05)	0.29(0.08)	0.59**(0.05)**	0.14**(0.06)**
		Non	MV	0.01(0.08)	1.00(0.04)	0.01(0.07)	1.00(0.03)	0.30(0.05)	0.59(0.03)	0.15(0.03)
		Équilibré	IFM	0.00(0.09)	0.99(0.04)	0.01(0.08)	1.00(0.03)	0.30(0.05)	0.59(0.03)	0.15(0.04)
	50	*Équilibré*	MV	0.00(0.08)	1.00(0.04)	0.00(0.07)	1.00(0.03)	0.30(0.05)	0.59**(0.03)**	0.15**(0.04)**
			IFM	0.00(0.08)	1.00(0.04)	0.00(0.07)	1.00(0.03)	0.30(0.05)	0.59**(0.03)**	0.15**(0.04)**
0.6(0.81)		*Non*	MV	0.01(0.18)	0.99(0.08)	0.01(0.17)	0.99(0.06)	0.29(0.10)	0.81(0.03)	0.14(0.07)
		Équilibré	IFM	0.00(0.20)	0.98(0.08)	0.00(0.18)	0.99(0.07)	0.27(0.10)	0.81(0.03)	0.12(0.07)
	10	*Équilibré*	MV	0.01(0.18)	0.99(0.07)	0.01(0.16)	0.99(0.06)	0.29(0.10)	0.81**(0.03)**	0.14**(0.07)**
			IFM	0.00(0.18)	0.98(0.08)	0.00(0.17)	0.98(0.07)	0.27(0.10)	0.81**(0.03)**	0.13**(0.07)**
		Non	MV	0.01(0.13)	0.99(0.06)	0.01(0.12)	1.00(0.05)	0.29(0.08)	0.81(0.03)	0.15(0.06)
		Équilibré	IFM	0.01(0.14)	0.99(0.06)	0.00(0.13)	0.99(0.06)	0.29(0.08)	0.81(0.03)	0.14(0.06)
	20	*Équilibré*	MV	0.00(0.13)	1.00(0.06)	0.00(0.12)	1.00(0.05)	0.30(0.08)	0.81**(0.03)**	0.15**(0.05)**
			IFM	0.00(0.13)	0.99(0.06)	0.00(0.13)	0.99(0.06)	0.29(0.08)	0.81**(0.03)**	0.14**(0.06)**
		Non	MV	0.00(0.08)	1.00(0.04)	0.00(0.08)	1.00(0.03)	0.30(0.05)	0.81(0.01)	0.15(0.04)
		Équilibré	IFM	0.00(0.09)	0.99(0.04)	0.00(0.09)	0.99(0.04)	0.30(0.05)	0.81(0.02)	0.15(0.04)
	50	*Équilibré*	MV	0.01(0.08)	1.00(0.04)	0.01(0.07)	1.00(0.03)	0.30(0.05)	0.81**(0.01)**	0.15**(0.03)**
			IFM	0.00(0.08)	1.00(0.04)	0.00(0.08)	1.00(0.03)	0.30(0.05)	0.81**(0.02)**	0.15**(0.04)**

Les deux écarts types, IFM et MV, pour δ_3 sont presque toujours égaux. Pour δ_2 celui pour MV est en général plus petit IFM. En effet, $R^{'}$ est plus proche de 1 que R pour les valeurs de ρ_2, ρ_3 et n choisies, voir les figures 3.5 et 3.6. Globalement, le tableau 3.2 donne des résultats conforment avec les écarts-types théoriques calculés à la section 3.4.3.

Dans la sous-section 3.4.2 et le cas particulier 3, nous calculons les matrices asymptotiques de variance-covariance théoriques de l'estimateur du vecteur (δ_2, δ_3) dans le cas où les trois copules $\left(C_{1,n}^{(1)}\right)$, $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$ sont toutes normales et les marginales sont des lois normales et connues. Les formules théoriques des estimateurs de variances des paramètres $\delta_2 = \rho_2$ et $\delta_3 = \rho_3$ proviennent des équations (3.39) et (3.38) lorsque le nombre de grappes tend vers l'infini. En utilisant ces deux équations et pour n fixé, nous calculons les écarts-types théoriques en fonction de la méthode d'estimation lorsque le nombre de grappes est fini lorsque les marges sont connues.

En utilisant la méthode du maximum de vraisemblance, nous avons

$$\begin{aligned}
\mathbb{V}(\hat{\rho}_2) &= \frac{1}{m} \cdot \frac{(1-\rho_2^2)^2(\sim)}{(\sim)(\oplus) + \frac{(\sim)[**]}{(1-\rho_3)(\diamond)} - 2n(n-1)\rho_2^2\rho_3^2}, \\
\mathbb{V}(\hat{\rho}_3) &= \frac{2(1-\rho_3)(\diamond)}{n(n-1)m} \left[\frac{(1-\rho_3)(\diamond)(\oplus) + [**]}{(\sim)(\oplus) + \frac{(\sim)[**]}{(1-\rho_3)(\diamond)} - 2n(n-1)\rho_2^2\rho_3^2} \right],
\end{aligned}$$

où $(\oplus)$, $(\diamond)$, $(\sim)$ et $[**]$ proviennent des équations (3.36) et (3.37).

En utilisant la méthode IFM généralisée, nous avons

$$\mathbb{V}(\tilde{\rho}_2) = \frac{(1-\rho_2^2)^2}{m(\oplus)}, \ \mathbb{V}(\tilde{\rho}_3) = \frac{2(1-\rho_3)^2(\diamond)^2}{n(n-1)m(\sim)} \left\{ 1 + \frac{2n(n-1)\rho_2^2\rho_3^2}{(\oplus)(\sim)} \right\}.$$

Par ailleurs, nous rappelons que les calculs des écarts-types Monte Carlo à partir des simulations se font dans le scénario ***Équilibré*** et les résultats se trouvent dans le tableau 3.2 (en gras dans le tableau).

Le tableau 3.3 présente les résultats des calculs théoriques des écarts-types que nous comparons aux écarts-types issues de la méthode Monte-Carlo. Les écarts-types théoriques sont plus faibles que ceux issus de la simulation. De plus, l'écart diminue lorsque nous passons de $m = 10$ à $m = 50$.

TABLEAU 3.3 – Écarts-types des estimateurs $\hat{\delta}_2(\tilde{\delta}_2)$ et $\hat{\delta}_3(\tilde{\delta}_3)$ avec entre crochets les écarts-types Monte-Carlo lorsque $C^{(2)}$ est une copule *normale*.

$\tau_2(\delta_2)$	$m(n)$	*Méthode*	*Écart-type de* $\hat{\delta}_2[\tilde{\delta}_2]$	*Écart-type de* $\hat{\delta}_3[\tilde{\delta}_3]$
0.4(0.59)	10(30)	MV	0.02[0.05]	0.04[0.07]
		IFM	0.02[0.06]	0.06[0.07]
	20(15)	MV	0.02[0.04]	0.05[0.06]
		IFM	0.03[0.05]	0.05[0.06]
	50(18)	MV	0.01[0.03]	0.03[0.04]
		IFM	0.01[0.03]	0.03[0.04]
0.6(0.81)	10(30)	MV	0.01[0.03]	0.04[0.07]
		IFM	0.01[0.03]	0.05[0.07]
	20(15)	MV	0.01[0.03]	0.05[0.05]
		IFM	0.01[0.03]	0.05[0.06]
	50(18)	MV	0.01[0.01]	0.03[0.03]
		IFM	0.01[0.02]	0.03[0.04]

Nous présentons le biais relatif des estimateurs de variances, calculés à partir de la formule (3.47) dans le cas où la copule $C^{(2)}$ est *normale*.

Dans le cas où toutes les copules sont normales et les marginales sont normales, nous nous retrouvons dans le cas standard du modèle de Battese *et al.* (1988). Pour le cas spécifique du paramètre μ_1, les variances des estimateurs $\tilde{\mu}_1$ et $\hat{\mu}_1$ issus des méthodes IFM généralisée et maximum de vraisemblance respectivement dans le cas de grappes *Équilibré* sont évaluées à l'aide des formules

$$\mathbb{V}_{IFM}(\tilde{\mu}_1) = \frac{\sigma_1^2}{n \cdot m}, \qquad \mathbb{V}_{MV}(\hat{\mu}_1) = \frac{\sigma_1^2}{n \cdot m}\left\{1 + (n-1)\rho_1\right\}\cdot \tag{3.48}$$

Le résultat nous permet de conclure facilement que les variances issues des deux méthodes diffèrent d'un facteur multiplicatif de $1 + (n-1)\rho_1$. En conséquence, plus le nombre d'individus dans la grappe est grand, plus la méthode IFM généralisée sous-estime la variance de l'estimateur. L'estimateur de la variance des paramètres construits à partir de la méthode IFM généralisée n'est pas convergent comme en témoignent les biais relatifs rapportés dans le tableau 3.4.

TABLEAU 3.4 – Biais relatif des estimateurs de la variance des paramètres du modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule *normale*.

$\tau_2(\delta_2)$	m	*Types*	*Méthode*	μ_1	σ_1	μ_1	σ_2	δ_1	δ_2	δ_3
0.4(0.59)		*Non*	MV	< 1	19	-5	11	3	8	4
	10	*Équilibré*	IFM	-91	-74	-88	-58	-79	-69	-36
		Équilibré	MV	-4	22	5	18	6	-1	22
			IFM	-91	-72	-85	-52	-98	-68	-37
	20	*Non*	MV	< 1	5	-2	9	-7	3	-8
		Équilibré	IFM	-83	-61	-77	-39	-83	-59	-44
		Équilibré	MV	2	7	-3	6	5	3	8
			IFM	-81	-58	-76	-38	-54	-55	-60
	50	*Non*	MV	11	5	13	10	-6	2	7
		Équilibré	IFM	-87	-68	-82	-47	-68	-66	-30
		Équilibré	MV	4	10	7	7	8	$< 10^{-2}$	-3
			IFM	-84	-61	-77	-42	-60	-61	-84
0.6(0.81)		*Non*	MV	< 1	32	-1	23	4	5	5
	10	*Équilibré*	IFM	-92	-74	-90	-67	-91	-77	-28
		Équilibré	MV	4	31	2	29	8	12	17
			IFM	-90	-71	-89	-64	-98	-70	-37
	20	*Non*	MV	-3	9	$< 10^{-2}$	6	-2	7	13
		Équilibré	IFM	-83	-59	-80	-51	-57	-65	-71
		Équilibré	MV	2	15	-2	2	5	3	16
			IFM	-82	-57	-79	-53	-89	-66	-75
	50	*Non*	MV	3	15	1	19	12	11	5
		Équilibré	IFM	-87	-67	-85	-56	-65	-70	-34
		Équilibré	MV	6	10	8	12	6	8	1
			IFM	-85	-62	-82	-51	-62	-68	-94

Pour la suite de cette étude Monte-Carlo, nous passons à un autre choix de la copule $C^{(2)}$.

3.5.2 Cas où $C^{(2)}$ est la copule de *Clayton*

Nous nous intéressons par la suite au cas où la copule $C^{(2)}$ est une copule de *Clayton*.

TABLEAU 3.5 – Espérances des estimateurs et leurs écarts-types Monte Carlo entre parenthèses, pour le modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule de *Clayton*.

$\tau_2(\delta_2)$	m	*Types*	*Méthode*	$\mu_1 = 0$	$\sigma_1 = 1$	$\mu_1 = 0$	$\sigma_2 = 1$	$\delta_1 = 0.31$	δ_2	$\delta_3 = 0.16$
0.4(1.33)		*Non*	MV	0.00(0.13)	0.99(0.05)	0.00(0.11)	0.99(0.05)	0.29(0.07)	1.35(0.21)	0.15(0.05)
	10	*Équilibré*	IFM	0.00(0.14)	0.98(0.06)	0.01(0.12)	0.99(0.06)	0.29(0.08)	1.32(0.23)	0.14(0.06)
		Équilibré	MV	0.01(0.17)	0.99(0.07)	0.00(0.14)	0.99(0.06)	0.29(0.09)	1.34(0.26)	0.14(0.06)
			IFM	0.01(0.18)	0.98(0.08)	0.01(0.14)	0.99(0.06)	0.27(0.10)	1.32(0.29)	0.13(0.06)
	20	*Non*	MV	0.00(0.13)	0.99(0.05)	0.00(0.11)	0.99(0.05)	0.29(0.07)	1.34(0.21)	0.15(0.05)
		Équilibré	IFM	0.00(0.14)	0.99(0.06)	0.00(0.12)	0.99(0.06)	0.29(0.08)	1.32(0.23)	0.14(0.06)
		Équilibré	MV	0.00(0.13)	1.00(0.05)	0.00(0.11)	1.00(0.05)	0.30(0.07)	1.35(0.22)	0.15(0.06)
			IFM	0.00(0.13)	0.99(0.06)	0.00(0.11)	0.99(0.05)	0.29(0.08)	1.34(0.24)	0.14(0.06)
	50	*Non*	MV	0.00(0.08)	1.00(0.03)	0.00(0.07)	1.00(0.03)	0.30(0.04)	1.34(0.13)	0.15(0.04)
		Équilibré	IFM	0.00(0.09)	0.99(0.04)	0.00(0.08)	1.00(0.04)	0.30(0.05)	1.33(0.16)	0.15(0.04)
		Équilibré	MV	0.00(0.08)	1.00(0.03)	0.00(0.07)	1.00(0.03)	0.30(0.04)	1.34(0.13)	0.15(0.03)
			IFM	0.00(0.08)	1.00(0.04)	0.00(0.07)	1.00(0.03)	0.30(0.05)	1.34(0.14)	0.15(0.04)
0.6(3)		*Non*	MV	0.00(0.16)	1.00(0.06)	0.00(0.14)	0.99(0.06)	0.29(0.09)	3.05(0.51)	0.14(0.07)
	10	*Équilibré*	IFM	-0.01(0.20)	0.98(0.08)	-0.01(0.17)	0.98(0.07)	0.27(0.11)	2.99(0.59)	0.13(0.07)
		Équilibré	MV	0.01(0.15)	0.99(0.06)	0.00(0.14)	0.99(0.06)	0.29(0.09)	3.04(0.50)	0.14(0.07)
			IFM	0.01(0.18)	0.98(0.08)	0.01(0.16)	0.99(0.07)	0.27(0.11)	2.97(0.57)	0.13(0.07)
	20	*Non*	MV	0.00(0.11)	0.99(0.05)	0.00(0.11)	0.99(0.05)	0.30(0.07)	3.02(0.41)	0.15(0.06)
		Équilibré	IFM	0.00(0.13)	0.99(0.06)	0.00(0.12)	0.99(0.06)	0.29(0.08)	2.98(0.45)	0.14(0.06)
		Équilibré	MV	0.00(0.12)	1.00(0.05)	0.00(0.12)	1.00(0.05)	0.30(0.07)	3.05(0.44)	0.15(0.06)
			IFM	0.00(0.14)	0.99(0.06)	0.00(0.13)	0.99(0.06)	0.29(0.08)	3.00(0.47)	0.14(0.06)
	50	*Non*	MV	0.00(0.07)	1.00(0.03)	0.00(0.07)	1.00(0.03)	0.31(0.04)	3.02(0.26)	0.15(0.04)
		Équilibré	IFM	0.00(0.10)	0.99(0.04)	0.00(0.09)	1.00(0.04)	0.30(0.05)	3.00(0.30)	0.15(0.04)
		Équilibré	MV	0.00(0.07)	1.00(0.03)	0.00(0.07)	1.00(0.03)	0.31(0.04)	3.01(0.24)	0.15(0.03)
			IFM	0.00(0.09)	1.00(0.04)	0.00(0.08)	1.00(0.03)	0.30(0.05)	2.99(0.28)	0.15(0.03)

Le tableau C.1 de l'annexe C donne les résultats du biais relatif des estimateurs de variances lorsque $C^{(2)}$ est une copule de *Clayton*.

3.5.3 Cas où $C^{(2)}$ est une copule de *Khoudraji*

Nous considérons cette fois-ci que $C^{(2)}$ est une copule de *Khoudraji* de paramètres δ_2. Plus précisément, de la définition 1.9, nous considérons que C_1 est une copule d'indépendance et C_2 est une copule normale de paramètre κ. Nous avons donc $\delta_2 = (\kappa, \kappa_1, \kappa_2)$.

Dans le but de faciliter la convergence de l'algorithme pour maximiser les vraisemblances, nous utilisons la reparamétrisation de l'équation (3.45) pour les trois paramètres κ, κ_1 et κ_2. Nous choisissons toujours deux valeurs de corrélation de $\tau_2 \in \{0.4, 0.6\}$. Pour $\tau_2 = 0.4$, nous choisissons les paramètres $(\kappa, \kappa_1, \kappa_2) = (0.68, 0.82, 0.96)$. Pour $\tau_2 = 0.6$, nous choisissons les paramètres $(\kappa, \kappa_1, \kappa_2) = (0.81, 0.97, 0.89)$. Nous présentons les résultats dans le tableau 3.6. Le tableau C.2, à l'annexe C.4 donne le résultat des biais relatifs des estimateurs de la variance des paramètres. Nous notons une tendance opposée à celle des simulations précédentes du fait que la méthode IFM est plus précise que la méthode du maximum de vraisemblance pour les paramètres κ_1 et κ_2.

TABLEAU 3.6 – Espérance des estimateurs et leurs écarts-types entre parenthèses, pour le modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule de *Khoudraji*.

τ_2	m	*Types*	*Méthode*	μ_1	σ_1	μ_1	σ_2	δ_1	$\kappa = 0.68$	$\kappa_1 = 0.82$	$\kappa_2 = 0.96$	δ_3
0.4		*Non*	MV	0.00(0.19)	0.99(0.08)	0.00(0.15)	0.99(0.05)	0.28(0.10)	0.69(0.06)	0.82(0.10)	0.95(0.05)	0.14(0.07)
	10	*Équilibré*	IFM	0.00(0.20)	0.98(0.08)	0.00(0.16)	0.98(0.06)	0.27(0.11)	0.67(0.06)	0.83(0.08)	0.96(0.04)	0.13(0.07)
		Équilibré	MV	0.00(0.18)	0.99(0.08)	0.00(0.15)	0.99(0.05)	0.28(0.10)	0.68(0.06)	0.82(0.10)	0.95(0.06)	0.14(0.07)
			IFM	0.00(0.18)	0.98(0.08)	0.00(0.15)	0.99(0.06)	0.27(0.11)	0.67(0.06)	0.82(0.08)	0.95(0.05)	0.14(0.07)
		Non	MV	0.00(0.13)	0.99(0.08)	0.00(0.12)	0.99(0.06)	0.29(0.08)	0.68(0.06)	0.82(0.10)	0.95(0.07)	0.14(0.06)
	20	*Équilibré*	IFM	0.00(0.14)	0.99(0.08)	0.00(0.12)	0.99(0.07)	0.29(0.09)	0.67(0.06)	0.83(0.08)	0.95(0.06)	0.14(0.06)
		Équilibré	MV	0.01(0.14)	0.99(0.06)	0.01(0.11)	0.99(0.05)	0.29(0.07)	0.68(0.06)	0.82(0.10)	0.95(0.05)	0.15(0.06)
			IFM	0.01(0.14)	0.99(0.06)	0.01(0.11)	0.99(0.05)	0.29(0.08)	0.67(0.06)	0.83(0.07)	0.96(0.03)	0.15(0.06)
		Non	MV	0.00(0.08)	1.00(0.04)	0.00(0.06)	1.00(0.03)	0.30(0.05)	0.68(0.03)	0.82(0.06)	0.96(0.03)	0.15(0.04)
	50	*Équilibré*	IFM	0.00(0.09)	0.99(0.04)	0.00(0.07)	1.00(0.03)	0.30(0.05)	0.68(0.03)	0.82(0.05)	0.96(0.02)	0.15(0.04)
		Équilibré	MV	0.00(0.08)	1.00(0.04)	0.00(0.07)	1.00(0.03)	0.31(0.05)	0.68(0.03)	0.82(0.06)	0.95(0.04)	0.15(0.04)
			IFM	0.00(0.08)	1.00(0.04)	0.00(0.07)	1.00(0.03)	0.30(0.05)	0.67(0.03)	0.83(0.04)	0.96(0.03)	0.15(0.04)

τ_2	m	*Types*	*Méthode*	μ_1	σ_1	μ_1	σ_2	δ_1	$\kappa = 0.81$	$\kappa_1 = 0.97$	$\kappa_2 = 0.89$	δ_3
0.6		*Non*	MV	0.01(0.17)	1.00(0.07)	-0.01(0.15)	1.00(0.06)	0.30(0.08)	0.87(0.03)	0.97(0.03)	0.89(0.05)	0.15(0.08)
	10	*Équilibré*	IFM	0.00(0.19)	0.99(0.08)	-0.01(0.17)	0.99(0.07)	0.27(0.12)	0.86(0.03)	0.97(0.03)	0.89(0.04)	0.15(0.08)
		Équilibré	MV	0.01(0.17)	1.01(0.08)	0.01(0.17)	1.01(0.06)	0.31(0.09)	0.87(0.02)	0.97(0.03)	0.89(0.05)	0.14(0.07)
			IFM	0.01(0.18)	1.00(0.09)	0.01(0.18)	0.99(0.07)	0.29(0.11)	0.87(0.02)	0.97(0.03)	0.89(0.05)	0.13(0.07)
		Non	MV	0.01(0.12)	1.00(0.06)	0.01(0.12)	1.00(0.05)	0.31(0.08)	0.87(0.02)	0.97(0.02)	0.89(0.04)	0.15(0.06)
	20	*Équilibré*	IFM	0.00(0.13)	0.99(0.06)	0.00(0.12)	0.99(0.06)	0.29(0.08)	0.87(0.02)	0.97(0.03)	0.89(0.02)	0.14(0.07)
		Équilibré	MV	-0.01(0.13)	1.00(0.06)	0.00(0.11)	1.00(0.05)	0.31(0.07)	0.87(0.02)	0.97(0.03)	0.89(0.05)	0.15(0.06)
			IFM	-0.02(0.14)	0.99(0.06)	-0.01(0.12)	0.99(0.06)	0.30(0.08)	0.87(0.02)	0.97(0.02)	0.89(0.05)	0.14(0.06)
		Non	MV	0.00(0.08)	1.00(0.04)	-0.01(0.07)	1.00(0.03)	0.32(0.04)	0.87(0.01)	0.97(0.01)	0.89(0.03)	0.15(0.04)
	50	*Équilibré*	IFM	-0.01(0.10)	0.99(0.04)	-0.01(0.10)	0.99(0.03)	0.31(0.06)	0.87(0.01)	0.97(0.01)	0.89(0.01)	0.15(0.04)
		Équilibré	MV	-0.01(0.06)	1.00(0.03)	0.00(0.06)	1.00(0.03)	0.31(0.05)	0.87(0.01)	0.97(0.01)	0.89(0.02)	0.15(0.03)
			IFM	-0.01(0.08)	1.00(0.04)	0.00(0.08)	0.99(0.04)	0.31(0.06)	0.87(0.01)	0.97(0.01)	0.89(0.02)	0.15(0.03)

Discussion de l'ensemble des résultats obtenus de l'étude Monte-Carlo

Au terme des résultats de simulation et de la présentation des tableaux, nous formulons les conclusions suivantes sur le modèle de 2-copule échangeable.

- Lorsque $C^{(2)}$ est une copule *normale*, de *Clayton* ou de *Khoudraji*, les espérances des estimateurs de paramètres sont quasiment les mêmes en utilisant les deux méthodes.
- La méthode du maximum de vraisemblance est "meilleure" que la méthode IFM en termes de précision des estimateurs. Les estimateurs de la variance par la méthode IFM généralisée ne sont pas convergents. En témoigne les tableaux 3.4, C.1 et C.2 portant sur les biais relatifs. Cependant, lorsque $C^{(2)}$ est une copule de *Khoudraji*, la méthode IFM semble un peu plus précise que la méthode MV concernant l'estimation des paramètres κ_1 et κ_2.
- La précision des estimations croît au fur et à mesure que le nombre de grappes croît passant de $m = 10$, $m = 20$ et $m = 50$.
- La précision des estimations ne change pas lorsque nous considérons le cas de grappes *Équilibré* et *Non Équilibré*.
- La complexité de la copule $C^{(2)}$ n'impacte pas significativement l'écart-type de l'estimateur du paramètre de la famille de copules $\left(C_{1,n}^{(3)}\right)$ ou celle de la famille de copules échangeables $\left(C_{1,n}^{(1)}\right)$.

Nous avons jugé important de ne pas d'alourdir la présentation des résultats en élargissant le choix des familles de copules échangeables $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$.

3.6 Conclusion

Dans ce chapitre, nous montrons comment on peut partir des données en grappes pour sélectionner les différents éléments impliqués dans la modélisation avec une 2-copule échangeable. Pour faire cette sélection, nous expliquons dans un premier temps le choix les lois marginales F et G et des copules $\left(C_{1,n}^{(1)}\right)$, $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$. Par la suite, nous montrons l'estimation des paramètres associés par deux méthodes : la méthode du maximum de vraisemblance et la méthode IFM généralisée. Ces méthodes conduisent à des estimateurs sans biais, asymptotiquement normaux. Dans un deuxième temps, nous comparons les deux méthodes d'estimation en termes de matrice variance-covariance puis terminons par l'étude des propriétés échantillonnales des paramètres à partir d'une simulation Monte-Carlo.

Dans la suite de cette thèse, le chapitre 4 donne une application du modèle de 2-copule échangeable sur des données réelles.

Chapitre 4

Modélisation des données avec un modèle de 2-copule échangeable

4.1 Mise en contexte

L'éducation constitue un socle essentiel sur lequel reposent les systèmes évolutifs. Elle permet de former la génération future en instituant des idéaux et des normes apprises dans un contexte de compétences pluridimensionnelles. Pour ce faire, l'état doit mettre en œuvre des programmes qui tendent à mieux outiller aussi bien les apprenants que les formateurs pour une satisfaction morale, intellectuelle et sociale viable et qui leur permettent d'interagir dans leur milieu de vie. Les collèges et lycées sont des lieux où l'apprentissage débute pour les touts petits. Les objectifs finaux sont en perpétuelle amélioration par rapport à ceux fixés au départ et il faut une adaptation rapide et conséquente. Pour ce faire, il s'agit de se baser sur les caractéristiques individuelles et les informations démographiques pour analyser et corriger les éventuels problèmes puis d'orienter et les solutionner pour mieux appréhender leur mécanisme.

Notre travail dans ce chapitre est de tirer profit des informations portant sur les apprenants lors de leur passage au primaire pour faire des recommandations sur les processus d'amélioration à la fin du curcus d'une part et d'autre part, de discuter des possibilités d'amélioration. Précisément, nous disposons des observations sur les notes en mathématiques des élèves au cours de la 4e et 7e année de leur curcus scolaire et nous définissons deux objectifs spécifiques :

— modéliser la note en 7e année en fonction de la note en 4e année en utilisant le modèle de 2-copule échangeable ;

— Utiliser le modèle construit pour prédire la note en 7e année en mathématique sachant la note en 4e année ;

— comparer le modèle construit aux modèles issus des régressions linéaire et non paramé-

trique.

Pour ce faire, nous structurons le chapitre en trois parties (1) la première partie décrit les données et l'ajustement du modèle linéaire mixte ; (2) la deuxième partie détaille le choix des éléments entrant dans la construction du modèle de 2-copule échangeable et l'estimation des paramètres associés à ce modèle et (3) la troisième fait une comparaison du modèle avec le modèle linéaire mixte de la première partie et avec le modèle de la régression non-paramétrique.

4.2 Description des données de l'échantillon

Les données proviennent de 48 écoles primaires où ils ont fait le suivi des élèves pendant trois années (de 8 à 11 ans) de leur cursus d'étude entre les années 2012 et 2015 au centre de Londres, dans le cadre du «Junior School Project» (JSP). Les élèves de chaque école ont subi des examens en mathématiques au cours de leur quatrième année (huitième anniversaire) et leur dernière année du primaire (onzième anniversaire). Ainsi, il a été sélectionné un certain nombre d'élèves par école, variant d'une école à une autre et au total $N = 728$ élèves dans l'échantillon d'étude. Du livre de Goldstein (2011, page 15), les données proviennent d'une base plus complète utilisée dans l'article Mortimore *et al.* (1988). Deux variables sont mesurées sur chacun des individus à savoir :

— la variable `math1` qui est la note sur 40, obtenue par l'élève au cours de la quatrième année du primaire ;

— la variable `math3` qui est la note sur 40, obtenue par l'élève au cours de la septième année du primaire.

En résumé, les données se présentent sous la forme de grappes où la grappe est l'école et les individus de la grappe sont les élèves. Nous présentons une esquisse de deux écoles sur la figure 4.1.

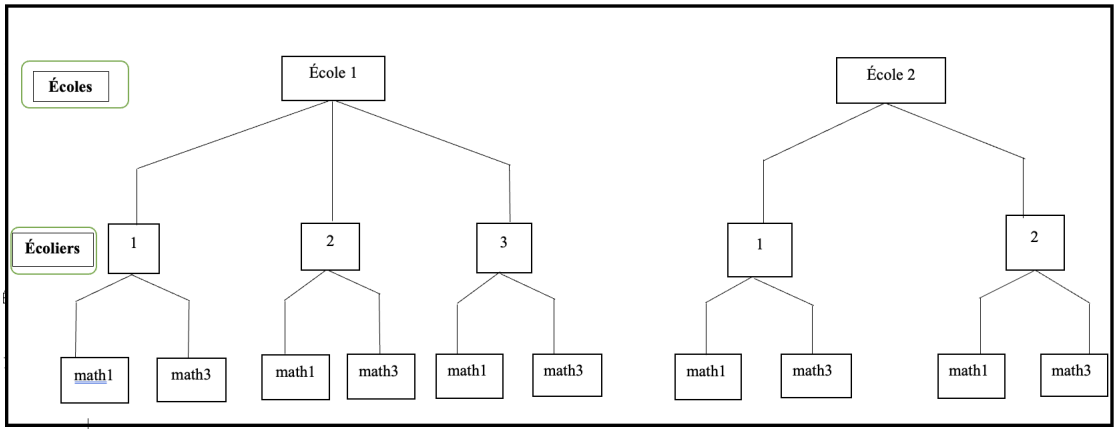


FIGURE 4.1 – Présentation hiérarchique partielle des données de deux écoles.

Les écoles (notés ici École 1 et École 2) représentent les écoles et les individus (notés ici 1, 2 et 3) sont les élèves dont le nombre varie en fonction de l'école.

Pour ces données, nous souhaitons modéliser la variable `math3` en fonction de la variable `math1` en tenant compte de l'effet-école (effet grappe). Au vu de la structure des données, nous notons la corrélation entre les notes des élèves d'une même école et celle entre deux notes (`math1` et `math3`) d'un même élève. Nous tentons de répondre aux préoccupations suivantes formulées sous forme d'hypothèses.

- Il existe une variation significative des performances scolaires en mathématiques des élèves au fil du temps, tant au sein des élèves que des écoles.
- Les performances scolaires des élèves en mathématique suivent une tendance non linéaire au fil du temps.
- Il existe une relation significative qui permet de prédire la note en septième année sachant la note en quatrième année et l'école de provenance de l'individu.

De nombreuses méthodes de modélisation de ces types de données sont possibles comme celle de modèle multiniveau décrit dans les chapitres 2 et 3 du livre de Goldstein (2011) ou la méthode de la régression copule, voir Noh *et al.* (2013) ou la régression non paramétrique, voir Cleveland *et al.* (1992). Nous proposons ici une méthode s'apparentant à la régression copule et vérifiant l'hypothèse d'échangeabilité au sein d'une école, notion définie dans le chapitre 2. Plus spécifiquement, nous ajustons le modèle de 2-copule échangeable aux données.

Nous proposons une analyse descriptive sommaire des variables impliquées qui se résume dans le tableau 4.1 et *EC* est l'écart-type.

TABLEAU 4.1 – Quelques statistiques descriptives des variables `math1` et `math3`.

Variable	*Moyenne*	*Médiane*	*EC*	*Asymétrie*	*Aplatissement*
`math1`	25.97	26	6.96	$-0.35(< 0.05)$	$2.62(< 0.05)$
`math3`	30.72	32	6.61	$-1.1(< 0.05)$	$3.89(< 0.05)$

Les valeurs négatives (-0.35 et -1.11) et la vapeur de p-valeur (< 0.05), nous permettent de conclure sur l'asymétrie de la distribution de la note en quatrième année et en septième année. Le test d'Agostino permet de confirmer que la normalité ne puisse pas être acceptée. Nous adoptons les notations suivantes

— y_{ji}, la note en septième année pour un individu i de l'école j et

— x_{ji} la note en quatrième année pour un individu i de l'école j.

Dans la section qui suit, nous proposons la modélisation linéaire mixte pour ces données et par la suite le modèle de 2-copule.

4.3 Modélisation par un modèle linéaire mixte

Le premier niveau est constitué des individus, le deuxième niveau, les écoles. Le modèle linéaire mixte considéré où la variable `math1` est une variable explicative s'énonce par :

$$y_{ji} = \beta_0 + u_{0j} + \beta_1 x_{ji} + e_{ji},\ i = 1, \ldots, n_j,\ \ j = 1, \ldots, 48, \tag{4.1}$$

où n_j est le nombre d'individus dans l'école j et les hypothèses suivantes sont

$$u_{0j} \sim \mathcal{N}(0, \sigma^2_{u_0}), \qquad e_{ji} \sim \mathcal{N}(0, \sigma^2_e)\cdot$$

L'effet aléatoire de la grappe j est u_{0j} et e_{ji} est l'erreur expérimentale ou de prédiction. Le modèle de l'équation (4.1) décrit une relation pour chacune des m écoles. Les paramètres estimés du modèle, en utilisant le logiciel *R*, se présentent dans le tableau 4.2.

TABLEAU 4.2 – Paramètres estimés et erreurs types entre parenthèses du modèle multiniveau de l'équation (4.1).

Effet	*Paramètres*	*Estimation (Erreur type)*
Effet fixe		
	β_0	13.94(0.71)
	β_1	0.65(0.03)
Effet aléatoire		
	σ^2_u	3.28(1.81)
	σ^2_e	19.80(4.45)

La corrélation intraclasse pour les erreurs expérimentales est notée $\widehat{ICC} = 0.14$. Ces résultats sont obtenus aussi par Goldstein (2011, page 29). À partir de ce modèle, nous prédisons la note en septième année pour de nouveaux individus dont nous connaissons la note en quatrième année en tenant compte de l'école.

L'approche de la régression linéaire mixte est optimale lorsque l'hypothèse de normalité de la variable `math3` est vérifiée. Pour ce modèle linéaire mixte de l'équation (4.1), nous notons plusieurs limites liées principalement à la prédiction.

— La prédiction de la note en septième année (Y) connaissant la note en quatrième (X) peut être en dehors de l'intervalle $[0, 40]$, voir la droite en couleure noire sur la figure 4.11. Ceci vient de la difficulté du modèle à tenir compte de la bornitude des variables.

— Le modèle linéaire mixte construit prédit des grandes valeurs de Y pour de grandes valeurs de X. Une grande sensibilité du modèle aux grandes valeurs. Ceci se confirme sur les deux droites de la figure 4.11.

— Le prédicteur construit à partir du modèle linéaire mixte est linéaire. Dans le cas de présence de courbure dans les données, ce modèle ne permet pas d'établir des liens non linéaires entre Y et X.

Dans la section suivante, nous décrivons les différentes étapes de la procédure de 2-copule échangeable que nous proposons pour modéliser les données. Le modèle est intéressant pour prendre en compte les cas où la normalité des résidus n'est pas acquise. Pour un traitement plus raffiné des données avec le 2-copule échangeable, nous changeons l'échelle des variables `math1` et `math3`. Pour faire la prédiction avec la 2-copule échangeable, nous cherchons à revenir à l'intervalle $[0, 40]$.

4.4 Modélisation par un modèle de 2-copule échangeable

Dans cette section, nous expliquons le choix des différents éléments entrant dans le cadre de la construction d'un modèle de 2-copule échangeable sur les notes en mathématiques. Pour faire référence au chapitre 2, le nombre de grappes (ici les écoles) est $m = 48$ et chaque école j dispose de n_j élèves. Nous supposons qu'il existe une 2-copule échangeable, voir (2.17) permettant de modéliser les données. Dans le cas précis et discuté dans cette section, le modèle échangeable défini par l'équation (2.17) se réduit à

$$\begin{aligned} c_{2,n}\left\{(u_1,v_1),\ldots,(u_n,v_n)\right\} &= c_{1,n}^{(1)}(u_1,\ldots,u_n;\delta_1)\prod_{i=1}^{n}c^{(2)}(u_i,v_i;\delta_2) \\ &\quad c_{1,n}^{(3)}\left\{C_{2|1}(v_1|u_1),\ldots,C_{2|1}(v_n|u_n);\delta_3\right\} \end{aligned} \tag{4.2}$$

Les fonctions $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ appartiennent à des familles de copules échangeables en dimension n. La copule $C^{(2)}$ est une copule bivariée et $C_{2|1}$ se définit par

$$C_{2|1}(v|u) = \partial C^{(2)}(u,v;\delta_2)/\partial u, \text{ pour } u,v \in [0,1]\cdot \tag{4.3}$$

La densité de la distribution jointe des variables aléatoires dans une grappe à n individus s'écrit

$$\begin{aligned} f_{2,n}\left\{(x_1,y_1),\ldots,(x_n,y_n)\right\} &= c_{1,n}^{(1)}\left\{F(x_1;\alpha),\ldots,F(x_n;\alpha);\delta_1\right\}\times \\ &\quad \prod_{i=1}^{n}\left[f(x_i;\alpha)g(y_i;\beta)c^{(2)}\left\{F(x_i;\alpha),G(y_i;\beta);\delta_2\right\}\right]\times \\ &\quad c_{1,n}^{(3)}\left[C_{2|1}\left\{G(y_1;\beta)|F(x_1;\alpha)\right\},..,C_{2|1}\left\{G(y_n;\beta)|F(x_n;\alpha)\right\};\delta_3\right] \end{aligned} \tag{4.4}$$

En résumé, cinq éléments sont à déterminer suivant deux points successifs :

- l'identification des lois marginales pour les notes de la quatrième année et de la septième année notées respectivement F et G ;
- la détermination de la copule échangeable associée à la note de la quatrième année au sein d'une même école d'une part notée $\left(C_{1,n}^{(1)}\right)$ et la copule associée aux notes de la

quatrième année et de la septième année notée $C^{(2)}$ d'autre part puis de déboucher sur la copule échangeable $\left(C_{1,n}^{(3)}\right)$ associée à la dépendance résiduelle dans une école.

Par la suite, nous utilisons le modèle de 2-copule échangeable pour prédire la note en septième année sur de nouveaux élèves dont nous connaissons la note en quatrième année d'école. Nous expliquons dans les prochaines sections, le choix des différents éléments entrant dans la construction du modèle de l'équation (4.2).

4.4.1 Choix des fonctions de répartitions marginales F et G

Dans cette section, nous ignorons la structure de grappe des observations c'est-à-dire les observations à l'intérieur d'une école sont indépendantes d'un élève à un autre. Nous rappelons que les lois F et G sont associées aux variables `math1` et `math3` respectivement. Nous transformons les notes de mathématiques en quatrième et septième année en des observations appartenant à l'intervalle $]0, 1[$. Nous divisons la note en mathématiques par 41 pour éviter d'atteindre la borne 1 pour les valeurs différentes de 0. Les observations transformées s'utilisent pour la suite de l'ajustement. La première étape est de construire l'histogramme et la densité pour chacune des variables dans le but de postuler des modèles paramétriques candidats. Nous présentons les graphiques présentant un histogramme des données et une courbe de la densité paramétrique de la loi bêta à deux paramètres, voir 1.6, postulée par exemple. Ce graphique se fait pour avoir un effet visuel d'une proposition de loi candidates en guise d'illustration.

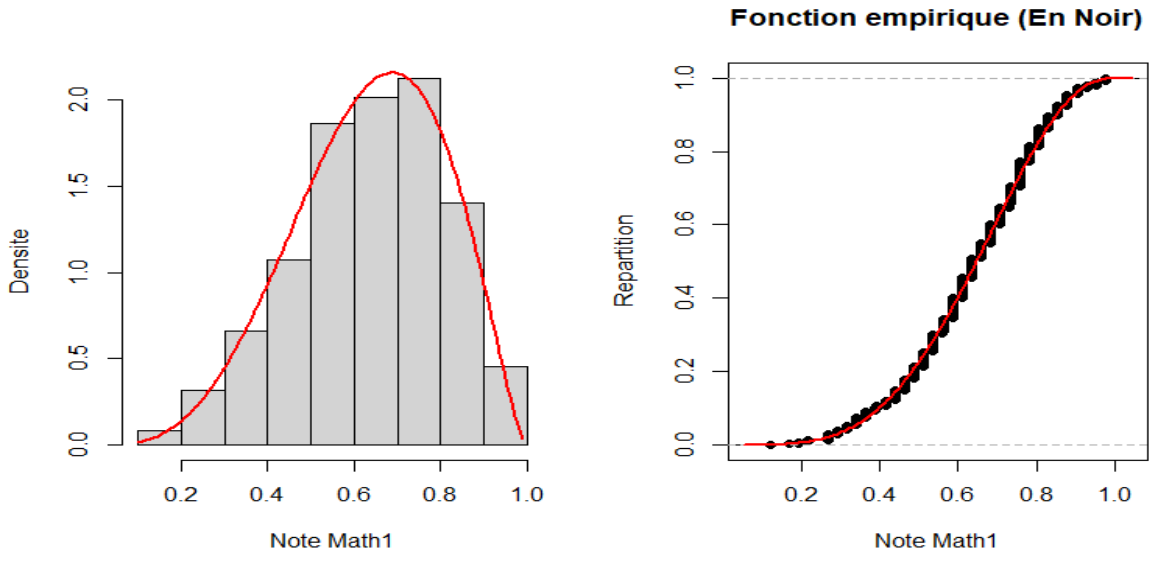


FIGURE 4.2 – Histogramme de la note `math1` et de la courbe de la loi bêta de paramètres $\tilde{\alpha}_1 = 4.38(0.22)$ et $\tilde{\beta}_1 = 2.52(0.12)$.

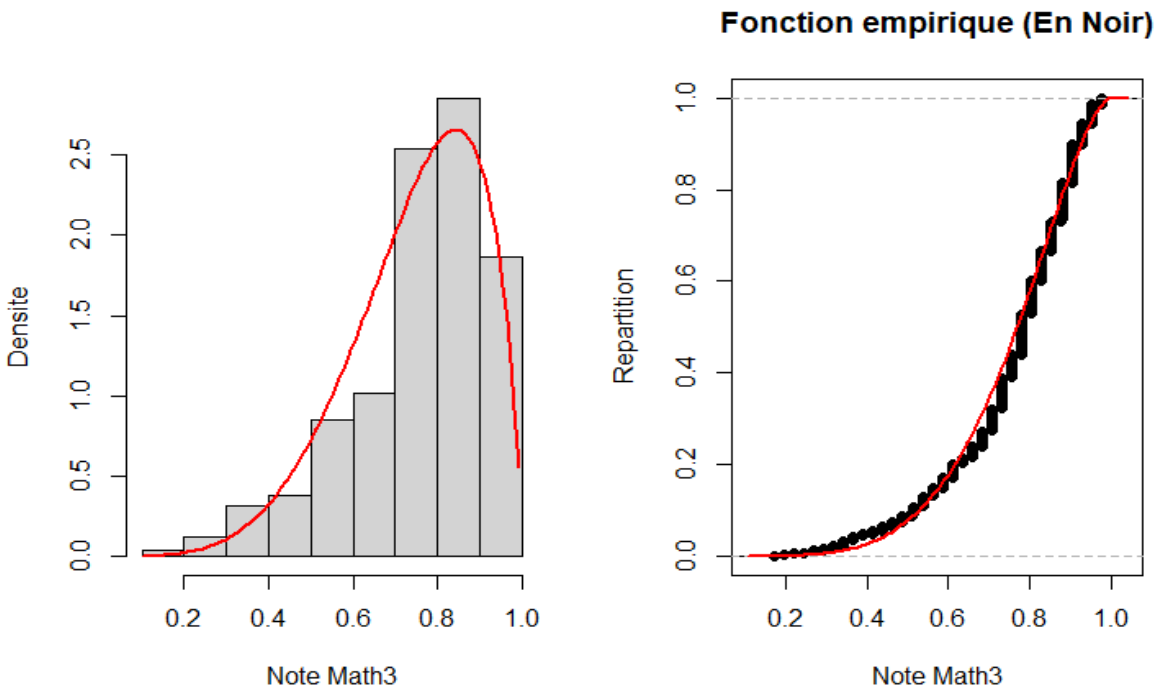


FIGURE 4.3 – Histogramme de la note `math3` et de la courbe de la loi bêta de paramètres $\tilde{\alpha}_2 = 5.23(0.27)$ et $\tilde{\beta}_2 = 1.78(0.08)$.

Les figures 4.2 et 4.3 nous confirment que les données se répartissent de manière asymétrique, plus concentrée à droite (confirmant l'analyse descriptive de la première partie, section 4.2). Nous suggérons pour les deux variables (`math1` et `math3`) comme lois candidates asymétriques : une loi bêta de première espèce, la loi bêta de deuxième espèce et la loi bêta à quatre paramètres qui semblent "adapter" à la circonstance. D'autres lois asymétriques sont moins réussies que ces trois proposées. L'estimation des paramètres se base sur la maximisation de la vraisemblance. Par exemple en supposant une loi bêta à deux paramètres pour la variable `math1`, les paramètres s'estiment grâce au package `betareg` (Cribari-Neto et Zeileis, 2010) par maximisation de la vraisemblance avec la fonction `fitdist`. La log-vraisemblance dans ce cas s'écrit

$$\mathcal{L} = -\sum_{j=1}^{m} n_j \log\{B(\alpha,\beta)\} + (\alpha-1)\sum_{j=1}^{m}\sum_{i=1}^{n_j} \log(x_{ji}) + (\beta-1)\sum_{j=1}^{m}\sum_{i=1}^{n_j} \log(1-x_{ji}), \qquad (4.5)$$

où pour α et β sont positifs, $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. Nous faisons quelques rappels sur la loi bêta élargie à trois paramètres que nous utilisons dans cet ajustement.

Définition 4.1. ***Distribution bêta à trois paramètres***

Une variable aléatoire Y suit une loi bêta généralisée à 3 paramètres α, β et λ noté $\mathcal{BG}3(\alpha, \beta, \lambda)$ si sa densité de probabilité f s'écrit

$$f(y; \alpha, \beta, \lambda) = \frac{\lambda^{\alpha}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\{1-(1-\lambda)y\}^{\alpha+\beta}}, \quad 0 \leq y \leq 1\cdot \tag{4.6}$$

• *Si $\lambda = 1$ alors la distribution $\mathcal{BG}3(\alpha, \beta, \lambda)$ se réduit à la distribution bêta à deux paramètres de la définition (1.16) qu'on note $\mathcal{B}(\alpha, \beta)$.*

• *Si $X \sim \mathcal{B}(\alpha, \beta)$ alors la variable aléatoire $Y = \frac{X}{\lambda+(1-\lambda)X}$ suit la loi $\mathcal{BG}3(\alpha, \beta, \lambda)$.*

La maximisation de la vraisemblance s'applique aux densités d'autres lois candidates pour obtenir les estimés des paramètres. Pour les lois candidates proposées, nous résumons, les paramètres estimés, les erreurs types et les coefficients AIC, résumé dans le tableau 4.3 pour chacune des deux variables. Les erreurs types, dans ce cas, s'appellent ***pseudo-erreur type*** parce qu'elles proviennent d'une pseudo-vraisemblance.

TABLEAU 4.3 – Estimation des paramètres des lois marginales candidates : la loi bêta $\mathcal{B}$ de première espèce, la loi bêta à trois paramètres $\mathcal{BG}3$.

Variable	*Loi*	*Estimés paramètres*	*Pseudo-erreur type*	*AIC*
math1	$\mathcal{B}$	**(4.38,2.52)**	**(0.22,0.12)**	**-548.88**
	$\mathcal{GB}3$	$(4.96, 2.38, 1.23)$	$(0.89, 0.22, 0.35)$	-547.46
math3	$\mathcal{B}$	$(5.24, 1.79)$	$(0.28, 0.08)$	-789.57
	$\mathcal{GB}3$	**(2.35,3.35,0.18)**	**(0.24,0.47,0.04)**	**-826.85**

Du tableau 4.3, nous sélectionnons la loi bêta à deux paramètres pour la distribution de la variable `math1` et dont les paramètres estimés sont $\tilde{\alpha}_1 = 4.38$ et $\tilde{\beta}_1 = 2.52$. Pour la variable `math3`, la loi bêta à trois paramètres $\tilde{\alpha}_3 = 2.35$, $\tilde{\beta}_3 = 3.35$ et $\tilde{\lambda} = 0.18$ semble être le "meilleur" ajustement. À partir des lois paramétriques postulées, nous calculons les pseudo-observations (estimations provenant des pseudo-vraisemblances) $\tilde{u}_{ji}$ et $\tilde{v}_{ji}$ donnés par

$$\tilde{u}_{ji} = \mathcal{B}(x_{ji}; \tilde{\alpha}_1, \tilde{\beta}_1), \qquad \tilde{v}_{ji} = \mathcal{GB}3(y_{ji}; \tilde{\alpha}_3, \tilde{\beta}_3, \tilde{\lambda})\cdot \tag{4.7}$$

Les pseudo-observations servent d'observations pour identifier les autres éléments du modèle de 2-copule échangeable. Nous identifions par la suite la copule bivariée $C^{(2)}$ associée aux notes de la quatrième et la septième année en utilisant les lois marginales.

4.4.2 Choix des deux familles de copules échangeables et de la copule bivariée

Une analyse préliminaire des données et des tests de diagnostic sont nécessaires pour sélectionner les copules impliquées. Cette sous-section se subdivise en trois parties. La première partie explique comment nous déterminons la copule $C^{(2)}$ associée à la dépendance entre les notes de quatrième année (`math1`) et septième année (`math3`). La deuxième partie explique la procédure de détermination de la copule échangeable $\left(C_{1,n}^{(1)}\right)$ et la dernière partie se consacre à la détermination de la copule échangeable $\left(C_{1,n}^{(3)}\right)$. Le choix de chaque copule se fait en deux étapes : une première étape graphique d'identification pour proposer des modèles candidats et une étape analytique d'identification pour faire un choix à partir de l'AIC calculé sur les copules des candidats.

Partie 1 : Détermination de la copule bivariée $C^{(2)}$ associé à (X, Y)

Nous construisons un nuage de points des variables uniformes (pseudo-observations) provenant respectivement des deux variables `math1` et `math3` de l'équation (4.7).

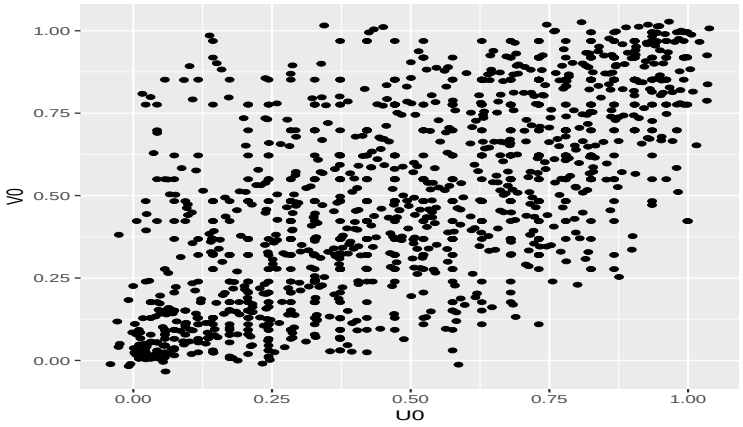


FIGURE 4.4 – Nuage de points des variables uniformes (pseudo-observations) $\tilde{u}$ (U_0) et $\tilde{v}$ (V_0) provenant de la note en quatrième année et de celle en septième année.

Nous calculons les coefficients de corrélation conditionnelle en utilisant la méthode suggérée par Joe (2014) pour vérifier la forme de la dépendance caudale dans le carré unité. En effet, nous subdivisons le carré [0,1] en 4 cadrans et la corrélation de chaque partie se calcule puis est comparée deux à deux. Premièrement, nous calculons le tau de Kendall pour les valeurs plus grande ou plus petite que $1/2$ pour U_0 et V_0 simultanément. Pour calculer, le tau de Kendall sur les données, nous utilisons la formule (1.22). Sous le logiciel R, nous utilisons la fonction

tau simplement. Ainsi, nous obtenons

$$\hat{\tau}_1 = \text{cor}(U_0, V_0 | U_0 \geq 1/2, V_0 \geq 1/2) = 0.36(0.01),$$

et

$$\hat{\tau}_2 = \text{cor}(U_0, V_0 | U_0 \leq 1/2, V_0 \leq 1/2) = 0.25(0.03)\cdot$$

Les erreurs types de ces corrélations se calculent en utilisant la formule de l'erreur type proposée par Hald (2007, page 143). Les valeurs obtenues des corrélations sont légèrement différentes ($\hat{\tau}_2 > \hat{\tau}_1$) et confirment une asymétrie radiale des données c'est-à-dire la corrélation des données dans le coin inférieur gauche n'est pas la même dans le coin supérieur droit, mais tout de même important. Ce qui nous conduit à ne pas considérer les copules qui sont réflexives. De plus, les corrélations des parties supérieures gauche et inférieure droit sont calculées par

$$\hat{\tau}_3 = \text{cor}(U_0, V_0 | U_0 \leq 1/2, V_0 \geq 1/2) = -0.005(0.03)\cdot$$

et

$$\hat{\tau}_4 = \text{cor}(U_0, V_0 | U_0 \geq 1/2, V_0 \leq 1/2) = 0.11(0.03)\cdot$$

La valeur de $\hat{\tau}_4$ plus grand que $\hat{\tau}_3$, négatif, nous permet encore de confirmer que les copules à proposer ne sont pas échangeables. En considérant l'asymétrie des données et connaissant le support $[0, 1]$ des lois marginales ajustées plus haut, nous postulons comme copules candidates, les copules asymétriques. À partir de la définition 1.9, nous considérons dans ce cas, la copule de *Khoudraji* notée $C^{(2)}$ dont l'expression est

$$C^{(2)}(u, v; \kappa, \kappa_1, \kappa_2) = u^{1-\kappa_1} v^{1-\kappa_2} C_0(u^{\kappa_1}, v^{\kappa_2}; \kappa), \quad \kappa_1, \kappa_2, u, v \in [0, 1], \tag{4.8}$$

où C_0 est une copule symétrique de vecteur de paramètres κ et $\kappa_1 \neq \kappa_2$.

Pour choisir les copules candidates C_0, nous ajustons les copules symétriques ordinaires à (U, V) pour sélectionner les copules bivariées maximisant la pseudo-vraisemblance. Ceci nous permet de choisir trois copules : la copule `normale`, la copule de `gumbel` et la copule `BB1`, à deux paramètres de l'équation (1.9) et leurs copules de survie. Plusieurs études utilisent souvent les copules asymétriques définies d'une manière particulière à partir de l'équation (4.8). Nous définissons trois copules particulières à partir de l'équation (4.8) et considérant les trois copules choisies au départ.

- La copule nommée `Khoudraji 1` provient de l'équation (4.8) où C_0 est la copule normale de paramètres κ. Pour cette copule, nous estimons 3 paramètres. En particulier, si $\kappa_1 = \kappa_2 = 1$, cette copule `Khoudraji 1` est exactement la copule normale.

- La copule nommée `Khoudraji 2` provient de l'équation (4.8) où C_0 est la copule de Gumbel de paramètres κ. Pour cette copule, nous estimons 3 paramètres.

- La copule nommée `Khoudraji 3` provient de l'équation (4.8) où C_0 est la copule BB1 et κ est un vecteur de dimension 2 dans ce cas. Pour cette copule, nous estimons 4 paramètres.

Nous considérons d'autres copules candidates comme la copule bêta bivariée à trois paramètres, voir la définition 1.6 et la copule de chi-deux à trois paramètres, voir Quessy *et al.* (2016). Ces deux copules ne donnent pas des résultats satisfaisants.

L'ajustement de copules bivariées sous le logiciel R, se fait avec les packages `Copula` et `VineCopula` en utilisant les fonctions définies comme `fitCopula` et `BiCopEst` pour certaines familles de copules paramétriques. Les paramètres estimés et les *pseudo-erreur type* associés aux copules candidates se présentent dans le tableau 4.4.

TABLEAU 4.4 – Paramètres estimés, pseudo-erreur type et l'AIC en fonction du modèle de copule suggéré pour $C^{(2)}$.

Copule $C^{(2)}$	*Paramètres estimés*	*Pseudo-erreur type*	*AIC*
Normale	0.68	0.17	-458
Gumbel	1.8	0.09	-400.4
Gumbel survie	1.88	0.02	-457.4
BB1	(0.6,1.44)	(0.09,0.06)	-455.8
BB1 Survie	(0.2,0.71)	(0.07,0.07)	-465.6
Khoudraji 1	$(0.75, 0.98, 0.87)$	$(0.02, 0.01, 0.04)$	-469
Khoudraji 1 survie	**(0.78,0.82,0.97)**	**(0.02,0.04,0.02)**	**-476.4**
Khoudraji 2	$(1.88, 0.98, 0.94)$	$(0.1, 0.01, 0.05)$	-397.8
Khoudraji 2 survie	$(2.03, 0.88, 1)$	$(0.11, 0.05, 0.06)$	-460.2
Khoudraji 3	$(1.1, 1.41, 0.96, 0.86)$	$(0.25, 0.09, 0.04, 0.04)$	-461.2
Khoudraji 3 survie	$(0.24, 1.84, 0.88)$	$(0.12, 0.1, 0.05, 0.05)$	-467

Au vu des résultats précédents du tableau 4.4, nous pouvons sélectionner la copule $C^{(2)}$, comme la copule de *Khoudraji 1 survie* à trois paramètres $(\tilde{\kappa}_2, \tilde{\kappa}_1, \tilde{\kappa}_2) =$ **(0.78,0.82,0.97)** où C_0 est une copule normale de paramètre 0.78.

Par ailleurs, nous traçons, à la figure 4.5 les courbes de niveau de certaines copules candidates pour observer graphiquement la courbe de prédiction à partir de quelques copules en utilisant des valeurs de paramètres estimés. Nous rappelons que la marginale F de la variable `math1` est une loi $\mathcal{B}$ et celle G de la variable `math3` est une loi $\mathcal{GB}$. La courbe en couleur verte est la courbe de la régression linéaire simple, pente et ordonnée à l'origine de la variable `math3` en la variable `math1`. Les éléments de la figure s'expliquent par :

— La droite de la régression linéaire, pente et ordonnée à l'origine de la variable `math3` en la variable `math1` est une droite d'équation `math3`$= 0.35\times$ `math1` $+ 0.65$ (courbe de couleur verte). La courbe de régression se rapporte à $[0, 1]$.

— La courbe en couleur rouge est la courbe de la régression copule par l'équation

$$\mathbb{E}(Y|x) = \int_0^1 G^{-1}(u;\tilde{\alpha}_3,\tilde{\beta}_3,\tilde{\lambda})c^{(2)}\left\{F(x;\tilde{\alpha}_1,\tilde{\beta}_1),v;\tilde{\kappa},\tilde{\kappa}_1,\tilde{\kappa}_2\right\}dv\cdot \tag{4.9}$$

Cette intégrale s'approxime par une somme de Riemann. La prédiction se rapporte à $[0,1]$ en utilisant les lois marginales. C'est une alternative à la régression linéaire mixte.

— les lignes de niveau représentent la densité conjointe de X et Y évaluée au point $\left(F^{-1}(u;\tilde{\alpha}_1,\tilde{\beta}_1),G^{-1}(v;\tilde{\alpha}_3,\tilde{\beta}_3,\tilde{\lambda})\right)$ par

$$f\left\{F^{-1}(u;\tilde{\alpha}_1,\tilde{\beta}_1),\tilde{\alpha}_1,\tilde{\beta}_1\right\}g\left\{G^{-1}(v;\tilde{\alpha}_3,\tilde{\beta}_3,\tilde{\lambda}),\tilde{\alpha}_3,\tilde{\beta}_3,\tilde{\lambda}\right\}c^{(2)}(u,v;\tilde{\kappa},\tilde{\kappa}_1,\tilde{\kappa}_2)\cdot \tag{4.10}$$

Cette densité s'évalue en 500×500 points de $[0,1]\times[0,1]$.

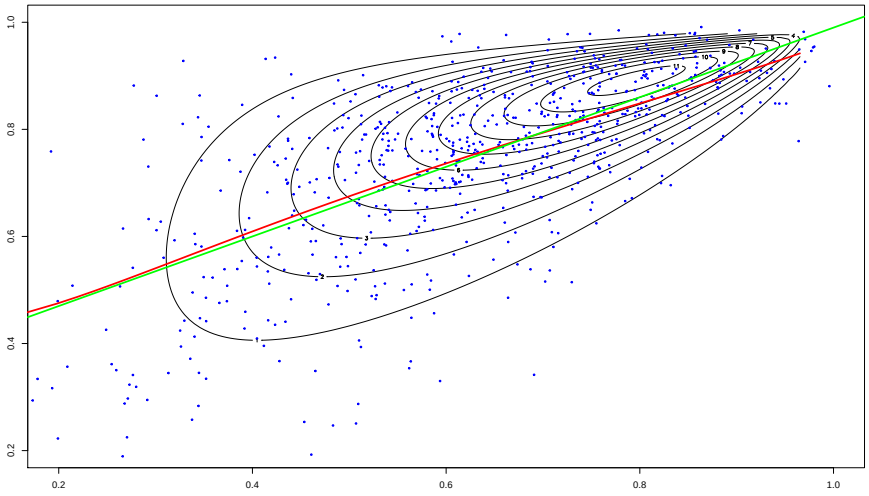


FIGURE 4.5 – Courbes de niveau de *Khoudraji 1* avec les deux courbes de régression(rouge et la verte).

Le programme du code `R` de la procédure d'obtention de la figure 4.5 se trouve à l'annexe D.1.

Pour la copule de *Khoudraji 2 survie*, nous présentons aussi la figure.

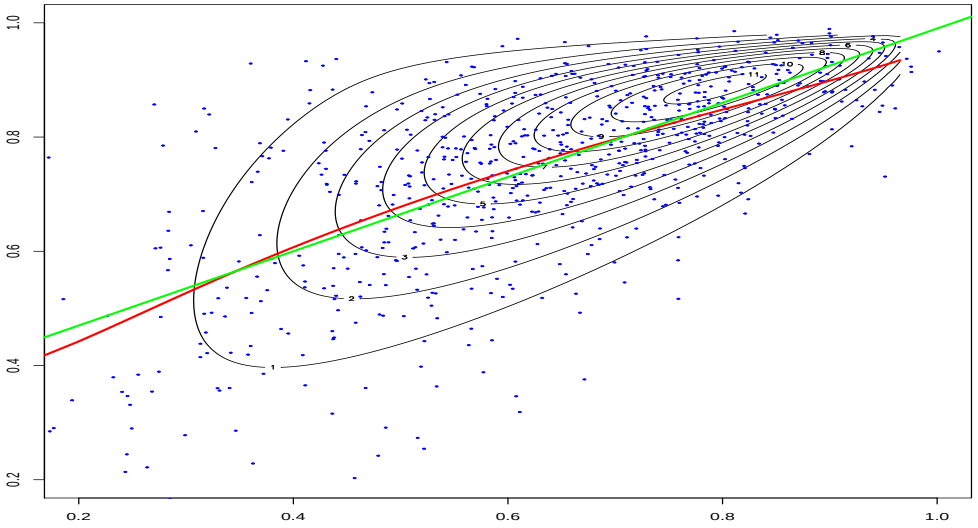


FIGURE 4.6 – Courbes de niveau de *Khoudraji 1 survie* avec les deux courbes de régression.

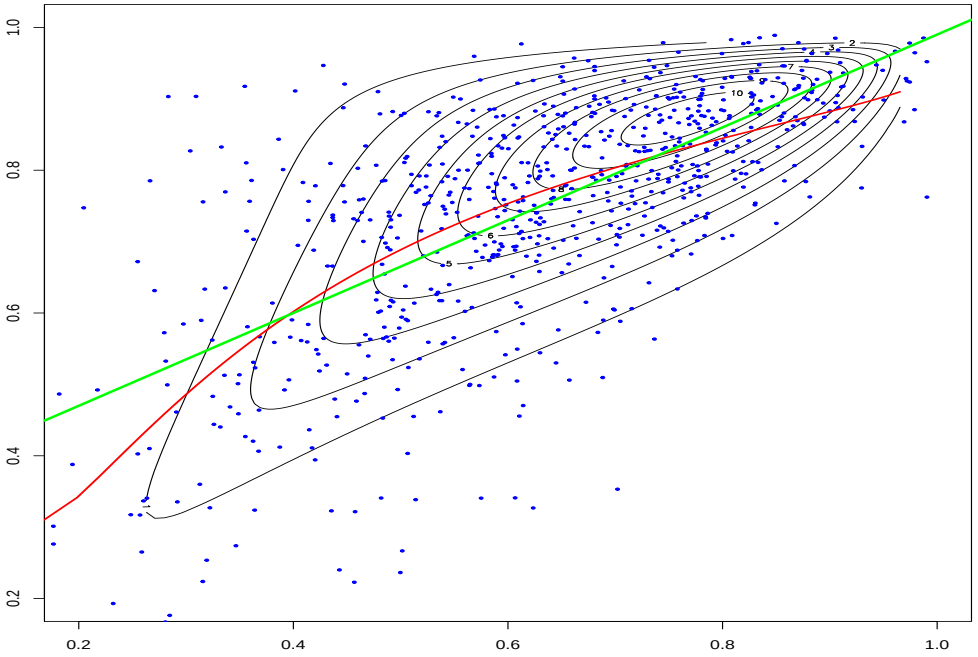


FIGURE 4.7 – Courbes de niveau de *Khoudraji 2 survie* avec les deux courbes de régression.

La même procédure de l'annexe D.1 se répète pour obtenir les figures 4.7 et 4.6 en changeant juste la copule $C^{(2)}$.

Pour confirmer l'adéquation de la copule de *Khoudraji 1 survie* aux données, nous simulons 1000 observations qui suivent cette copule. Nous obtenons le graphique de la figure 4.8 que nous comparons au graphe construit à la figure 4.4.

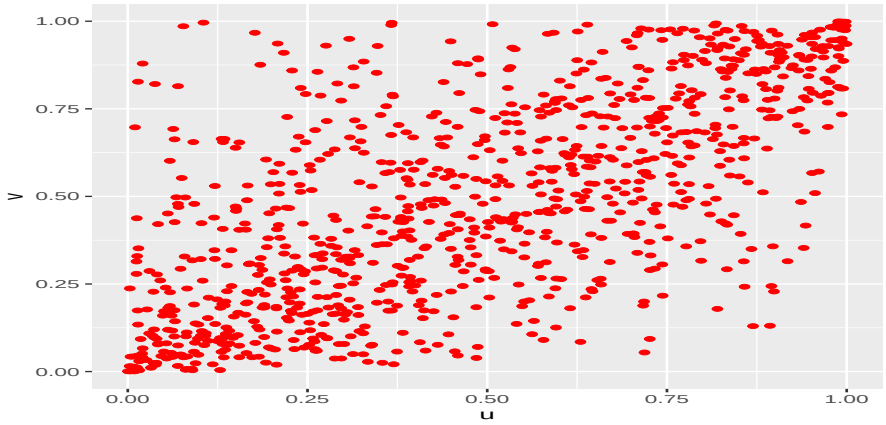


FIGURE 4.8 – Graphique des données simulées suivant la copule asymétrique `Khoudraji 1 survie` de paramètres $(\tilde{\kappa}, \tilde{\kappa}_1, \tilde{\kappa}_2) = (0.78, 0.82, 0.97)$ du tableau 4.4.

De cette comparaison entre les graphiques 4.4 et 4.8, nous confirmons l'adéquation de la copule choisie.

Nous passons maintenant au choix de la famille de copule échangeable $\left(C_{1,n}^{(1)}\right)$ en utilisant l'approche descriptive présentée au chapitre 3 et nous inspirant de Rivest *et al.* (2016) puis confirmée avec une approche analytique.

Partie 2 : Identification de la famille de copule échangeable $\left(C_{1,n}^{(1)}\right)$

Nous recherchons la famille de copule échangeable $\left(C_{1,n}^{(1)}\right)$ liée aux données de la note en quatrième année dans la même école. Nous calculons le tau de Kendall sur les données `math1` dans toutes les écoles. Le tau de Kendall échangeable se présente dans une thèse de l'Université Laval écrite et comportant l'article Romdhani *et al.* (2014a). Le tau de Kendall échangeable est relativement équivalent au tau de Kendall habituel. Pour la variable `math1`, la valeur du tau de Kendall échangeable est $\hat{\tau}_E = 0.04$ avec l'erreur type $s.e = 0.01$. Nous présentons sur la figure 4.9, le graphique observé (à partir des données brutes) et les graphiques théoriques (à partir des données simulées), obtenus à partir de copules candidates (Clayton, normale, etc.).

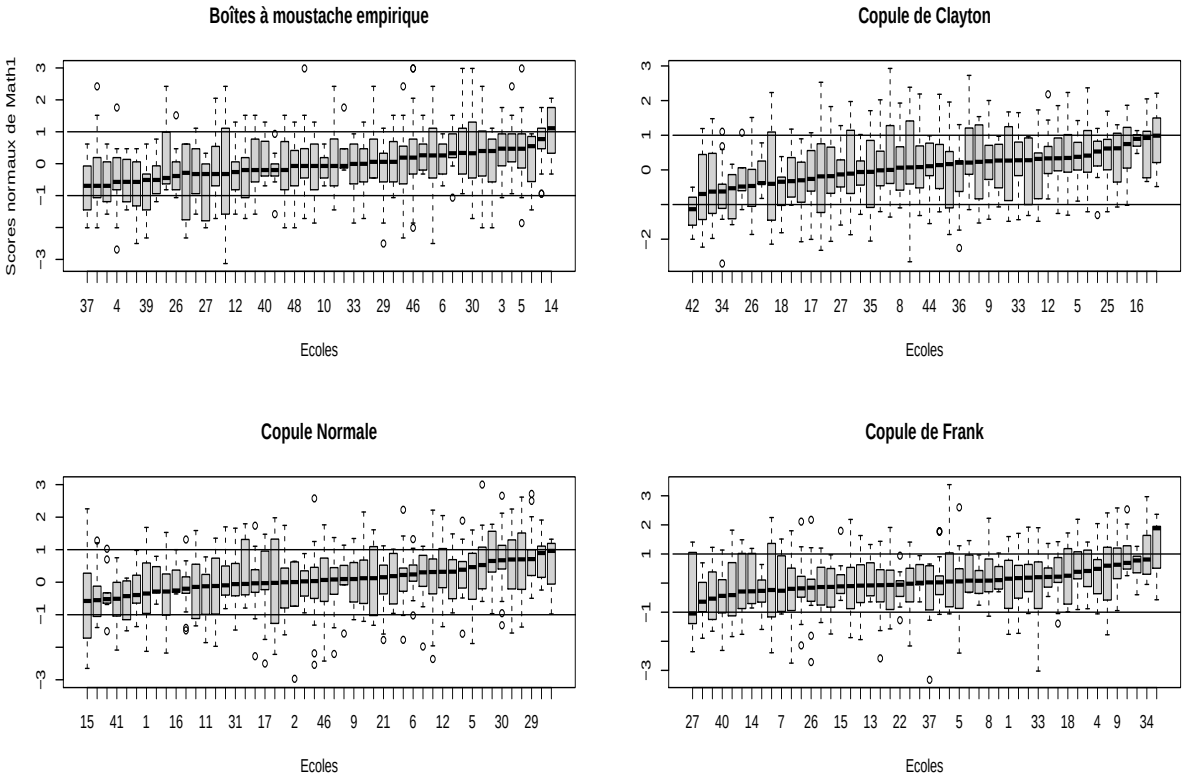


FIGURE 4.9 – Boîtes à moustache des scores normaux des notes en année 4 (graphique observé) des différentes écoles avec trois familles de copules échangeables candidates.

Sur le graphique 4.9, la boîte à moustache empirique provient des données observées tandis que les autres boîtes à moustaches proviennent des données simulées suivant les copules spécifiées. Nous estimons les paramètres des copules échangeables candidates et les coefficients AIC de chaque copule en maximisant la pseudo-vraisemblance.

TABLEAU 4.5 – Paramètres estimés, pseudo-erreur type et l'AIC de chaque choix pour la famille de copule échangeable $\left(C_{1,n}^{(1)}\right)$.

Candidate $C_{1,n}^{(1)}$	*Paramètre estimé*	*Pseudo-erreur type*	*AIC*
Frank	0.22	0.13	-5.58
Frank-Survie	0.35	0.15	-11.31
Gumbel	1.04	0.02	-5.02
Gumbel-Survie	1.05	0.02	-9.62
Clayton	0.06	0.02	-10.33
Clayton-Survie	0.05	0.02	-9.01
Normale	**0.06**	**0.02**	**-13.05**

Du graphique 4.9 et des résultats des différentes estimations de l'AIC du tableau 4.5, nous déduisons que la copule normale serait un bon modèle pour modéliser la dépendance entre les notes de mathématiques en quatrième année. Ceci signifie que la copule $\left(C_{1,n}^{(1)}\right)$ est la copule normale échangeable, prise comme ajustement.

Pour la suite de la construction du modèle, nous recherchons la famille de copule échangeable $\left(C_{1,n}^{(3)}\right)$ par la même méthodologie que précédemment. Cette copule est associée aux pseudo-observations résiduelles, obtenues à partir de la copule $C^{(2)}$ que nous définissons dans la partie 3.

Partie 3 : Identification de la famille de copule échangeable $\left(C_{1,n}^{(3)}\right)$

Elle se consacre à la détermination de la famille de copule notée $\left(C_{1,n}^{(3)}\right)$ associé aux résidus. Premièrement, nous calculons les pseudo-observations résiduelles w_{ji}. En utilisant la formule $w_{ji} = 1 - \partial C^{(2)}(1-u_{ji}, 1-v_{ji}; \kappa, \kappa_1, \kappa_2)/\partial u_{ji}$ et de l'équation (4.8) de la copule, nous avons

$$\begin{aligned} w_{ji} &= 1 + (\kappa_1 - 1)\left(1-u_{ji}\right)^{-\kappa_1}\left(1-v_{ji}\right)^{1-\kappa_2} C_0\left\{\left(1-u_{ji}\right)^{\kappa_1}, \left(1-v_{ji}\right)^{\kappa_2}; \kappa\right\} \\ &- \kappa_1 (1-v_{ji})^{1-\kappa_2} \Phi\left[\frac{\Phi^{-1}\left\{(1-v_{ji})^{\kappa_2}\right\} - \kappa\Phi^{-1}\left\{(1-u_{ji})^{\kappa_1}\right\}}{\sqrt{1-\kappa^2}}\right], \end{aligned} \tag{4.11}$$

pour $i = 1, \ldots, n_j$ et $j = 1, \ldots, m$.

Les valeurs estimées $(\hat{\kappa}, \hat{\kappa}_1, \hat{\kappa}_2) = (0.78, 0.82, 0.97)$ remplacent les paramètres $(\kappa, \kappa_1, \kappa_2)$ et u_{ji}, v_{ji} respectivement par les pseudo-observations $\tilde{u}_{ji}$ et $\tilde{v}_{ji}$ de l'équation (4.7).

La procédure précédente (utilisée dans la partie 4.4.2), faite pour le choix de la copule $\left(C_{1,n}^{(1)}\right)$ s'applique en recherchant cette fois-ci la copule échangeable qui modélise la dépendance entre les pseudo-observations résiduelles dans une grappe. Nous calculons le tau de Kendall échangeable qui donne $\tau_E^{'} = 0.16 (s.e = 0.03)$. Nous construisons la bôite à moustache des pseudo-observations résiduelles que nous comparons avec les simulations des copules théoriques échangeables. La figure 4.10 y donne un aperçu des boîtes à moustaches.

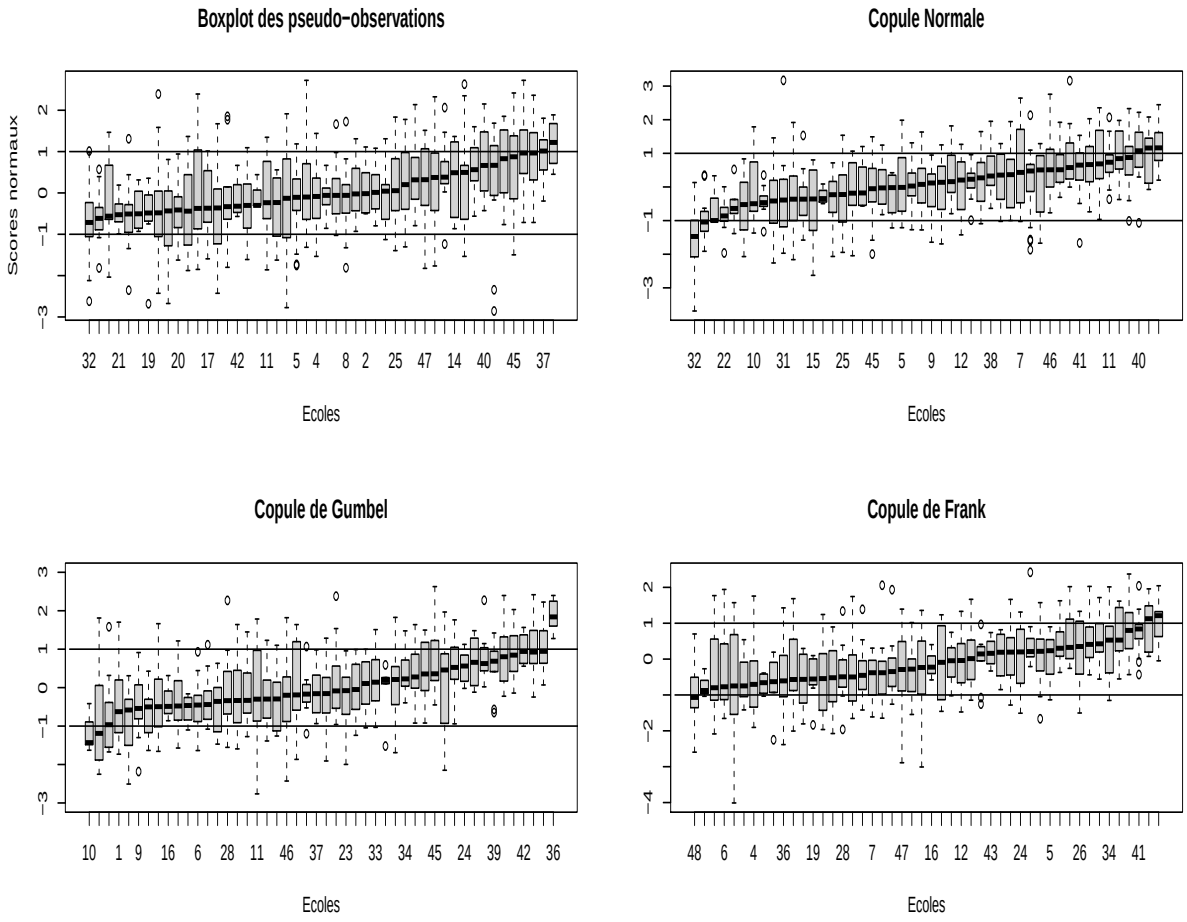


FIGURE 4.10 – Boxplot des scores normaux des pseudo-observations des différentes écoles avec ceux des copules candidates.

De ce graphique, nous ne pouvons dire avec précision la copule qui s'adapte aux données. Nous estimons les paramètres associés puis calculons le coefficient AIC associé pour les copules candidates.

TABLEAU 4.6 – Paramètres estimés, valeur de l'AIC et pseudo-erreur type pour le choix de la copule $\left(C_{1,n}^{(3)}\right)$

Candidate $\left(C_{1,n}^{(3)}\right)$	*Paramètres estimés*	*Pseudo-erreur type*	*AIC*
Frank	0.92	0.17	-70.61
Gumbel	1.1	0.02	-65.78
Normale	0.16	0.03	**-80.11**

Nous avons essayé aussi les copules de survie échangeable qui s'ajustent moins bien. Selon le critère AIC pour les différentes copules précédemment postulées, nous pouvons alors tirer la conclusion que la copule normale échangeable s'ajuste mieux aux pseud-obervations résiduelles. Ceci termine le choix des différents éléments du modèle de 2-copule échangeable.

En résumé des différentes étapes précédentes, nous avons :

- les lois marginales F et G sont respectivement la distribution bêta de paramètres (α_1, β_1) et une loi bêta de 3 paramètres $(\alpha_3, \beta_3, \lambda)$;
- la famille de copule échangeable $\left(C_{1,n}^{(1)}\right)$ est la *copule normale échangeable* de paramètre δ_1 en dimension n ;
- la copule bivariée $C^{(2)}$ est la copule *Khoudraji 1 survie* de paramètres κ, κ_1 et κ_2 ;
- la famille de copule échangeable $\left(C_{1,n}^{(3)}\right)$ est la *copule normale échangeable* de paramètre δ_3 en dimension n.

Les paramètres du modèle sont $\theta = (\alpha_1, \beta_1, \alpha_3, \beta_3, \lambda, \delta_1, \kappa, \kappa_1, \kappa_2, \delta_3)$. Dans la suite de l'analyse, nous estimons les paramètres du modèle de 2-copule échangeable en maximisant la log-vraisemblance globale.

4.4.3 Estimation des paramètres du modèle de 2-copule échangeable

Le modèle postulé contient au total dix (10) paramètres. Nous les estimons en maximisant la log-vraisemblance globale donnée par

$$\mathcal{L}(\theta; x, y) = \mathcal{L}_1(\alpha_1, \beta_1; x) + \mathcal{L}_2(\alpha_3, \beta_3, \lambda; y) + \mathcal{L}_3(\theta; x, y), \tag{4.12}$$

où $x = (x_{ji})_{j=1}^{m}$, $y = (y_{ji})_{j=1}^{m}$, $i = 1, \ldots, n_j$ et la fonction $\mathcal{L}_1$ provient de l'équation (4.5) avec α, β, respectivement remplacés par α_1 et β_1. La fonction $\mathcal{L}_2$ est

$$\begin{aligned}\mathcal{L}_2(\alpha_3, \beta_3, \lambda; y) &= \alpha_3 \log(\lambda) \sum_{j=1}^{m} n_j - \sum_{j=1}^{m} n_j \log\{B(\alpha_3, \beta_3)\} + (\alpha_3 - 1) \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log(y_{ji}) \\ &+ (\beta_2 - 1) \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log(1 - y_{ji}) + (\alpha_3 + \beta_3) \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log\{1 - (1 - \lambda) y_{ji}\}.\end{aligned}$$

Pour la fonction $\mathcal{L}_3$, elle se définit par

$$\begin{aligned}\mathcal{L}_3(\theta; x, y) &= \sum_{j=1}^{m} \log\left\{c_{1,n_j}^{(1)}\left(u_{j1}, \ldots, u_{jn_j}; \delta_1\right)\right\} + \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log\left\{c^{(2)}(u_{ji}, v_{ji}; \kappa, \kappa_1, \kappa_2)\right\} \\ &+ \sum_{j=1}^{m} \log\left\{c_{1,n_j}^{(3)}\left(w_{j1}, \ldots, w_{jn_j}; \delta_3\right)\right\},\end{aligned}$$

où w_{ji} provient de l'équation (4.11) et $u_{ji} = \mathcal{B}(x_{ji}; \alpha_1, \beta_1)$, $v_{ji} = \mathcal{GB}3(y_{ji}; \alpha_3, \beta_3, \lambda)$. Le programme **R** du calcul de la log-vraisemblance de l'équation (4.12) se trouve en annexe D.2.

À partir des données, nous résumons, dans le tableau 4.7, les estimations des paramètres issues de la maximisation de cette log-vraisemblance.

TABLEAU 4.7 – Paramètres estimés par la méthode du maximum de vraisemblance globale du modèle avec les erreurs types associées à chaque paramètre

Élément	*Estimé*	*Erreur type*
Marginale F (loi $\mathcal{B}$)	$(\hat{\alpha}_1, \hat{\beta}_1) = (4.36, 2.56)$	$(0.24, 0.13)$
Marginale G (loi $\mathcal{GB}3$)	$(\hat{\alpha}_3, \hat{\beta}_3, \hat{\lambda}) = (2.32, 3.48, 0.17)$	$(0.21, 0.46, 0.04)$
$C_{1,n}^{(1)}$ *: Copule normale*	$\hat{\delta}_1 = 0.06$	0.02
$C^{(2)}$ *: Khoudraji 1 survie*	$(\hat{\kappa}, \hat{\kappa}_1, \hat{\kappa}_2) = (0.78, 0.82, 0.97)$	$(0.01, 0.03, 0.01)$
$C_{1,n}^{(3)}$ *: Copule normale*	$\hat{\delta}_3 = 0.16$	0.02

Les paramètres estimés permettent de reconstruire le modèle avec la 2-copule échangeable que nous utilisons pour faire de nouvelles prédictions de la note en septième année. Dans la suite de notre travail, nous traçons les courbes de prédiction à partir du modèle obtenu et le comparons au modèle linéaire mixte, le modèle de régression copule et le modèle de régression non-paramétrique. Nous utilisons les lois marginales ajustées et les copules sélectionnées.

4.5 Prédiction de nouvelles observations en utilisant le modèle de 2-copule échangeable

Dans cette section, nous construisons les courbes de prédiction de la note en septième année d'un élève à l'aide du modèle de 2-copule échangeable et comparée avec la prédiction faite avec un modèle linéaire mixte ou avec un modèle de régression copule. En particulier, pour une école j, dans laquelle nous connaissons les données observées $\mathbf{x}_j$ et $\mathbf{y}_j$ (les notes en quatrième et en septième) de n_j individus et nous voulons prédire la note Y en septième année du (n_j+1) ième individu sachant la note x en quatrième année pour le même individu selon le modèle choisi.

Nous rappelons que F est la loi bêta de paramètres $(\hat{\alpha}_1, \hat{\beta}_1) = (4.36, 2.56)$ et que G est la loi bêta de paramètres $(\hat{\alpha}_3, \hat{\beta}_3, \hat{\lambda}) = (2.32, 3.48, 0.17)$. Nous rappelons aussi que les copules $\left(C_{1,n}^{(1)}\right)$ et $\left(C_{1,n}^{(3)}\right)$ sont toutes deux, la copule normale échangeable. À partir de l'équation (2.31), nous

calculons la loi conditionnelle h_x de Y sachant x et $\mathbf{x}_j$ et $\mathbf{y}_j$ est

$$\begin{aligned} h_x(y) &= g(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})c^{(2)}\left\{F(x;\hat{\alpha}_1,\hat{\beta}_1),G(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda});\hat{\kappa},\hat{\kappa}_1,\hat{\kappa}_2\right\}\times \\ &\quad \frac{\exp\left\{-\frac{1}{2\hat{\sigma}_j^2}(\Phi^{-1}\left[C_{2|1}\left\{G(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\}\right]-\hat{\mu}_j)^2\right\}}{\sqrt{2\pi}\hat{\sigma}_j\phi\left(\Phi^{-1}\left[C_{2|1}\left\{G(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\}\right]\right)}, \end{aligned} \tag{4.13}$$

où $x,y\in]0,1[$ et ϕ est la densité de la loi normale centrée réduite. Les estimations $\hat{\mu}_j$ et $\hat{\sigma}_j^2$ proviennent de l'équation (4.14) par

$$\hat{\mu}_j=\frac{n_j\hat{\delta}_3\overline{w}_{n_j}}{1+(n-1)\hat{\delta}_3},\qquad \hat{\sigma}_j^2=\frac{(1-\hat{\delta}_3)(1+n_j\hat{\delta}_3)}{1+(n_j-1)\hat{\delta}_3}, \tag{4.14}$$

et

$$\overline{w}_{n_j}=\frac{1}{n_j}\sum_{i=1}^{n_j}\Phi^{-1}\left[C_{2|1}\left\{G(y_{ji};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})|F(x_{ji};\hat{\alpha}_1,\hat{\beta}_1)\right\}\right], \tag{4.15}$$

où nous rappelons que $\mathbf{x}_j=(x_{j1},\ldots,x_{jn_j})^T$ et $\mathbf{y}_j=(y_{j1},\ldots,y_{jn_j})^T$ et

$$\begin{aligned} C_{2|1}(v|u) &= 1+(\hat{\kappa}_1-1)(1-u)^{-\hat{\kappa}_1}(1-v)^{1-\hat{\kappa}_2}C_0\left\{(1-u)^{\hat{\kappa}_1},(1-v)^{\hat{\kappa}_2}\right\} \\ &- \hat{\kappa}_1(1-v)^{1-\hat{\kappa}_2}\Phi\left[\frac{\Phi^{-1}\left\{(1-v)^{\hat{\kappa}_2}\right\}-\hat{\kappa}\Phi^{-1}\left\{(1-u)^{\hat{\kappa}_1}\right\}}{\sqrt{1-\hat{\kappa}^2}}\right], \end{aligned} \tag{4.16}$$

est une sorte de résidu associé au vecteur d'observation $(u,v)=\left\{F(x;\hat{\alpha}_1,\hat{\beta}_1),G(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})\right\}$. L'inverse de $C_{2|1}$ se note $C_{2|1}^{-1}$.

La densité conditionnelle dépend de la valeur prise par la note `math1` des unités observées dans l'école via $\hat{\mu}_j$ et $\hat{\sigma}_j$. Par transformation, la fonction de répartition H_x de Y sachant x et $\mathbf{x}_j$ et $\mathbf{y}_j$, associée à h_x est

$$H_x(y)=\Phi\left\{\frac{\Phi^{-1}\left[C_{2|1}\left\{G(y;\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda})|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\}\right]-\hat{\mu}_j}{\hat{\sigma}_j}\right\}, \tag{4.17}$$

puis nous obtenons les quantiles conditionnels par

$$q_Y(x;\tau)=G^{-1}\left[C_{2|1}^{-1}\left\{\Phi\left(\hat{\mu}_j+\hat{\sigma}_j\Phi^{-1}(\tau)\right)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right],\ 0<\tau<1\cdot \tag{4.18}$$

L'expression du prédicteur de Y, valeur de la note en septième année du (n_j+1)ième individu sachant x, $\mathbf{x}_j$ et $\mathbf{y}_j$ s'inspirant de l'équation (2.32) est

$$\mathbb{E}\left(Y|x,\mathbf{x}_j,\mathbf{y}_j\right) = \int_{\mathbb{R}} G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\hat{\sigma}_j t)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right]\phi(t)dt\cdot \tag{4.19}$$

Une méthode d'approximation de cette intégrale est la méthode de Gauss-Hermite. L'approximation de Gauss-Hermite consiste à calculer la valeur de la fonction à l'intérieur de l'intégrale pour K points t_i. L'intégrale est alors approximée par une somme pondérée de ces évaluations. Donc nous obtenons

$$\int_{\mathbb{R}} G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\sqrt{2}\hat{\sigma}_j z)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right]e^{-z^2}dz \quad \approx$$
$$\sum_{i=1}^{K} w_i G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\sqrt{2}\hat{\sigma}_j t_i)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right],$$

voir Steen *et al.* (1961). Le vecteur (t_i, w_i) se calcule par la fonction `gauss.quad` du package `statmod` du logiciel `R`. Le point t_i appartient à l'ensemble des nombres réels $\mathbb{R}$ et $w_i = \exp(-t_i^2)$. Enfin, en faisant le changement de variable $z = t/\sqrt{2}$ et le fait que $\phi(t) = e^{-t^2/2}$, nous avons

$$\mathbb{E}\left(Y|x,\mathbf{x}_j,\mathbf{y}_j\right) \quad \approx \quad \sum_{i=1}^{K}\frac{w_i}{\sqrt{\pi}}G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\sqrt{2}\hat{\sigma}_j t_i)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right]\cdot \quad (4.20)$$

Une évaluation de cette approximation sous le logiciel `R` se présente en annexe D.3.

Pour faire la comparaison de la 2-copule échangeable avec les autres modèles, nous construisons des courbes de prédiction en fonction de l'école d'appartenance des élèves. Nous donnons un bref résumé sur la variable $\overline{w}_{n_j}$ dans le tableau 4.8 en fonction d'une sélection d'écoles. Dans ce tableau, les écoles sont classées en ordre croissant selon la $\ll$ moyenne du quantile normale des résidus $\gg$, $\hat{\mu}_j$. L'entier n_j est la taille de l'école (grappe) dans l'échantillon de départ.

TABLEAU 4.8 – Statistiques sommaires des résidus dans chaque école sélectionnée en fonction de la moyenne des résidus.

École	*Taille* n_j	$\Phi(\overline{w}_{n_j}/\sqrt{\hat{\delta}_3})$	$\hat{\mu}_j$	$\hat{\sigma}_j$
`Numéro 1`	38	0.033	-0.643	0.927
`Numéro 2`	49	0.088	-0.488	0.925
`Numéro 3`	23	0.319	-0.153	0.932
`Numéro 4`	24	0.884	0.392	0.932
`Numéro 5`	27	0.901	0.432	0.930
`Numéro 6`	31	0.990	0.788	0.929

Nous passons à une comparaison du prédicteur du modèle de 2-copule échangeable, équation (4.19) avec celui obtenu de trois autres approches en utilisant spécifiquement les écoles citées dans le tableau 4.8 pour faciliter l'illustration.

4.5.1 Comparaison de la méthode de 2-copule échangeable et de la méthode de régression linéaire mixte

Nous comparons ici le prédicteur obtenu par la méthode de 2-copule échangeable, voir l'équation (4.19) avec le prédicteur obtenu par la méthode de régression linéaire mixte, voir l'équation (4.1). Ainsi, nous construisons les courbes des valeurs prédites par les deux méthodes accompagnées du nuage de points des valeurs observées. Pour ces courbes de prédiction, la légende de couleurs se présente comme suit.

— les courbes en bleue et noire trait plein sont respectivement les courbes de la régression utilisant la 2-copule échangeable pour deux écoles.

— les courbes en bleue et noire pointillées sont les droites de la régression linéaire mixte.

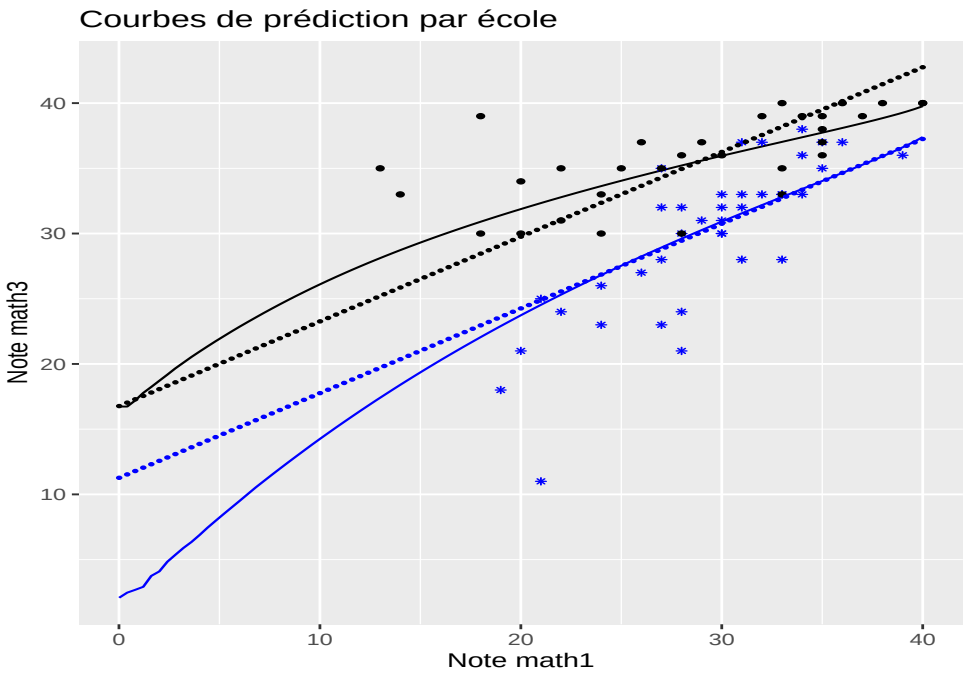


FIGURE 4.11 – Prédiction pour deux écoles `Numéro 1` (en bleue) et `Numéro 6`(en noire) de la note en 7e année des élèves par le modèle de 2-copule échangeable (courbe en bleue et en noire) et le modèle linéaire mixte (les deux droites en pointillés).

Pour ces graphiques, nous constatons globalement que le modèle issu de la modélisation 2-copule est l'ajustement d'une courbe de régression à chaque école alors que les régressions linéaires mixtes sont des droites parallèles. Nous présentons spécifiquement la comparaison des prédictions du modèle de 2-copule échangeable avec celles d'un le modèle de régression linéaire mixte par la suite.

4.5.2 Comparaison de la méthode de 2-copule échangeable et de la méthode de régression copule

Une autre manière de prédire la note en septième année est d'utiliser la copule $C^{(2)}$ associée au vecteur aléatoire (X, Y) et les lois marginales associées F et G. Ainsi, à partir de l'équation (1.28), nous avons

$$\mathbb{E}(Y|x) = \int_0^1 G^{-1}(t; \hat{\alpha}_3, \hat{\beta}_3, \hat{\lambda}) c^{(2)} \left\{ t, F(x; \hat{\alpha}_1, \hat{\beta}_1); \hat{\kappa}, \hat{\kappa}_1, \hat{\kappa}_2 \right\} dt. \tag{4.21}$$

Le prédicteur ne dépend pas bien sûr de $\mathtt{x}_j$ et $\mathtt{y}_j$. Le prédicteur ne tient pas compte de l'appartenance de la valeur prédite à une grappe fixée. Pour construire la courbe de régression, nous considérons l'école `Numéro 2` et l'école `Numéro 4`. Ces deux écoles sont choisies pour avoir un échantillon de taille élevée dans le but de faire une régression non-paramétrique. La régression non-paramétrique utilise les données de chaque école pour construire sa courbe de prédiction.

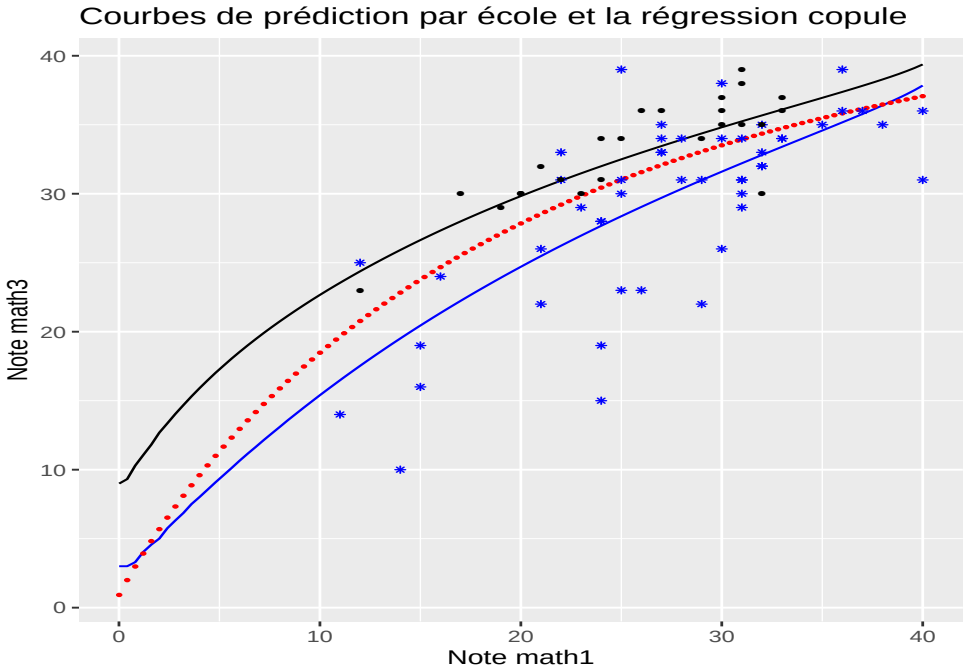


FIGURE 4.12 – Prédiction, pour deux écoles `Numéro 1` (courbe en bleue) et `Numéro 4` (courbe en noire) de la note en septième année par le modèle échangeable (courbe en bleue et noire) et le modèle de régression copule (courbe en rouge).

4.5.3 Comparaison de la méthode de 2-copule échangeable et de la méthode de régression non paramétrique

Nous utilisons ici la méthode de régression non paramétrique de la fonction *ggplot*. Nous construisons la courbe de la régression non paramétrique en utilisant l'ensemble des données dans une grappe et le modèle de 2-copule échangeable par la même occasion. Nous considérons deux écoles ayant un grand nombre d'élèves pour avoir un ajustement raisonnable avec la régression non paramétrique qui utilise la fonction *loess*. Les intervalles de confiance se calculent pour la régression non-paramétrique en utilisant le quantile de la loi de Student avec un niveau de 95%.

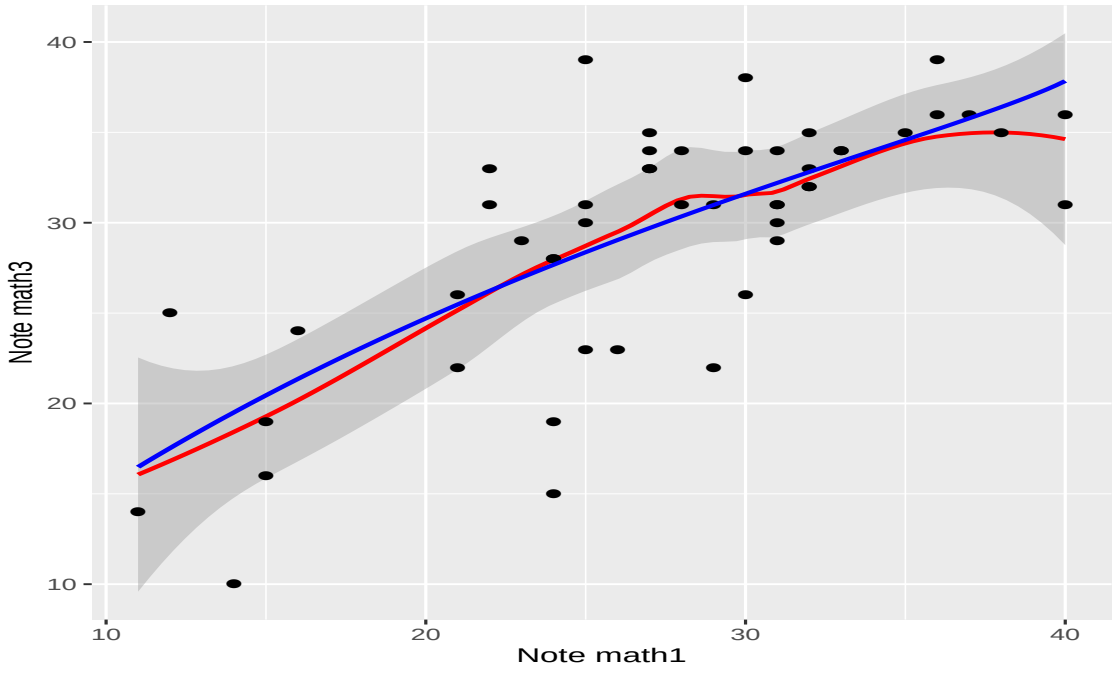
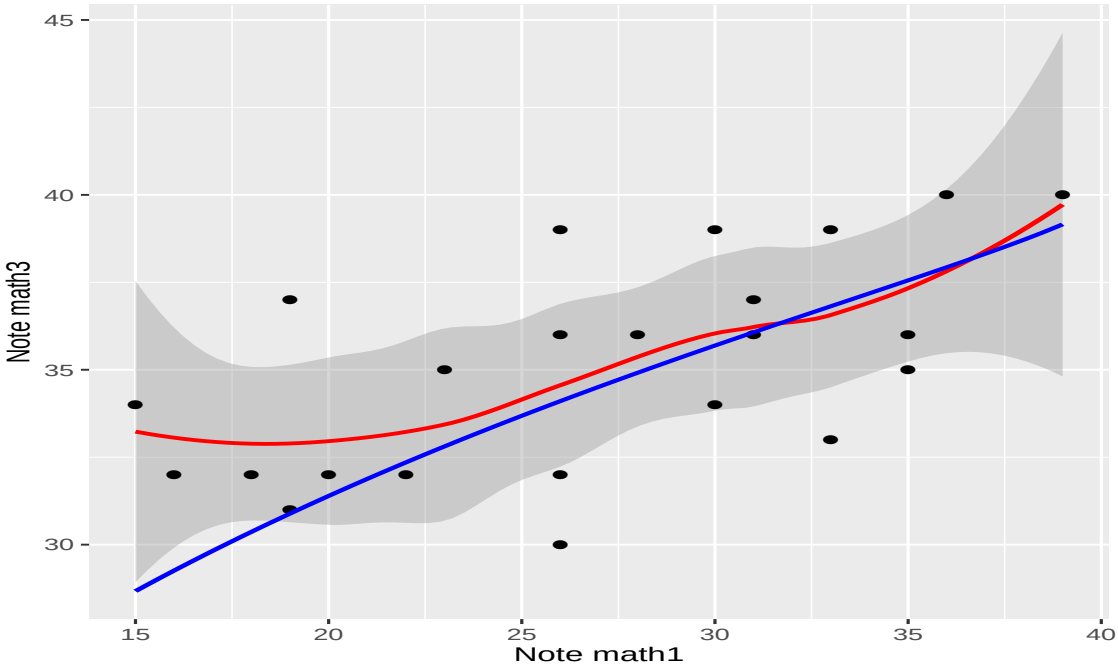


FIGURE 4.13 – Prédiction de la note en septième année de l'école `Numéro 2` en utilisant la 2-copule échangeable (courbe en bleue) et le modèle de régression non-paramétrique (courbe en rouge).

Le graphique de l'école `Numéro 3` se présente sur la figure.



!h

FIGURE 4.14 – Prédiction de la note en septième année de l'école `Numéro 3` en utilisant la 2-copule échangeable (courbe en bleue) et le modèle de régression non-paramétrique (courbe en rouge).

En somme, les graphes permettent de mettre en exergue le fait que la courbe de prédiction par le modèle de 2-échangeable décrit "correctement" les données. Autrement dit, les courbes de prédictions sont fonction de l'école. Nous pouvons dire globalement que le modèle prédictif est satisfaisant au regard des résultats obtenus et justifie amplement le modèle alternatif que nous proposons dans cette thèse.

4.6 Étude approfondie de la précision de la prédiction avec le modèle de 2-copule échangeable

Dans cette section, nous étudions la variance de prédiction en utilisant le modèle de 2-copule échangeable construit. Nous rappelons que nous utilisons les ajustements des lois marginales et les trois copules sélectionnées dans les sections précédentes.

Pour prédire Y dans une grappe j connaissant x, $\mathbf{x}_j$ et $\mathbf{y}_j$, nous utilisons la formule de l'équation

(4.20). La variance de cette prédiction à partir de l'équation (2.33) donne

$$\begin{aligned}\mathbb{V}\left(Y|x,\mathbf{x}_j,\mathbf{y}_j\right) &= \int_{\mathbb{R}}\left\{G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\hat{\sigma}_j t)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right]\right\}^2\phi(t)dt \\ &- \left\{\mathbb{E}\left(Y|x,\mathbf{x}_j,\mathbf{y}_j\right)\right\}^2. \end{aligned} \tag{4.22}$$

L'expression de cette variance s'approxime en utilisant la méthode de Gauss-Hermite par

$$\begin{aligned}\mathbb{V}\left(Y|x,\mathbf{x}_j,\mathbf{y}_j\right) \approx & \sum_{i=1}^{K}\frac{w_i}{\sqrt{\pi}}\left(G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\sqrt{2}\hat{\sigma}_j t_i)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\}\right]\right)^2 - \\ & \left(\sum_{i=1}^{K}\frac{w_i}{\sqrt{\pi}}G^{-1}\left[C_{2|1}^{-1}\left\{\Phi(\hat{\mu}_j+\sqrt{2}\hat{\sigma}_j t_i)|F(x;\hat{\alpha}_1,\hat{\beta}_1)\right\};\hat{\alpha}_3,\hat{\beta}_3,\hat{\lambda}\right]\right)^2,\end{aligned}$$

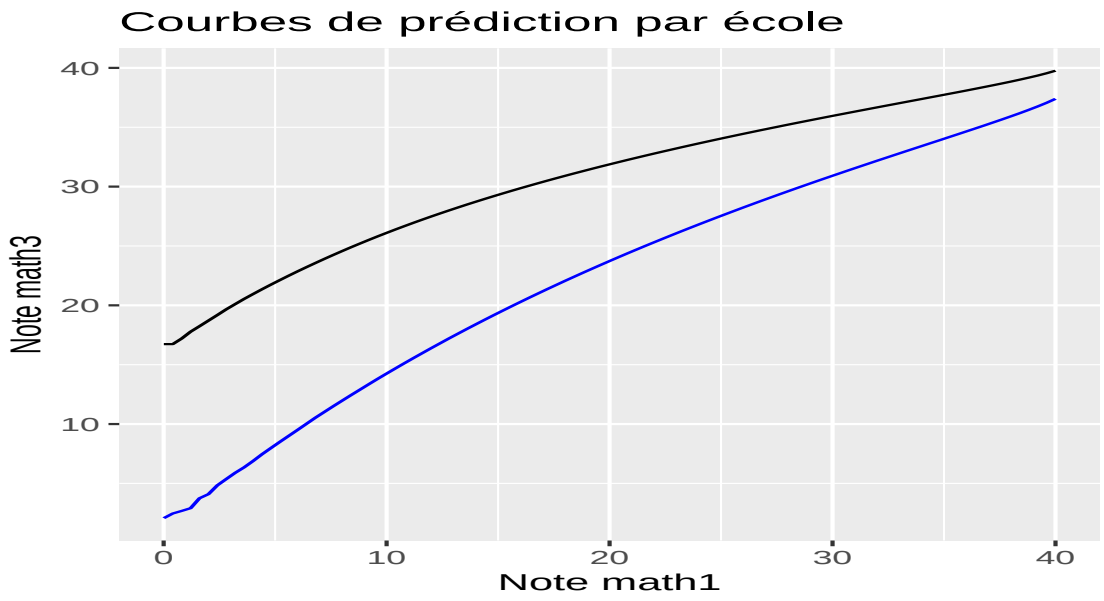
où nous rappelons que K est le nombre de points utilisé par l'approximation de Gauss-Hermite.

Avant de présenter les courbes de l'écart-type de prédiction en fonction de l'école, nous présentons une description des variables `math1` et `math3`. Nous choisissons les écoles `Numéro 1` et `Numéro 6`.

TABLEAU 4.9 – Sommaire de quelques statistiques descriptives des variables `math1` et `math3` par école.

École		*Minimum*	*Médiane*	*Moyenne*	*Maximum*	*Écart-type*
`Numéro 1`						
	`math1`	19	30	29.21	39	4.70
	`math3`	11	31.5	29.82	38	6.04
`Numéro 6`						
	`math1`	13	29	28.52	40	7.57
	`math3`	30	36	35.84	40	3.34

Nous construisons des courbes de l'écart-type de prédiction de la note `math3` (note en septième année) pour des élèves de deux écoles connaissant la valeur de la note `math1`(note en quatrième année).



H

FIGURE 4.16 – Prédiction de la note en 7e année des élèves appartenant aux écoles `Numéro 1` (couleur bleue) et `Numéro 6` (couleur noire) en utilisant le modèle de 2-copule échangeable.

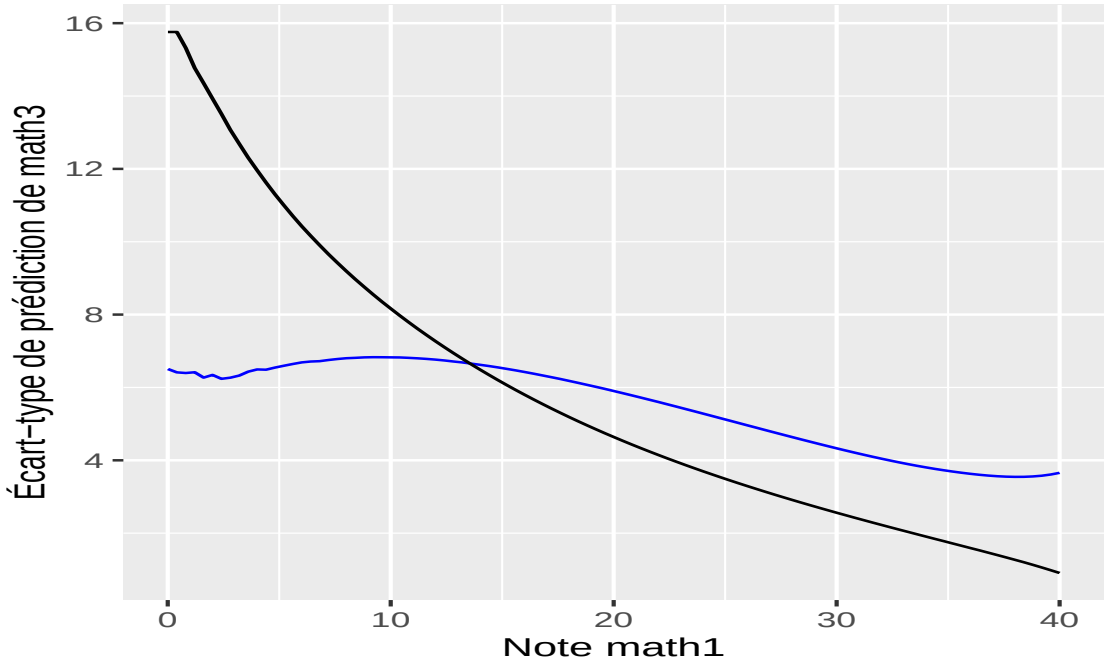


FIGURE 4.15 – Écart-type de prédiction de la note en 7e année des élèves appartenant aux écoles `Numéro 1` (couleur bleue) et `Numéro 6` (couleur noire) en utilisant le modèle de 2-copule échangeable.

En rappelant que la prédiction dépend intrinsèquement de l'école, la conclusion clairement émise est celle de la dépendance de la variance de prédiction de l'école d'appartenance de l'unité. Ceci confirme l'écriture de l'équation (4.22) en fonction des variables $\hat{\mu}_j$ et $\hat{\sigma}_j$. Spécifiquement, les notes observées de `math3` dans l'école `Numéro 1` appartiennent à [11,38] alors qu'elles appartiennent à [30,40] dans l'école `Numéro 6`. La précision des notes dans l'école `Numéro 6` est forte plus que celle de l'école `Numéro 1`, voir figure 4.15. Nous notons aussi que plus, nous sommes proche de la note maximale, plus la prédiction est précise.

Pour les écoles du tableau 4.8, les valeurs $\hat{\sigma}_j$ sont sensiblement les mêmes et donc la variation de l'équation (4.13) dépend de $\hat{\mu}_j$. Les fonctions de densité et de répartition de la variable `math3` connaissant `math1`. Dans notre cas, x et y représentent respectivement `math1`/41 et `math3`/41. Nous présentons pour les écoles `Numéro 1` et `Numéro 6`, la distribution conditionnelle de l'équation (4.13) pour $(\hat{\mu}_j, \hat{\sigma}_j) = (-0.643, 0.927)$ et $(\hat{\mu}_j, \hat{\sigma}_j) = (0.788, 0.929)$ respectivement. La densité conditionnelle dépend de la valeur de `math1` fixée au départ. Nous fixons les valeurs de la note `math1` comme étant le minimum, la médiane et le maximum des observations de l'école à partir du tableau 4.9. Ce choix reste le même pour les deux écoles dans le but de faire une comparaison.

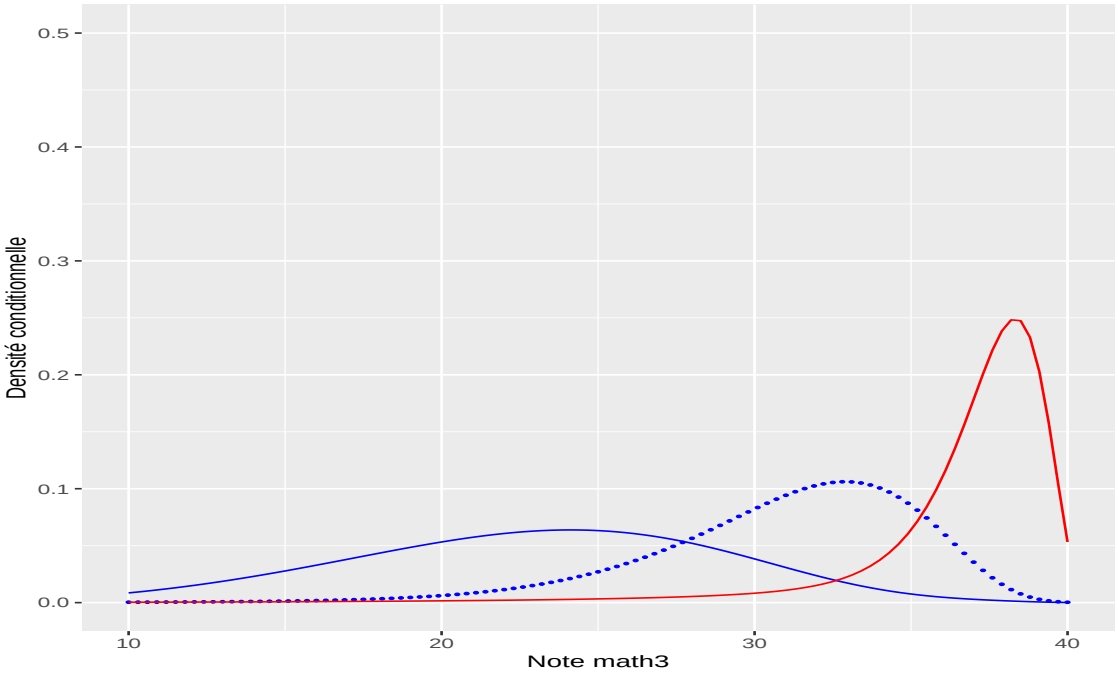
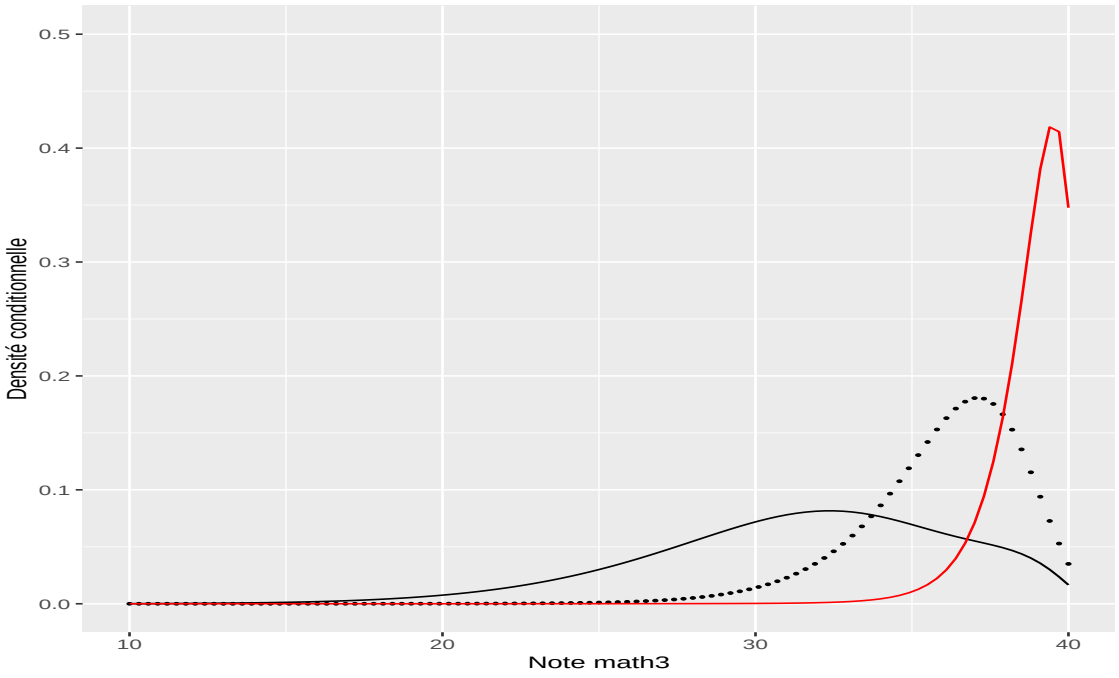


FIGURE 4.17 – Densité conditionnelle de l'équation (4.13) pour `math1`=19 (courbe en trait plein), `math1`=30 (courbe en pointillé) et `math1`=39 (courbe en rouge) en utilisant le modèle de 2-copule échangeable pour l'école `Numéro 1`.



H

FIGURE 4.18 – Densité conditionnelle de l'équation (4.13) pour `math1`=19 (courbe en trait plein), `math1`=30 (courbe en pointillé) et `math1`=39 (courbe en rouge) en utilisant le modèle de 2-copule échangeable pour l'école `Numéro 6`.

Nous constatons que les courbes de la densité conditionnelle sont des fonctions unimodales et atteignent un pic pour une valeur donnée avant de décroitre rapidement pour des valeurs proches de la note maximale (40) pour `math3`. De plus, le graphique est asymétrique comme les graphiques des figures 4.2 et 4.3 portant sur les lois marginales F et G. L'asymétrie est plus prononcée pour de grandes valeurs de `math1`. Les graphiques montrent une grande dispersion des données de l'école `Numéro 1` (figure 4.17) par rapport à celle à `Numéro 6` (figure 4.18). Cette dernière conclusion se confirme par le fait que la courbe de l'écart-type de l'école `Numéro 1` (courbe en bleue) est au-dessus de la courbe de l'école `Numéro 6`, voir figure 4.15. Enfin, nous remarquons que le modèle est très flexible et permet de construire, pour chaque valeur de `math1` une courbe de densité conditionnelle.

4.7 Conclusion sur la construction de la 2-copule échangeable

Pour analyser les données portant sur les notes en mathématiques de quatrième et de septième année, obtenues par des écoliers, nous utilisons trois modèles : le modèle de 2-copule échangeable, le modèle de régression copule et le modèle linéaire mixte. Précisément, pour la

2-copule échangeable, nous avons choisi les cinq éléments F, G, $\left(C_{1,n}^{(1)}\right)$, $C^{(2)}$ et $\left(C_{1,n}^{(3)}\right)$ entrant dans sa forme explicite puis estimée les paramètres associés. Pour quantifier la différence entre les différents modèles, nous visualisons graphiquement les courbes de prédiction. Ces courbes montrent le caractère curviligne et adaptable du modèle prédictif basé sur la 2-copule échangeable. La courbe de prédiction dans ce cas s'ajuste aux données en fonction de l'école et qui confirme la flexibilité du modèle de 2-copule échangeable.

Conclusion

En modélisation multidimensionnelle, plusieurs obstacles se posent et diverses approches existent. La formulation des modèles flexibles et facilement maniables constitue donc un enjeu. Dans le cas particulier de données en grappes provenant de plusieurs sites, l'ajustement d'un modèle linéaire mixte utilisant une hypothèse de normalité est très restrictif. Ceci se vérifie dans le domaine de la finance et des assurances où généralement, les données sont reparties de manière asymétrique. Dans le cas de notre thèse, nous proposons une nouvelle forme de modélisation pour ces données, utilisant une d-copule échangeable, disposant des propriétés dites de compatibilité, d'échangeabilité et fermeture. Le modèle proposé fait intervenir des copules et particulièrement des copules échangeables. Par la suite, nous utilisons le modèle pour faire de la prédiction. Enfin, nous montrons que notre modèle généralise le modèle de Battese *et al.* (1988), en considérant un cas particulier.

Pour le modèle proposé, nous avons expliqué la procédure d'ajustement et deux méthodes d'estimation des paramètres. L'une des méthodes d'estimation est une généralisation de la méthode IFM décrite par Joe (1996). Nous comparons la méthode IFM généralisée et la méthode du maximum de vraisemblance par une étude de Monte-Carlo. Nous concluons que, généralement, la méthode du maximum de vraisemblance est plus précise que la méthode IFM généralisée. Plus particulièrement, nous faisons un calcul théorique des variances des estimateurs de paramètres du modèle de 2-copule échangeable dans un exemple. Finalement, nous mettons en œuvre la procédure d'ajustement d'un modèle de 2-copule échangeable sur des données qui confirment sa flexibilité.

Le modèle de 2-copule proposé dans cette thèse s'ajuste aux variables continues. Il est souhaitable que dans les travaux ultérieurs qu'il soit étendu à tout type de variables sans restriction et l'implémenter sur un logiciel statistique.

Annexe A

Matériel supplémentaire associé au chapitre 1

Cette annexe donne les résultats pour

- le calcul de la distribution conditionnelle $C_{2|1}$, présenté au tableau 1.3 à partir d'une copule bêta et
- le calcul du prédicteur $\mathbb{E}(Y|X_1)$ en utilisant la méthode de la régression avec copule de la section 1.6 dans le cas bivarié.

A.1 Calcul théorique de la fonction de répartition conditionnelle $C_{2|1}$ associé à une copule bêta bivariée et de son inverse

Détermination de $C_{2|1}$ et son inverse à partir de la copule bêta

La fonction de répartition de la loi bêta de paramètres p et q est notée $B_{p,q}$ et son inverse $B_{p,q}^{-1}$. On note $c(\cdot; p_1, p_2, q)$, la densité de la copule bêta bivariée obtenue à partir de l'équation (1.18) pour $d = 2$, de paramètres p_1, p_2 et q. Pour $u \in [0,1]$ et fixé, on a par définition

$$C_{2|1}(z|u) = \int_0^z c(u, v; p_1, p_2, q)dv, \ z \in [0,1]. \tag{A.1}$$

En posant $w = B_{p_2,q}^{-1}(v)/\left\{1 - B_{p_2,q}^{-1}(v)\right\}$, nous avons $v = B_{p_2,q}\left(w/1+w\right)$. Ainsi donc

$$dv \quad = \quad \frac{1}{(1+w)^2} B'_{p_2,q}\left(\frac{w}{1+w}\right) dw = \frac{\Gamma(p_2+q)}{\Gamma(p_2)\Gamma(q)} \frac{w^{p_2-1}}{(1+w)^{p_2+q}} dw,$$

où $B'_{p_2,q}$ est la densité de la loi bêta de paramètres p_2 et q, voir équation (1.16). En posant $z_\nu = B^{-1}_{p_2,q}(z)/\left\{1 - B^{-1}_{p_2,q}(z)\right\}$, nous avons un développement de l'équation (A.1) par

$$\begin{aligned} C_{2|1}(z|u) &= \frac{\Gamma(p_1+p_2+q)}{\Gamma(p_1+q)\Gamma(p_2)} \left\{1 - B^{-1}_{p_2,q}(u)\right\}^{-(p_1+q)} \int_0^{z_\nu} \frac{w^{p_2-1}}{\left\{w + \frac{1}{1-B^{-1}_{p_2,q}(u)}\right\}^{p_1+p_2+q}} dw \\ &= \frac{\Gamma(p_1+p_2+q)}{\Gamma(p_1+q)\Gamma(p_2)} \int_0^{z_\nu\left\{1-B^{-1}_{p_2,q}(u)\right\}} \frac{w^{p_2-1}}{(1+w)^{p_1+p_2+q}} dw. \end{aligned}$$

En posant $y = w/(1+w)$, $dw = 1/(1-y)^2 dy$ donc

$$C_{2|1}(z|u) = \int_0^\alpha \frac{\Gamma(p_1+p_2+q)}{\Gamma(p_1+q)\Gamma(p_2)} y^{p_2-1}(1-y)^{p_1+q-1} dy, \tag{A.2}$$

où α est défini par

$$\alpha = \frac{z_\nu \left\{1 - B^{-1}_{p_1,q}(u)\right\}}{1 + z_\nu \left\{1 - B^{-1}_{p_1,q}(u)\right\}}.$$

Or la fonction f définie par

$$f(y; p_1, p_2, q) = \frac{\Gamma(p_2+p_1+q)}{\Gamma(p_2)\Gamma(p_1+q)} y^{p_2-1}(1-y)^{p_1+q-1}, \quad y \in [0,1],$$

est la densité de la loi bêta de paramètres p_2 et p_1+q. En conséquence, la fonction conditionnelle (A.2) est donnée par

$$C_{2|1}(z|u) = B_{p_2,p_1+q}\left[\frac{B^{-1}_{p_2,q}(z)\left\{1 - B^{-1}_{p_1,q}(u)\right\}}{1 - B^{-1}_{p_1,q}(u)B^{-1}_{p_2,q}(z)}\right]. \tag{A.3}$$

De cette fonction de répartition conditionnelle, en posant $t = C_{2|1}(v|u)$ nous avons l'équation

$$B^{-1}_{p_2,p_1+q}(t) = \frac{B^{-1}_{p_2,q}(v)\left\{1 - B^{-1}_{p_1,q}(u)\right\}}{1 - B^{-1}_{p_1,q}(u)B^{-1}_{p_2,q}(v)}, \tag{A.4}$$

et en la résolvant nous obtenons

$$v = B_{p_2,q}\left\{\frac{B^{-1}_{p_2,p_1+q}(t)}{1 - B^{-1}_{p_1,q}(u) + B^{-1}_{p_1,q}(u)B^{-1}_{p_2,p_1+q}(t)}\right\}.$$

D'où l'inverse de la fonction de répartition conditionnelle de l'équation (A.3) est

$$C^{-1}_{2|1}(z|u) = B_{p_2,q}\left\{\frac{B^{-1}_{p_2,p_1+q}(z)}{1 - B^{-1}_{p_1,q}(u) + B^{-1}_{p_1,q}(u)B^{-1}_{p_2,p_1+q}(z)}\right\}. \tag{A.5}$$

Ceci conclut les résultats de $C_{2|1}$ et $C^{-1}_{2|1}$, obtenus et présentés dans le tableau 1.3 par rapport à la copule bêta bivariée.

A.2 Évaluation du prédicteur construit au chapitre 1, section 1.6 par la régression avec copule pour deux copules spécifiques

L'expression du prédicteur par la régression avec copule dans le cas bivarié est

$$\mathbb{E}(Y|X_1) = \int_0^1 G_0^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\} du_0. \tag{A.6}$$

La fonction G_0 est la fonction de répartition de la loi normale de moyenne μ et d'écart-type σ.

Cas 1 : nous considérons que la copule associée à (X_1, Y) est la copule de Clayton de paramètre θ.

Considérons que C_θ est la copule de Clayton de paramètre θ et de l'équation (A.6), nous avons

$$\begin{aligned}\mathbb{E}(Y|X_1) &= \int_0^1 G_0^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\} du_0 \\ &= \mu + \sigma \int_0^1 \Phi^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\} du_0\end{aligned}$$

En posant $t = u_0^{-\theta}$, nous avons $u_0 = t^{-1/\theta}$ et

$$\int_0^1 \Phi^{-1}(u_0)c\left\{u_0, F_1(X_1)\right\} du_0 = \int_1^\infty \Phi^{-1}(t^{-1/\theta}) f_T(t)dt,$$

où la fonction f_T est la densité de la variable T, définie par

$$f_T(t) = \left(\frac{\theta+1}{\theta}\right) F_1(X_1)^{-\theta-1}\left\{t + F_1(X_1)^{-\theta} - 1\right\}^{-1/\theta-2},\ t > 1.$$

Au final, nous avons le résultat de l'équation (1.29) défini par

$$\mathbb{E}(Y|X_1) = \mu + \sigma\mathbb{E}\left\{\Phi^{-1}(T^{-1/\theta})\right\}.$$

Cas 2 : nous considérons que la copule associée à (X_1, Y) est la copule FGM de paramètre θ.

Nous partons de l'équation (A.6) et on obtient

$$\mathbb{E}(Y|X_1) = \mu + \sigma \int_0^1 \Phi^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\} du_0,$$

$$\begin{aligned}\int_0^1 \Phi^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\} du_0 &= \int_0^1 \Phi^{-1}(u_0)\left[1 + \theta\left\{1 - 2F_1(X_1)\right\}(1 - 2u_0)\right] du_0 \\ &= \int_0^1 \Phi^{-1}(u_0)du_0 + \theta\left\{1 - 2F_1(X_1)\right\}\int_0^1 \Phi^{-1}(u_0)du_0 \\ &- 2\theta\left\{1 - 2F_1(X_1)\right\}\int_0^1 \Phi^{-1}(u_0)u_0 du_0,\end{aligned}$$

Or nous avons les résultats

$$\int_0^1 \Phi^{-1}(u_0)du_0 = \int_{\mathbb{R}} t\phi(t)dt = 0, \quad \int_0^1 \Phi^{-1}(u_0)u_0du_0 = \int_{\mathbb{R}} t\phi(t)\Phi(t)dt = E_2/2 = \frac{1}{2\sqrt{\pi}},$$

E_2 est donné par l'équation (B.3). En conséquence, nous avons

$$\int_0^1 \Phi^{-1}(u_0)c\left\{u_0, F_1(X_1); \theta\right\}du_0 \quad = \quad -\theta\left\{1 - 2F_1(X_1)\right\} \times \frac{1}{\sqrt{\pi}}.$$

D'où le prédicteur de l'équation (1.30) est défini par

$$\mathbb{E}(Y|X_1) = \mu + \frac{\sigma\theta}{\sqrt{\pi}}\left\{2F_1(X_1) - 1\right\}.$$

Annexe B

Matériel supplémentaire associé au chapitre 2

L'application du modèle échangeable de l'équation (2.10) dans le cas de régression pour $d = 2$ à $n = 2$ et $(u_1, u_2, v_1, v_2) \in [0,1]^4$ est

$$\begin{aligned} c_{2,2}\left\{(u_1,v_1),(u_2,v_2)\right\} &= c_{13}^{(1)}(u_1,u_2) \times c_{12}^{(2)}(u_1,v_1) \times c_{12}^{(2)}(u_2,v_2) \\ &\quad \times c_{24;13}^{(3)}\left\{C_{2|1}(v_1|u_1), C_{2|1}(v_2|u_2)\right\}, \end{aligned} \tag{B.1}$$

La sous-section 2.7.3 présente l'évaluation du prédicteur pour quelques copules particulières toujours en dimension 2.

B.1 Modèle échangeable construit à l'aide de la copule bêta et modèle multiniveau satisfaisant l'hypothèse d'indépendance conditionnelle partielle

Sous les hypothèses formuler à l'exemple 2.6, de la section 2.3, les copules impliquées dans le modèle défini dans l'équation (B.1) sont alors

- C_{13} est la copule bêta de paramètres p, p et q ;
- C_{12} est la copule bêta de paramètres p, r et q ;
- $C_{24;13}$ est la copule bêta de paramètres r, r et $p+q$.

Justifions que la copule C_{13} est la copule bêta de paramètres p, p et q

La densité de probabilité du vecteur aléatoire (A, B_1, B_2) est

$$f(a, b_1, b_2) = \frac{a^{q-1} b_1^{p-1} b_2^{p-1}}{\Gamma(q)\,\Gamma(p)^2} \exp\{-a - b_1 - b_2\}, \quad a > 0,\ b_1 > 0,\ b_2 > 0. \tag{B.2}$$

On pose $x_1 = b_1/(a + b_1)$ et $x_2 = b_2/(a + b_2)$. Le Jacobien associé au vecteur aléatoire (A, X_1, X_2), où X_1 et X_2 proviennent de l'équation (2.11) est $J = a^2/x_1^2 x_2^2$. La densité de probabilité g du vecteur aléatoire (A, X_1, X_2) est donc

$$g(a, x_1, x_2) = \frac{a^{2p+q-1}}{\Gamma(q)\,\Gamma(p)^2} \frac{x_1^{p-3}}{(1-x_1)^{p-1}} \frac{x_2^{p-3}}{(1-x_2)^{p-1}} \exp\left[-a\left\{1 + \left(\frac{x_1}{1-x_1}\right) + \left(\frac{x_2}{1-x_2}\right)\right\}\right].$$

En faisant une intégration par rapport à a, la densité de probabilité h du vecteur aléatoire (X_1, X_2) est

$$h(x_1, x_2) = \frac{\Gamma(2p+q)}{\Gamma(p)^2} \left(\frac{x_1}{1-x_1}\right)^{p-1} \left(\frac{x_2}{1-x_2}\right)^{p-1} \left(1 + \frac{x_1}{1-x_1} + \frac{x_2}{1-x_2}\right)^{-2p-q},$$

où $0 < x_1, x_2 < 1$. En conséquence, la densité de la copule associée au vecteur aléatoire (X_1, X_2) s'écrit

$$c(u_1, u_2; p, q) = \frac{\Gamma(q)\Gamma(2p+q)}{\Gamma(p+q)^2} \frac{(1+z_1)^{p+q}(1+z_2)^{p+q}}{(1+z_1+z_2)^{2p+q}},$$

où z_1 et z_2 sont définis par

$$z_1 = \frac{B_{p,q}^{-1}(u_1)}{1 - B_{p,q}^{-1}(u_1)}, \quad z_2 = \frac{B_{p,q}^{-1}(u_2)}{1 - B_{p,q}^{-1}(u_2)}.$$

et $B_{p,q}^{-1}(x)$ est l'inverse de la fonction de répartition de la loi bêta de paramètres p et q, évaluée au point x. La copule associée au vecteur aléatoire (X_1, X_2) est donc la copule bêta de paramètres p, p et q qui est bien sûr échangeable. Elle est obtenue à partir de l'équation (1.18) pour $p_1 = p_2 = p$ et $q = q$.

Justifions que la copule C_{12} est la copule bêta de paramètres p, r et q

En posant $D = B_1 + A_1$, nous avons

$$X_1 = \frac{B_1/A}{B_1/A + 1}, \quad Y_1 = C_1/A.$$

La copule du vecteur aléatoire (X_1, Y_1) est la même que celle associée au vecteur $(B_1/A, C_1/A)$ et en utilisant le fait qu'on a

$$B_1/A \sim \mathcal{GB}2(p, q) \ \text{ et } \ C_1/A \sim \mathcal{GB}2(r, q),$$

la copule associée est donc une copule bêta de paramètres p, r et q.

Justifions que la copule $C_{24;13}$ est la copule bêta de paramètres r, r et $p+q$

La copule $C_{24;13}$ est la copule associée à (Y_1, Y_2) sachant (X_1, X_2). Nous avons les résultats suivants.

- Sachant X_1, la variable aléatoire Y_1 suit à une constante près la loi $\mathcal{GB}2$ de paramètres r et $p+q$;
- Sachant X_2, la variable aléatoire Y_2 suit à une constante près la loi $\mathcal{GB}2$ de paramètres r et $p+q$.

En conséquence, la copule associée à $(Y_1, Y_2|X_1, X_2)$ est la copule bêta de paramètre r, r et $p+q$.

B.2 Calcul de la prédiction avec quelques copules particulières pour $C^{(2)}$ et $C^{(3)}$

Cette section donne les preuves des résultats obtenus dans l'exemple de la section 2.7.5. Nous rappelons que les copules $C^{(2)}$ et $C^{(3)}$ sont des copules FGM de paramètres θ_2 et θ_3 respectivement. En utilisant l'équation (2.26), le prédicteur de Y_2 sachant X_1, Y_1, X_2 est donné par

$$\mathbb{E}(Y_2|X_1, Y_1, X_2) = \int_{\mathbb{R}} y g_0(y) A_0(y, X_1, Y_1, X_2) dy,$$

où A_0 est

$$A_0(y, X_1, Y_1, X_2) = c^{(2)}\left\{F_0(X_2), G_0(y); \theta_2\right\} c^{(3)}\left[F(G_0(Y_1)|F_0(X_1), F\left\{G_0(y)|F_0(X_2)\right\}; \theta_3\right].$$

La prédiction dépend des copules $C^{(2)}$ et $C^{(3)}$. Le prédicteur est donc

$$\begin{aligned}\mathbb{E}(Y_2|X_1, Y_1, X_2) &= \frac{\theta_2\left\{2\Phi(X_2)-1\right\}}{\sqrt{\pi}} + \theta_3\left[1 - 2\Phi(Y_1)\left\{1+\theta_2\left(1-\Phi(Y_1)\right)\left(1-2\Phi(X_1)\right)\right\}\right] \\ &\times \left[-\frac{1}{\sqrt{\pi}} + \theta_2\left\{2\Phi(X_2)-1\right\}\left\{\frac{1}{\sqrt{\pi}} + 0.2\theta_2(1-2\Phi(X_2))\right\}\right].\end{aligned}$$

Justification du résultat de prédiction

En notant E_n, l'espérance du maximum de n variables aléatoires indépendantes et identiquement distribuées suivant la loi normale standard. Nous avons les résultats suivants en nous inspirant du lien https://math.stackexchange.com/questions/473229/expected-value-of-maximum-and-minimum-of-n-normal-random-variables,

$$\int_{\mathbb{R}} y\Phi'(y)\Phi^{n-1}(y)dy = \frac{E_n}{n}, \quad n \in \mathbb{N}^*.$$

En particulier, nous avons

$$E_1 = 0,\ E_2 = 1/\sqrt{\pi},\ E_3 = 3/2\sqrt{\pi}\, E_4 = 3\pi^{-3/2}\cos^{-1}(-1/3). \tag{B.3}$$

La fonction A_0 peut être explicitée par

$$\begin{aligned} A_0(y; X_1, Y_1, X_2) &= 1 + \theta_3\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\left\{1 - C_{2|1}(v_2|u_2)\right\} + \theta_2(1-2u_2)\left\{1 - 2G_0(y)\right\} \\ &+ \theta_2\theta_3(1-2u_N)\left\{1 - 2G_0(y)\right\}\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\left[1 - 2F\left\{G_0(y)|u_2\right\}\right], \end{aligned}$$

et $u_1 = F_0(X_1)$, $u_2 = F_0(X_2)$ et $v_1 = G_0(Y_1)$. En utilisant l'hypothèse que les marges F_0 et G_0 sont des lois normales centrées réduites, nous posons :

$$\mathbb{E}(Y_2|X_1, Y_1, X_2) = B_1 + B_2 + B_3, \tag{B.4}$$

où les fonctions B_1, B_2 et B_3 sont calculées séparément.

Pour la fonction B_1 de l'équation (B.4), nous avons

$$\begin{aligned} B_1 &= \theta_2(1-2u_2)\int_{\mathbb{R}} y\Phi^{'}(y)\left\{1 - 2\Phi(y)\right\}dy \\ &= \theta_2(1-2u_2)\left\{\int_{\mathbb{R}} y\Phi^{'}(y)dy - 2\int_{\mathbb{R}} y\Phi^{'}(y)\Phi(y)dy\right\} \\ &= \theta_2(1-2u_2)\left(E_1 - E_2\right)\cdot \end{aligned}$$

En conséquence, l'expression de B_1, donne

$$B_1 = \frac{-\theta_2\left\{1 - 2\Phi(X_2)\right\}}{\sqrt{\pi}}\cdot \tag{B.5}$$

Pour la fonction B_2 de l'équation (B.4), nous avons

$$\begin{aligned} B_2 &= \theta_3\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\int_{\mathbb{R}} y\Phi^{'}(y)\left[1 - 2F\left\{\Phi(y)|u_2\right\}\right]dy \\ &= \theta_3\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\left[-2\theta_2(1-u_2)\left\{\int_{\mathbb{R}} y\Phi^{'}(y)\Phi(y)dy - \int_{\mathbb{R}} y\Phi^{'}(y)\Phi^2(y)dy\right\}\right] \\ &\quad -2\theta_3\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\int_{\mathbb{R}} y\Phi^{'}(y)\Phi(y)dy \\ &= \theta_3\left\{1 - 2C_{2|1}(v_1|u_1)\right\}\left\{-2\theta_2(1-2u_2)(E_2/2 - E_3/3) - 2\times E_2/2\right\}, \end{aligned}$$

d'où

$$B_2 = \frac{-\theta_3\left[1 - 2F\left\{\Phi(Y_1)|\Phi(X_1)\right\}\right]}{\sqrt{\pi}}\cdot \tag{B.6}$$

On pose d'abord $\alpha = \theta_2\theta_3(1-2u_2)\{1-2F(v_1|u_1)\}$ et le calcul de B_3 de l'équation (B.4) donne

$$\begin{aligned} B_3 &= \theta_2\theta_3(1-2u_2)\{1-2F(v_1|u_1)\}\int_{\mathbb{R}} y\Phi^{'}(y)\{1-2\Phi(y)\}\{1-2F(\Phi(y)|u_2)\} \\ &= \alpha\left[\frac{1}{\sqrt{\pi}}+4\int_{\mathbb{R}} y\Phi^{'}(y)\{1-2\Phi(y)\}\Phi(y)dy\right] \\ &+ 4\alpha\theta_2(1-2u_2)\int_{\mathbb{R}} y\Phi^{'}(y)\Phi(y)\{1-2\Phi(y)\}\{1-\Phi(y)\}dy \\ &= \alpha\left[\frac{1}{\sqrt{\pi}}+4\left\{\frac{E_2}{2}-2\frac{E_3}{3}+\theta_2(1-2u_2)(E_2/2-E_3+E_4/2)\right\}\right] \end{aligned}$$

d'où B_3 est

$$B_3 = \alpha\left[-\frac{1}{\sqrt{\pi}}+4\theta_2(1-\Phi(X_2))\left\{-\frac{1}{\sqrt{\pi}}+\frac{3\pi^{-3/2}}{2}\cos^{-1}(-1/3)\right\}\right] \tag{B.7}$$

De tout ce qui précède, des équations (B.5), (B.6) et (B.7) intégrées dans (B.4), nous obtenons l'expression du prédicteur $\mathbb{E}(Y_2|X_1,Y_1,X_2)$ de l'équation (2.36).

B.3 Construction de prédicteur à l'aide du modèle échangeable dans le cas de copules normales

La prédiction de Y_0 sachant x_0 et $\mathbf{z}_1=(x_1,y_1)^T$, à partir de l'équation (2.34) est

$$\mathbb{E}_{\rho_2}(Y_0|x_0,\mathbf{z}_1)=\rho_2 x_0+\frac{\rho_3}{\sqrt{1-\rho_2^2}}(y_1-\rho_2 x_1),\ \ \mathbb{V}_{\rho_2}(Y_0|x_0,\mathbf{z}_1)=(1-\rho_2^2)(1-\rho_3)\cdot$$

En utilisant le résultat que si $C^{(2)}$ une copule normale bivariée de paramètre ρ_2 et du tableau 1.3, nous avons

$$C_{2|1}^{-1}\{\Phi(\mu_0+\sigma_0 t)|\Phi(x_0)\}=\Phi\left\{\sqrt{1-\rho_2^2}(\mu_0+\sigma_0 t)+\rho_2 x_0\right\},$$

et de l'équation (2.30), nous avons :

$$\overline{w}_n = \frac{1}{n}\sum_{i=1}^{n}\Phi^{-1}\left[C_{2|1}\{G(y_i)|F(x_i)\}\right]=\frac{1}{\sqrt{1-\rho_2^2}}(\overline{y}_n-\rho_2\overline{x}_n)\cdot \tag{B.8}$$

Le résultat final du prédicteur noté m est

$$m(x_0)=\mathbb{E}_{\rho_2}(Y_0|x_0,\mathbf{z}_1,\ldots,\mathbf{z}_n)=\rho_2 x_0+\frac{n\rho_3}{1+(n-1)\rho_3}(\overline{y}_n-\rho_2\overline{x}_n)\cdot \tag{B.9}$$

En supposant que toutes les copules sont normales, l'équation (B.9) correspond à celui du modèle de Battese *et al.* (1988) de l'équation (1.8), chapitre 1 où les correspondances suivantes sont établies.

$$\rho_2\equiv\beta,\qquad \rho_3\equiv\frac{\sigma_\nu^2}{\sigma_\varepsilon^2+\sigma_\nu^2}\cdot$$

La variance conditionnelle dans (2.33) pour $C^{(2)}$, comme une copule normale est

$$\mathbb{V}_{\rho_2}\left(Y_0|x_0,\mathbf{z}_1,\ldots,\mathbf{z}_n\right)=\mathbb{E}_{\rho_2}\left(Y_0^2|x_0,\mathbf{z}_1,\ldots,\mathbf{z}_n\right)-\left\{m(x_0)\right\}^2, \tag{B.10}$$

où nous avons

$$\mathbb{E}_{\rho_2}\left(Y_0^2|x_0,\mathbf{z}_1,\ldots,\mathbf{z}_n\right)=\rho_2^2x_0^2+\underbrace{2\rho_2x_0\sqrt{1-\rho_2^2}\mu_0}_{A}+\underbrace{\left(1-\rho_2^2\right)\left(\mu_0^2+\sigma_0^2\right)}_{B},$$

et en remplaçant $m(x_0)$, μ_0 et σ_0^2 par les équations (B.9), (2.29) et en utilisant la valeur de $\overline{w}_n$ de l'équation (B.8),

$$\underbrace{2\rho_2x_0\sqrt{1-\rho_2^2}\mu_0}_{A} \;=\; 2\rho_2x_0\sqrt{1-\rho_2^2}\cdot\frac{n\rho_3\overline{w}_n}{1+(n-1)\rho_3}=\frac{2n\rho_2\rho_3x_0}{1+(n-1)\rho_3}\cdot(\overline{y}_n-\rho_2\overline{x}_n),$$

et

$$\begin{aligned}\underbrace{\left(1-\rho_2^2\right)\left(\mu_0^2+\sigma_0^2\right)}_{B} &= \left(1-\rho_2^2\right)\mu_0^2+\left(1-\rho_2^2\right)\sigma_0^2\\ &= \frac{n^2\rho_3^2(\overline{y}_n-\rho_2\overline{x}_n)^2}{\left\{1+(n-1)\rho_3\right\}^2}+\frac{\left(1-\rho_2^2\right)(1-\rho_3)(1+n\rho_3)}{1+(n-1)\rho_3}.\end{aligned}$$

Nous rappelons aussi que

$$\begin{aligned}\left\{m(x_0)\right\}^2 &= \left[\rho_2x_0+\frac{n\rho_3}{1+(n-1)\rho_3}\cdot(\overline{y}_n-\rho_2\overline{x}_n)\right]^2\\ &= \rho_2^2x_0^2+\frac{n^2\rho_3^2}{\left\{1+(n-1)\rho_3\right\}^2}\cdot(\overline{y}_n-\rho_2\overline{x}_n)^2+\frac{2n\rho_2\rho_3x_0}{1+(n-1)\rho_3}\cdot(\overline{y}_n-\rho_2\overline{x}_n).\end{aligned}$$

En remplaçant les résultats précédents, nous obtenons

$$\mathbb{V}_{\rho_2}\left(Y_0|x_0,\mathbf{z}_1,\ldots,\mathbf{z}_n\right)=\frac{(1-\rho_2^2)(1-\rho_3)(1+n\rho_3)}{1+(n-1)\rho_3}.$$

L'expression du prédicteur dans le cas où nous avons un individu dans la grappe et on cherche à prédire pour le deuxième individu.

$$\mathbb{E}_{\rho_2}\left(Y_0|x_0,\mathbf{z}_1\right)=\rho_2x_0+\rho_3(y_1-\rho_2x_1),\;\;\mathbb{V}_{\rho_2}\left(Y_0|x_0,\mathbf{z}_1\right)=(1-\rho_2^2)(1-\rho_3^2).$$

B.4 Évaluation du prédicteur à l'aide de la 2-copule échangeable

Cette annexe présente d'autres figures de la suite de la section 2.7.3 pour différents types de copule $C^{(2)}$.

Nous présentons, dans un premier temps, le programme R d'obtention des courbes de prédiction et dans un deuxième temps, les graphiques des courbes de prédiction. Il permet de calculer (2.34) en utilisant l'approximation de Gauss-Hermite expliqué.

```
###-----------Fonction de prédiction de l'équation (2.29)----------#
#       Nom de la fonction : Predict_Gauss_Herm
#-------------------------------------------------------------------#

#-----------------Charger les packages-----------------------------#

library(copula)
library(VineCopula)
library(statmod)

Predict_Gauss_Herm<-Vectorize(function(x,phiz,n0=20,
                                  familic2,tau2,tau3=0.1,K)
{

#-------------Principales fonctions utilisées--------------------#

   # gauss.quad  : calcul des points d'évaluation de Gauss-Hermite
   # BiCopHinv1  : calcule v tel que w=F(v|u)
   # BiCopTau2Par: calcul le paramètre de la copule sachant tau


  #-------------------------ENTRÉES---------------------------#
  # @x la valeur de X associée la prédiction Y
  # @K est le nombre de nœuds de l'algorithme Gauss-Hermit
  # @phi\textbf{z} le quantile de la moyenne
  # @n0 la taille de la grappe
  # @familic2 la famille de la copule c2 (1: normale, 3 : Clayton,
  #                         4 : Gumbel, 5 : Frank et 6 : Joe
  # @tau2 corrélation associée à c2
  # @tau3 corrélation associée à c3


  #-------------------------SORTIE----------------------------#
  # @E(Y/x,phiz)

   #--Calcul des paramètres de c2 et c3 à partir de tau2 et tau3
```

```
  delta2<-BiCopTau2Par(family =familic2, tau = tau2)
  rho3<-BiCopTau2Par(family =1, tau = tau3)

 #***********-----Function de prédiction-------**************

    quadra<-gauss.quad(K,kind="hermite")
    mu0<-n0*rho3*qnorm(phiz)/(1+(n0-1)*rho3)
    sigma0<-sqrt((1-rho3)*(1+n0*rho3)/(1+(n0-1)*rho3))
    u<-pnorm(q=x)  ### loi marginale F0
    z<-pnorm(q=mu0+sqrt(2)*sigma0*quadra$nodes)  ##z=Phi(..)

  ##---Calcul de l'inverse de la conditionnelle F_c---#
    Fvinv<-BiCopHinv1(u1=rep(u,K), u2=z, family=familic2,
                 par=delta2, par2 = 0)

    return(sum(quadra$weights*qnorm(Fvinv))/sqrt(pi))
},"x")
```

B.4.1 Cas d'une copule normale pour $C^{(2)}$

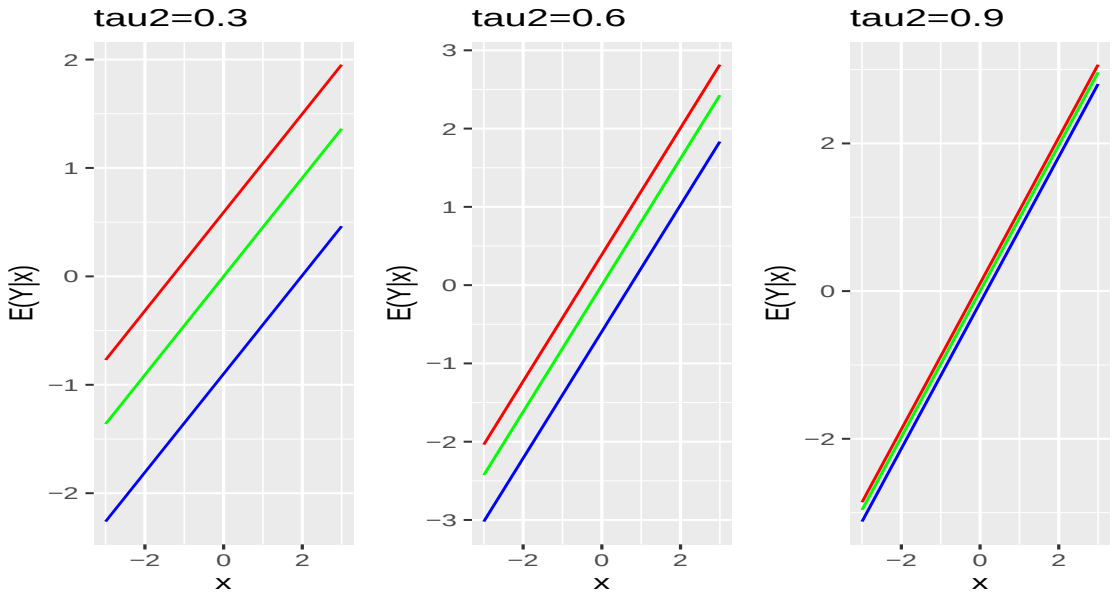


FIGURE B.1 – Courbes de prédiction de $\mathbb{E}(Y|x)$ avec une copule normale en fonction de l'importance des résidus $\overline{w}_n = -0.84$ (courbe bleue), $\overline{w}_n = 0$ (courbe verte) et $\overline{w}_n = 0.84$ (courbe rouge) avec taille $n = 20$.

B.4.2 Cas d'une copule de Gumbel pour $C^{(2)}$

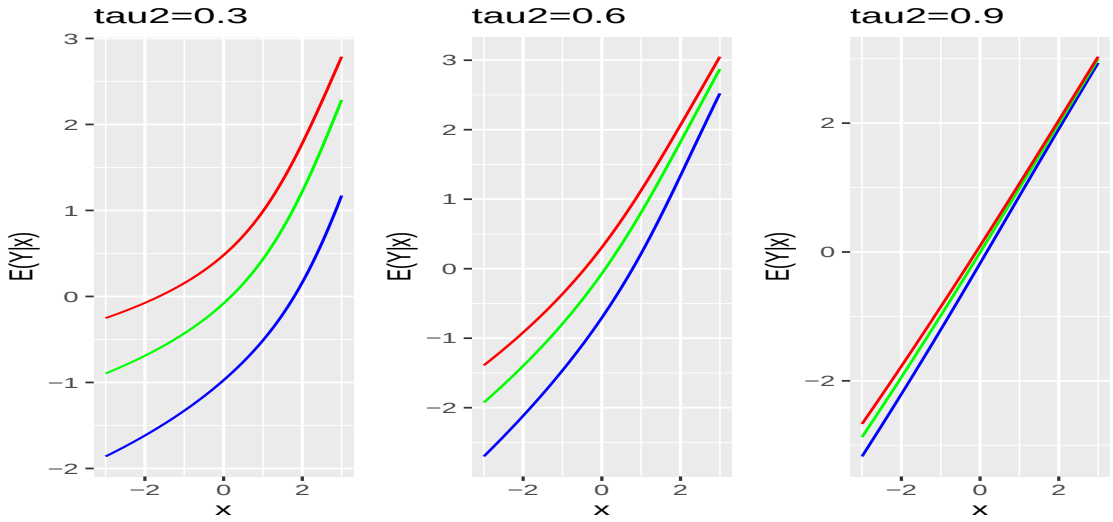


FIGURE B.2 – Courbes de prédiction de $\mathbb{E}(Y|x)$ avec une copule de Gumbel en fonction de l'importance des résidus $\overline{w}_n = -0.84$ (courbe bleue), $\overline{w}_n = 0$ (courbe verte) et $\overline{w}_n = 0.84$ (courbe rouge) avec taille $n = 20$.

B.4.3 Cas d'une copule de Joe pour $C^{(2)}$

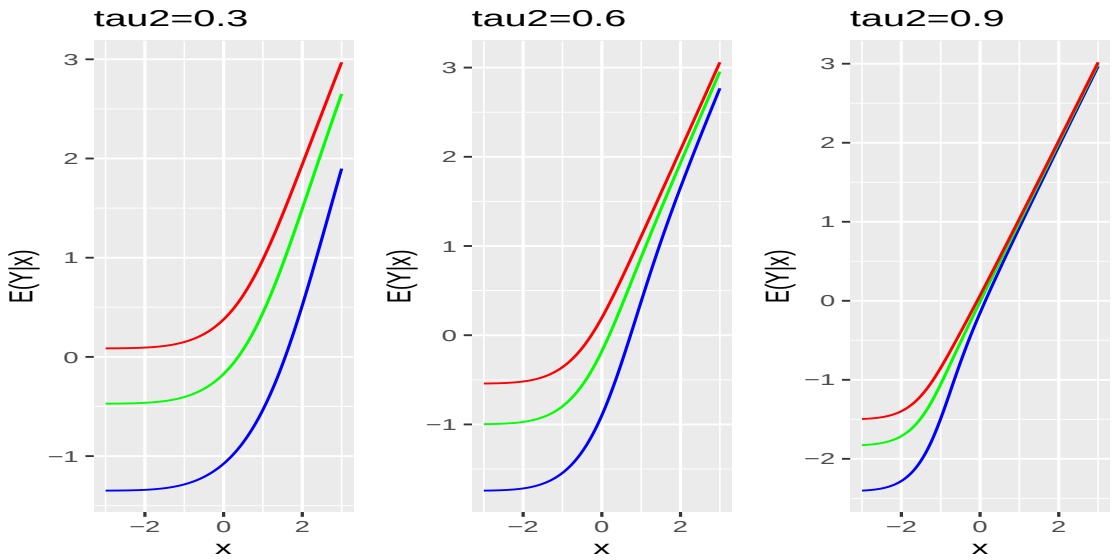


FIGURE B.3 – Courbes de prédiction de $\mathbb{E}(Y|x)$ avec une copule de Joe en fonction de l'importance des résidus $\overline{w}_n = -0.84$ (courbe bleue), $\overline{w}_n = 0$ (courbe verte) et $\overline{w}_n = 0.84$ (courbe rouge) avec taille $n = 20$.

B.5 Calcul de l'inverse de la matrice de corrélation échangeable

Le but de cette section est de donner des indications sur le calcul de l'inverse de la matrice de variance-covariance d'un vecteur aléatoire échangeable et utilisé dans le corollaire 2.1. On cherche l'inverse de la matrice de taille $nd \times nd$ où n est le nombre d'individus et d est le nombre de mesures sur chacun. Soient Σ_w et Σ_b deux matrices de covariances $d \times d$. La matrice de variance-covariance de l'équation (2.1) de taille $nd \times nd$ peut se décomposer par

$$V = \begin{pmatrix} \Sigma_w + \Sigma_b & \Sigma_b & \dots & \Sigma_b \\ \Sigma_b & \ddots & \Sigma_w + \Sigma_b & \Sigma_b \\ \Sigma_b & \dots & \Sigma_b & \Sigma_w + \Sigma_b \end{pmatrix} = \begin{pmatrix} \Sigma_w & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \Sigma_w \end{pmatrix} + \begin{pmatrix} \Sigma_b & \dots & \Sigma_b \\ \Sigma_b & \ddots & \Sigma_b \\ \Sigma_b & \dots & \Sigma_b \end{pmatrix}.$$

Son inverse a, pour une certaine matrice A d'ordre $d \times d$, la forme particulière suivante,

$$V^{-1} = \begin{pmatrix} \Sigma_w^{-1} & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \Sigma_w^{-1} \end{pmatrix} + \begin{pmatrix} A & \dots & A \\ A & \ddots & A \\ A & \dots & A \end{pmatrix}. \tag{B.11}$$

En résolvant l'équation $VV^{-1} = I$ on trouve que $A = -(\Sigma_w + n\Sigma_b)^{-1}\Sigma_b\Sigma_w^{-1}$, ce qui donne une forme explicite pour l'inverse de la matrice de covariances $nd \times nd$.

Inversion matricielle particulière

Dans le cas spécifique où nous écrivons la matrice de corrélation V s'écrit de la forme $V = \begin{pmatrix} A & B^T \\ B & C \end{pmatrix}$ une matrice partitionnée en matrices blocs avec A inversible. L'inverse de la matrice A est donnée par $V^{-1} = \begin{pmatrix} X & Y^T \\ Y & Z \end{pmatrix}$, où X, Y et Z sont données par :

$$\begin{aligned} X &= A^{-1} + A^{-1}B(C - B^T A^{-1} B)B^T A^{-1}, \\ Y &= -A^{-1}B(C - B^T A^{-1} B)^{-1}, \\ Z &= (C - B^T A^{-1} B)^{-1} \end{aligned}$$

Ce résultat est un cas particulier de l'inversion de matrice bloc présenté par Lu et Shiou (2002).

Annexe C

Matériel supplémentaire associé au chapitre 3

Dans cette annexe, nous démontrons la propriété 3.1 et nous utilisons cette propriété pour calculer la variance asymptotique des estimateurs proposés pour le modèle de 2-copule échangeable dans un exemple précis.

C.1 Démonstration de la propriété 3.1 sur les composantes de la fonction score de la méthode IFM généralisée

On rappelle les cinq éléments de la fonction score sur l'estimation des paramètres de la 2-copule échangeable par la méthode IFM généralisée par les équations (3.13), (3.14) et (3.16). Nous démontrons les résultats de la propriété 3.1. Pour les besoins de simplicité, nous posons $W_i = C_{2|1}\{G_0(Y_i;\beta)|F_0(X_i;\alpha)\}$, $w_i = C_{2|1}\{G_0(y_i;\beta)|F_0(x_i;\alpha)\}$ et pour une fonction $f(.;\theta)$ dépendant principalement d'un vecteur de paramètres θ, nous utilisons la notation $\partial f(\theta)/\partial\theta^T = \dot{f}$. Nous considérons dans toute la suite, pour une question de simplicité, le cas particulier $n = 2$.

• Pour le calcul de la covariance entre $\psi_1^{(n)}$ et $\psi_3^{(n)}$, on a

$$\begin{aligned}
\mathrm{cov}(\psi_1^{(n)}, \psi_3^{(n)}) &= \mathrm{cov}\left[\sum_{i=1}^{n} \frac{\partial \log\{f_0(X_i;\alpha)\}}{\partial \alpha^T}, \frac{\partial \log\left\{c_{1,n}^{(1)}\{F_0(X_1;\alpha), F_0(X_2;\alpha);\delta_1\}\right\}}{\partial \delta_1^T}\right] \\
&= \sum_{i=1}^{n} \mathrm{cov}\left[\frac{\partial \log\{f_0(X_i;\alpha)\}}{\partial \alpha^T}, \frac{\partial \log\left\{c_{1,n}^{(1)}\{F_0(X_1;\alpha), F_0(X_2;\alpha);\delta_1\}\right\}}{\partial \delta_1^T}\right] \\
&= n\mathbb{E}\left\{\frac{\dot{f}_0(X_1;\alpha)}{f_0(X_1;\alpha)} \times \frac{\dot{c}_{1,n}^{(1)}\{F_0(X_1;\alpha), F_0(X_2;\alpha);\delta_1\}}{c_{1,n}^{(1)}\{F_0(X_1;\alpha), F_0(X_2;\alpha)\}}\right\} = n\mathbb{E}\left\{\frac{\dot{f}_0}{f_0} \times \frac{\dot{c}_{1,n}^{(1)}}{c_{1,n}^{(1)}}\right\},
\end{aligned}$$

or nous calculons

$$\begin{aligned}
\mathbb{E}\left\{\frac{\dot{f}_0}{f_0} \times \frac{\dot{c}_{1,n}^{(1)}}{c_{1,n}^{(1)}}\right\} &= \int_{\mathbb{R}^2} \frac{\dot{f}_0(x_1;\alpha)}{f_0(x_1;\alpha)} \frac{\dot{c}_{1,n}^{(1)}\{F_0(x_1;\alpha), F_0(x_2;\alpha);\delta_1\}}{c_{1,n}^{(1)}\{F_0(x_1;\alpha), F_0(x_2;\alpha);\delta_1\}} f_0(x_1;\alpha) \times f_0(x_2;\alpha) \\
&\times \; c_{1,n}^{(1)}\{F_0(x_1;\alpha), F_0(x_2;\alpha);\delta_1\}\, dx_1 dx_2 \\
&= \int_{\mathbb{R}} \dot{f}_0(x_1;\alpha)\left\{\int_{\mathbb{R}} f_0(x_2;\alpha)\dot{c}_{1,n}^{(1)}\{F_0(x_1;\alpha), F_0(x_2;\alpha);\delta_1\}\, dx_2\right\} dx_1,
\end{aligned}$$

or nous avons

$$\int_{\mathbb{R}} f_0(x_2;\alpha)\dot{c}_{1,n}^{(1)}\{F_0(x_1;\alpha), F_0(x_2;\alpha);\delta_1\}\, dx_2 = \frac{\partial}{\partial \delta_1^T}\left[\int_0^1 c_{1,n}^{(1)}\{F_0(x_1;\alpha), z;\delta_1\}\, dz\right] = 0,$$

donc nous avons $\mathrm{cov}\left(\psi_1^{(n)}, \psi_3^{(n)}\right) = 0$.

• Pour la covariance entre $\psi_4^{(n)}$ et $\psi_5^{(n)}$

$$\begin{aligned}
\mathrm{cov}(\psi_4^{(n)}, \psi_5^{(n)}) &= \mathrm{cov}\left[\sum_{j=1}^{n} \frac{\partial \log\left[c^{(2)}\{F_0(X_i;\alpha), G_0(Y_i;\beta);\delta_2\}\right]}{\partial \delta_2^T}, \frac{\partial \log\left\{c_{1,n}^{(3)}(W_1, W_2;\delta_3)\right\}}{\partial \delta_3^T}\right] \\
&= n\mathbb{E}\left[\frac{\partial \log\left[c^{(2)}\{F_0(X_1;\alpha), G_0(Y_1;\beta);\delta_2\}\right]}{\partial \delta_2^T}, \frac{\partial \log\left\{c_{1,n}^{(3)}(W_1, W_2;\delta_3)\right\}}{\partial \delta_3^T}\right].
\end{aligned}$$

Nous avons aussi

$$\begin{aligned}
\mathbb{E}\left\{\frac{\dot{c}^{(2)}}{c^{(2)}}\times\frac{\dot{c}^{(3)}_{1,n}}{c^{(3)}_{1,n}}\right\} &= \int_{\mathbb{R}^4}\frac{\dot{c}^{(2)}\left\{F_0(x_1;\alpha),G_0(y_1;\beta);\delta_2\right\}}{c^{(2)}\left\{F_0(x_1;\alpha),G_0(y_1;\beta);\delta_2\right\}}\\
&\times\ f_0(x_1;\alpha)g_0(y_1;\beta)f_0(x_2;\alpha)g_0(y_2;\beta)c^{(1)}_{1,n}\left\{F_0(x_1;\alpha),F_0(x_2;\beta);\delta_1\right\}\\
&\times\ \frac{\dot{c}^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},C_{2|1}\left\{G_0(y_2;\beta)|F_0(x_2;\alpha)\right\};\delta_3\right]}{c^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},C_{2|1}\left\{G_0(y_2;\beta)|F_0(x_2;\alpha)\right\};\delta_3\right]}\\
&\times\ c^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},C_{2|1}\left\{G_0(y_2;\beta)|F_0(x_2;\alpha)\right\};\delta_3\right]\\
&\times\ c^{(2)}\left\{F_0(x_1;\alpha),G_0(y_1;\beta);\delta_2\right\}\times c^{(2)}\left\{F_0(x_2;\alpha),G_0(y_2;\beta);\delta_2\right\}\\
&\times\ dy_1dy_2dx_1dx_2\\
&= \int_{\mathbb{R}^4}\dot{c}^{(2)}\left\{F_0(x_1;\alpha),G_0(y_1;\beta);\delta_2\right\}\\
&\times\ c^{(2)}\left\{F_0(x_2;\alpha),G_0(y_2;\beta);\delta_2\right\}\times c^{(1)}_{1,n}\left\{F_0(x_1;\alpha),F_0(x_2;\beta);\delta_1\right\}\\
&\times\ \dot{c}^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},C_{2|1}\left\{G_0(y_2;\beta)|F_0(x_2;\alpha)\right\};\delta_3\right]\\
&\times\ f_0(x_1;\alpha)g_0(y_1;\beta)\times f_0(x_2;\alpha)g_0(y_2;\beta)dy_1dy_2dx_1dx_2\\
&= \int_{\mathbb{R}^3}f_0(x_1;\alpha)f_0(x_2;\alpha)c^{(1)}_{1,n}\left\{F_0(x_1;\alpha),F_0(x_2;\beta);\delta_1\right\}\\
&\times\ \left(\int_{\mathbb{R}}\dot{c}^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},z;\delta_3\right]dz\right)\\
&\times\ g_0(y_1;\beta)\dot{c}^{(2)}\left\{F_0(x_1;\alpha),G_0(y_1;\beta);\delta_2\right\}dy_1dx_1dx_2,
\end{aligned}$$

et

$$\int_{\mathbb{R}}\dot{c}^{(3)}_{1,n}\left[C_{2|1}\left\{G_0(y_1;\beta)|F_0(x_1;\alpha)\right\},z;\delta_3\right]dz=0,$$

donc $\mathbb{E}\left\{\frac{\dot{c}^{(2)}}{c^{(2)}}\times\frac{\dot{c}^{(3)}_n}{c^{(3)}_n}\right\}=0$ et nous concluons que $\operatorname{cov}\left(\psi^{(n)}_4,\psi^{(n)}_5\right)=0$.

En répétant les calculs précédents faits pour les autres covariances, nous obtenons les résultats de la propriété 3.1.

C.2 Comparaison des méthodes IFM généralisée et méthode de maximisation globale sur un cas particulier du modèle de 2-copule échangeable

Dans la section 3.4.2, nous calculons les matrices de variance-covariance asymptotiques associées à (δ_2,δ_3) provenant des méthodes de maximum de la vraisemblance globale et d'IFM généralisée. Nous présentons ici la procédure d'obtention de ces résultats.

Justification des résultats des matrices des équations (3.34) et (3.35)

Nous calculons la variance asymptotique associée au vecteur de paramètres (δ_2, δ_3). Nous rappelons que $A = \mathbb{E}\left\{\frac{\partial \psi_5^{(n)}}{\partial \delta_2}\right\}$ et $B = \mathbb{E}\left\{\frac{\partial^2 \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2^2}\right\}$. Nous réduisons A par

$$A = \mathbb{E}\left[\frac{1}{c_{1,n}^{(3)}(\delta_2)} \cdot \frac{\partial \dot{c}_{1,n}^{(3)}(\delta_2)}{\partial \delta_2} - \frac{\dot{c}_{1,n}^{(3)}(\delta_2)}{c_{1,n}^{(3)}(\delta_2)} \cdot \frac{\partial \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2}\right] = -\mathbb{E}\left\{\psi_5^{(n)} \cdot a_3^{(n)}\right\}, \tag{C.1}$$

car $\mathbb{E}\left\{\frac{1}{c_{1,n}^{(3)}(\delta_2)} \cdot \frac{\partial \dot{c}_{1,n}^{(3)}(\delta_2)}{\partial \delta_2}\right\} = 0$ où $a_3^{(n)} = \frac{\partial \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2}$.

De même, B donne

$$B \;=\; -\mathbb{E}\left\{\frac{\partial \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2} \cdot \frac{\partial \log\left\{c_{1,n}^{(3)}(\delta_2)\right\}}{\partial \delta_2}\right\} = -\mathbb{E}\left\{a_3^{(n)2}\right\}. \tag{C.2}$$

• **Calcul de la matrice de variance-covariance asymptotique associée** (δ_2, δ_3)
On rappelle que l'expression de la matrice de Godambe utilisée ici provient de l'équation (3.30).

L'espérance du gradient du vecteur score $\left(\psi_4^{(n)}, \psi_5^{(n)}\right)$ est

$$\mathbb{E}\left\{\nabla_\theta(\psi_4^{(n)}, \psi_5^{(n)})\right\} = \begin{pmatrix} \mathbb{E}\left\{\frac{\partial \psi_4^{(n)}}{\partial \delta_2}\right\} & 0 \\ \mathbb{E}\left\{\frac{\partial \psi_5^{(n)}}{\partial \delta_2}\right\} & \left\{\frac{\partial \psi_5^{(n)}}{\partial \delta_3}\right\} \end{pmatrix}.$$

Or nous avons

$$\mathbb{E}\left\{\frac{\partial \psi_4^{(n)}}{\partial \delta_2}\right\} = -\mathbb{E}\left(\psi_4^{(n)2}\right), \mathbb{E}\left\{\frac{\partial \psi_5^{(n)}}{\partial \delta_3}\right\} = -\mathbb{E}\left(\psi_5^{(n)2}\right).$$

Pour $\mathbb{E}(\psi_4^{(n)2}) \neq 0$ et $\mathbb{E}(\psi_5^{(n)2}) \neq 0$, nous avons aussi

$$D = \begin{pmatrix} \mathbb{E}(\psi_4^{(n)2}) & 0 \\ -A & \mathbb{E}(\psi_5^{(n)2}) \end{pmatrix}, \;\; D^{-1} = \frac{1}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})} \begin{pmatrix} \mathbb{E}(\psi_5^{(n)2}) & 0 \\ A & \mathbb{E}(\psi_4^{(n)2}) \end{pmatrix}. \tag{C.3}$$

De plus, en partant de l'équation (3.30), on a :

$$\mathbb{E}\left\{(\psi_4^{(n)}, \psi_5^{(n)})^T(\psi_4^{(n)}, \psi_5^{(n)})\right\} = \begin{pmatrix} \mathbb{E}(\psi_4^{(n)2}) & \mathbb{E}(\psi_4^{(n)}\psi_5^{(n)}) \\ \mathbb{E}(\psi_4^{(n)}\psi_5^{(n)}) & \mathbb{E}(\psi_5^{(n)2}) \end{pmatrix}.$$

Or $\mathbb{E}(\psi_4^{(n)} \cdot \psi_5^{(n)}) = 0$ (confère propriété (3.1)) alors V_0 de l'équation (3.31) donne

$$V_0 = \begin{pmatrix} \mathbb{E}(\psi_4^{(n)2}) & 0 \\ 0 & \mathbb{E}(\psi_5^{(n)2}) \end{pmatrix}. \tag{C.4}$$

Nous calculons enfin la matrice de Godambe 2×2 de l'équation (3.30) à partir des équations (C.3) et (C.4) par

$$\begin{aligned}\mathcal{G}(\delta_2,\delta_3) &= \frac{1}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})}\begin{pmatrix}\mathbb{E}(\psi_5^{(n)2}) & 0\\ A & \mathbb{E}(\psi_4^{(n)2})\end{pmatrix}\begin{pmatrix}\mathbb{E}(\psi_4^{(n)2}) & 0\\ 0 & \mathbb{E}(\psi_5^{(n)2})\end{pmatrix}\\ &\times \frac{1}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})}\begin{pmatrix}\mathbb{E}(\psi_5^{(n)2}) & A\\ 0 & \mathbb{E}(\psi_4^{(n)2})\end{pmatrix},\end{aligned}$$

donc

$$\begin{aligned}\mathcal{G}(\delta_2,\delta_3) &= \begin{pmatrix}\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})^2 & A\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})\\ A\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2}) & A^2\mathbb{E}(\psi_4^{(n)2})+\mathbb{E}(\psi_4^{(n)2})^2\mathbb{E}(\psi_5^{(n)2})\end{pmatrix}\\ &\times \frac{1}{\left\{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})\right\}^2}\end{aligned}$$

qui nous conduit logiquement au résultat de l'équation (3.34).

• **Calcul de la matrice de variance-covariance asymptotique associée à (δ_2,δ_3) de la méthode du maximum de vraisemblance.**

Pour la méthode de maximisation de la vraisemblance globale, nous avons : $\left(\Psi_4^{(n)},\Psi_5^{(n)}\right) = \left(\psi_4^{(n)}+a_3^{(n)},\psi_5^{(n)}\right)$ où $a_3^{(n)}$ provient de l'équation (3.26).

L'espérance du gradient de la fonction score dans ce cas s'écrit

$$\mathbb{E}\left\{\nabla_\theta(\Psi_4^{(n)},\Psi_5^{(n)})\right\} = \begin{pmatrix}\mathbb{E}\left\{\frac{\partial\psi_4^{(n)}}{\partial\delta_2}\right\}+\mathbb{E}\left[\frac{\partial^2\log\left\{c_n^{(3)}(\delta_2)\right\}}{\partial\delta_2^2}\right] & \mathbb{E}\left\{\frac{\partial\psi_5^{(n)}}{\partial\delta_2}\right\}\\ \mathbb{E}\left\{\frac{\partial\psi_5^{(n)}}{\partial\delta_2}\right\} & \mathbb{E}\left\{\frac{\partial\psi_5^{(n)}}{\partial\delta_3}\right\}\end{pmatrix}.$$

La matrice d'information de Fisher est

$$\mathcal{I}(\delta_2,\delta_3) = \begin{pmatrix}\mathbb{E}\left(\psi_4^{(n)2}\right)-B & -A\\ -A & \mathbb{E}\left(\psi_5^{(n)2}\right)\end{pmatrix}.$$

En supposant que $\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right)-B\mathbb{E}\left(\psi_5^{(n)2}\right)-A^2\neq 0$, en inversant cette matrice, nous obtenons le résultat attendu de l'équation (3.35).

Nous procédons donc à la comparaison des deux matrices variance-covariance asymptotiques. Nous posons que la différence des deux matrices de variances-covariances asymptotiques s'écrit

$$\mathcal{G}(\delta_2,\delta_3)-\mathcal{I}^{-1}(\delta_2,\delta_3) = \begin{pmatrix}d_{11} & d_{12}\\ d_{21} & d_{22}\end{pmatrix}. \tag{C.5}$$

Les 4 éléments de la matrice de l'équation (C.5) sont donc

$$\begin{aligned} d_{11} &= \frac{1}{\mathbb{E}(\psi_4^{(n)2})} - \frac{\mathbb{E}\left(\psi_5^{(n)2}\right)}{\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2} \\ &= \frac{\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2 - \mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right)}{\mathbb{E}(\psi_4^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} \\ &= \frac{-B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2}{\mathbb{E}(\psi_4^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]}. \end{aligned}$$

$$\begin{aligned} d_{12} &= \frac{A}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})} - \frac{A}{\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2} \\ &= \frac{A\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - AB\mathbb{E}\left(\psi_5^{(n)2}\right) - A^3 - A\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right)}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} \\ &= \frac{-A\left\{B\mathbb{E}\left(\psi_5^{(n)2}\right) + A^2\right\}}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} = d_{21}. \end{aligned}$$

$$\begin{aligned} d_{22} &= \frac{\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right)^2 + A^2\mathbb{E}\left(\psi_5^{(n)2}\right)}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})^3} - \frac{\mathbb{E}\left(\psi_4^{(n)2}\right) - B}{\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2} \\ &= \frac{-A^2B\mathbb{E}\left(\psi_5^{(n)2}\right)^2 - A^4\mathbb{E}\left(\psi_5^{(n)2}\right)}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})^3\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} \\ &= \frac{-A^2\left\{B\mathbb{E}\left(\psi_5^{(n)2}\right) + A^2\right\}}{\mathbb{E}(\psi_4^{(n)2})\mathbb{E}(\psi_5^{(n)2})^2\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]}. \end{aligned}$$

En conséquence, nous avons

$$\begin{aligned} \mathcal{I}^{-1}(\delta_2,\delta_3) - \mathcal{G}(\delta_2,\delta_3) &= \frac{-\left\{B\mathbb{E}\left(\psi_5^{(n)2}\right) + A^2\right\}}{\mathbb{E}(\psi_4^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} \\ &\quad \times \begin{pmatrix} 1 & \frac{A}{\mathbb{E}\left(\psi_5^{(n)2}\right)} \\ \frac{A}{\mathbb{E}\left(\psi_5^{(n)2}\right)} & \frac{A^2}{\mathbb{E}\left(\psi_5^{(n)2}\right)^2} \end{pmatrix}. \end{aligned}$$

Pour tout vecteur $u = (x, y)^T \in \mathbb{R}^2$, nous avons

$$\begin{aligned} u^T \left\{\mathcal{I}^{-1}(\delta_2, \delta_3) - \mathcal{G}(\delta_2, \delta_3)\right\} u &= \frac{-\left\{B\mathbb{E}\left(\psi_5^{(n)2}\right) + A^2\right\}}{\mathbb{E}(\psi_4^{(n)2})\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right]} \\ &\times \left(x + \frac{Ay}{\mathbb{E}(\psi_5^{(n)2})}\right)^2. \end{aligned}$$

Or $\mathbb{E}(\psi_4^2)\left[\mathbb{E}\left(\psi_4^{(n)2}\right)\mathbb{E}\left(\psi_5^{(n)2}\right) - B\mathbb{E}\left(\psi_5^{(n)2}\right) - A^2\right] > 0$ et $B\mathbb{E}\left(\psi_5^{(n)2}\right) + A^2 > 0$, nous avons donc $u^T \left\{\mathcal{G}(\delta_2, \delta_3) - \mathcal{I}^{-1}(\delta_2, \delta_3)\right\} u > 0$.

C.3 Calculs pour l'évaluation de la matrice de variance-covariance asymptotique des méthodes IFM et MV dans le cas où toutes les copules sont normales

Cette annexe présente les résultats utilisés pour calculer les matrices obtenues à la section 3.4.3.

En utilisant les notations des équations (3.36) et (3.37), nous avons les résultats suivants.

La variance de $\psi_4^{(n)}$ est

$$\mathbb{V}\left\{\psi_4^{(n)}\right\} = \frac{(\oplus)}{(1-\rho_2^2)^2}. \tag{C.6}$$

La variance de $\psi_5^{(n)}$ est

$$\mathbb{V}\left\{\psi_5^{(n)}\right\} = \frac{n(n-1)(\sim)}{2(1-\rho_3)^2(\diamond)^2}. \tag{C.7}$$

Le calcul de A, de l'équation (C.1) donne

$$A = \frac{-n(n-1)\rho_2\rho_3}{(1-\rho_2^2)(1-\rho_3)(\diamond)}. \tag{C.8}$$

Le calcul de B, de l'équation (C.2) est

$$B = -\frac{[**]}{(1-\rho_3)(1-\rho_2^2)^2(\diamond)}. \tag{C.9}$$

Quelques rappels importants

Nous notons $\Sigma(\rho, n)$, la matrice carrée de dimension n dont les termes de la diagonale sont égales à 1 et les termes hors de la diagonale sont ρ. On rappelle que J_n est la matrice carrée d'ordre n où tous les éléments sont 1 et I_n est la matrice unité d'ordre n et $\Sigma(\rho, n) = (1-\rho)I_n + \rho J_n$. La dérivée matricielle et utilisant la propriété, nous avons

$$\frac{\partial \Sigma(\rho, n)}{\partial \rho} = J_n - I_n, \quad \frac{\partial \Sigma^{-1}(\rho, n)}{\partial \rho} = -\Sigma^{-1}(\rho, n)(J_n - I_n)\Sigma^{-1}(\rho, n). \tag{C.10}$$

En utilisant l'équation (B.11), où on prend $\Sigma_w = 1 - \rho$ et $\Sigma_b = \rho$, l'inverse de la matrice $\Sigma(\rho, n)$ pour $\rho \neq 1$ et $\rho \neq \frac{-1}{n-1}$ est

$$\Sigma^{-1}(\rho, n) = \frac{1}{1-\rho}\left\{I_n - \frac{\rho}{1+(n-1)\rho} J_n\right\}.$$

Par calcul simple, nous avons aussi

$$\Sigma^{-1}(\rho, n)(J_n - I_n)\Sigma^{-1}(\rho, n) = \frac{1}{(1-\rho)^2}\left\{\frac{1+(n-1)\rho^2}{\{1+(n-1)\rho\}^2} J_n - I_n\right\}. \tag{C.11}$$

Soient X_1, X_2, X_3 et X_4 quatre vecteurs aléatoires centrés de lois normales multivariées telles que les matrices de corrélation sont définies par $\mathbb{E}\left(X_i X_j^T\right) = \Sigma_{ij}$. Notons M_1 et M_2 deux matrices telles que $X_1^T M_1 X_2$ et $X_3^T M_2 X_4$ soient calculables. Nous avons les résultats suivants :

$$\text{Cov}\left(X_1^T M_1 X_2, X_3^T M_2 X_4\right) = Tr(M_1^T \Sigma_{13} M_2 \Sigma_{42}) + Tr(M_1^T \Sigma_{14} M_2^T \Sigma_{32}), \tag{C.12}$$

où Tr est l'application trace. Ce calcul d'espérances des formes bilinéaires se base sur des formules de Ghazal (2000). En particulier, nous avons

$$\mathbb{E}\left(X_1^T M_1 X_1\right) = Tr(M_1 \Sigma_{11}), \ \ \mathbb{V}\left(X_1^T M_1 X_1\right) = 2Tr\left\{(M_1\Sigma_{11})^2\right\}. \tag{C.13}$$

Justification de chaque résultat des équations (C.6), (C.7), (C.8) et (C.9)

• Présentation du contexte de calcul dans le cas normal

Les log-vraisemblances $\mathcal{L}_2$ et $\mathcal{L}_3$ des équations respectives (3.7) et (3.10), dans le cas particulier où toutes les copules sont normales et les marginales, des lois normales centrées réduites sont

$$\mathcal{L}_2 = -\frac{n}{2}\log(1-\rho_2^2) - \sum_{i=1}^{n}\left\{\frac{x_i^2 - 2\rho_2 x_i y_i + y_i^2}{2(1-\rho_2^2)}\right\},$$

et

$$\mathcal{L}_3 = -\frac{1}{2}\log\{1+(n-1)\rho_3\} - \frac{n-1}{2}\log(1-\rho_3) - \frac{1}{2}\underline{t}_n^T\left\{\Sigma^{-1}(\rho_3, n) - I_n\right\}\underline{t}_n,$$

où $\underline{t}_n = (t_1, \ldots, t_n)^T$ et $t_i = \Phi^{-1}\left[C_{2|1}\left\{\Phi(y_i)|\Phi(x_i)\right\}\right] = (y_i - \rho_2 x_i)/\sqrt{1-\rho_2^2}$. Le vecteur $\underline{t}_n$ suit une loi normale de moyenne 0 et de matrice de variance-covariance $\Sigma(\rho_3, n)$. Nous avons aussi

$$\mathbb{E}(\underline{x}_n \underline{t}_n^T) = \mathbf{0}_{n\times n}, \quad \text{et} \quad \mathbb{E}(\underline{x}_n \underline{x}_n^T) = \Sigma(\rho_1, n)\cdot \tag{C.14}$$

La fonction $\psi_4^{(n)} = \partial\mathcal{L}_2/\partial\rho_2$ est

$$\psi_4^{(n)} = \frac{n\rho_2}{1-\rho_2^2} + \sum_{i=1}^{n}\left\{\frac{(1+\rho_2^2)x_i y_i - \rho_2(x_i^2+y_i^2)}{(1-\rho_2^2)^2}\right\},$$

soit

$$\psi_4^{(n)} = \frac{n\rho_2}{1-\rho_2^2} + \sum_{i=1}^{n}\left\{\frac{\sqrt{1-\rho_2^2}x_i t_i - \rho_2 t_i^2}{(1-\rho_2^2)}\right\}. \tag{C.15}$$

Pour la fonction $\psi_5^{(n)} = \partial\mathcal{L}_3/\partial\rho_3$ et en utilisant les équations (C.10), nous avons

$$\psi_5^{(n)} = \frac{1-n}{2\left\{1+(n-1)\rho_3\right\}} + \frac{n-1}{2(1-\rho_3)} + \frac{1}{2}\underline{t}_n^T\Sigma^{-1}(\rho_3, n)\left\{\frac{\partial\Sigma(\rho_3, n)}{\partial\rho_3}\right\}\Sigma^{-1}(\rho_3, n)\underline{t}_n.$$

En posant $A_0 = \Sigma^{-1}(\rho_3, n)(J_n - I_n)\Sigma^{-1}(\rho_3, n)$, nous avons

$$\psi_5^{(n)} = \frac{1-n}{2\left\{1+(n-1)\rho_3\right\}} + \frac{n-1}{2(1-\rho_3)} + \frac{1}{2}\underline{t}_n^T A_0 \underline{t}_n. \tag{C.16}$$

Pour la fonction $a_3^{(n)} = \partial\mathcal{L}_3/\partial\rho_2$, on a

$$a_3^{(n)} = \frac{1}{1-\rho_2^2}\underline{t}_n^T\left\{\Sigma^{-1}(\rho_3, n) - I_n\right\}\underline{z}_n,$$

avec $\underline{z}_n = (z_1, \ldots, z_n)^T$ et $z_i = (x_i - \rho_2 y_i)/\sqrt{1-\rho_2^2}$. Nous réécrivons $\underline{z}_n$ par $\underline{z}_n = \sqrt{1-\rho_2^2}\underline{x}_n - \rho_2\underline{t}_n$ et en posant $B_0 = \Sigma^{-1}(\rho_3, n) - I_n$, nous avons

$$a_3^{(n)} = \frac{1}{\sqrt{1-\rho_2^2}}\underline{t}_n^T B_0 \underline{x}_n - \frac{\rho_2}{1-\rho_2^2}\underline{t}_n^T B_0 \underline{t}_n. \tag{C.17}$$

• **Calcul de la variance de $\psi_4^{(n)}$**

Pour la variance de $\psi_4^{(n)}$, nous avons

$$\begin{aligned}
\mathbb{V}\left\{\psi_4^{(n)}\right\} &= \mathbb{V}\left[\frac{n\rho_2}{1-\rho_2^2} + \sum_{i=1}^{n}\left\{\frac{\sqrt{1-\rho_2^2}x_i t_i - \rho_2 t_i^2}{(1-\rho_2^2)}\right\}\right] \\
&= \frac{n}{(1-\rho_2^2)^2}\mathbb{V}\left(\sqrt{1-\rho_2^2}x_1 t_1 - \rho_2 t_1^2\right) \\
&+ \frac{n(n-1)}{(1-\rho_2^2)^2}\mathrm{cov}\left(\sqrt{1-\rho_2^2}x_1 t_1 - \rho_2 t_1^2, \sqrt{1-\rho_2^2}x_2 t_2 - \rho_2 t_2^2\right)\cdot
\end{aligned}$$

Par ailleurs, la variance de $\mathcal{X}_i = \sqrt{1-\rho_2^2}x_i t_i - \rho_2 t_i^2$ donne

$$\mathbb{V}\left(\sqrt{1-\rho_2^2}x_i t_i - \rho_2 t_i^2\right) = \mathbb{V}\left\{\mathbb{E}(\mathcal{X}_i|x_i)\right\} + \mathbb{E}\left\{\mathbb{V}(\mathcal{X}_i|x_i)\right\} = (1+\rho_2^2),$$

et pour $i \neq j$, nous avons

$$\begin{aligned}\text{cov}\left(\mathcal{X}_i, \mathcal{X}_j\right) &= (1-\rho_2^2)\mathbb{E}(x_i x_j t_i t_j) + \rho_2^2\left\{\mathbb{E}(t_i^2 t_j^2)\right\} \\ &= (1-\rho_2^2)\rho_1\rho_3 + 2\rho_2^2\rho_3^2.\end{aligned}$$

En conclusion, nous avons le résultat

$$\mathbb{V}\left\{\psi_4^{(n)}\right\} = \frac{n(1+\rho_2^2) + n(n-1)\left\{(1-\rho_2^2)\rho_1\rho_3 + 2\rho_2^2\rho_3^2\right\}}{(1-\rho_2^2)^2}.$$

• **Calcul de la variance de $\psi_5^{(n)}$.**

Le calcul de la variance est :

$$\mathbb{V}\left\{\psi_5^{(n)}\right\} = \mathbb{V}\left(\frac{1}{2}\underline{t}_n^T A_0 \underline{t}_n\right) = \frac{1}{2}Tr\left\{\left(A_0\Sigma(\rho_3, n)\right)^2\right\},$$

en appliquant l'équation (C.13). Nous rappelons à partir de l'équation (C.11) que

$$A_0\Sigma(\rho_3, n) = \frac{1}{(1-\rho_3)^2}\left\{\frac{1+(n-1)\rho_3^2}{\left\{1+(n-1)\rho_3\right\}^2}J_n\Sigma(\rho_3, n) - \Sigma(\rho_3, n)\right\}. \tag{C.18}$$

En remplaçant les matrices et en appliquant la fonction trace, nous avons

$$\mathbb{V}\left\{\psi_5^{(n)}\right\} = \frac{n(n-1)\left\{1+(n-1)\rho_3^2\right\}}{2(1-\rho_3)^2\left\{1+(n-1)\rho_3\right\}^2}. \tag{C.19}$$

• **Calcul de A et B des équations (C.1) et (C.2)**

Pour le calcul de A, nous avons :

$$\begin{aligned}A = -\mathbb{E}\left\{\psi_5^{(n)} a_3^{(n)}\right\} &= -\text{cov}\left\{\psi_5^{(n)}, a_3^{(n)}\right\} \\ &= -\text{cov}\left\{\frac{1}{2}\underline{t}_n^T A_0 \underline{t}_n, \frac{1}{\sqrt{1-\rho_2^2}}\underline{t}_n^T B_0 \underline{x}_n - \frac{\rho_2}{1-\rho_2^2}\underline{t}_n^T B_0 \underline{t}_n^T\right\} \\ &= -\frac{\text{cov}\left(\underline{t}_n^T A_0 \underline{t}_n, \underline{t}_n^T B_0 \underline{x}_n\right)}{2\sqrt{1-\rho_2^2}} + \frac{\rho_2 \text{cov}\left(\underline{t}_n^T A_0 \underline{t}_n, \underline{t}_n^T B_0 \underline{t}_n\right)}{2(1-\rho_2^2)},\end{aligned}$$

or en utilisant le résultat de l'équation (C.12), nous avons

$$\text{cov}\left(\underline{t}_n^T A_0 \underline{t}_n, \underline{t}_n^T B_0 \underline{x}_n\right) = 0, \quad \text{cov}\left(\underline{t}_n^T A_0 \underline{t}_n, \underline{t}_n^T B_0 \underline{t}_n\right) = 2Tr\left\{A_0\Sigma(\rho_3, n)B_0\Sigma(\rho_3, n)\right\},$$

dans le cas particulier $X_1 = X_2 = X_3 = X_4 = t_n$ et $M_1 = A_0$, $M_2 = B_0$. Donc, nous avons

$$A = \frac{\rho_2 Tr\left\{A_0\Sigma(\rho_3,n)B_0\Sigma(\rho_3,n)\right\}}{1-\rho_2^2}. \tag{C.20}$$

Nous rappelons que $B_0\Sigma(\rho_3,n) = I - \Sigma(\rho_3,n)$ et

$$\begin{aligned} A_0\Sigma(\rho_3,n)B_0\Sigma(\rho_3,n) &= \Sigma^{-1}(\rho_3,n)(J_n - I_n)\left\{I_n - \Sigma(\rho_3,n)\right\} \\ &= \Sigma^{-1}(\rho_3,n)J_n - \Sigma^{-1}(\rho_3,n)J_n\Sigma(\rho_3,n) - \Sigma^{-1}(\rho_3,n) + I_n, \end{aligned}$$

or nous avons

$$Tr\left\{\Sigma^{-1}(\rho_3,n)J_n\Sigma(\rho_3,n)\right\} = n,\ \ Tr\left\{\Sigma^{-1}(\rho_3,n)\right\} = \frac{n}{1+(n-1)\rho_3} + \frac{n(n-1)\rho_3}{(1-\rho_3)\left\{1+(n-1)\rho_3\right\}},$$

$$Tr\left\{\Sigma^{-1}(\rho_3,n)J_n\right\} = \frac{n}{1+(n-1)\rho_3}.$$

Par conséquent, le résultat attendu est

$$A = \frac{-n(n-1)\rho_2\rho_3}{(1-\rho_2^2)(1-\rho_3)\left\{1+(n-1)\rho_3\right\}}. \tag{C.21}$$

Pour le calcul de B, nous avons

$$\begin{aligned} B = -\mathbb{E}\left\{a_3^{(n)}\cdot a_3^{(n)}\right\} &= -\mathbb{E}\left\{\left(\frac{1}{\sqrt{1-\rho_2^2}}\underline{t}_n^T B_0\underline{x}_n - \frac{\rho_2}{1-\rho_2^2}\underline{t}_n^T B_0\underline{t}_n\right)^2\right\} \\ &= \frac{-(1-\rho_2^2)\mathbb{V}\left(\underline{t}_n^T B_0\underline{x}_n\right) - \rho_2^2\mathbb{V}\left(\underline{t}_n^T B_0\underline{t}_n\right)}{(1-\rho_2^2)^2}, \end{aligned}$$

or en appliquant l'équation (C.12), pour $X_1 = X_3 = t_n$, $X_2 = X_4 = x_n$, $M_1 = M_2 = B_0$, nous avons

$$\mathbb{V}\left(\underline{t}_n^T B_0\underline{t}_n\right) = 2Tr\left\{\left(B_0\Sigma(\rho_3,n)\right)^2\right\} = 2n(n-1)\rho_3^2,$$

$$\mathbb{V}\left(\underline{t}_n^T B_0\underline{x}_n\right) = \mathrm{Cov}\left(\underline{t}_n^T B_0\underline{x}_n, \underline{t}_n^T B_0\underline{x}_n\right) = 2Tr\left\{B_0\Sigma(\rho_3,n)B_0\Sigma(\rho_1,n)\right\},$$

et par développement $B_0\Sigma(\rho_3,n)B_0\Sigma(\rho_1,n) = \Sigma^{-1}(\rho_3,n)\Sigma(\rho_1,n) - 2\Sigma(\rho_1,n) - \Sigma(\rho_3,n)\Sigma(\rho_1,n)$ et

$$Tr\left\{\Sigma(\rho_3,n)\Sigma(\rho_1,n)\right\} = n\left\{1+(n-1)\rho_1\rho_3\right\},\ \ Tr\left\{J_n\Sigma(\rho_1,n)\right\} = n\left\{1+(n-1)\rho_1\right\},$$

puis nous avons

$$Tr\left\{B_0\Sigma(\rho_3,n)B_0\Sigma(\rho_1,n)\right\} = \frac{1}{1-\rho_3}\left\{n - \frac{n\rho_3\left\{1+(n-1)\rho_1\right\}}{\left\{1+(n-1)\rho_3\right\}}\right\} - n\left\{1-(n-1)\rho_1\rho_3\right\},$$

donc nous obtenons finalement

$$B = -\frac{2n(n-1)\rho_3^2}{(1-\rho_2^2)^2}\left[\frac{\rho_2^2(1-\rho_3)\left\{1+(n-1)\rho_3\right\} + (1-\rho_2^2)\left\{1+(n-2)\rho_1 - (n-1)\rho_1\rho_3\right\}}{(1-\rho_3)\left\{1+(n-1)\rho_3\right\}}\right].$$

C.4 Présentation du programme *R* de l'étude Monte-Carlo du modèle de 2-copule échangeable

```
####################ÉTUDE MONTE-CARLO PAR SIMULATION###################
#
# Objectifs :- 1 :Simuler des données suivant la 2-copule échangeable #
#            - 2 :Estimer les paramètres du modèle                     #
#            - 3 :Répéter plusieurs fois                               #
#######################################################################


##---------Charger les packages-------------------------------#

library(copula)  ## Important pour les fonctions sur les copules
library(VineCopula)   ## Construction des copules bivariées
library(fitdistrplus)  ## Ajustement de la loi marginale

#-------------Principales fonctions utilisées--------------------#

   # BiCopHfunc1 : calcul F(v|u)
   # BiCopHinv1  : calcule v tel que w=F(v|u)
   # BiCopPDF    : densité d'une copule bivariée
   # dCopula     : densité d'une copule
   # BiCopTau2Par: calcul le paramètre de la copule sachant tau
   # rCopula     : simulation des données suivant une copule
   # optim       : optimisation d'une fonction
   # dnorm       : densité d'une loi normale
   # BiCopEst    : estimation des paramètres d'une copule bivariée
   # nlm         : optimisation d'une fonction
   # fitdist     : estimation des paramètres d'une distribution

###----ÉTAPE 1 : SIMULATION DES DONNÉES----------------###

 #---Simul_data fonction qui permet de simuler des données
 #   suivant la 2-copule échangeable en utilisant la section 2.5

Simul_data<-function(mu1,sigma1,mu2,sigma2,tau1,tau2,tau3,familic2,m,n)
```

```
   {

##-----------ENTRÉES-----------------------
# mu1 et sigma1 sont les paramètres de la marginale F0 normale
# mu2 et sigma2 sont les paramètres de la marginale G0 normale
# tau1 : tau de Kendall de la copule c1 (normal paramètre delta1)
# tau2 : tau de Kendall  de la copule c2 (familic2 paramètre delta2)
# familic2 est le nombre représentant la copule c2 : 1=normale
#   3=Clayton 4=Gumbel et 5=Frank, etc.
# tau3 : tau de Kendall de la copule C3(normal paramètre delta3)
# m : le nombre de grappes
# n : la taille des grappes


##-------------SORTIES--------------------------
#matrice de colonnes X,Y et grappes nommée data_sim


##----Calcul des valeurs des paramètres des copules.
delta1<-BiCopTau2Par(family =1, tau = tau1)  #pour la copule C1
delta2<-BiCopTau2Par(family = familic2, tau = tau2) #pour la copule C2
delta3<-BiCopTau2Par(family =1, tau = tau3)  #pour la copule C3

grappes<-rep(1:m,each=n)
data_sim<-matrix(0,nrow=n*m,ncol=3)
data_sim[,3]<-grappes


for(j in 1:m){
  ind<-which(data_sim[,3]==j)
  u<-rCopula(1,normalCopula(delta1,dim=length(ind)))
  w<-rCopula(1,normalCopula(delta3,dim=length(ind)))
  v<-BiCopHinv1(u1=u[1,],u2=w[1,],BiCop(family=familic2, par=delta2))


  data_sim[ind,1]<-qnorm(u[1,],mean=mu1,sd=sigma1)  #Calcul de x
  data_sim[ind,2]<-qnorm(v,mean=mu2,sd=sigma2)      #Calcul de y
```

```
  }
  return(data.frame(data_sim))

}

###----ÉTAPE 2 : Fonction à maximiser de la vraisemblance globale----#
#
#-------------------------------------------------------------------#

 ##--Objectif : La log-vraisemblance globale de l'équation (3.42)

Vrai_simul<-function(data,familic2,para,m){

  ##-----------ENTREE-------------------##
  #--@@data constituée de 3 colonnes (X,Y,grappe)
  #    @colonne 1: le vecteur X
  #    @colonne 2: le vecteur Y
  #    @colonne 3: les grappes
  #--familic2 : famille de la copule c2 prenant des valeurs numériques
  #--para : les paramètres de dimension 7()

  ##-----------SORTIE-------------------##
  #La log-vraisemblance globale équation (3.42) (chapitre 3)


  mu1<-para[1] ; sigma1<-para[2] #paramètres de la marginale normale F0

  mu2<-para[3] ; sigma2<-para[4] #paramètres de la marginale normale G0

  eta1<-para[5] ; delta1<-exp(eta1)/(1+exp(eta1)) #paramètre copule c1

  delta2<-para[6] #paramètre copule c2
  eta3<-para[7] ; delta3<-exp(eta3)/(1+exp(eta3)) #paramètre copule c3
  #delta3<-para[7] ;

  u<-pnorm(data[,1],mean=mu1,sd=sigma1)  #tranformer en uniforme
  v<-pnorm(data[,2],mean=mu2,sd=sigma2)
```

```
  l1<--sum(dnorm(data[,1],mean=mu1,sd=sigma1,log=TRUE)) #densité f0

  l2<--sum(dnorm(data[,2],mean=mu2,sd=sigma2,log=TRUE)) #densité g0

  ##-------partie III-1 de la copule c1-----#
  l01<-0
  for(j in 1:m){
    ind1<-which(data[,3]==j)
    lnC1<- -dCopula(u[ind1],normalCopula(delta1,dim=length(ind1)),log=TRUE)
    l01<-l01+lnC1
       }

  ##-------pour la copule c2----------------#
   l02<--sum(log(BiCopPDF(u1=u, u2=v,family=familic2,par=delta2)))

    ##-------pour la copule c3----------------#

       #---calcul de w--------#
   w<-BiCopHfunc1(u1=u, u2=v,family=familic2, par=delta2)

  l03<-0
  for(j in 1:m){
    ind<-which(data[,3]==j)
    lnC3<- -dCopula(w[ind],normalCopula(delta3,dim=length(ind)),log=TRUE)
    l03<-l03+lnC3
      }

  return(l1+l2+l01+l02+l03)
}


###----ÉTAPE 3--RÉPÉTER B fois les étapes 1 et 2-----------------##
#
#-----------------------------------------------------------------#

 # ÉTAPE 3.1 : Fonction échangeable pour la copule c1 ou c3----#

Norm_Ech<-function(data,eta){
```

```
  ##--------------ENTREES-----------##
  #data de 2 colonnes : - valeurs contenues dans (0,1)
                       #- les classes associées
  #eta : est la réparametrisation de rho

  ##--------------SORTIE------------##
   #la log-vraisemblance de la copule normale échangeable

  rho<-exp(eta)/(1+exp(eta))

  l<-0;
  for(j in 1: max(data[,2])){
    ind<-which(data[,2]==j)
    lnC<--dCopula(data[,1][ind],normalCopula(rho,length(ind)),log=TRUE)
    l<-l+lnC
  }
  return(l)
}

  # ÉTAPE 3.2 : Construction de la fonction de répétition

funct_estimation<-function(B,mu1_0,sigma1_0,mu2_0,sigma2_0,
                           tau1_0,tau2_0,tau3_0,familic2,m,n){

  ##---Cette fonction part des vraisemblances du paramètre pour estimer---#
  #        par deux méthodes différentes IFM et MV à partir d'une data
  #                simuler à l'étape 1
  #
  ##-------------------------------------------------------------------#

  #------------------ENTRÉES----------------------------------------#
  ## @B est le nombre de simulations
  # mu1_0 et sigma1_0 sont les paramètres de la marginale F0 normale
  # mu2_0 et sigma2_0 sont les paramètres de la marginale G0 normale
  # tau1_0 : tau de Kendall de la copule c1 (ici normal paramètre delta1)
  # tau2_0 : tau de Kendall  de la copule c2 (familic2 paramètre delta2)
  # familic2 est le nombre représentant la copule c2 : 1=normale 3=Clayton
  #    4=Gumbel et 5=Frank, etc.
```

```
# tau3_0 : tau de Kendall de la copule c3(ici normal paramètre delta3)
#m : le nombre de grappes
#n : la taille des grappes équilibrée

##-----------------SORTIE-------------------------------------
  #B estimations des 7 paramètres du modèle par MV et IFM

##-----Calcul des paramètres de base--------------#

delta1_0<-BiCopTau2Par(family =1, tau = tau1_0)
delta2_0<-BiCopTau2Par(family = familic2, tau = tau2_0)
delta3_0<-BiCopTau2Par(family =1, tau = tau3_0)

eta1_0<-log(delta1_0/(1-delta1_0))
eta3_0<-log(delta3_0/(1-delta3_0))

data_estim<-matrix(0,nrow=B,ncol=14) ## data des 14 estimations de para
data_estim_error<-matrix(0,nrow=B,ncol=14)  ## erreurs paramètres

theta0<-c(mu1_0,sigma1_0,mu2_0,sigma2_0,eta1_0,delta2_0,eta3_0) ##

for(i in 1:B){
  data_SIM<-Simul_data(mu1_0=mu1_0,sigma1_0=sigma1_0,mu2_0=mu2_0,
                   sigma2_0=sigma2_0,tau1_0=tau1_0,tau2_0=tau2_0,
                   tau3_0=tau3_0,familic2=familic2,m=m,n=n)

  ##-----------------Estimation par la méthode MV--------------#
  #
  #-----------------------------------------------------------#
  resu_MV<-optim(par=theta0,Vrai_simul,familic2=familic2,
                        data=data_SIM, hessian =TRUE)

  data_estim[i,1:7]<-resu_MV$par ##estimer MV des 7 paramètres
  data_estim_error[i,1:7]<-sqrt(diag(solve(resu_MV$hessian)))

  ##-----------------Estimation par la méthode IFM -------------#
  #
  #-----------------------------------------------------------#
```

```
x<-data_SIM[,1] ; y<-data_SIM[,2]

para1<-summary(fitdist(x,"norm")) #Paramètres de F0
para2<-summary(fitdist(y,"norm")) #Paramètres de G0
mu1_IFM<-para1$estimate[[1]]
sigma1_IFM<-para1$estimate[[2]]
mu2_IFM<-para2$estimate[[1]]
sigma2_IFM<-para2$estimate[[2]]
data_estim[i,8:9]<-c(mu1_IFM,sigma1_IFM)
data_estim[i,10:11]<-c(mu2_IFM,sigma2_IFM)

##---Erreurs associées au 4 paramètres-------#
data_estim_error[i,8:9]<-c(para1$sd[[1]],para1$sd[[2]])
data_estim_error[i,10:11]<-c(para2$sd[[1]],para2$sd[[2]])

#----Calcul de u et v-------------------------#
u<-pnorm(x,mean=mu1_IFM,sd=sigma1_IFM)  ##--calcul de u--
v<-pnorm(y,mean=mu2_IFM,sd=sigma2_IFM)  ##--calcul de v--

#---Estimation de eta1,delta2 et eta3 de IFM--#
##Le code des fonctions échangeables Norm_Ech1 (Étape 3.1)

   ##--Estimation de eta1 par IFM avec son erreur type----
resu_eta1<-nlm(Norm_Ech,p=eta1_0,data=cbind(u,data_SIM[,3]),
               hessian=TRUE)

data_estim[i,12]<-resu_eta1$estimate ##paramètre estimé

data_estim_error[i,12]<-sqrt(1/resu_eta1$hessian) ##sd de eta1

   ##--Estimation de delta2 par IFM avec son erreur-type---
resu_delta2<- BiCopEst(u1=u, u2=v, family =familic2,
                       method = "mle",se=TRUE)  ## estimation de c2

delta2_IFM<-resu_delta2$par
data_estim[i,13]<-delta2_IFM
data_estim_error[i,13]<-resu_delta2$se
```

```
      ##--Estimation de eta3 par IFM avec son erreur-type--------
    w<-BiCopHfunc1(u1=u, u2=v,family=familic2, par=delta2_IFM)

    resu_eta3<-nlm(Norm_Ech,p=resu_MV$par[7],data=cbind(w,data_SIM[,3]),
                   hessian=TRUE)
    data_estim[i,14]<-resu_eta3$estimate

    data_estim_error[i,14]<-sqrt(1/resu_eta3$hessian)  ##sd de eta3
  }
  ##----Faire ressortir les résultats d'estimation et la moyenne-----
  return(list(data_estim,round(colMeans(data_estim),2),
              data_estim_error,round(colMeans(data_estim_error),2)))
}
```

Nous présentation les résultats des biais relatifs des estimateurs de variance lorsque $C^{(2)}$ est la copule de *Clayton* ou est une copule de *Khoudraji*.

TABLEAU C.1 – Biais relatif des estimateurs de la variance des paramètres du modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule de *Clayton*.

$\tau_2(\delta_2)$	m	*Types*	*Méthode*	μ_1	σ_1	μ_1	σ_2	δ_1	δ_2	δ_3
0.4(0.59)		*Non*	MV	1	9	8	16	7	11	13
	10	*Équilibré*	IFM	-90	-74	-84	-61	-99	-80	-74
		Équilibré	MV	-4	23	5	19	6	$< 10^{-2}$	22
			IFM	-91	-72	-85	-53	-98	-69	-37
		Non	MV	-1	5	-2	5	-7	5	-8
	20	*Équilibré*	IFM	-83	-61	-77	-41	-83	-58	-44
		Équilibré	MV	-1	7	1	8	4	2	6
			IFM	-82	-60	-74	-43	-95	-71	-70
		Non	MV	11	$< 10^{-2}$	13	$< 10^{-2}$	-6	$< 10^{-2}$	7
	50	*Équilibré*	IFM	-87	-71	-82	-40	-68	-60	-30
		Équilibré	MV	2	10	2	$< 10^{-2}$	4	8	3
			IFM	-85	-57	-78	-45	-61	-72	-29
0.6(0.81)		*Non*	MV	11	27	17	24	4	30	9
	10	*Équilibré*	IFM	-90	-73	-87	-68	-98	-86	-42
		Équilibré	MV	4	31	2	29	-8	12	17
			IFM	-90	-71	-89	-64	-98	-70	-37
		Non	MV	-2	11	$< 10^{-2}$	4	-2	17	13
	20	*Équilibré*	IFM	-83	-60	-80	-52	-57	-57	-71
		Équilibré	MV	$< 10^{-2}$	8	-3	12	$< 10^{-2}$	6	9
			IFM	-84	-59	-80	-47	-89	-71	-75
		Non	MV	3	15	2	22	12	$< 10^{-2}$	5
	50	*Équilibré*	IFM	-88	-69	-86	-54	-65	-67	-34
		Équilibré	MV	14	11	13	25	4	19	1
			IFM	-85	-57	-81	-50	-61	-82	-83

TABLEAU C.2 – Biais relatif des estimateurs de la variance des paramètres du modèle de 2-copule échangeable lorsque $C^{(2)}$ est la copule de *Khoudraji*.

$\tau_2(\delta_2)$	m	*Types*	*Méthode*	μ_1	σ_1	μ_1	σ_2	δ_1	κ	κ_1	κ_2	δ_3
0.4(0.59)		*Non*	MV	-17	-4	-4	19	-1	9	5	69	-10
	10	*Équilibré*	IFM	-92	-75	-87	-55	-77	8	74	162	-50
		Équilibré	MV	-9	$< 10^{-2}$	-8	12	-6	5	-4	46	-13
			IFM	-90	-72	-86	-55	-76	5	50	112	-52
		Non	MV	-3	-32	-7	-43	-5	-11	-4	-6	3
	20	*Équilibré*	IFM	-84	-73	-78	-64	-64	3	65	29	-28
		Équilibré	MV	-9	13	-1	2	10	$< 10^{-2}$	1	77	5
			IFM	-83	-53	-75	-36	-56	35	116	370	-30
		Non	MV	10	16	26	$< 10^{-2}$	19	16	7	71	2
	50	*Équilibré*	IFM	-87	-66	-79	-50	-63	45	50	273	-38
		Équilibré	MV	11	1	-1	26	3	37	21	133	-17
			IFM	-83	-62	-78	-28	-61	262	174	513	49
0.6(0.81)		*Non*	MV	-75	-73	-75	-68	-66	-70	-69	-76	-80
	10	*Équilibré*	IFM	-97	-92	-96	-90	-93	-90	-76	-63	-86
		Équilibré	MV	17	23	-4	19	27	9	16	-4	-8
			IFM	-90	-78	-90	-70	-79	-50	-31	-49	-51
		Non	MV	17	26	8	20	16	4	10	50	5
	20	*Équilibré*	IFM	-81	-57	-77	-52	-60	142	32	269	-47
		Équilibré	MV	6	16	28	17	32	4	3	-9	-11
			IFM	-82	-58	-76	-47	-57	45	323	-22	-36
		Non	MV	33	-17	10	23	12	6	71	4	2
	50	*Équilibré*	IFM	-88	-80	-89	-47	-74	-131	1113	594	-37
		Équilibré	MV	71	23	40	10	13	27	94	38	21
			IFM	-82	-69	-81	-59	-73	-49	13	-32	-22

Annexe D

Programme *R* de certains résultats du chapitre 4

Les applications et résultats obtenus pour certaines sections du chapitre 4 sont présentés dans cette annexe. Nous répartissons le programme en deux sections :

— La première section donne le programme R pour construire les courbes de prédiction à l'aide de copules candidates.

— la deuxième partie est celle liée à l'analyse des données liées aux notes en mathématiques par le biais du modèle de 2-copule échangeable.

D.1 Présentation du programme *R* d'obtention des graphiques des courbes de niveau

Le programme R présenté ici permet d'obtenir la figure **??**.

```
##-----------Charger les packages-----------------------##
library(copula) ##Pour construire les copules

##-------------Principales fonctions utilisées-----------------##
     #@khoudrajiCopula : Construction de copules de Khoudraji
     #@integrate       : Intégration d'une fonction
     #@Vectorize       : Utilisation de la fonction sur un vecteur

##---------Paramètres estimés de F0----------------------#
```

```
 alpha1<-4.36 ; beta1<-2.56

##---------Paramètres estimés de G0----------------------#
 alpha3<-2.32 ; beta3<-3.48 ; lambda<-0.17

#------Paramètres de la copule c2----------------------#
 delta2<-0.78; kappa1<-0.82 ; kappa2<-0.97

#------Paramètres de la copule c2----------------------#
 delta3<-0.16


#--Fonction quantile de la loi bêta à trois paramètres----#
qbeta3<-function(x,alpha,beta,lambda)
{ x1<-qbeta(x,alpha,beta)
  x2<-x1/(lambda+x1*(1-lambda))
  x2}

#--Fonction de densité de la loi beta à trois paramètres----#
dbeta3<-function(x,alpha,beta,lambda){(lambda/(1-(1-lambda)*
         x)^2)*dbeta(lambda*x/(1-(1-lambda)*x),alpha,beta)}

#--Nombre de points pour évaluation-----------#
N<-500

########### Koudraji 1 ###################################

#---Construction de la copule de koudraji1----------------------#
  #c1 : copule d'indépendance
  #c2 : copule normale
koudraji1<-khoudrajiCopula(copula1 =indepCopula(),copula2 =
        normalCopula(delta2) ,shape = c(kappa1,kappa2))

#--Évaluation de c(u,v) de la copule de Koudraji 1-----------------#
  #--(u,v) sont choisies dans [0,1]x[0,1] par (i/(N+1),i/(N+1))

resu<-matrix(0,N,N)   ## Contenant les résultats de c(u,v)
for (i in (1:N))
```

```
{
  for (j in (1:N)){
    resu[i,j]=dCopula(c(i/(N+1),j/(N+1)),koudraji1)
  }
}
#-----Calcul de f(u,v) =f(u)*f(v)*c(u,v) : équation (4.9)-----------#

resu1<-diag(dbeta(qbeta((1:N)/(N+1),alpha1,beta1),alpha1,beta1))%*%
       resu%*%diag(dbeta3(qbeta3((1:N)/(N+1),alpha3,beta3,lambda),
       alpha3,beta3,lambda))

#-------Tracer des lignes de niveau (u,v,c(u,v))---------------------#
contour(x=qbeta((1:N)/(N+1),alpha1,beta1),y=qbeta3((1:N)/(N+1),alpha3,
    beta3,lambda), z=resu1,nlevels=10, xlim=c(0.2,1),ylim=c(0.2,1))

#---Ajout des points observés sur le graphe-----------------------#
points(jitter(math1,amount=.03),jitter(math3,amount=.03),
     pch=20, cex=0.5,col="blue")

#---Approximation de la régression copule-------------------------#
   # On calcule E(Y|X=x) pour les N valeurs de x
pred<-rep(0,N)
for ( i in 1:N){
  pred[i]<-sum(resu[i,]*qbeta3((1:N)/(N+1),alpha3,beta3,
    lambda))/sum(resu[i,])
}

#---Construction de la courbe en rouge--------------------------##
points(qbeta((1:N)/(N+1),shape1=alpha1, shape2=beta1),pred,
       type="line", col="red",lwd=2)

#---Construction de la droite de régression linéaire (verte)------##
   u0<-pbeta(math1,alpha1,beta1)
   v0<-pbeta(lambda*math3/(1-(1-lambda)*math3),alpha3,beta3)
   lineaire_reg<-lm(v0~u0)    ##régression linéaire
   abline(a=coef(a)[[2]],b=coef(a)[[1]], col="green")
```

Nous présentons par la suite, la fonction de prédiction pour régression copule.

```
###-----------------Fonction de prédiction : équation (4.8)--------##
#                                                                     #
#          Nom de la fonction : Predict_global_reg_cop_kou           #
#                                                                     #
#---------------------------------------------------------------------##


Predict_global_reg_cop_kou<-Vectorize(FUN=function(x,delta2,kappa1,
                            kappa2, alpha1,beta1,alpha3,beta3,lambda)
{

    ##----------------ENTRÉES-------------------------------------##
      #@x la valeur de la nouvelle note math1
      #@alpha1, @beta1 paramètres de F0
      #@alpha3, @beta3,lambda paramètres de G0
      #@delta2, @kappa1 et @kappa2 paramètres de c2 (khoudraji)
      #@copula1 : Indépendance ; @copula2 : Normale

    ##---------------SORTIE----------------------------------------##
      #la valeur de E(math3|x)

    #----Construction de la copule de Khoudraji--------------#
    Khoudraji1<-khoudrajiCopula(copula1 = indepCopula(),
                 copula2 =normalCopula(delta2,dim = 2),
                                  shapes = c(kappa1,kappa2))

    Integrale <- integrate(
      f = Vectorize(
        FUN = function(t) {
          x<-x/41 #---transformation des données sur [0,1]
          z<-pbeta(q=x,alpha1,beta1)  #---- loi marginale F0
          return(qbeta3_inv(x=t,alpha3,beta3,lambda)*
                   dCopula(c(t,z),Khoudraji1))
        },
        vectorize.args = "t"),lower=0,upper=1)$value
```

```
    return(41*Integrale) #---pour avoir une note entre [0,40]
 },  vectorize.args = "x")
```

D.2 Présentation du programme *R* du calcul de la log-vraisemblance

```
#** Fonction de répartition F de la loi beta (alpha,beta,lambda)*****

qbeta3<-function(x,alpha,beta,lambda){
  x2<-qbeta(lambda*x/(1-x*(1-lambda)),alpha,beta) x2}

#**********Inverse de qbeta3    ************************************

qbeta3_inv<-function(x,alpha,beta,lambda){
  x2<-qbeta(x,alpha,beta)/(lambda+qbeta(x,alpha,beta)*(1-lambda)) x2}

#**** Fonction de densité de la loi beta (alpha,beta,lambda)******

dbeta3<-function(x,alpha,beta,lambda){(lambda/(1-(1-lambda)*x)^2)*
    dbeta(lambda*x/(1-(1-lambda)*x),alpha,beta)}




###############cas général de l'équation (4.11)   *****************

gen_model<-function(data,para){

  #***** @data de trois colonnes : School, math1 et math3 ****
  #
  ##------Assignation des paramètres de chaque modèle------###
  alpha1<-para[1] ; beta1<-para[2]

  alpha3<-para[3]  ; beta3<-para[4] ; lambda<-para[5]
```

```
delta1<-para[6]

delta2<-para[7] ; kappa1<-para[8] ; kappa2<-para[9]

delta3<-para[10]
m<-max(data$School)


##-----Construction des pseudo-observations U0 et V0----------#
u0<-pbeta(math1,alpha1,beta1)
v0<-pbeta(lambda*math3/(1-(1-lambda)*math3),alpha3,beta3)

##--------partie I de la log-vraisemblance----------------------##

l1<- -sum(log(dbeta(math1,alpha1,beta1)))

##------partie II de la log-vraisemblance------------#

l2<- -sum(log(dbeta3(x=math3,alpha3,beta3,lambda)))

##----Vraisemblance de la partie l3=l01+l02+l03------------------##

l01<-0
for(i in 1:m){

  ##-------partie III-1 liée à la copule c1------#
  ind<-which(data$School==i)
  lnC1<--dCopula(u0[ind],normalCopula(delta1,
            dim=length(ind)),log=TRUE)
  l01<-l01+lnC1
}
  ##--partie III-2 liée à la copule c2-----------
koudraji<-khoudrajiCopula(copula2 =normalCopula(delta2),
                          shapes= c(kappa1,kappa2))

l02<- -sum(dCopula(matrix(cbind(1-u0,1-v0),ncol=2),koudraji,log=TRUE))

  ##-------partie III-3 liée à la copule c3-------
```

```
      ##--calcul des w0------------

  w0<-1+(kappa1-1)*(1-u0)^(-kappa1)*(1-v0)^(1-kappa2)*
  pCopula(matrix(cbind((1-u0)^kappa1,(1-v0)^kappa2),ncol=2),
     normalCopula(delta2))-kappa1*(1-v0)^(1-kappa2)*
      BiCopHfunc1(u1=(1-u0)^kappa1,u2=(1-v0)^kappa2,
       BiCop(family =1,par=delta2)) ##family=1 (equation 4.7)


       ##--calcul de la pseudo-vraisemblance avec la copule c3------
  l03<-0
  for(j in 1:m){
    ind1<-which(data$School==j)
    lnC3<- -dCopula(w0[ind1],normalCopula(delta3,
              dim=length(ind1)),log=TRUE)
    l03<-l03+lnC3
  }

  return(l1+l2+l01+l02+l03)
}
```

D.3 Présentation du programme *R* de la fonction de la prédiction du modèle de 2-copule échangeable

L'annexe présente le programme R pour l'évaluation de l'équation (4.19).

```
#-------------Fonction de prédiction de l'équation (4.12)----------------#
#                                                                         #
#         Nom de la fonction : Predict_2copule_gen                        #
#-------------------------------------------------------------------------#

##-----------------Charger des packages-----------------------------#
library(statmod)  #--important pour Gauss-Hermit
library(VineCopula)  #---Important pour les fonctions conditionnelles
```

```
library(copula)    #--Important pour les copules
library(GoFKernel) #--Important pour l'inversion de fonction


##-----------------Construction de la fonction-------------------------#


Predict_2copule_gen<-Vectorize(function(x,data,j,alpha1,beta1,alpha3,
                      beta3,lambda,delta2,kappa1,kappa2,delta3,K)
{
      # Note : c1, c2, c3 sont trois copules de 2-copule échangeable

      ##----------------ENTRÉES-------------------------------------##
      #@x la valeur de la nouvelle note en quatrième année (math1)
      #@data contenant : - La note en 4ième année (math1)
      #                  - La note en 7ième année (math3)
      #                  - School(j) est l'école des élèves
      #@j est la grappe dans laquelle se trouve @x
      #@alpha1, @beta1 : paramètres de F0
      #@alpha3, @beta3,lambda : paramètres de G0
      #@delta2, @kappa1 et @kappa2 : paramètres de c2 (khoudraji)
      #@delta3 paramètre de la copule c3
      #@K est le nombre de subdivisions Gauss-Hermite


      ##---------------SORTIE----------------------------------------##
      #la valeur de E(math3|x,j,data)

      ##---------------FONCTIONS UTILISÉES--------------------------##
      # khoudraji  : construis les copules de khoudraji
      # gauss.quad : calcule le point (ti,wi) utile pour Gauss-Hermite


         #********--Remplacement de math1 et math3--***************#
             # pour briser les égalités ------#
     math1<-(data$math1[which(data$School==j)]+
        rnorm(length(data$math1[which(data$School==j)]),sd=0.0001))/41;
     math3<-(data$math3[which(data$School==j)]+
```

```
    rnorm(length(data$math3[which(data$School==j)]),sd=0.0001))/41;

     #***************--calcul de U0 et V0--********************#

  u0<-pbeta(math1,alpha1,beta1)
  v0<-pbeta(lambda*math3/(1-(1-lambda)*math3),alpha3,beta3)

      #*************--------calcul W0------**************#

  w0<-1+(kappa1-1)*(1-u0)^(-kappa1)*(1-v0)^(1-kappa2)*
     dCopula(matrix(cbind((1-u0)^kappa1,(1-v0)^kappa2),ncol=2),
     normalCopula(delta2))-kappa1*(1-v0)^(1-kappa2)*pnorm(
      (qnorm((1-v0)^kappa2)-delta2*qnorm((1-u0)^kappa1))/
                                      sqrt(1-delta2^2))


   #***********---Function de prediction---**************

  quadra<-gauss.quad(K,kind="hermite") #
  n0<-length(w0)
  mu0<-n0*delta3*mean(qnorm(w0))/(1+(n0-1)*delta3)
  sigma0<-sqrt((1-delta3)*(1+n0*delta3)/(1+(n0-1)*delta3))
  x<-x/41   #--Ramener la valeur dans [0,1]
  u<-pbeta(q=x,alpha1,beta1)  #-- loi marginale F0
  z<-pnorm(q=mu0+sqrt(2)*sigma0*quadra$nodes)  #--loi normale

 ##----Inverse de la fonction conditionnelle F_c-------------#
 F_vu<-Vectorize(function(v){
     a<-1+(kappa1-1)*(1-u)^(-kappa1)*(1-v)^(1-kappa2)*
        pCopula(cbind((1-u)^kappa1,
                 (1-v)^kappa2),normalCopula(delta2,dim=2))-
        kappa1*(1-v)^(1-kappa2)*pnorm((qnorm((1-v)^kappa2)-delta2*
                        qnorm((1-u)^kappa1))/sqrt(1-delta2^2))
    return(a)
  },c("v"))
Inv_F_c<-inverse(F_vu,lower=0,upper=1)  #--Inverse de F_vu
Fvinv<-c()
for(j in 1:K){ Fvinv[j]<-Inv_F_c(z[j])}
```

```
   return(41*sum(quadra$weights*qbeta3_inv(x=Fvinv,alpha3,
                beta3,lambda))/sqrt(pi))

},"x")
```

Bibliographie

Aas, K., Czado, C., Frigessi, F., et Bakken, H. (**2009**). "Pair-copula constructions of multiple dependence", *Insurance Mathematics and Economics* **44**, 182–198, URL https://doi.org/10.1016/j.insmatheco.2007.02.001.

Acar, E. F., Azimaee, P., et Hoque, M. E. (**2019**). "Predictive assessment of copula models", *Canadian Journal of Statistics* **47**, 8–26.

Albert, P. S. (**2012**). "A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification", *Statistics in Medicine* **31**, 143–154.

Aldous, D. J. (**1985**). "Exchangeability and related topics", *In École d'Été de Probabilités de Saint-Flour XIII—1983* 1–198.

Atique, F. et Attoh-Okine, N. (**2016**). "Using copula method for pipe data analysis", *Construction and Building Materials* **106**, 140—-148.

Battese, G. E., Harter, R. M., et Fuller, W. A. (**1988**). "An error-components model for prediction of county crop areas using survey and satellite data", *Journal of the American Statistical Association* **83**, 28–36.

Bedford, T. et Cooke, R. M. (**2002**). "Vines a new graphical model for dependent random variables", *Annal of Statistics* **30**, 1031–1068.

Bernard, C. et Czado, C. (**2015**). "Conditional quantiles and tail dependence", *Journal of Multivariate Analysis* **138**, 104–126.

Brahim, B., Fatah, B., et Djabrane, Y. (**2018**). "Copula conditional tail expectation for multivariate financial risks", *Arabe Journal of Mathematical Sciences* **24**, 82–100.

Bray, J. H. et Maxwell, S. E. (**1985**). *Multivariate Analysis of Variance* (*Sage Publications*).

Brown, H. et Prescott, R. (**2014**). *Applied Mixed Models in Medicine* (*John Wiley and Sons*).

Casella, G. et Berger, R. L. (**2002**). *Statistical Inference* (Duxbury/Thomson Learning).

Chang, B. (**2019**). "Vine copulas : dependence structure learning, diagnostics, and applications to regression analysis", Ph.D. thesis, University of British Columbia, URL https://open.library.ubc.ca/collections/ubctheses/24/items/1.0379699.

Chang, B. et Joe, H. (**2020**). "Copula diagnostics for asymmetries and conditional dependence", *Journal of Applied Statistics* **47**, 1587–1615.

Chen, X. et Fan, Y. (**2006**). "Estimation of copula-based semiparametric time series models", *Journal of Econometrics* **130**, 307–335.

Cherubini, U., Luciano, E., et Vecchiato, W. (**2014**). *Copula Methods in Finance* (John Wiley).

Cleveland, W. S., Grosse, E., et Shyu, W. M. (**1992**). *Local regression models* (Chapter 8 of Statistical Models in S. eds Chambers JM and Hastie TJ, Wadsworth Brooks/Cole).

Cockriel, W. M. et McDonald, J. B. (**2018**). "Two multivariate generalized beta families", *Communications in Statistics - Theory and Methods* **47**, 5688–5701.

Cooke, R. M., Joe, H., et Chang, B. (**2019**). "Vine copula regression for observational studies", *A Journal of the German Statistical Society* **104**, 141–167.

Cousineau, D., Brown, S., et Heathcote, A. (**2004**). "Fitting distributions using maximum likelihood : Methods and packages", *Behavior Research Methods, Instruments, and Computers* **36**, 742–756.

Crane, G. J. et Van der Hoek, J. (**2008**). "Conditional expectation formulae for copulas", *Australian and New Zealand Journal of Statistics* **50**, 53–67.

Cribari-Neto, F. et Zeileis, A. (**2010**). "Beta regression in R", *Journal of Statistical Software* **34**, 1–24, URL http://www.jstatsoft.org/v34/i02/.

Czado, C. (**2019**). *Analyzing dependent data with vine copulas :a practical guide with R*, 2nd edition (Springer).

Czado, C. et Nagler, T. (**2022**). "Vine copula based modeling", *Annual Review of Statistics and Its Application* **9**, 453–477.

Das, A. (**2015**). "Estimation of urinary arsenic exposure using copula-based regression : a case study of west bengal", *Environmental Modeling and Assessment* **20**, 159–167.

Delignette-Muller, M. L. et Dutang, C. (**2015**). "Fitdistrplus : an r package for fitting distributions", *Journal of Statistical Software* **64**.

Diggle, P. J., Liang, K. Y., et Zeger, S. L. (**1994**). "Analysis of longitudinal data", *Oxford University Press* **30**.

Durante, F. et Okhrin, O. (**2015**). "Estimation procedures for exchangeable marshall copulas with hydrological application", *Stochastic Environmental Research and Risk Assessment* **29**, 205–226.

Feller, W. (**1972**). *An Introduction to Probability Theory and Its Applications*, 2nd edition (Wiley).

Frahm, G., Junker, M., et Szimayer, A. (**2003**). "Elliptical copulas : applicability and limitations", *Statistics and Probability Letters* **3**, 275–286.

Genest, C., Nikoloulopoulos, A. K., Rivest, L.-P., et Fortin, M. (**2013**). "Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas", *Brazilian Journal of Probability and Statistics* **27**, 265–284.

Genest, C. et Rivest, L.-P. (**1993**). "Statistical inference procedures for bivariate archimedean copula", *Journal of the American Statistical Association* **88**, 1034–1043.

Ghazal, G. A. (**2000**). "Recurrence formula for expectations of products of bilinear forms and expectations of bilinear forms and random matrices", *Statistics and Probability Letters* **48**, 1—9.

Ghiselli, R. (**2013**). "Exchangeable copulas", *Fuzzy Sets and Systems* **220**, 88–90.

Goldstein, H. (**2011**). *Multilevel Statistical Models*, 2nd edition (Wiley).

Graf, M., Marin, J., et Molina, I. (**2018**). "A generalized mixed model for skewed distributions applied to small area estimation", **28**, 565–597, URL https://doi.org/10.1007/s11749-018-0594-2.

Grover, K., Acar, E. F., et Torabi, M. (**2020**). "Copula-based predictions in small area estimation", *Canadian Journal of Statistic* **48**.

Hald, A. (**2007**). *A history of parametric statistical inference from Bernoulli to fisher, 1713–1935* (Springer), URL https://doi.org/10.1007/978-0-387-46409-1.

Han, D., Tan, K. S., et Weng, C. (**2017**). "Vine copula models with glm and sparsity", *Communications in Statistics - Theory and Methods* **46**.

Hill, B. M. (**2003**). "Statistical robustness", in *Encyclopedia of Physical Science and Technology (Third Edition)*, edited by A. M. Robert, third edition edition, 821–833 (Academic Press, New York), URL https://www.sciencedirect.com/science/article/pii/B0122274105007286.

Hoang, Q., Khandelwal, P., et Ghosh, S. (**2019**). "Robust predictive model using copulas", *Data-Enabled Discovery and Applications* **3**, 1–11.

Hobaek Haff, I. (**2013**). "Parameter estimation for pair-copula constructions", *Bernoulli* **19**, 462—491.

Huang, F. L. (**2020**). "Manova : a procedure whose time has passed ?", *Gifted Child Quarterly* **64**, 56–60.

Joe, H. (**1996**). "Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters", *Lecture Notes-Monograph Series* **28**, 120–141.

Joe, H. (**1997**). *Multivariate Models and Dependence Concepts* (Chapman and Hall).

Joe, H. (**2005**). "Asymptotic efficiency of the two-stage estimation method for copula-based models", *Journal of Multivariate Analysis* **94**, 401–419.

Joe, H. (**2014**). *Dependence Modeling with Copulas* (Chapman and Hall).

Joe, H. et Kurowicka, D. (**2011**). "Dependence modeling : vine copula handbook", *World Scientific* .

Joe, H. et Xu, J. (**1996**). "The estimation method of inference functions for margins for multivariate models", *Technical Report, Department of Statistics,University of British Columbia* 1.

Karian, Z. A. et Dudewicz, E. J. (**2000**). *Fitting statistical distributions : the generalized lambda distribution and generalized bootstrap methods* (Chapman and Hall).

Kraemer, N., Brechmann, E., Silvestrini, D., et Czado, C. (**2013**). "Total loss estimation using copula-based regression models", *Insurance : Mathematics and Economics* **53**, 829–839.

Kumar, P., S. M. (**2007**). "Copula based prediction models : an application to an aortic regurgitation study", *BMC Medical Research Methodology* **7**, 1–9.

Kurowicka, D. et Cooke, R. (**2006**). *Uncertainty Analysis with High Dimensional Dependence Modelling* (Chichester).

Leong, Y. K. et Valdez, E. A. (**2005**). "Claims prediction with dependence using copula models", *Insurance : Mathematics and Economics* .

Li, C.-Q., Firouzi, A., et Yang, W. (**2016**). "Prediction of total cost of construction project with dependent cost items", *Journal of Construction Engineering and Management* **142**.

Lu, T.-T. et Shiou, S.-H. (**2002**). "Inverses of 2×2 block matrices. computers and mathematics with applications", *Dependence Modeling* **43**, 119–129.

Mai, J. F. et Scherer, M. (**2012**). *Simulating copulas : stochastic models, sampling algorithms, and applications* (Quantitative Finance).

McCulloch, C. E. et Searle, S. R. (**2001**). *Generalized, linear, and mixed models* (John Wiley et Sons), URL https://doi.org/10.1007/978-3-319-64221-5.

McNeil, A. J. et Nêslehovà, J. (**2009**). "Multivariate archimedean copulas, d-monotone functions and l1-norm symmetric distribution", *The Annals of Statistics* **37**, 3059–3097.

Menard, E. et Raoult, J. P. (**1978**). "Indépendance conditionnelle et uniformité pour les lois fortes des grands nombres dans les espaces de banach", *Z. Wahrscheinlichkeitstheorie verw Gebiete* **41**, 193—204.

Mortimore, P., Sammons, P., Stoll, L., Stoll, L., Lewis, D., et Ecob, R. (**1988**). *School Matters* (Wells).

Mukherjee, S., Lee, Y., Kim, J., Jang, J., et Park, J. (**2018**). "Construction of bivariate asymmetric copulas", *Fuzzy sets and systems* **25**, 217–234, URL https://doi.org/10.29220/CSAM.2018.25.2.217.

Nelsen, R. B. (**2006**). *An Introduction to Copulas* (Springer).

Nelsen, R. B. (**2017**). *Copulas and Dependence Models with Applications : Contributions in Honor of Roger B. Nelsen.* (Springer).

Noh, H., Ghouch, A. E., et Bouezmarni, T. (**2013**). "Copula-based regression estimation and inference", *Journal of the American Statistical Association* **108**, 676–688.

Panagiotelis, A., Czado, C., et Joe, H. (**2012**). "Pair copula constructions for multivariate discrete data", *Journal of the American Statistical Association* **107**, 1063–1072.

Prenen, L., Braekers, R., et Duchateau, L. (**2017**). "Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size", *Journal of the Royal Statistical Society* **79**, 483–505.

Quessy, J.-F., Rivest, L.-P., et Toupin, M.-H. (**2015**). "Semi-parametric pairwise inference methods in spatial models based on copulas", *Spatial Statistics* **14**, 472–490, URL https://doi.org/10.1016/j.spasta.2015.08.002.

Quessy, J.-F., Rivest, L.-P., et Toupin, M.-H. (**2016**). "On the family of multivariate chi-square copula", *Journal of Multivariate Analysis* **152**, 40–60.

R Development Core Team (**2008**). *R :*, R Foundation for Statistical Computing, Vienne, Autriche.

Rivest, L., Verret, F., et Baillargeon, S. (**2016**). "Unit level small area estimation with copulas", *The Canadian Journal of Statistics / La revue Canadienne de statistique* **44**, 397–415.

Romdhani, H., Lakhal-Chaieb, L., et Rivest, L.-P. (**2014**a). "An exchangeable kendall's tau for clustered data", *The Canadian Journal of Statistics* **42**, 384–403.

Romdhani, H., Lakhal-Chaieb, L., et Rivest, L.-P. (**2014**b). "Kendall's tau for hierarchical data", *Journal of Multivariate Analysis* **128**, 210–225.

Schepsmeier, U., Stöber, J., Brechmann, E. C., Graeler, B., Nagler, T., et Erhardt, T. (**2018**). *VineCopula : Statistical Inference of Vine Copulas*, volume 2, URL https://CRAN.R-project.org/package=VineCopula, r package version 2.1.8.

Schucany, W. R., Parr, W. C., et Boyer, J. E. (**1978**). "Correlation structure in farlie-gumbel-morgenstern distributions", *Biometrika* 650–653.

Shemyakin, A. et Kniazev, A. (**2017**). *Introduction to bayesian estimation and copula models of dependence* (John Wiley Sons).

Shi, P., Feng, X., et Boucher, J.-P. (**2011**). "Multilevel modeling of insurance claims using copulas", *Annals of Applied Statistics* **10**, 834–863.

Shi, P. et Lee, G. Y. (**2022**). "Copula regression for compound distributions with endogenous covariates with applications in insurance deductible pricing", *Journal of the American Statistical Association* 1–10.

Shi, P. et Zhang, W. (**2011**). "A copula regression model for estimating firm efficiency in the insurance industry", *Journal of Applied Statistics* **38**, 2271–2287.

Sklar, A. (**1959**). "Fonctions de répartition à n dimensions et leurs marges", *Publications de l'Institut de Statistique de L'Université de Paris 8* 229–231.

Smith, K., Lamb, K., et Henson, R. (**2020**). "Making meaning out of manova : Using descriptive discriminant analysis for multivariate post hoc testing in gifted education research", *Gifted Child Quarterly* 41–55.

Smith, M. S. et Klein, N. (**2021**). "Bayesian inference for regression copulas", *Journal of Business Economic Statistics* **39**, 712–728.

Spanhel, F. et Kurz, M. S. (**2019**). "Simplified vine copula models : approximations based on the simplifying assumption", *Journal of Statistics* **13**, 712–728.

Steen, N. M., Byrne, G. D., et Gelbard, E. M. (**1961**). "Gaussian quadratures for the integrals", *Journal of Association Computation Machine* 21–40.

Stefanski, L. A. et Boos, D. D. (**2002**). "The calculus of m-estimation", *The American Statistician* 29–38.

Stöber, J., Joe, H., et Czado, C. (**2013**). "Simplified pair copula constructions—limitations and extensions", *Journal of Multivariate Analysis* **119**, 101–118.

Su, C.-L. et Lin, F.-C. (**2019**). "Analysis of clustered failure time data with cure fraction using copula", *Statistics in Medicine* **38**, 3961–3973.

Su, C.-L., Neslehova, J. G., et Wang, W. (**2019**). "Modelling hierarchical clustered censored data with the hierarchical kendall copula", *Canadian Journal of Statistics* **47**, 182–203.

Su, L. (**2004**). "*Nonparametric tests for conditional independence* ", Ph.D. thesis.

Taniguchi, M. et Hirukawa, J. (**2012**). "Generalized information criterion", *Journal of Time Series Analysis* **33**, 287–297.

Thomopoulos, N. T. (**2017**). *Statistical Distributions : Applications and Parameter Estimates* (Springer).

Thomopoulos, N. T. (**2018**). *Probability Distributions : with Truncated, Log and Bivariate Extensions* (Springer).

Tsiatis, A. A. (**2006**). *Semiparametric Theory and Missing Data* (Springer), URL https://doi.org/10.1007/0-387-37345-4.

Vaz de Melo Mendes, B. et Aiube, C. (**2011**). "Copula based models for serial dependence", *International Journal of Managerial Finance* **7**, 68–82, URL https://ideas.repec.org/a/ffe/journl/v8y2011i2p89-111.html.

Verbeke, G. et Molenberghs, G. (**2000**). *Linear mixed models for longitudinal data* (Springer).

Wang, H. J., Feng, X., et Dong, C. (**2019**). "Copula-based quantile regression for longitudinal data", *Statistica Sinica* **29**, 245–264.

Wang, K. et Shan, W. (**2020**). "Copula and composite quantile regression-based estimating equations for longitudinal data", *Annals of the Institute of Statistical Mathematics* **73**, 441—-455.

Wu, F., Valdez, E. A., et Sherris, M. (**2007**). "Simulating exchangeable multivariate archimedean", *Communications in Statistics-Simulation and Computation* **36**, 1019–1034.

Xue-Kun Song, P. (**2000**). "Multivariate dispersion models generated from gaussian copula", *Scandinavian Journal of Statistics* **27**, 305–320.

Yang, X., Frees, E. W., et Zhang, Z. (**2011**). "A generalized beta copula with applications in modeling multivariate long-tailed data", *Insurance Mathematics and Economics* **49**, 265–284.

Zhang, W., Wang, J., Qian, F., et Chen, Y. (**2020**). "A joint mean-correlation modeling approach for longitudinal zero-inflated count data", *Journal of Probability and Statistics* **34**, 35–50.