UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LE MODÈLE DE RÉGRESSION QUANTILE BINAIRE À BASE D'UNE COPULE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

GILLIS DELMAS TCHOUANGUE DINKOU

OCTOBRE 2020

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.10-2015). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Tout seul on ne peut arriver à bâtir. Qu'il me soit permis d'adresser mes sincères remerciements et d'exprimer ma gratitude au Seigneur Dieu qui m'a pourvu en santé et en intelligence pour mener à terme ces études. Je tiens à remercier aussi mes chers parents qui par leurs sacrifices je suis ici, ainsi que mes frères, mes soeurs et ma conjointe pour le soutien indéfectible et toute l'affection inconditionnelle que vous m'avez toujours apportés.

Je souhaite également remercier tous les organismes et particuliers qui m'ont été d'une grande aide financière, sans quoi, j'aurai rencontré des difficultés dans le paiement de mes frais de scolarité : ESF (Études Sans Frontières), STATQAM, la Fondation de l'UQÀM sans oublier mon directeur de mémoire.

Mes remerciements vont encore à l'endroit de mon directeur de mémoire, Dr. Karim Oualkacha, qui a bien voulu accepter de diriger ce chef d'oeuvre et qui, malgré ses occupations, a soutenu mes efforts jusqu'au bout. Il s'est révélé réellement présent et précieux et n'a ménagé aucun effort pour que ce travail puisse voir le jour.

En dernier lieu je remercie l'ensemble du personnel du département de mathématiques, surtout les enseignants de statistiques pour avoir participé à ma formation académique ; mes collègues, surtout ceux de mon bureau pour la bonne ambiance qu'ils ont toujours fait régner autour de moi et les conseils prodigués quand les recherches semblaient difficiles ; mes tuteurs et la chorale Marie Reine Des Âpotres qui ont été d'un soutien indéfectible dans ma réussite. Merci pour vos sempiternels encouragements.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Dans le domaine de la santé, la variable d'intérêt est le plus souvent binaire et provient d'une variable latente continue. Bien que la régression logistique soit une méthode de référence utilisée pour expliquer la relation entre une variable binaire et ses déterminants, elle ne permet de modéliser la relation qu'au centre de la distribution de la variable latente. (Kordas, 2006), pour pallier ce problème, a proposé le modèle de régression quantile binaire linéaire. Or, en pratique, la relation entre la variable latente et ses déterminants n'est pas toujours linéaire. Le modèle que nous proposons étend celui de Kordas au modèle de régression quantile binaire à base d'une copule. La copule rend le modèle plus réaliste en ce sens qu'elle est capable de modéliser de façon appropriée la relation entre la variable d'intérêt et ses déterminants. Dans notre travail, nous estimons les paramètres de façon semi-paramétrique et nous calculons de façon paramétrique la probabilité de succès. À l'aide des simulations, nous comparons le modèle proposé aux modèles linéaires généralisés sous le prisme de la classification. Nous ajustons enfin notre modèle à des données réelles de méthylation de l'ADN.

Mots clés : variable latente, régression logistique, régression quantile, copule, modèle linéaire généralisé

ABSTRACT

In health framework, several variables of interest are binary. Although logistic regression are gold standard methods used to explain the relationship between a binary outcome and its determinants, such approaches rely only on the mean of the outcome to the predictors. Thus, they well characterize the centre of the outcome distribution, but they might fail to capture covariates' effects in the tails of the response distribution. (Kordas, 2006) proposed the linear binary quantile regression model to overcome this problem. However, the relationship between the variables is not always linear in practice. To face this limit, the model we propose extends Kordas' model to the copula-based binary quantile regression model. We show how to obtain the estimators and parametric estimate of success probability. Using simulations, the proposed model is compared to generalized linear models in term of classification. We illustrate the use of our methodology via DNA methylation data analysis.

Keywords : latent variable, logistic regression, quantile regression, copula, Generalized Linear Model

INTRODUCTION

Un des objectifs principaux des statisticiens est d'établir la relation/association entre une variable d'intérêt ou variable réponse, Y (exemple : la présence ou l'absence d'une maladie en épidémiologie), et ses déterminants, X (ou prédicteurs ou covariables, par exemple : des marqueurs génétiques). Le modèle standard pour ce type d'analyse est la régression linéaire, qui modélise la moyenne de Y conditionnelle à X. L'origine du mot régression vient de Sir Françis Galton. En 1885, travaillant sur l'hérédité, il chercha à expliquer la taille des fils (Y) en fonction de celle des pères (X) et de là naquit le modèle de régression linéaire. Bien qu'essentiel en terme de sa capacité à apporter des informations, il s'avère que ce modèle présente des limites. En effet, non seulement la dépendance/association entre Y et X en pratique n'est pas forcément linéaire, mais aussi il pourrait avoir plus d'informations utiles dans les queues qu'au centre de la distribution de Y [1].

La régression quantile se profile comme une bonne alternative à la régression linéaire pour remédier à ces limites inhérentes à la moyenne. Elle a été formalisée par Koenker et Basset (Koenker et Bassett Jr, 1978) comme une extension de la régression linéaire standard au lien entre les prédicteurs X et les quantiles conditionnels de Y sachant X. Elle permet d'avoir une description plus précise de la distribution d'une variable conditionnelle à ses déterminants, comparativement à la régression linéaire qui ne se focalise que sur la moyenne conditionnelle. Dans le cas d'une variable réponse continue, (Noh *et al.*, 2015) utilisent les copules dans

1. La moyenne n'est donc plus un outil optimal pour décrire la relation entre Y et X.

un nouveau modèle de régression quantile afin de mieux capter la dépendance [2] entre Y et X, et de ce fait, résoudre le problème de linéarité, une des limites de la régression linéaire standard.

Dans le cas d'une variable réponse binaire [3], (Kordas, 2006) montre que la régression quantile binaire a des avantages sur la régression linéaire binaire (régression logistique). Cet avantage est d'autant plus probant que l'échantillon est non balancé. Un exemple de jeu de données non balancé serait un cas de maladie rare qui affecte moins de 1% de la population.

Dans ce travail, nous adaptons l'approche de Noh (Noh *et al.*, 2015) au modèle de Kordas (Kordas, 2006) en modélisant la dépendance entre Y et X par l'intermédiaire d'une copule, ce qui conduit à un tout nouveau modèle, le modèle de régression quantile binaire à base d'une copule. L'inclusion de la copule dans le modèle de Kordas pour capter la relation entre X et Y rend le modèle plus réaliste. En effet, le modèle de Kordas suppose la relation entre X et Y linéaire, ce qui n'est toujours pas le cas empiriquement. La copule vient donc capter la relation, quelle qu'elle soit, entre la variable réponse et ses déterminants. Le chapitre 1 est essentiellement constitué de rappels des propriétés des quantiles et porte sur le modèle de Kordas. La théorie des copules, outil indispensable pour comprendre et étudier notre modèle - le modèle de régression quantile binaire à l'aide d'une copule - fera l'objet du chapitre 2. Le chapitre 3 constitue essentiellement notre apport personnel. On y développe la théorie derrière le modèle proposé et on l'applique à quelques copules. Enfin, dans le chapitre 4, on évalue la performance

---

2. Ladite dépendance pourrait être linéaire ou pas.

3. Une variable binaire ou dichotomique dans la littérature scientifique est une variable qualitative qui ne peut prendre que deux modalités souvent codées par 0 ou 1, 1 pour atteint d'une maladie et 0 pour non-atteint de la maladie en ce qui concerne le diagnostique d'une maladie.

de notre méthode comparée au modèle de régression logistique par des études de simulation et d'analyse des données réelles de méthylation de l'ADN.

CHAPITRE I

LE MODÈLE DE KORDAS

Le modèle de Kordas est une extension des travaux de (Horowitz, 1992) aux quantiles autres que la médiane. Ses travaux portaient sur l'estimation du maximum de la fonction score d'un modèle de régression binaire à la médiane. Pour mieux aborder le modèle de Kordas (le modèle de régression quantile binaire lissé), nous avons besoin de certains outils. Dans ce qui suit, tout ce qui sera developpé sera axé sur le strict nécessaire pour la compréhension de ce modèle (Kordas, 2006).

1.1 Introduction à la régression quantile

La régression quantile est un outil développé pour étudier la distribution d'une variable d'intérêt étant donné un ensemble de covariables. Il pourrait dans ce sens être vu comme une extension et un complément à la méthode des moindres carrés qui estime au centre de la distribution par l'intermédiaire de la moyenne conditionnelle. La régression quantile est une méthode semi-paramétrique car elle permet d'effectuer de l'inférence sur les quantiles conditionnels sans reposer sur une quelconque forme paramétrique de la distribution des termes de l'erreur.

L'idée de la régression quantile voit le jour en 1757 lorsqu'elle est introduite par Boscovich. Il a été le premier à démontrer que la médiane d'une variable est aussi la solution au problème de minimisation d'une somme de déviations absolues.

La méthode de régression quantile est restée longtemps dans l'ombre et inutilisée à cause des difficultés de programmation causées par la fonction valeur absolue qui n'est pas partout différentiable. Elle l'est restée ainsi jusqu'à l'introduction de la programmation linéaire dans les années 1950, car en 1805, la méthode des moindres carrés introduite par Legendre, du fait de sa simplicité, devient la plus connue et la plus utilisée dans les problèmes d'inférence statistique. La régression quantile refait néanmoins surface dans les années 70 avec Koenker (Koenker et Bassett Jr, 1978) et devient populaire dans le domaine de l'économie appliquée à la fin des années 90 à cause de sa capacité à tenir compte de l'hétérogénéité dans les données.

1.2 Quantiles et fonction de quantiles

Soit Y une variable aléatoire continue. Sa distribution de probabilité est généralement décrite par sa densité de probabilité et sa fonction de répartition (*cdf* de l'anglais « cumulative distribution function »).

Définition 1.2.1. ***(Fonction de répartition)***
La cdf de Y est définie par

$$F_Y(y) = P(Y \leq y), \quad y \in \mathbb{R}.$$

Une autre quantité très utile, souvent utilisée pour explorer la distribution de la variable aléatoire Y est la fonction de quantiles.

Définition 1.2.2. ***(Quantile)***
Considérons pour $y \in \mathbb{R}$ la limite à gauche de la cdf de Y, $F_Y(y^-) := \lim_{t \uparrow y} F_Y(t)$. Pour $\tau \in (0,1)$ donné, la valeur de y telle que

$$F_Y(y^-) = P(Y < y) \leq \tau \quad et \quad F_Y(y) = P(Y \leq y) \geq \tau$$

est le quantile d'ordre τ de la distribution de Y. Un quantile est donc une valeur qui divise la distribution en deux régions complémentaires.

Il existe une sorte de relation inverse entre les quantiles et les valeurs de la *cdf*. Cette relation n'est en général pas la fonction inverse puisque la fonction de distribution d'une variable aléatoire n'est pas toujours une bijection et donc pas toujours inversible. À la Figure 1.1, on voit clairement que la loi de Poisson n'est pas bijective et donc n'admet pas d'inverse. C'est pour cette raison qu'il est important et nécessaire de définir un inverse généralisé d'une fonction de distribution.

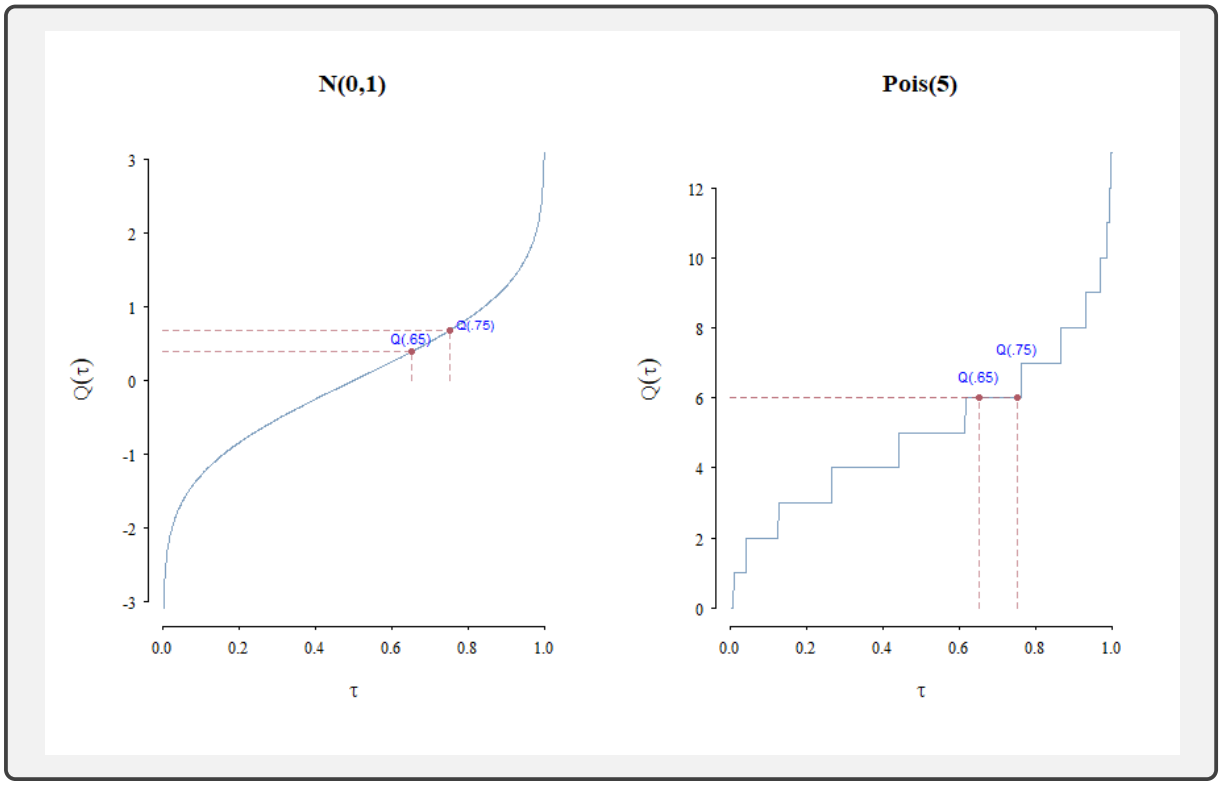

Figure 1.1: Fonction de quantiles d'une loi normale standard et d'une loi de poisson de paramètre 5.

Définition 1.2.3. ***(Fonction de quantiles)***

1. *On définit une fonction de quantiles par*

$$Q_Y(\tau) = F_Y^{-1}(\tau) := \inf\{y \in \mathbb{R} : F_Y(y) \geq \tau\}, \quad \tau \in (0,1),$$

   où F_Y^{-1} est l'inverse généralisé de la cdf de Y, F_Y.

2. *On définit aussi le quantile d'ordre τ comme la valeur $Q_Y(\tau)$ telle que*

$$P(Y \leq Q_Y(\tau)) = \tau, \quad \tau \in (0,1).$$

   La plupart des quantiles définis de la sorte n'existent pas pour des distributions discrètes.

La médiane par exemple est obtenue pour $\tau = .5$, c'est-à-dire $Q_Y(.5)$.
La fonction quantile $\tau \mapsto Q_Y(\tau)$ est croissante, continue à gauche et satisfait pour tout $a > 0$ et b

$$Q_{(aY+b)}(\tau) = aQ_Y(\tau) + b, \quad \tau \in (0,1). \tag{1.1}$$

On n'a pas en général $Q_{Y+Z}(\tau) = Q_Y(\tau) + Q_Z(\tau)$, où Y et Z sont deux variables aléatoires.

Du point de vue optimisation, le τ-ième quantile d'une variable aléatoire Y peut également être obtenu en minimisant l'espérance d'une certaine fonction de perte appelée « check function » définie ci-après.

Définition 1.2.4. ***(Fonction de perte ρ_τ)***

La fonction de perte ρ_τ dont la minimisation de l'espérance conduit au τ-ième quantile d'une variable aléatoire est définie par

$$\rho_\tau(u) = \big(\tau - \mathbb{1}(u < 0)\big)u = \begin{cases} \tau|u| & si\ u \geq 0, \\ (1-\tau)|u| & si\ u \leq 0. \end{cases} \tag{1.2}$$

Elle est linéaire, asymétrique et illustrée par la Figure 1.2 pour différentes valeurs de τ.

De plus, pour tout $\tau \in (0,1)$, la fonction de perte ρ_τ est convexe et vérifie

(i) $\rho_\tau(0) = 0, \quad \lim_{|u|\to\infty} \rho_\tau(u) = \infty.$

(ii) Elle est lipshitzienne de constante

$$L_\tau = \max\{\tau, 1-\tau\},$$

c'est-à-dire

$$|\rho_\tau(u) - \rho_\tau(v)| \leq L_\tau |u - v|, \quad \forall u, v \in \mathbb{R}.$$

(iii) En outre,

$$\min\{\tau, 1-\tau\}|u| \leq \rho_\tau(u) \leq |L_\tau||u|,$$

pour tout $u \in \mathbb{R}$.

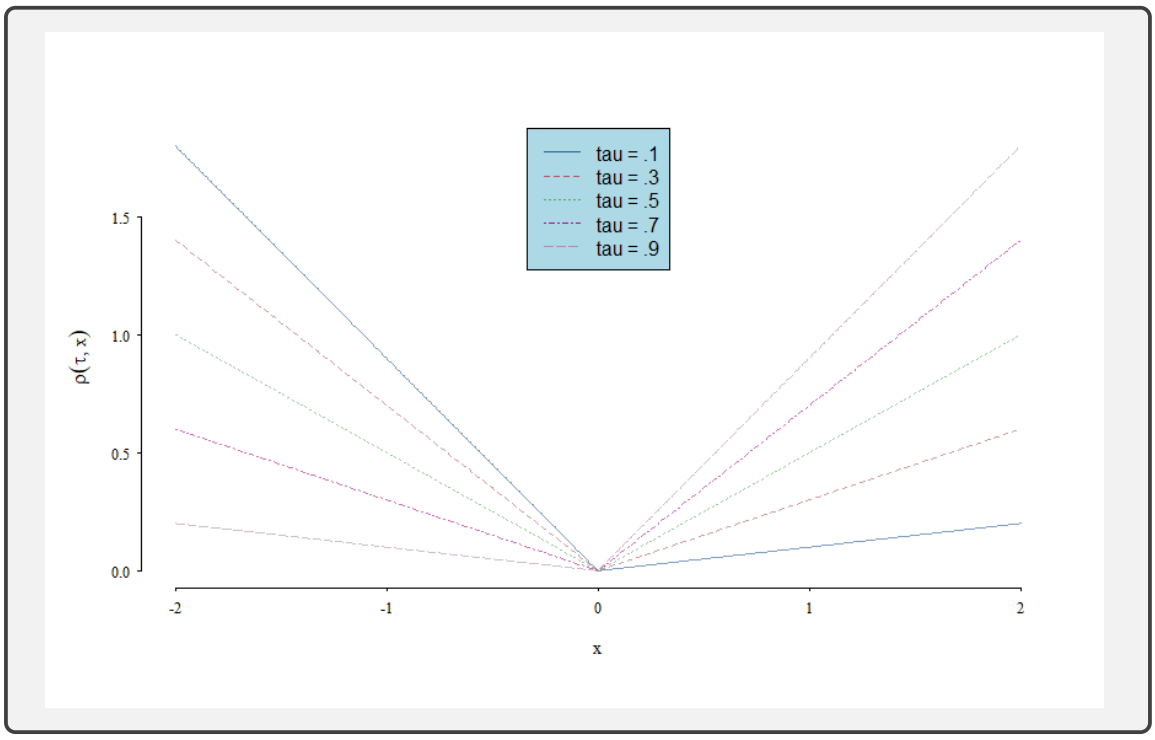

Figure 1.2: Fonction de perte ρ_τ pour différentes valeurs de $\tau \in \{.1, .3, .5, .7, .9\}$.

Proposition 1.2.1. *La fonction de quantiles est solution au problème de minimisation*

$$Q_Y(\tau) = \arg\min_{\theta} E[\rho_\tau(Y - \theta)].$$

Cette propriété est démontrée en appendice à la section A.1.1.

Remarque. *Ce minimum n'est pas nécessairement unique car il pourrait exister plusieurs solutions à l'équation $F_Y(\hat{\theta}) = \tau$. Lorsque la variable aléatoire Y n'est pas continue, l'équation $F_Y(\hat{\theta}) = \tau$ pourrait ne pas avoir de solution et même dans ce cas, $Q_Y(\tau)$ est un minimum de $E_Y[\rho_\tau(Y - \theta)]$.*

Corollaire 1.2.2. *Soit X et Y deux variables aléatoires. La fonction de quantiles conditionnels est solution au problème de minimisation*

$$Q_{Y|X=\boldsymbol{x}}(\tau) = \arg\min_{\theta} E_{Y|X=\boldsymbol{x}}[\rho_\tau(Y - \theta)].$$

Nous avons à présent assez d'outils sur les quantiles pour aborder sereinement la régression quantile linéaire.

1.3 La régression quantile linéaire (Koenker, 2005)

1.3.1 Le modèle

Soit $Y_1, \ldots, Y_n$ un échantillon de variables aléatoires i.i.d. (indépendantes et identiquement distribuées) de distribution inconnue. Soit X le vecteur de taille $p \times 1$ ne contenant pas l'ordonnée à l'origine. On souhaiterait établir une relation linéaire entre le quantile conditionnel d'ordre τ de l'échantillon aléatoire $Y_1, \ldots, Y_n$ étant donné le vecteur $X = \boldsymbol{x}$ en supposant que la fonction quantile a la forme

$$m(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau, \quad \tau \in (0, 1), \tag{1.3}$$

où $\boldsymbol{\beta}_\tau$ est un vecteur de p paramètres. L'équation (1.3) représente le modèle de régression quantile linéaire.

Du Corollaire 1.2.2, nous avons

$$Q_{Y|X=\boldsymbol{x}}(\tau) = \arg\min_{\boldsymbol{\beta}_\tau} E_{Y|X=\boldsymbol{x}}[\rho_\tau(Y - m(\boldsymbol{x}))].$$

1.3.2 Motivations

Dans cette section, nous présentons les raisons qui ont motivé le modèle de régression quantile linéaire. Supposons que $X = (1, X_2, \ldots, X_p)$.

Motivation 1 : Mesure des effets hétérogènes

Les effets des covariables sur la variable réponse ne sont en général pas les mêmes pour tous les individus. Ceci n'est pas pris en compte dans les régressions linéaires standard qui supposent que les résidus sont distribués de façon homogène.

Considérons par exemple le modèle de translation-échelle

$$Y = X^\top \boldsymbol{\beta} + (X^\top \boldsymbol{\gamma})\varepsilon,$$

où ε est indépendant de X, de moyenne nulle et $X^\top \boldsymbol{\gamma} \geq 0$. C'est un modèle hétéroscédastique car la variance des résidus n'est pas constante. D'après l'équation (1.1), on a pour tout $\tau \in (0, 1)$,

$$Q_{Y|X=\boldsymbol{x}}(\tau) = \boldsymbol{x}^\top(\boldsymbol{\beta} + \boldsymbol{\gamma} Q_{\varepsilon|X=\boldsymbol{x}}(\tau)) = \boldsymbol{x}^\top(\boldsymbol{\beta} + \boldsymbol{\gamma} Q_\varepsilon(\tau)).$$

On déduit de l'équation (1.3) [1] que $\boldsymbol{\beta}_\tau = \boldsymbol{\beta} + \boldsymbol{\gamma} Q_\varepsilon(\tau)$. Dans le modèle de type « translation-échelle », en supposant $E(\varepsilon) = 0$, l'estimateur des moindres carrés,

1. Cette équation peut encore s'écrire comme $Q_{Y|X=\boldsymbol{x}}(\tau) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau$.

$\boldsymbol{\beta}_{OLS}$, est tel que $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}$. La méthode des moindres carrés ne capture donc pas tous les effets de X qui ne sont pas les mêmes aux différents quantiles de la variable réponse.

Motivation 2 : Robustesse aux valeurs aberrantes et aux queues épaisses (d'Haultfoeuille et Givord, 2014)

(i) **Robustesse aux queues épaisses**

Dans le même esprit, considérons un modèle linéaire

$$Y = X^\top \boldsymbol{\beta} + \varepsilon,$$

avec ε indépendant de X. Si ε suit une distribution symétrique autour de zéro (la loi normale par exemple), on peut estimer $\boldsymbol{\beta}$ soit par OLS [2], soit par la régression médiane. On préfère le plus souvent utiliser la régression médiane lorsque la distribution de ε a une queue épaisse. En effet, si $E(|\varepsilon|) = \infty$, l'estimateur des moindres carrés n'est pas convergent alors que la médiane est toujours définie. On peut même montrer que l'estimateur issu de la régression médiane dans ce cas est convergent. Cette notion est très utile en finance, assurance, économie, etc.

(ii) **Robustesse aux valeurs aberrantes**

Nous aimerons par exemple effectuer de l'inférence sur une variable Y^* mais ce qu'on observe est plutôt les données contaminées $Y = CX^\top \boldsymbol{\alpha} + (1-C)Y^*$, où $Y^* = X^\top \boldsymbol{\beta} + \varepsilon$, C est une variable latente non observée qui vaut 1 si les données sont contaminées et 0 sinon. On supppose que $p = \Pr(C = 1)$ est petit et $X^\top \boldsymbol{\alpha}$ est très grand.

Si on considère le modèle linéaire $E(Y^*|X = \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$, alors au lieu d'estimer $\boldsymbol{\beta}$, les moindres carrés estiment plutôt $(1-p)\boldsymbol{\beta} + p\boldsymbol{\alpha}$. Le biais $p(\boldsymbol{\alpha} - \boldsymbol{\beta})$

2. Ordinary Least Squared ou les moindres carrés ordinaires.

peut être grand, même si p est petit.

Si par contre on considère le modèle quantile $Q_{Y^*|X=\boldsymbol{x}}(\tau) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau$, alors par définition de Y^* et du quantile τ,

$$\begin{aligned}
\tau &= P(Y \leq X^\top \boldsymbol{\beta}_\tau | X = \boldsymbol{x}) \\
&= P(C = 1 | X = \boldsymbol{x}) P(Y \leq X^\top \boldsymbol{\beta}_\tau | X = \boldsymbol{x}, C = 1) \\
&\quad + P(C = 0 | X = \boldsymbol{x}) P(Y \leq X^\top \boldsymbol{\beta}_\tau | X = \boldsymbol{x}, C = 0) \\
&= pP(X^\top \boldsymbol{\alpha} < X^\top \boldsymbol{\beta}_\tau | X = \boldsymbol{x}) + (1-p) P(X^\top \boldsymbol{\beta} + \varepsilon \leq X^\top \boldsymbol{\beta}_\tau | X = \boldsymbol{x}, C = 0) \\
&= (1-p) P(\varepsilon \leq \boldsymbol{x}^\top (\boldsymbol{\beta}_\tau - \boldsymbol{\beta})),
\end{aligned}$$

où la dernière égalité résulte de l'hypothèse selon laquelle $X^\top \boldsymbol{\alpha}$ est très grand, donc supérieur à $X^\top \boldsymbol{\beta}_\tau$ et ε est indépendant de (C, X). Il s'ensuit pour tout $\boldsymbol{x}$, $\boldsymbol{x}^\top (\boldsymbol{\beta}_\tau - \boldsymbol{\beta}) = Q_\varepsilon \left(\frac{\tau}{1-p} \right)$. On en déduit que $\boldsymbol{\beta}_{k,\tau} = \boldsymbol{\beta}_k$ pour tout $k > 1$ et $\beta_{1,\tau} = \beta_1 + Q_\varepsilon \left(\frac{\tau}{1-p} \right)$. L'estimateur de l'effet de X_k $(k > 1)$ obtenu par une régression quantile vaut bien $\boldsymbol{\beta}_k$. La présence des valeurs aberrantes n'affecte donc pas les résultats de la régression quantile, sauf les coefficients de la constante.

Si on suppose le cas général [3] $Y^* = X^\top \boldsymbol{\beta}_U$ (où U est indépendant de (C, X), de loi uniforme sur $[0, 1]$) telle que la fonction $u \mapsto \boldsymbol{x}^\top \boldsymbol{\beta}_u$ est strictement croissante pour tout $\boldsymbol{x}$, en suivant le même raisonnement que précédemment, on obtient pour tout $\boldsymbol{x}$ l'équation

$$\tau = (1-p) P(\boldsymbol{x}^\top \boldsymbol{\beta}_U < \boldsymbol{x}^\top \boldsymbol{\beta}_\tau). \tag{1.4}$$

L'unique solution vérifiant (1.4) est $\hat{\boldsymbol{\beta}}_\tau = \boldsymbol{\beta}_{\frac{\tau}{1-p}}$ si $\frac{\tau}{1-p} \in (0, 1)$ et le paramètre identifié . Cet estimateur ne dépend pas de $\boldsymbol{\alpha}$ et est assez proche de $\boldsymbol{\beta}_\tau$ lorsque la proportion des valeurs aberrantes p est faible. De plus,

3. Dans ce modèle, U peut s'interpréter comme une composante individuelle inobservée qui positionne l'individu dans la distribution de Y^*.

s'il existe des composantes de $\boldsymbol{\beta}_\tau$ indépendants de τ (effets homogènes), la contamination n'affectera pas les estimations.

1.3.3 Estimation des paramètres du modèle

Définition de l'estimateur (version empirique)

Supposons que

$$Q_{Y|X=\boldsymbol{x}}(\tau) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau, \quad \tau \in (0,1).$$

Alors, d'après la Proposition 1.2.1,

$$\hat{\boldsymbol{\beta}}_\tau \in \arg\min_{\boldsymbol{\beta}_\tau} \; E[\rho_\tau(Y - \boldsymbol{x}^\top \boldsymbol{\beta}_\tau)|X = \boldsymbol{x}].$$

La version empirique de l'estimateur $\hat{\boldsymbol{\beta}}_\tau$ est donnée par l'équation

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\tau &\in \arg\min_{\boldsymbol{\beta}_\tau} \; \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau) \\ &= \arg\min_{\boldsymbol{\beta}_\tau} \; \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau), \end{aligned} \tag{1.5}$$

où $\{(y_i, x_i), i = 1, \ldots, n\}$ est un échantillon observé et ρ_τ la fonction de perte définie à l'équation (1.2).

Remarque. *À la médiane, c'est-à-dire à $\tau = 1/2$, résoudre le problème de minimisation (1.5) équivaut à résoudre le problème*

$$\arg\min_{\boldsymbol{\beta}_\tau} \; \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau|.$$

La solution à ce problème est appelée l'estimateur de LAD, « Least Absolute Deviation ».

Le problème de régression quantile peut être résolu de façon efficiente en utilisant des algorithmes de programmation linéaire (Koenker, 2005). La méthode est implémentée dans `R` avec la fonction `rq`.

1.3.4 Propriétés d'équivariance de $\hat{\boldsymbol{\beta}}_\tau$

Le paramètre estimé du modèle de régression quantile linéaire (1.5) vérifie les propriétés d'équivariance énoncées dans (Koenker et Bassett Jr, 1978) au Théorème 3.2. Ainsi, si $\boldsymbol{X}$ est la matrice de design de dimension $n \times p$ et Y la variable d'intérêt, alors pour tout $a \geq 0$ et $\tau \in (0,1)$, nous avons

1. « Scale Equivariance » :
   $\hat{\boldsymbol{\beta}}_\tau(aY, \boldsymbol{X}) = a\hat{\boldsymbol{\beta}}_\tau(Y, \boldsymbol{X}) \quad \text{et} \quad \hat{\boldsymbol{\beta}}_{1-\tau}(-aY, \boldsymbol{X}) = -a\hat{\boldsymbol{\beta}}_\tau(Y, \boldsymbol{X})\,;$
2. « Regression Shift » : pour tout $\boldsymbol{\gamma} \in \mathbb{R}^p$,
   $$\hat{\boldsymbol{\beta}}_\tau(Y + \boldsymbol{X}\boldsymbol{\gamma}, \boldsymbol{X}) = \hat{\boldsymbol{\beta}}_\tau(Y, \boldsymbol{X}) + \boldsymbol{\gamma};$$
3. « Reparameterization of Design » : pour toute matrice régulière $\boldsymbol{A}$ de dimension $p \times p$,
   $$\hat{\boldsymbol{\beta}}_\tau(Y, \boldsymbol{X}\boldsymbol{A}) = \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}_\tau(Y, \boldsymbol{X}).$$

D'autre part, si g est une fonction croissante et continue à gauche sur $\mathbb{R}$, alors nous avons la **propriété d'équivariance par transformation monotone**

$$Q_{g(Y)}(\tau) = g(Q_Y(\tau)).$$

Cette propriété est démontrée en appendice à la section A.1.2 et permet de construire à la section qui suit, lorsque $g(x) = \mathbb{1}\{x > 0\}$, le modèle de Kordas.

1.4 Régression quantile binaire linéaire (Kordas, 2006)

1.4.1 Le modèle

Considérons le modèle

$$\begin{aligned} Y^* &= X^\top \boldsymbol{\beta} + U \\ Y &= \mathbb{1}\{Y^* \geq 0\}, \end{aligned} \tag{1.6}$$

où Y^* est une variable latente continue (non observable) de distribution supposée normale, Y une variable indicatrice observable, X un vecteur de taille $p \times 1$ de variables explicatives observables, $\boldsymbol{\beta}$ un vecteur de paramètres de taille $p \times 1$ et U un bruit non observable. Le modèle de régression quantile linéaire latent (Koenker et Bassett Jr, 1978) a pour expression

$$Q_{Y^*|X=\boldsymbol{x}}(\tau) := F^{-1}_{Y^*|X=\boldsymbol{x}}(\tau) = \boldsymbol{x}^\top \boldsymbol{\beta}_\tau, \quad \tau \in (0,1), \tag{1.7}$$

où $Q(.)$ et $F(.)$ sont respectivement le quantile et la *cdf* conditionnels de la variable latente Y^*. Comme on n'observe pas la variable Y^*, (1.6) n'est plus utile. Le modèle de régression quantile binaire est obtenu en appliquant à l'équation (1.7) la propriété d'équivariance par transformation monotone avec $g(x) = \mathbb{1}\{x \geq 0\}$. On obtient ainsi le modèle de score maximum (Manski, 1985)

$$Q_{Y|X=\boldsymbol{x}}(\tau) = \mathbb{1}\{\boldsymbol{x}^\top \boldsymbol{\beta}_\tau \geq 0\}. \tag{1.8}$$

D'après la Proposition 1.2.1, l'estimateur de la régression quantile binaire linéaire (1.8) est la solution au problème de minimisation

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1\}}{\arg\min} \left\{ L_{n\tau}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\}) \right\}, \tag{1.9}$$

où $\|.\|$ est la norme euclidienne. Notons que l'échelle du paramètre $\boldsymbol{\beta}$ n'est pas identifiée car si $X^\top \boldsymbol{\beta} > 0$, alors $aY^* = aX^\top \boldsymbol{\beta} > 0$ pour tout $a > 0$. Pour normaliser l'échelle, deux méthodes sont couramment utilisées. La première (Manski, 1985) consiste à imposer la restriction $\|\boldsymbol{\beta}\| = 1$ comme dans (1.9) et la seconde (Horowitz, 1992), à fixer une seule coordonnée du vecteur $\boldsymbol{\beta}$, $|\boldsymbol{\beta}_k| = 1$, $k \in \{1, \ldots, p\}$.

1.4.2 Motivations

En plus d'avoir les mêmes motivations que celles de la régression quantile, le modèle de régression quantile binaire peut être très utile dans les situations où

le nombre d'observations est beaucoup plus grand dans un groupe que dans un autre. Par exemple, dans les cas des maladies rares, le nombre de cas est plus petit que le nombre de contrôles. Dans les études de chômage aussi, le nombre de travailleurs est bien plus grand que le nombre de chômeurs. Face à de telles situations, $\boldsymbol{x}_i^\top \boldsymbol{\beta}_{.5} < 0$ pour une grande portion d'individus dans l'échantillon et $\boldsymbol{x}_i^\top \boldsymbol{\beta}_{.5} > 0$ pour peu d'individus [4]. Dans ces cas, il est naturel de spécifier un modèle par rapport à un quantile autre que la médiane, afin de permettre davantage de croisements et d'établir l'identification.

La régression quantile binaire et la régression quantile binaire lissée sont des solutions aux problèmes d'optimisation assez difficiles qui nécessitent l'utilisation d'algorithmes coûteux. Ces difficultés émanent de ce que la fonction objective lissée est le plus souvent une fonction multimodale et celle non lissée constante par morceaux.

1.4.3 Estimation des coefficients

Manski (1985) a défini l'estimateur du score maximum du quantile d'ordre τ

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1\}}{\arg\max} \left\{ S_{n\tau}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big[y_i - (1-\tau)\big] \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} \right\} \tag{1.10}$$

qui est équivalent à l'estimateur de la régression quantile binaire (1.9).

Proposition 1.4.1. *Les problèmes d'optimisation (1.9) et (1.10) sont équivalents.*

Cette proposition est démontrée en appendice à la section A.1.3.

(Manski, 1985) a montré après avoir étudié les propriétés asymptotiques de l'estimateur $\hat{\boldsymbol{\beta}}_\tau$ qu'il est convergent sous des conditions faibles, mais très coûteux

4. Beaucoup plus de 1 que de 0 dans l'échantillon de la variable binaire.

en termes de calcul. La cause principale de ce désagrément proviendrait du fait que la fonction objective $S_{n\tau}(\boldsymbol{\beta})$ n'est pas différentiable [5]. (Horowitz, 1992), pour pallier ces problèmes, a proposé la fonction de score lissée et a défini l'estimateur du maximum du score comme la solution du problème

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1\}}{\arg\max} \left\{ S_{n\tau}(\boldsymbol{\beta}, h_n) = \sum_{i=1}^{n} \big[y_i - (1-\tau)\big] K\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau}{h_n}\right) \right\}, \tag{1.11}$$

où $K(.)$ est un noyau et h_n une suite de réels positifs (non aléatoires) qui converge vers 0 quand n est grand, encore appelé le paramètre de lissage du noyau. La nouvelle fonction objective est continue et différentiable en $\boldsymbol{\beta}$ et le problème converge rapidement au quantile d'ordre .5, du moins plus vite que le problème (1.9). Kordas vient étendre ce problème à une famille de quantiles autres que la médiane.

Remarque. ***(Relation entre l'équation (1.10) et le taux d'erreur de classification)***

Le problème de maximisation (1.10) entretient une relation étroite avec le classifieur $\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\}$. *En effet, on a la relation*

$$\begin{aligned} S_{n\tau}(\boldsymbol{\beta}) &= \tau \sum_{i=1}^{n} \mathbb{1}\{y_i = 1\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} - (1-\tau) \sum_{i=1}^{n} \mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} \\ &= \tau TP - (1-\tau)FP, \end{aligned}$$

où TP *et* FP *sont respectivement le taux de vrais positifs et le taux de faux positifs du classifieur* $\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\}$. *Clairement, nous avons*

$$S_{n\tau}(\boldsymbol{\beta}) \approx \begin{cases} TP & \textit{quand } \tau \to 1, \\ -FP & \textit{quand } \tau \to 0. \end{cases}$$

Ainsi maximiser $S_{n\tau}$ *revient à maximiser* TP *quand* τ *est grand et minimiser* FP *quand* τ *est petit.*

---

5. Elle n'est en effet pas différentiable à cause de la fonction indicatrice.

Après avoir estimé les paramètres, (Kordas, 2006) veut calculer la probabilité de succès conditionnelle qui n'admet pas d'expression analytique. Pour le faire, il va l'approximer en calculant les estimateurs des intervalles qui contiennent les probabilités de succès conditionnelles de chaque individu.

1.4.4 Calcul de la probabilité de succès conditionnelle

Dans cette partie, on calcule la probabilité de succès conditionnelle telle que définie par (Kordas, 2006).

Soit $P_{1|X} = P(Y = 1|X) := \int_{\mathbb{R}} \mathbb{1}\{X^\top \boldsymbol{\beta} + U \geq 0\} dF_{U|X}$ la probabilité de succès conditionnelle étant donné X. Partant de la définition du quantile, nous avons $P(Y^* \geq X^\top \boldsymbol{\beta}_\tau | X) = 1 - \tau$, ce qui implique pour tout X,

$$X^\top \boldsymbol{\beta}_\tau \quad \begin{matrix} \leqq \\ > \end{matrix} \quad 0 \iff P_{1|X} \quad \begin{matrix} \leqq \\ > \end{matrix} \quad 1 - \tau,$$

relation qui en terme de la variable binaire observée peut encore s'écrire sous la forme

$$P\left(Y = 1 \middle| X^\top \boldsymbol{\beta}_\tau \quad \begin{matrix} \leqq \\ > \end{matrix} \quad 0 \right) \quad \begin{matrix} \leqq \\ > \end{matrix} \quad 1 - \tau. \tag{1.12}$$

L'équation (1.12) est équivalente au système d'équations

$$P(Y = 1|X = \boldsymbol{x}) \begin{cases} > 1 - \tau & \text{si } \boldsymbol{x}^\top \boldsymbol{\beta}_\tau > 0 \ , \\ = 1 - \tau & \text{si } \boldsymbol{x}^\top \boldsymbol{\beta}_\tau = 0 \ , \\ < 1 - \tau & \text{si } \boldsymbol{x}^\top \boldsymbol{\beta}_\tau < 0 \ . \end{cases} \tag{1.13}$$

Il en découle que, un quantile, disons la médiane, sépare l'échantillon en deux groupes : ceux qui sont en dessous de la médiane conditionnelle et ceux qui y sont au dessus. Afin d'obtenir une image plus complète, il serait donc naturel de regarder sur une famille de quantiles estimés et d'obtenir une caractérisation

en termes de quantiles conditionnels. Ainsi, un changement de variables $U \mapsto F^{-1}_{U|X}(\tau|X)$ conduit à

$$\begin{aligned} P_{1|X} &= \int_{\mathbb{R}} \mathbb{1}\{X^\top \boldsymbol{\beta} + U \geq 0\} dF_{U|X}(U|X) \\ &= \int_0^1 \mathbb{1}\{X^\top \boldsymbol{\beta} + F^{-1}_{U|X}(\tau|X) \geq 0\} dF_{U|X}(F^{-1}_{U|X}(\tau|X)) \\ &= \int_0^1 \mathbb{1}\{X^\top \boldsymbol{\beta} + F^{-1}_{U|X}(\tau|X) \geq 0\} d\tau \\ &= \int_0^1 \mathbb{1}\{X^\top \boldsymbol{\beta} + Q_{U|X}(\tau) \geq 0\} d\tau \\ &= \int_0^1 \mathbb{1}\{X^\top \boldsymbol{\beta}_\tau \geq 0\} d\tau, \end{aligned}$$

qui désigne la probabilité de succès en fonction du processus quantile $\boldsymbol{\beta}_\tau, \tau \in (0,1)$. Mais compte tenu de la difficulté dans le calcul de l'estimation de l'ensemble du processus $\boldsymbol{\beta}_\tau$, (Kordas, 2006) convient d'approximer plutôt cette probabilité conditionnelle sur une grille de quantiles estimés. Il procède comme suit : après avoir estimé le vecteur de paramètres $\hat{\boldsymbol{\beta}}_\tau$ sur la grille de quantiles $\theta = \{\tau_1, \ldots, \tau_m : \tau_1 < \cdots < \tau_m\}$, on détermine la quantité

$$\hat{\tau}_i = \arg\min_{\tau \in \theta} \{\tau : \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau \geq 0\}, \quad i = 1, \ldots, n.$$

Un estimateur de l'intervalle qui contient la probabilité de succès conditionnelle de l'individu i est donnée par l'équation

$$\hat{P}_{i,1|X_i=\boldsymbol{x}_i}(\theta) = [1 - \hat{\tau}_i, 1 - \hat{\tau}_{i,-1}),$$

où $\hat{\tau}_{i,-1}$ est le quantile qui précède $\hat{\tau}_i$.

Dans ce qui suit, vous trouverez l'algorithme qui sert à calculer cette probabilité de succès conditionnelle ainsi qu'une figure illustrative.

Algorithme 1 Estimateur de l'intervalle de la probabilité de succès conditionnelle de chaque individu

a. Estimer avec la méthode de Kordas le vecteur de paramètres $\hat{\boldsymbol{\beta}}_\tau$ sur la grille de quantiles $\theta = \{\tau_1, \ldots, \tau_m : \tau_1 < \cdots < \tau_m\}$;

b. Calculer pour chaque individu i la quantité

$$\hat{\tau}_i = \arg\min_{\tau \in \theta} \{\tau : \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau \geq 0\}, \quad i = 1, \ldots, n;$$

c. Un estimateur de l'intervalle qui contient la probabilité de succès conditionnelle de l'individu i est

$$\hat{P}_{i,1|X_i=\boldsymbol{x}_i}(\theta) = [1 - \hat{\tau}_i, 1 - \hat{\tau}_{i,-1}),$$

où $\hat{\tau}_{i,-1}$ est le quantile qui précède directement $\hat{\tau}_i$;

d. De plus, si $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau \geq 0$ pour tout i, alors l'intervalle est

$$[\tau_m, 1);$$

si par contre $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau < 0$ pour tout i, l'intervalle est

$$(0, \tau_1].$$

Par exemple, admettons que l'on effectue des estimations sur la grille de quantiles $\theta = \{0.05, 0.1, \ldots, 0.95\}$ et que pour l'individu i, $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau$ est positif si $\tau \geq 0.4$ et négatif si $\tau < 0.4$. Alors, $\hat{\tau}_i = 0.40$ et $\hat{P}_{i,1|X_i}(\theta) = [0.60, 0.65)$. Si $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau$ est positif pour tous les quantiles, alors, la probabilité de succès conditionnelle de l'individu i appartient à l'intervalle $[.95, 1)$. Si par contre $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau$ est négatif pour tous les quantiles, cette probabilité est dans l'intervalle $(0, 0.05]$.

1.5 Les variantes de la méthode de Kordas

Nous présentons dans cette section deux variantes de la méthode de Kordas.

1.5.1 Une variante de la méthode de Kordas

La fonction ρ_τ - définie par l'équation (1.2) - n'est pas différentiable en 0. Une variante de la méthode de Kordas consiste à la lisser à l'origine par une autre fonction continûment différentiable sur tout son support, puis de résoudre le problème d'optimisation obtenu par l'une des méthodes adéquates connues.

Tout récemment, afin de lisser la fonction ρ_τ à l'origine, (Jennings *et al.*, 1996) et (Aravkin *et al.*, 2014) ont respectivement proposé les fonctions $\rho_{\tau,c}$ et $\rho_{\tau,k}$ définies par

$$\rho_{\tau,c}(u) = \begin{cases} (\tau-1)(u+.5c) & \text{si } u < -c, \\ .5(1-\tau)u^2/c & \text{si } -c \leq u < 0, \\ .5\tau u^2/c & \text{si } 0 \leq u < c, \\ \tau(u-.5c) & \text{si } c \leq u, \end{cases}$$

et

$$\rho_{\tau,k}(u) = \begin{cases} (\tau-1)|u| - \frac{k(1-\tau)^2}{2} & \text{si } u < -(1-\tau)k, \\ \frac{1}{2k}u^2 & \text{si } -(1-\tau)k \leq u < \tau k, \\ \tau|u| - \frac{k\tau^2}{2} & \text{si } \tau k \leq u. \end{cases}$$

Ces fonctions sont continûment différentiables, ont le même minimum que la fonction ρ_τ et possèdent de bonnes propriétés démontrées par (Mkhadri *et al.*, 2017) et énoncées comme suit :

(i) pour tout $u \in \mathbb{R}$,

$$|\rho_{\tau,c}(u) - \rho_\tau(u)| \leq \frac{c}{2}\sup(\tau, 1-\tau),$$

$$|\rho_{\tau,k}(u) - \rho_\tau(u)| \leq \frac{k}{2} \sup\left(\tau^2, (1-\tau)^2\right);$$

(ii) pour des valeurs de τ, c et k fixées, les fonctions $\rho_{\tau,c}$ et $\rho_{\tau,k}$ sont convexes ;

(iii) les fonctions $\rho_{\tau,c}$ et $\rho_{\tau,k}$ sont continûment différentiables et de gradients lipschitziens, c'est-à-dire

$$\left|\rho'_{\tau,c}(u) - \rho'_{\tau,c}(v)\right| \leq \frac{\sup(\tau, 1-\tau)}{c} |u - v|,$$

$$\left|\rho'_{\tau,k}(u) - \rho'_{\tau,k}(v)\right| \leq \frac{1}{k} |u - v|.$$

En utilisant ces fonctions de lissage, l'estimateur du paramètre $\boldsymbol{\beta}_\tau$ est la solution au problème d'optimisation

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1\}}{\arg\min} \left\{ L_{n\tau}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \rho_{\tau,*}\left(y_i - \mathbb{1}\{x_i^\top \boldsymbol{\beta}_\tau \geq 0\}\right) \right\},$$

où $\|.\|$ est la norme euclidienne et $* \in \{c, k\}$.

Ces deux fonctions lissées sont explorées au Chapitre 3 afin de résoudre le modèle de quantile binaire à base des copules.

1.5.2 Une autre variante de la méthode de Kordas

Cette autre alternative a été proposée par (Aristodemou *et al.*, 2019). Selon eux, l'estimateur de l'équation (1.5) peut être vu comme le τ-ième quantile du modèle de régression binaire linéaire général (Koenker et Bassett Jr, 1978), qui est obtenu en résolvant le problème d'optimisation

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:|\beta_1|=1\}}{\arg\min}\, \mathcal{R}(\boldsymbol{\beta}), \tag{1.14}$$

où

$$\mathcal{R}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(\tau) \left|y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\}\right|,$$

et

$$w_i(\tau) = \begin{cases} \tau & \text{si } y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} \geq 0, \\ 1-\tau & \text{si } y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} < 0. \end{cases} \tag{1.15}$$

Lorsque u et v sont binaires (0 ou 1), on a l'égalité $|u-v| = (u-v)^2$. Le problème d'optimisation (1.14) est donc équivalent au problème

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\{\boldsymbol{\beta}:|\beta_1|=1\}}{\arg\min} \sum_{i=1}^{n} w_i(\tau) \times \left(y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau \geq 0\} \right)^2,$$

où $\hat{\boldsymbol{\beta}}_\tau = (1, \hat{\boldsymbol{\beta}}_{\tau,-1}^\top)$, $\hat{\boldsymbol{\beta}}_{\tau,-1}^\top$ étant le vecteur $\hat{\boldsymbol{\beta}}_\tau$ privé de sa première composante. Comme Kordas, (Aristodemou *et al.*, 2019) vont obtenir un nouveau problème en lissant la fonction indicatrice avec le noyau gaussien de paramètre de lissage h_n :

$$\hat{\boldsymbol{\beta}}_{slaws}(\tau) = \underset{\{\boldsymbol{\beta}:|\beta_1|=1\}}{\arg\min} \sum_{i=1}^{n} w_i(\tau) \times \left(y_i - \boldsymbol{\Phi}\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau}{h_n} \right) \right)^2, \tag{1.16}$$

où $\boldsymbol{\Phi}$ est le noyau gaussien, h_n un paramètre de lissage, $\hat{\boldsymbol{\beta}}_{slaws}(\tau) = (1, \hat{\boldsymbol{\beta}}_{\tau,-1}^\top)$ tel que $\hat{\boldsymbol{\beta}}_{\tau,-1}^\top$ est le vecteur $\hat{\boldsymbol{\beta}}_{slaws}(\tau)$ privé de sa première composante, et

$$w_i(\tau) = \begin{cases} \tau & \text{si } y_i - \boldsymbol{\Phi}\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau}{h_n} \right) \geq 0, \\ 1-\tau & \text{si } y_i - \boldsymbol{\Phi}\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{\beta}_\tau}{h_n} \right) < 0. \end{cases} \tag{1.17}$$

L'**algortihme 2** qu'on peut retrouver dans (Aristodemou *et al.*, 2019) permet de résoudre le problème (1.16) avec la méthode de LAWS, « Non Linear Least Asymmetric Weighted squares ». Nous avons exploré la méthode sur le modèle de régression quantile binaire à base des copules, mais nous n'avons obtenu aucun résultat satisfaisant.

Algorithme 2 Régression quantile binaire linéaire à base de la méthode LAWS non linéaire.

a. Calculer le paramètre initial des paramètres $\boldsymbol{\beta}$ par la méthode des moindres carrés non linéaire ;

b. Calculer le paramètre initial des résidus $\varepsilon_i^0 = y_i - \Phi\left(\frac{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\tau}{h_n}\right)$;

c. Construire les poids $w_i^0(\tau)$ en utilisant l'équation (1.17) et estimer l'équation (1.16) à partir de la méthode de régression non linéaire pondérée des moindres carrés ;

d. Calculer les nouveaux estimateurs des résidus, $\varepsilon_i^1 = y_i - \Phi\left(\frac{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{slaws}(\tau)}{h_n}\right)$;

e. Mettre à jour les poids et obtenir $w_i^1(\tau)$ à partir de l'équation (1.17) ;

f. Estimer l'équation (1.16) à partir de la méthode de régression non linéaire pondérée des moindres carrés ;

g. Répéter les étapes d à f jusqu'à convergence.

CHAPITRE II

LA THÉORIE DES COPULES

2.1 Bref historique

Le terme copule vient du nom latin *copulæ* qui étymologiquement signifie lien, liaison, alliance ou union. Selon (Tibiletti, 1995), la notion des copules tire ses origines des travaux de Fréchet sur les espaces métriques probabilisés. En effet, l'article de (Fréchet, 1951) est un point de départ sur la notion de copule. Cependant, (Sklar *et al.*, 1959) ont été les premiers à formuler mathématiquement de façon rigoureuse le concept de copule dans un théorème qui porte son nom. La paternité des copules est ainsi consensuellement attribuée dans la littérature scientifique à Sklar. Le théorème de Sklar qui permet d'expliquer la relation entre une copule et un vecteur de variables aléatoires est incontournable dans la théorie des copules.

Toutefois, (Schweizer, 1991) souligne que les travaux de Wassily Hoeffding publiés en 1940 contiennent déjà de nombreux résultats de base sur les copules. (Nelsen, 2007) explique que les travaux de Hoeffding sont restés dans l'ombre, peu connus de la communauté scientifique en raison de leur parution dans une revue allemande à petit tirage au début de la première guerre mondiale. Et Fréchet est parvenu sans le savoir aux mêmes résultats une décennie plus tard.

2.2 Motivations

Le coefficient de corrélation linéaire de Pearson est un indicateur qui mesure la dépendance entre deux variables aléatoires. Cette mesure de dépendance, relativement facile à calculer vit dans le segment $[-1, 1]$. C'est un indicateur qui est performant lorsque la relation de dépendance est linéaire et l'univers considéré est gaussien. Dans ce cas, un coefficient de corrélation nul traduit l'indépendance entre les variables. Le coefficient de corrélation linéaire est un outil très utile pour les distributions provenant des familles elliptiques. Cependant, les praticiens se sont heurtés à plusieurs problèmes inhérents à son utilisation : un coefficient de corrélation nul pour des variables non gaussiennes impliquerait-il l'absence de dépendance entre celles-ci ? Comment mesurer la dépendance entre plus de deux variables, quand bien même celles-ci sont normales ? C'est à ce niveau que le coefficient de corrélation linéaire trouve ses limites.

Dans plusieurs domaines de la statistique, en occurence dans le cadre de notre travail, on n'utilise pas toujours le cas gaussien et on travaille la plupart du temps en dimension supérieure à deux. Un autre moyen de mesure de la dépendance entre des variables existe. Cette méthode - solution aux limites du coefficient de corrélation linéaire - permet d'agréger les lois marginales des variables aléatoires afin d'obtenir la loi jointe. Contrairement au coefficient de corrélation, elle est invariante par transformations strictement croissantes et est connue sous le nom de copule.

Les copules constituent un outil statistique qui permet de modéliser la dépendance entre des variables aléatoires. La fonction copule, comme nous le verrons plus tard, relie en effet la densité jointe aux densités marginales et contient ainsi toute l'information sur la structure de dépendance entre les variables aléatoires. Cette caractérisation de la copule est significativement importante dans l'étude

de la modélisation de la dépendance dans différents domaines de la statistique, en l'occurrence et non exhaustivement en finance, économie, biologie, génétique, actuariat pour n'en nommer quelques uns.

Dans la littérature, il existe plusieurs définitions des copules. Bien que les notations diffèrent selon les domaines, on emploie généralement une définition commune et on exploite les mêmes caractéristiques.

2.3 Définition et propriétés d'une copule

Dans l'introduction de son livre, (Nelsen, 2007) décrit les copules comme des *fonctions qui joignent des distributions multivariées à leurs lois marginales unidimensionnelles.*

(Embrechts, 2009) explique qu'une copule peut être définie de deux façons équivalentes :

- une copule est une fonction qui transforme un hypercube en un intervalle unitaire ;
- une copule est une *cdf* multivariée dont les lois marginales sont uniformes.

Il souligne qu'il est bon d'interpréter une copule comme la structure de dépendance d'une distribution jointe.

Quant à Sklar, *nous appellerons copule (à n dimensions) toute fonction C continue et non décroissante - au sens employé pour une cdf à n dimensions - définie sur le produit cartésien de n intervalles fermés $I = [0, 1]$ et satisfaisant aux conditions $C(0, \ldots, 0) = 0$ et $C(1, \ldots, 1, \mu, 1 \ldots, 1) = \mu$.*

Il découle de la définition de Sklar certaines propriétés de base sur les copules qui font office de définition d'une copule.

Remarque. *Pour faciliter la lecture, nous allons nous restreindre à la dimension*

$d = 2$ (Fathia, 2018), tous les résultats pouvant être généralisés en dimension $d \geq 2$ quelconque.

Définition 2.3.1. ***(Définition de la copule)***

On appelle copule 2-dimensionnelle ou copule bivariée toute fonction C de I^2 dans I possédant les propriétés suivantes :

(i) Pour tout u et v dans I, on a

$$\begin{cases} C(u,0) = C(0,v) = 0, \\ \\ C(u,1) = u \text{ et } C(1,v) = v. \end{cases}$$

(ii) La copule C est 2-croissante, c'est-à-dire pour tout $u_1, u_2, v_1, v_2 \in I$ tel que $u_1 \geq u_2$ et $v_1 \geq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Proposition 2.3.1. ***(Invariance fonctionnelle)***

Soit X et Y deux variables aléatoires de marges F et G et de copule $C_{X,Y}$. Si α et β sont deux fonctions strictement croissantes, alors

$$C_{\alpha(X),\beta(Y)} = C_{X,Y}.$$

On dit que la copule $C_{X,Y}$ est invariante par transformations strictement croissantes des variables aléatoires.

Cette proposition est démontrée en appendice à la section A.2.1.

Proposition 2.3.2. ***(Différentiabilité)***

Soit C une copule bivariée. Pour tout $v \in I$, la dérivée partielle de la copule par rapport à sa première composante $\dfrac{\partial C}{\partial u}$ existe presque sûrement pour tout $u \in I$ et vérifie

$$\frac{\partial}{\partial u} C(u,v) \in I.$$

Il en est de même de la dérivée partielle de la copule C par rapport à la deuxième composante v. En plus, les fonctions $u \to \dfrac{\partial}{\partial u}C(u,v)$ et $v \to \dfrac{\partial}{\partial v}C(u,v)$ sont croissantes presque partout sur I.

En résumé, les copules sont des fonctions non décroissantes, différentiables et donc continues. De plus, la structure de dépendance représentée par une copule est invariante par transformation strictement croissante et continue des marges.

2.3.1 Théorème de Sklar (1959)

Le théorème de Sklar est fondamental dans la théorie des copules. Il indique une façon plutôt simple d'analyser la structure de dépendance des distributions multivariées sans se soucier des distributions marginales et est à la base de la presque quasi totalité des applications des copules.

Proposition 2.3.3. ***(Théorème de Sklar)***
Soit H une cdf de dimension 2 dont les lois marginales sont F_1 et F_2. Il existe une copule $C : I^2 \to I$ telle que pour tout $x_i \in \mathbb{R},\ i = 1, 2$, on a

$$H(x_1, x_2) = C(F_1(x_1), F_2(x_2)). \tag{2.1}$$

Si F_1 et F_2 sont continues, alors la copule C est unique.

Pour démontrer le théorème de Sklar en appendice à la section A.2.2, nous utilisons le résultat suivant.

Remarque. *Soit X une variable aléatoire de cdf F. Désignons par F^{-1} l'inverse généralisé de F.*

1. *La fonction F^{-1} est croissante et continue à gauche. De plus, pour tout $x \in \mathbb{R}$ et $p \in (0, 1)$,*

$$F(x) \geq p \iff x \geq F^{-1}(p).$$

2. Si F est continue, alors la variable $F(X)$ est de loi uniforme sur $(0,1)$.

Corollaire 2.3.4. ***(Inverse du Théorème de Sklar)***
Soit H une cdf de dimension 2 dont les lois marginales sont F_1 et F_2. Alors, pour tout $u_i \in I,\ i = 1,2$, la copule associée à H a pour expression

$$C(u_1, u_2) \quad = \quad H(F_1^{-1}(u_1), F_2^{-1}(u_2)).$$

Ces théorèmes montrent comment construire un modèle bivarié en fonction des distributions marginales lorsque celles-ci sont connues et inversément.

Exemple. *On aimerait déterminer l'expression de la copule sous-jacente à la distribution logistique bivariée de Gumbel dont l'expression est*

$$H(x_1, x_2) = (1 + e^{-x_1} + e^{-x_2})^{-1}.$$

Les marges de cette distribution sont

$$\begin{cases} F_1(x_1) = \lim_{x_2 \to \infty} H(x_1, x_2) = (1 + e^{-x_1})^{-1}, \\ \\ F_2(x_2) = \lim_{x_1 \to \infty} H(x_1, x_2) = (1 + e^{-x_2})^{-1}. \end{cases}$$

Ces marges admettent pour inverses les fonctions

$$\begin{cases} F_1^{-1}(u) = -\ln\left(\dfrac{1}{u} - 1\right), \\ \\ F_2^{-1}(v) = -\ln\left(\dfrac{1}{v} - 1\right). \end{cases}$$

D'après le Corollaire 2.3.4, on a

$$\begin{aligned} C(u,v) &= H(F_1^{-1}(u), F_2^{-1}(v)) \\ &= \left\{1 + e^{\ln(\frac{1}{u} - 1)} + e^{\ln(\frac{1}{v} - 1)}\right\}^{-1} \\ &= \left\{\frac{1}{u} + \frac{1}{v} - 1\right\}^{-1} \\ &= \frac{uv}{u + v - uv}. \end{aligned}$$

2.3.2 Les bornes de Fréchet-Hoeffding

Hoeffding en 1940 et Fréchet une dizaine d'années plus tard en 1951 ont souligné les meilleures bornes possibles des fonctions de dépendance. Ces quantités qui portent le nom de bornes de Fréchet-Hoeffding sont aussi d'un grand intérêt dans l'étude des copules.

Définition 2.3.2. ***(Ordre partiel)***
Soit C_1 et C_2 deux copules. On dit que C_1 est plus petite que C_2 et on note $C_1 \prec C_2$ si

$$C_1(u, v) < C_2(u, v) \quad \forall u, v \in I.$$

L'ordre est partiel parce qu'on ne peut pas comparer toutes les copules entre elles. On a néanmoins toujours l'inégalité de Fréchet-Hoeffding, démontrée à la section A.2.3 de l'appendice, telle qu'énoncée comme suit.

Proposition 2.3.5. ***(L'inégalité de Fréchet-Hoeffding)***
L'inégalité suivante est appelée « the Fréchet-Hoeffding bounds inequality » et représente la borne supérieure et la borne inférieure d'une copule (Nelsen, 2007). Pour toute copule C, on a pour tout $u, v \in I$ l'inégalité

$$W := C^-(u, v) \leq C(u, v) \leq M := C^+(u, v), \tag{2.2}$$

où $W = \max(u + v - 1, 0)$ et $M = \min(u, v)$.

W et M sont des copules [1] et représentent respectivement la plus petite copule et la plus grande copule parmi la famille de toutes les copules.

La borne inférieure W d'une copule décrit une dépendance fonctionnelle négative parfaite. Parallèlement, la borne supérieure M représente une dépendance positive

1. En dimension supérieure à 2, la borne W n'est plus une *cdf*.

parfaite. Ces bornes permettent d'introduire le concept de concordance ou ordre partiel de concordance, car toutes les copules ne sont pas comparables deux à deux.

Avant d'aborder les notions de mesures d'association et de concordance, il est important de définir le concept de la densité d'une copule.

2.3.3 Densité d'une copule

Soit X et Y deux variables aléatoires continues de *cdf* respectives F et G, et de distribution conjointe H. Soit h la fonction de densité jointe de (X, Y), f et g celles des marges X et Y respectivement.

Définition 2.3.3. *La densité $c(F(x), G(y))$ associée à la copule $C(F(x), G(y))$ est définie par*

$$\begin{aligned} c(F(x), G(y)) &= \frac{\partial^2 C(F(x), G(y))}{\partial F(x) \partial G(y)} \\ &= \frac{h(x, y)}{f(x) g(y)}. \end{aligned} \tag{2.3}$$

On peut déduire du théorème de Sklar que

$$h(x, y) = f(x) g(y) c(F(x), G(y)).$$

Exemple. *Copule indépendante et copule produit*

1. *La densité de la copule indépendante définie par*

$$\pi := C^{\perp}(u, v) = uv \quad \forall u, v \in I,$$

a pour expression

$$\frac{\partial^2 C^{\perp}(u, v)}{\partial u \partial v} = 1.$$

2. *La densité bivariée d'un vecteur aléatoire (X, Y) dont la structure de dépendance est déterminée par la copule produit s'écrit $f(x, y) = f(x)f(y)$. La copule produit caractérise donc l'indépendance entre deux variables aléatoires.*

2.4 Mesures de concordance

Une définition précise de la notion de concordance est définie dans le livre de (Nelsen, 2007).

Définition 2.4.1. ***(Concordance)***
Soit (x_i, y_i) et (x_j, y_j) deux couples d'observations d'un vecteur (X, Y) de variables aléatoires continues. On dit que (x_i, y_i) et (x_j, y_j) sont en concordance si

$$\begin{cases} x_i < x_j, \\ \\ y_i < y_j, \end{cases} \quad ou \quad \begin{cases} x_j < x_i, \\ \\ y_j < y_i. \end{cases}$$

De façon semblable, on dit que (x_i, y_i) et (x_j, y_j) sont en discordance si

$$\begin{cases} x_i > x_j, \\ \\ y_i < y_j, \end{cases} \quad ou \quad \begin{cases} x_j < x_i, \\ \\ y_j > y_i. \end{cases}$$

Les mesures d'association les plus connues sont le tau [2] de Kendall et le rhô de Spearman. Ces mesures sont définies à l'aide du concept de concordance (et discordance).

---

2. Remarquer ici que « tau » c'est la lettre grecque et non le taux qui est un rapport de deux quantités.

Définition 2.4.2. *(**Mesure de concordance**)*

Une mesure d'association κ entre deux variables aléatoires continues X et Y de copule C est une mesure de concordance si elle vérifie les propriétés suivantes :

1. *κ est définie pour tout couple de variables aléatoires continues (X, Y) ;*
2. *$-1 \leq \kappa_{X,Y} \leq 1$, $\kappa_{X,-X} = -1$ et $\kappa_{X,X} = 1$;*
3. *κ est symétrique ;*
4. *$\kappa_{X,Y} = 0$ si les variables X et Y sont indépendantes ;*
5. *$\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$;*
6. *Si C_1 et C_2 sont deux copules telles que $C_1 \prec C_2$ alors $\kappa_{C_1} < \kappa_{C_2}$;*
7. *Si $\{(X_n, Y_n)\}$ est une suite de couples aléatoires continus de copule $\{C_n\}$, et si $\{C_n\}$ converge vers C, alors κ_{C_n} converge vers κ_C. En d'autres termes,*

$$\lim_{n\to\infty} \kappa_{C_n} = \kappa_C.$$

Définition 2.4.3. *(**Fonction de concordance**)*

Soit $\{(X_j, Y_j), j = 1, 2\}$ deux couples de variables aléatoires continues de distributions conjointes H_1 et H_2 et de marges respectives communes F et G. La fonction de concordance entre ces deux couples est définie par

$$Q = P\Big[\big((X_1 - X_2) - (Y_1 - Y_2)\big) > 0\Big] - P\Big[\big((X_1 - X_2) - (Y_1 - Y_2)\big) < 0\Big].$$

C'est la différence entre la probabilité de concordance et la probabilité de discordance.

Chaque distribution conjointe de variables aléatoires continues étant caractérisée par une copule unique, le théorème de Sklar permet d'établir une relation entre toute fonction de concordance et les copules associées.

Proposition 2.4.1. ***(Propriété de la fonction de concordance)***

Soit $\{(X_j, Y_j), j = 1, 2\}$ deux couples de variables aléatoires indépendants de distributions conjointes H_1 et H_2 et de marges respectives communes F et G. Soit

C_1 et C_2 les copules associées aux distributions H_1 et H_2 respectivement. Alors, nous avons

$$Q(C_1, C_2) = 4 \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) - 1.$$

On mesure généralement le lien entre deux variables par le coefficient de corrélation linéaire dit de Pearson. Cette mesure est effectuée sur les valeurs de ces variables. Cependant, il existe des situations pour lesquelles une mesure de corrélation sur les valeurs n'est pas adaptée. Si les variables sont par exemple ordinales, discrètes ou encore que les valeurs n'ont que peu d'importance, il nous reste les corrélations sur les rangs. Celles les plus connues et les plus utilisées en statistiques sont le rhô de Spearman et le tau de Kendall.

2.4.1 Rhô de Spearman

Le rhô de Spearman est basé sur la notion de concordance et de discordance. Soit $\{(X_j, Y_j), j = 1, 2, 3\}$ trois paires indépendantes de variables aléatoires de même distribution H dont les distributions des marges sont F et G, et la copule associée C.

Définition 2.4.4. ***(Rhô de Spearman théorique)***

Le rhô de Spearman est communément défini comme étant proportionnel à la différence entre la probabilité de concordance et celle de discordance des couples aléatoires (X_1, Y_1) et (X_2, Y_3) :

$$\rho_{X,Y} = 3P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) > 0\Big] - 3P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) < 0\Big],$$

laquelle équation peut facilement se réécrire, en utilisant la relation

$$P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) < 0\Big] = 1 - P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) > 0\Big],$$

sous la forme

$$\rho_{X,Y} = 6P\big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) > 0\big] - 3.$$

Comme H est la distribution de (X_1, Y_1) et π celle de (X_2, Y_3) (car les variables X_2 et Y_3 sont indépendantes) alors, on a d'après la Proposition 2.4.1 le résultat suivant :

Proposition 2.4.2. *Soit X et Y deux variables aléatoires continues de copule C. Le rhô de Spearman de ces deux variables est défini par*

$$\rho_{X,Y} = 12 \iint_{I^2} C(u, v) du dv - 3.$$

Cette proposition est démontrée à la section A.2.4 de l'appendice.

Définition 2.4.5. ***(Rhô de Spearman empirique)***
Soit un échantillon de taille n de données $\{(X_i, Y_i), i = 1, \ldots, n\}$. Le rhô de Spearman empirique noté ρ_n est défini par

$$\rho_n = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (R_i - S_i)^2,$$

où R_i est le rang de X_i et S_i celui de Y_i.

Remarque. *Le rhô de Spearman est un nombre compris entre -1 (classements inverses) et 1 (classements identiques), la valeur 0 indiquant que nos deux classements n'ont rien à voir l'un et l'autre. Les classements ici sont effectués en fonction des rangs de valeurs prises par les variables aléatoires.*

Une autre mesure utile du lien entre deux variables est le tau de Kendall. (Nie *et al.*, 1975) ont déclaré dans leur manuel que le tau était « plus significatif lorsque les données contenaient un grand nombre de rangs liés ».

2.4.2 Tau de Kendall

Le tau de Kendall dans la littérature concernant l'estimation des copules paramétriques est la mesure de dépendance la plus utilisée. dont l'une des raisons a été évoquée ci-dessus.

Définition 2.4.6. ***(Tau de Kendall théorique)***

Soit $\{(X_i, Y_i), i = 1, 2\}$ un échantillon de données bivariées d'un couple de variables aléatoires (X, Y) continues i.i.d. de cdf conjointe H. Le tau de Kendall du couple de variables aléatoires (X, Y), noté $\tau_{X,Y}$ est défini par

$$\tau_{X,Y} = P\Big[\big((X_1 - X_2) - (Y_1 - Y_2)\big) > 0\Big] - P\Big[\big((X_1 - X_2) - (Y_1 - Y_2)\big) < 0\Big].$$

Si C désigne la copule associée au couple (X, Y), alors le résultat suivant est la conséquence immédiate de la Proposition 2.4.1.

Proposition 2.4.3. *Soit X et Y deux variables aléatoires continues de copule C. Le tau de Kendall de ces variables est donné par*

$$\tau_{X,Y} := \tau_C = 4 \iint_{I^2} C(u, v) dC(u, v) - 1.$$

On peut aussi écrire

$$\tau_C = 4E\left[C(u, v)\right] - 1$$

puisque

$$\iint_{I^2} C(u, v) dC(u, v) = E[C(u, v)].$$

La démonstration de cette proposition est identique à celle sur le rhô de Spearman.

Définition 2.4.7. ***(Tau de Kendall empirique)***

Soit un échantillon de données bivariées de taille n d'un couple de variables aléatoires (X, Y) continues. On définit le tau de Kendall par

$$\tau_n = \frac{[\textit{Nombre de paires concordantes}] - [\textit{Nombre de paires discordantes}]}{\textit{Nombre total de paires}}.$$

Remarque. *Le tau de Kendall est un indicateur compris entre -1 et 1 et s'interprète comme un coefficient de Pearson : plus il tend vers 1, plus on est certain qu'il existe une corrélation positive et plus il est proche de -1, plus on peut supposer l'existence d'une dépendance négative. Si enfin le tau de Kendall empirique est enfin proche de 0, la probabilité qu'il n'existe aucune dépendance est significative.*

Dans la section qui suit, nous allons présenter de façon succinte les copules paramétriques les plus connues et les plus utilisées dans le cas bivarié, leurs propriétés, ainsi que leurs illustrations graphiques. Les deux principales familles de copules que nous allons explorer sont essentiellement les copules elliptiques au sein desquelles on trouve deux copules très utilisées (la copule gaussienne et la copule de Student) ; et les copules archimédiennes dont trois sont très connues (Frank, Clayton et Gumbel). On verra également comment calculer le paramètre de la copule à l'aide du tau de Kendall et du rhô de Spearman [3].

2.5 Copule archimédienne

Le statisticien canadien Christian Genest est associé aux copules archimédiennes. Il n'en est pas à l'origine, mais est le pionnier dans l'établissement d'une méthode d'analyse statistique de ces fonctions. Ses publications ont d'ailleurs contribué à la vulgarisation des copules archimédiennes.

Définition **2.5.1.** *Soit* $\phi : [0,1] \to [0,\infty)$ *une fonction telle que, pour tout* $u \in [0,1]$, $\phi(1) = 0$, $\phi'(u) < 0$ *et* $\phi''(u) > 0$. *Les copules archimédiennes sont définies par*

$$C(u,v) = \begin{cases} \phi^{-1}(\phi(u) + \phi(v)) & si\ \phi(u) + \phi(v) \leq \phi(0), \\ 0 & sinon. \end{cases}$$

La fonction ϕ *est appelée la fonction génératrice de la copule.*

---

3. Ce calcul est possible grâce à la relation qui existe entre le paramètre de la copule et ses principales mesures d'associations.

Les copules archimédiennes présentent de nombreuses propriétés intéressantes dont (Genest *et al.*, 1995) donnent une bonne caractérisation.

Proposition 2.5.1. *Soit C une copule archimédienne de générateur ϕ. Alors,*

- *C est symétrique, c'est-à-dire $C(u,v) = C(v,u)$ pour tout $u, v \in I$.*
- *C est associative, c'est-à-dire $C(u, C(v,w)) = C(C(u,v), w)$ pour tout $u, v, w \in I$.*

Proposition 2.5.2. *Soit X_1 et X_2 deux variables aléatoires de copule archimédienne de fonction génératrice ϕ. Alors le tau de Kendall est donné par*

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt.$$

Proposition 2.5.3. *La densité de la copule archimédienne bivariée de générateur ϕ deux fois différentiable et de paramètre θ est donnée par*

$$c_\theta(u,v) = -\frac{\phi'\big(C_\theta(u,v)\big)\phi'(u)\phi'(v)}{\phi'\big(C_\theta(u,v)\big)^3}, \quad \forall u, v \in I. \tag{2.4}$$

Nous allons maintenant présenter les principales familles de copules archimédiennes, les expressions de leur densité et la relation entre le paramètre de la copule et le tau de Kendall.

Les familles de copules archimédiennes seront principalement constituées entre autres des copules de Frank, de Clayton et de Gumbel.

2.5.1 La copule de Clayton

Expresssion de la copule

De générateur $\phi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1),\ \theta \in \mathbb{R}^+ - \{0\},$ la copule de Clayton a pour expression

$$C_\theta(u,v) = \big[\max(u^{-\theta} + v^{-\theta} - 1, 0)\big]^{-1/\theta},$$

qu'on pourrait simplifier sous la forme

$$C_\theta(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}, \quad \theta > 0.$$

Densité de la copule

On montre en utilisant l'équation (2.4) et après quelques developpements analytiques que sa fonction de densité est donnée par l'équation

$$c_\theta(u,v) = \frac{\partial^2 C_\theta(u,v)}{\partial u \partial v} = (\theta+1)(uv)^{-(\theta+1)}(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1+2\theta}{\theta}}, \quad \theta > 0.$$

Relation entre le paramètre de la copule et le tau de Kendall

Tel que souligné plus haut, il existe une relation étroite entre le tau de Kendall et le paramètre de la copule. Pour la copule de Clayton, cette relation est donnée par

$$\tau = \frac{\theta}{\theta + 2}.$$

On en déduit que

$$\theta = \frac{2\tau}{1-\tau}.$$

La Figure 2.1 représente la densité et la fontion de répartition (ainsi que leurs contours) de la copule de Clayton bivariée. On y observe que cette copule est asymétrique à gauche.

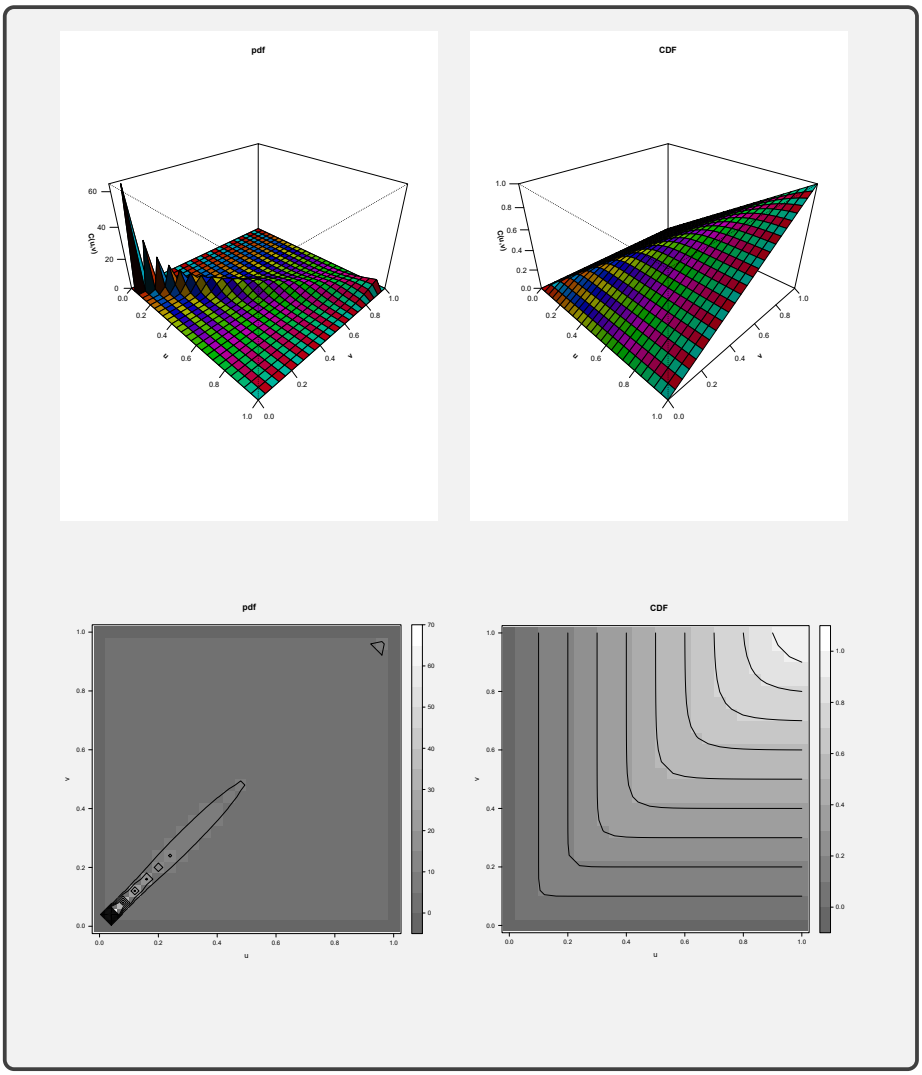

Figure 2.1: Copule de Clayton de paramètre $\theta = 4$. À gauche : densité et contours. À droite : fonction de répartition et contours.

2.5.2 La copule de Frank

Expresssion de la copule

La copule de Frank a pour générateur $\phi_\theta(t) = -\ln\left(\dfrac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$ avec $\theta \in \mathbb{R} - \{0\}$. Elle est définie par

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right).$$

Densité de la copule

On montre en utilisant l'équation (2.4) que la copule de Frank a pour densité

$$c_\theta(u,v) = \frac{\partial^2 C_\theta(u,v)}{\partial u \partial v} = \frac{\theta}{1-e^{-\theta}} L(T_\theta(u,v)) \frac{e^{-\theta(u+v)}}{T_\theta(u,v)},$$

avec

$$L(s) = \sum_{k=1}^{\infty} \frac{s^k}{k},$$

$$T_\theta(u,v) = \frac{(1-e^{-\theta u})(1-e^{-\theta v})}{1-e^{-\theta}},\ \theta \neq 0.$$

Relation entre le paramètre de la copule et le tau de Kendall

Le tau de Kendall correspond à la famille d'équations donnée par

$$\tau = 1 - \frac{4}{\theta} + \frac{4}{\theta^2} \int_0^\theta \frac{t}{e^t - 1} dt, \quad \theta \in \mathbb{R} - \{0\},$$

où θ est le paramètre de la copule.

La Figure 2.2 représente la densité et la fontion de répartition (ainsi que leurs contours) de la copule de Frank bivariée. Sur cette figure, on observe que la copule de Frank est symétrique.

2.5.3 La copule de Gumbel

Expresssion de la copule

La fonction génératrice de la copule de Gumbel est $\phi_\theta(t) = (-\ln t)^\theta$, avec $\theta \geq 1$ et $t \in (0,1)$. On vérifie facilement que

1. $\phi_\theta^{-1}(t) = e^{-t^{1/\theta}}$;
2. $\phi'_\theta(t) = -\frac{\theta}{t}(-\ln t)^{\theta-1}$ et $\phi'(t) < 0$;

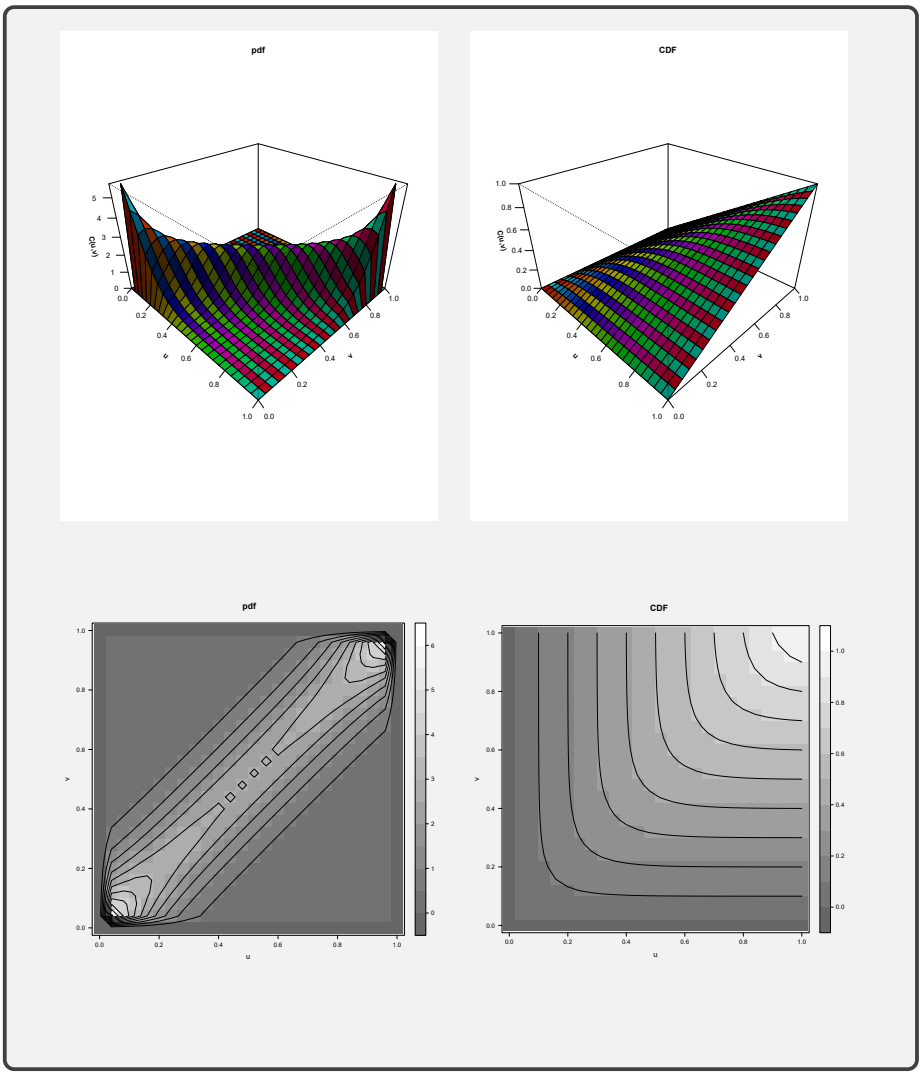

Figure 2.2: Copule de Frank de paramètre $\theta = 3$. À gauche : densité et contours. À droite : fonction de répartition et contours.

3. $\phi''_\theta(t) = \frac{\theta}{t^2}(\theta - 1 - \ln t)^{\theta-1}$ et $\phi''_\theta(t) \geq 0$.

Nous retrouvons la copule de Gumbel par la relation

$$C_\theta(u, v) = \phi_\theta^{-1}(\phi_\theta(u) + \phi_\theta(v)),$$

ce qui conduit à l'équation

$$C_\theta(u, v) = \exp\big[- \big((-\ln u)^\theta + (\ln v)^\theta\big)^{1/\theta}\big],\ \forall\theta \geq 1,\ \forall(u, v) \in I^2.$$

Densité de la copule

De l'équation (2.4), il découle

$$c_\theta(u,v) = -\frac{\frac{\theta}{C_\theta(u,v)^2}\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{\theta-2}{\theta}}\left[\theta - 1 + \left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{1}{\theta}}\right]\frac{-\theta}{u}(-\ln u)^{\theta-1}\frac{-\theta}{v}(-\ln v)^{\theta-1}}{\frac{\theta^3}{C_\theta(u,v)^3}\left[(-\ln u)^\theta + (-\ln v)^\theta\right]^{\frac{3(\theta-1)}{\theta}}}.$$

Après simplifications, on a la densité suivante

$$c_\theta(u,v) = C_\theta(u,v)\left[\phi_\theta(u) + \phi_\theta(v)\right]^{\frac{1}{\theta}-2}\left[\theta - 1 + \left(\phi_\theta(u) + \phi_\theta(v)\right)^{\frac{1}{\theta}}\right]\frac{\phi_{\theta-1}(u)\phi_{\theta-1}(v)}{uv},$$

pour tout $u, v \in I$ et $\theta \geq 1$.

Tau de Kendall

Pour une copule archimédienne, on a la relation

$$K_C(t) = t - \frac{\phi_\theta(t)}{\phi'_\theta(t)},\ t \in (0,1),\ \theta \geq 1.$$

Dans le cas de la copule de Gumbel, on a

$$K_C(t) = t - \frac{t\ln t}{\theta},\ t \in (0,1),\ \theta \geq 1.$$

Ainsi, sa densité est

$$K'_C(t) = 1 - \frac{1}{\theta} - \frac{\ln t}{\theta},\ t \in (0,1),\ \theta \geq 1.$$

Il s'ensuit que

$$\begin{aligned} E[C_\theta(U,V)] &= \int_0^1 tK'_C(t)dt = \left[\frac{t^2}{2}\left(1 - \frac{1}{\theta}\right)\right]_0^1 - \int_0^1 \frac{t\ln t}{\theta}dt \\ &= \frac{1}{2}\left(1 - \frac{1}{\theta}\right) - \frac{1}{\theta}\left(\left[\frac{t^2}{2}\ln t\right]_0^1 - \left[\frac{t^2}{4}\right]_0^1\right) \\ &= \frac{1}{2}\left(1 - \frac{1}{2\theta}\right). \end{aligned}$$

On déduit de la Proposition 2.4.3 que

$$\tau = \frac{\theta - 1}{\theta},\ \theta \geq 1.$$

On peut observer à la Figure 2.3 que la copule de Gumbel est asymétrique à droite.

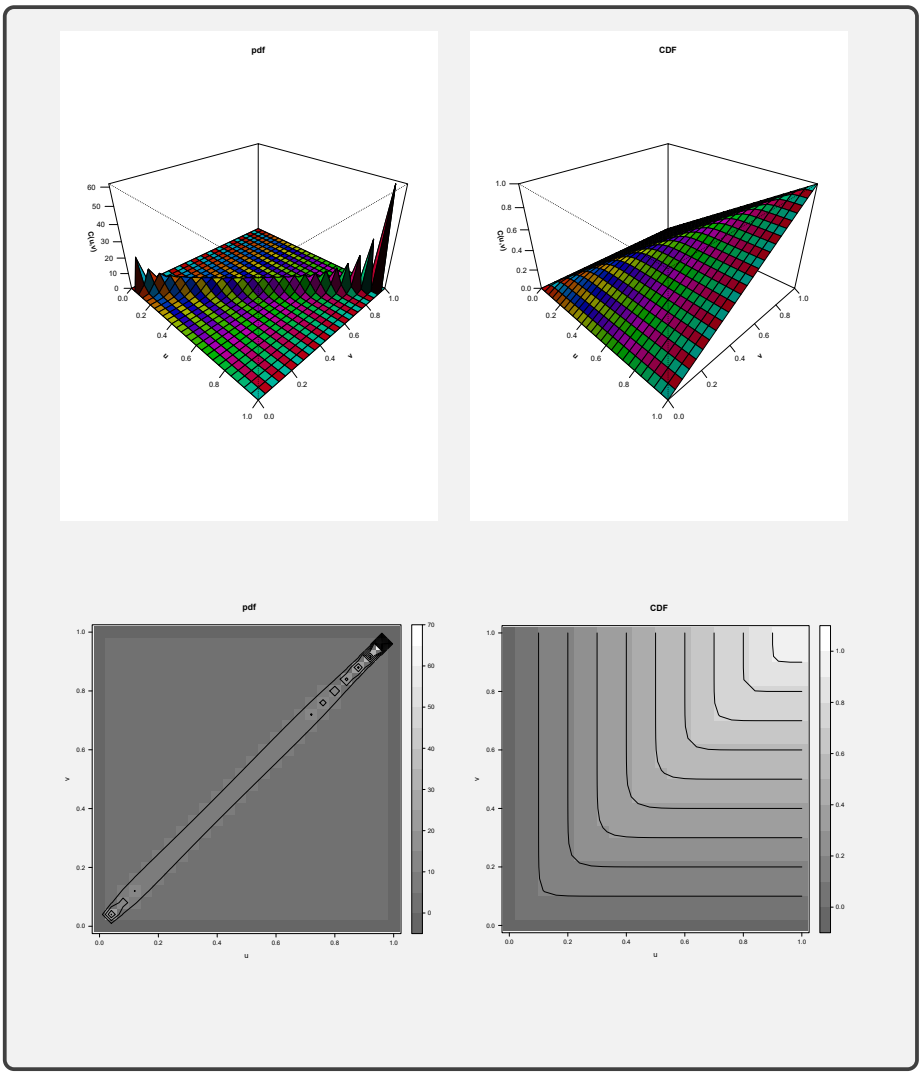

Figure 2.3: Copule de Gumbel de paramètre $\theta = 1.2$. À gauche : densité et contours. À droite : fonction de répartition et contours.

2.6 Les copules elliptiques

Les copules elliptiques sont définies à partir de la distribution elliptique grâce au théorème de Sklar. Les principales copules de cette famille sont la copule gaussienne et la copule de Student. Dans ce qui suit, nous donnons tour à tour quelques définitions et les propriétés des principales copules elliptiques.

Définition 2.6.1. ***(Distribution elliptique)***

Un vecteur aléatoire $X = (X_1, \ldots, X_d)^\top$ *suit une distribution elliptique s'il s'écrit sous la forme*

$$X = \boldsymbol{\mu} + R\boldsymbol{A}U,$$

avec $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$, U *un vecteur aléatoire uniforme sur la sphère unité de* $\mathbb{R}^d$, R *un vecteur aléatoire indépendant de* U *de distribution quelconque et* $\boldsymbol{A}$ *une matrice carrée de dimension* d *telle que* $\boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{A}^\top$ *est régulière*[4].

Définition 2.6.2. ***(Densité d'une distribution elliptique)***

Lorsqu'elle existe, la densité d'une distribution elliptique a pour équation

$$f(\boldsymbol{x}) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right),\ \boldsymbol{x} \in \mathbb{R}^d.$$

La fonction g *est appelée la fonction génératrice de densité et est définie de* $\mathbb{R}^+$ *dans* $\mathbb{R}$.

Définition 2.6.3. ***(Fonction caractéristique d'une distribution elliptique)***

Soit $X = (X_1, \ldots, X_d)$ *un vecteur aléatoire de distribution elliptique. La fonction caractéristique est donnée, pour* $t \in \mathbb{R}^d$, *par*

$$\begin{aligned} \phi_X(t) &= E[\exp(it^\top)X] \\ &= E[\exp it^\top(\boldsymbol{\mu} + R\boldsymbol{A}U)] \\ &= \exp it^\top \boldsymbol{\mu} g(t^\top \boldsymbol{\Sigma} t). \end{aligned}$$

2.6.1 La copule gaussienne

La copule gaussienne bivariée comme sus-mentionnée fait partie de la famille des copules elliptiques à un seul paramètre. Cette copule est intéressante car elle

4. Une matrice est régulière si elle est inversible.

est sous-jacente à la distribution normale bivariée et, lorsque les variables sont linéairement dépendantes, le paramètre de la copule est le coefficient de corrélation linéaire.

Expression de la copule

Soit

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), \quad \Phi(h) := \int_{-\infty}^{h} \phi(x)dx$$

la densité et la *cdf* d'une loi normale et

$$\phi_2(x, y; \rho) := \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}),$$

$$\Phi_2(h, k; \rho) := \int_{-\infty}^{h} \int_{-\infty}^{k} \phi_2(x, y; \rho)dxdy,$$

la densité et la *cdf* d'une loi normale bivariée de paramètre de corrélation $\rho \in (-1, 1)$. Alors, la copule normale bivariée de paramètre ρ est définie par

$$\begin{aligned} C_G(u, v) &= \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho) \\ &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right) dxdy. \end{aligned} \tag{2.5}$$

Densité de la copule

De l'expression de la copule gaussienne bidimensionnelle (2.5), on déduit sa densité dont l'expression est

$$c_G(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2 - 2\rho\Phi^{-1}(u)\Phi^{-1}(v)}{2(1-\rho^2)}\right).$$

Cette copule n'est adaptée ni pour modéliser l'interaction dans les queues (valeurs extrêmes), ni pour caractériser une dépendance non linéaire. La Figure 2.4 corrobore cette assertion. On peut en effet y observer que la copule gaussienne est symétrique.

Tau de Kendall

On montre que le tau de Kendall et le rhô de Spearman dans le cas de la copule normale ont pour expressions respectives

$$\tau = \frac{2}{\pi} \arcsin \rho_0, \tag{2.6}$$
$$\rho = \frac{6}{\pi} \arcsin \frac{\rho_0}{2},$$

où $\rho_0 \in (-1, 1)$ est le paramètre de la copule.

2.6.2 La copule de Student

La copule de Student bivariée ou t-copule représente la fonction de dépendance associée à la *cdf* d'une loi de Student bivariée. Elle est construite comme la copule gaussienne et est une copule elliptique à deux paramètres.

Expression de la copule

Soit $\rho \in (-1, 1)$ le coefficient de corrélation, T la *cdf* d'une loi de Student univariée standardisée de degrés de liberté $\kappa \in \mathbb{N} - \{0\}$ et $T_{\rho,\kappa}$ la *cdf* d'une loi de Student bivariée de matrice de corrélation associée à ρ et de degrés de liberté $\kappa \in \mathbb{N} - \{0\}$. Alors la copule de Student bivariée C_T est définie par

$$C_T(u, v; \rho, \kappa) = T_{\rho,\kappa}(T^{-1}(u), T^{-1}(v)), \quad \forall u, v \in (0, 1).$$

Densité de la copule

La densité de la copule de Student s'écrit

$$c_T(u, v; \rho, \kappa) = \rho^{-1/2} \frac{\Gamma\left(\frac{\kappa+2}{2}\right)\Gamma\left(\frac{\kappa}{2}\right)}{\left(\Gamma\left(\frac{\kappa+1}{2}\right)\right)^2} \left(1 + \frac{a^2 + b^2 - 2\rho ab}{\kappa(1-\rho)^2}\right)^{-\frac{\kappa+2}{2}} \left(1 + \frac{a}{\kappa}\right)^{-\frac{\kappa+2}{2}} \left(1 + \frac{b}{\kappa}\right)^{-\frac{\kappa+2}{2}},$$

avec $b = T^{-1}(u)$, $a = T^{-1}(v)$, $\Gamma : z \to \int_0^\infty t^{z-1} e^{-t} dt$, où $u, v \in (0, 1)$.

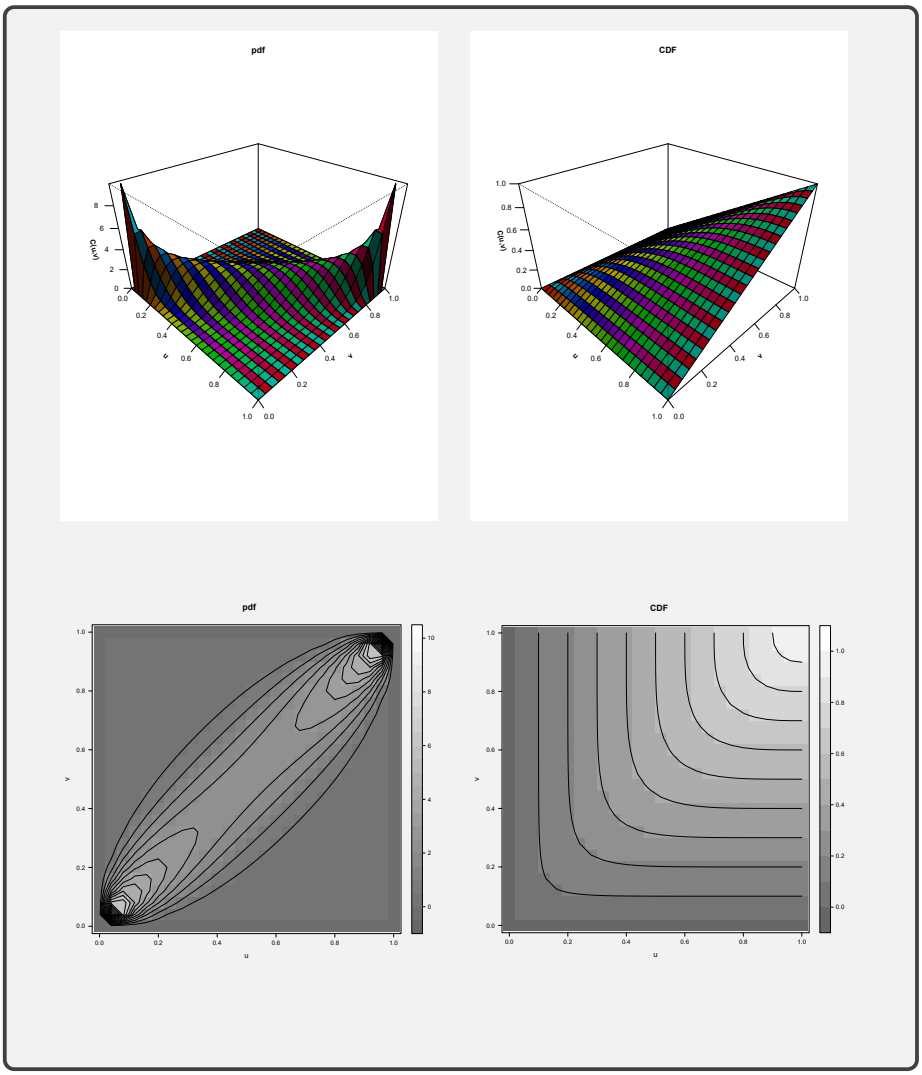

Figure 2.4: Copule gaussienne de paramètre $\theta = .7$. À gauche : densité et contours. À droite : fonction de répartition et contours.

Tau de Kendall

La copule de Student et la copule gaussienne ont exactement le même tau de Kendall donné par l'équation (2.6).

Tout comme la copule gaussienne, on observe à la Figure 2.5 que la copule de

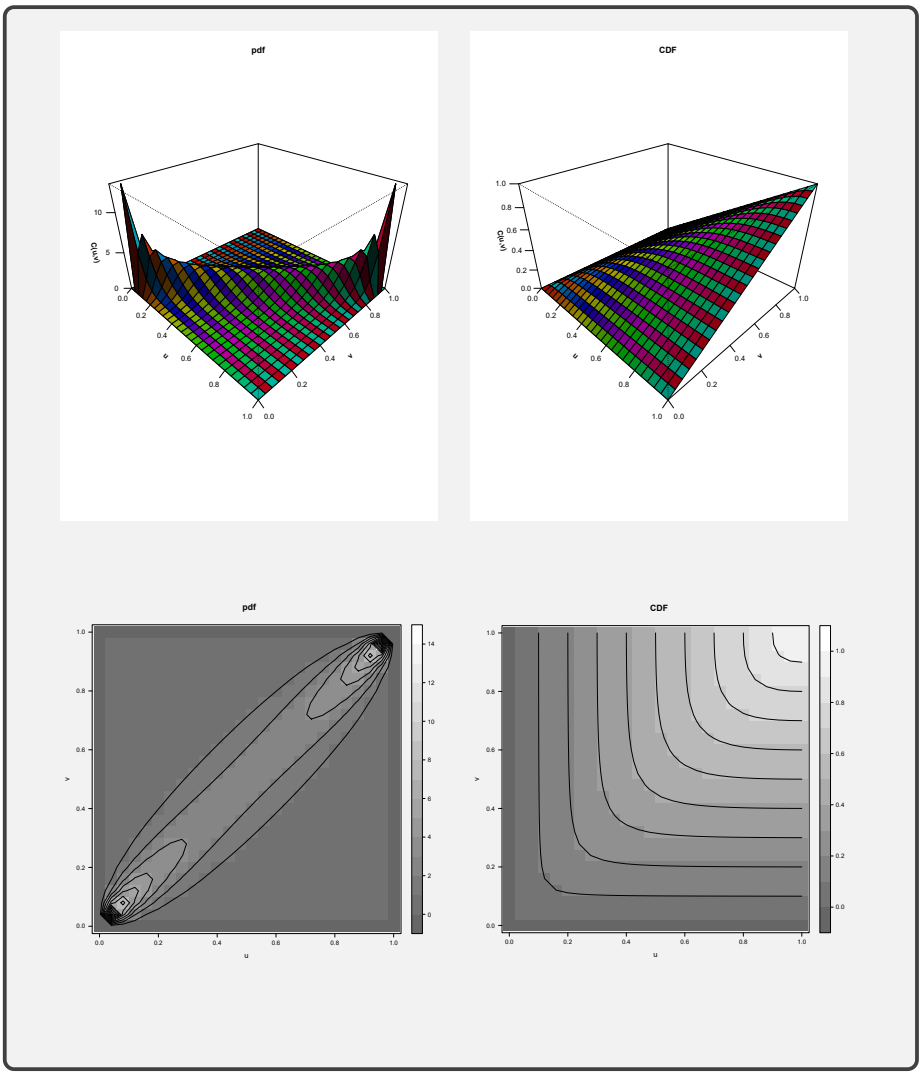

Figure 2.5: Copule de Student de paramètres $\theta = 0.8$ et $\kappa = 2$. À gauche : densité et contours. À droite : fonction de répartition et contours.

Student est symétrique.

Nous avons présenté les principales familles de copules ainsi que leurs caractéristiques. Dans ce qui suit, nous ferons de l'inférence sur celles-ci.

2.7 Estimation de la copule

Bien que la copule modélise complètement la dépendance entre les variables, elle reste en pratique inconnue, d'où la nécessité de l'estimer.

Supposons que l'on dispose d'un d-échantillon, $d \geq 2$, de taille n,

$$\boldsymbol{X} = \{(X_{1,i}, X_{2,i}, \ldots, X_{d,i})\}_{i=1}^{n}, \tag{2.7}$$

d'une copule appartenant à une famille paramétrique de paramètres possiblement multivariés θ, $\{C_\theta, \theta\}$ et que l'on souhaite estimer ce paramètre. Remarquons que si les fonctions de répartition marginales de $\{X_1, X_2, \ldots, X_d\}$ étaient connues et notées $F_1, \ldots, F_d$, alors on pourrait avoir recours aux méthodes d'inférence classiques de la statistique pour résoudre notre problème. Mais les marges ne sont malheureusement pas connues. Afin de pallier cette difficulté, principalement trois méthodes bien connues dans la littérature peuvent être utilisées : l'estimation paramétrique qui suppose que les marges de la copule appartiennent elles aussi à une famille paramétrique ; l'estimation semi-paramétrique où on estime les marges de façon non paramétrique. Dans les deux cas, on estime le paramètre de la copule après avoir estimé les marges. Si, enfin, on souhaite se passer de l'hypothèse selon laquelle la copule appartiendrait à une famille paramétrique, alors la troisième méthode estime la copule de façon non paramétrique.

2.7.1 Estimation paramétrique

L'approche paramétrique impose au préalable un modèle spécifique pour la copule et un autre pour les marges de la copule. Généralement, on utilise la méthode du maximum de vraisemblance complète (FML pour Full Maximum Likelihood) pour obtenir les estimateurs des paramètres (Shih et Louis, 1995) et (Joe, 1997). La méthode consiste à estimer conjointement les paramètres des marges et le(s) para-

mètre(s) de la copule. Cependant, la fonction de vraisemblance obtenue peut être compliquée à manipuler, voire impossible à calculer. Pour contourner ce problème, plusieurs auteurs - le tout premier a été (Joe et Xu, 1996) - ont proposé une méthode appelée IFM, pour « Inference Fonction for Margins » (Fonction d'inférence pour les marges) qui procède en deux étapes : la première étape consiste à estimer les paramètres des marges et la seconde à estimer ceux de la copule en utilisant la méthode du maximum de vraisemblance. La méthode IFM a le même inconvénient que la FML : elle dépend de l'hypothèse de la distribution des marges. L'approche IFM est la plus utilisée en raison de sa simplicité (Joe, 1997).

Méthode FML ou « Full Maximum Likelihood »

Lorsque la copule appartient à une famille paramétrique, on peut utiliser la méthode FML pour estimer les paramètres (Shih et Louis, 1995).
De l'égalité $F(x_1, \ldots, x_d) = C(F_1(x_1, \alpha_1), \ldots, F_n(x_d, \alpha_d); \theta)$, on aboutit, par dérivation, à la densité dont l'expression est

$$f(\boldsymbol{x}; \boldsymbol{\alpha}, \theta) = c(F_1(x_1, \alpha_1), \ldots, F_d(x_d, \alpha_d); \theta) \prod_{j=1}^{d} f_j(x_j, \alpha_j),$$

où f et f_j désignent respectivement la densité multivariée de $\boldsymbol{X}$ et univariée de la composante X_j, α_j le paramètre de la marge F_j pour tout $j \in \{1, \ldots, d\}$, c la densité de la copule définie à l'équation (2.3) et θ le paramètre de la copule.

Estimer les paramètres revient donc à maximiser la fonction de log-vraisemblance

$$l(\boldsymbol{\alpha}, \theta; \boldsymbol{x}) \quad = \quad \sum_{i=1}^{n} \log\left(c(F_1(x_{1,i}, \alpha_1), \ldots, F_d(x_{d,i}, \alpha_d); \theta)\right) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log(f_j(x_{j,i}; \alpha_j)). \quad (2.8)$$

L'estimateur FML est donc obtenu en maximisant la fonction définie en (2.8) par rapport à $\boldsymbol{\beta} = (\boldsymbol{\alpha}, \theta)$, c'est-à-dire

$$\hat{\boldsymbol{\beta}}_{FML} = \arg\max_{\boldsymbol{\beta} \in \Theta} l(\boldsymbol{\alpha}, \theta; \boldsymbol{x}),$$

où Θ est l'espace des paramètres. $\hat{\boldsymbol{\beta}}_{FML}$ est la solution du système d'équations

$$\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}, \frac{\partial l}{\partial \theta}\right) = \mathbf{0}^\top.$$

Proposition 2.7.1. *Sous certaines conditions de régularité, (Cherubini* et al.*, 2004) montrent à la page 154, que l'estimateur $\hat{\boldsymbol{\beta}}_{FML}$ existe, est convergent, asymptotiquement efficace et vérifie la propriété de normalité asymptotique*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{FML} - \boldsymbol{\beta}_0) \to \mathcal{N}(0, \boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)),$$

où $\boldsymbol{I}(\boldsymbol{\beta}_0)$ est l'information de Fisher et $\boldsymbol{\beta}_0$ la vraie valeur de $\boldsymbol{\beta}$.

La méthode FML est complexe car résoudre le problème (2.8) tant de façon analytique que computationnelle [5] s'avère être fastidieux. En plus du problème de coût en temps de calcul de cette méthode, l'estimation de la copule est sensible à une éventuelle mauvaise spécification des marges car celles-ci interviennent dans le calcul de la log-vraisemblance.

Méthode IFM ou « Inference Fonctions for Margins »

La méthode FML est fastidieuse. Il faut estimer simultanément le paramètre de la copule et les marges. Elle n'est par conséquent pas la mieux appropriée pour résoudre le problème d'estimation. (Joe et Xu, 1996) pour pallier ce problème ont proposé la méthode IFM pour « Inference Fonctions for Margins », qui permet d'estimer les paramètres séparément. Utilisant le fait que la fonction de log-vraisemblance (2.8) est une somme de deux fonctions (la première qui ne dépend que des marges et la seconde, des marges et de la copule), cette méthode procède en deux étapes :

5. Les calculs analytiques seraient en effet coûteux rendant le problème d'optimisation difficile à résoudre et l'optimisation numérique lente et complexe.

— elle estime les paramètres des marges en se servant du premier terme de (2.8) , c'est-à-dire

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \sum_{j=1}^{d} \log\left(f_j(x_{j,i}; \alpha_j)\right);$$

— elle estime ensuite les paramètres de la copule à partir du deuxième terme de (2.8), en tenant compte des paramètres des marges estimés à la première étape, c'est-à-dire

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log\left(c(F_1(x_{1,i}, \hat{\alpha}_1), \ldots, F_n(x_{d,i}, \hat{\alpha}_d); \theta)\right).$$

On retrouve cette méthode dans (Shih et Louis, 1995) sous la dénomination « two-stage parametric ML method » ainsi que dans « Inference Functions for Margins » de (Joe et Xu, 1996).

L'estimateur IFM obtenu à l'issue de ces deux étapes est

$$\hat{\boldsymbol{\beta}}_{IFM} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_d, \hat{\theta}).$$

Proposition 2.7.2. *(Joe, 1997)*
Sous certaines conditions de régularité, l'estimateur IFM vérifie la propriété de normalité asymptotique

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{IFM} - \boldsymbol{\beta}_0) \to \mathcal{N}(0, \mathcal{G}(\boldsymbol{\beta}_0)),$$

où $\mathcal{G}(\boldsymbol{\beta}_0)$ est la matrice d'information de Godambe.

Définition 2.7.1. ***(Matrice d'information de Godambe (Godambe, 1960))***
Posons

$$l_j = \sum_{i=1}^{n} \log\left(f_j(x_{j,i}; \alpha_j)\right), \quad j \in \{1, \ldots, d\},$$

et définissons la fonction score

$$U(\boldsymbol{\beta}) = \left(\frac{\partial l_1}{\partial \alpha_1}, \quad \ldots, \quad \frac{\partial l_d}{\partial \alpha_d}, \quad \frac{\partial l}{\partial \theta}\right).$$

Soit la matrice Hessienne ou la matrice de sensitivité

$$\boldsymbol{H}(\boldsymbol{\beta}) = E\big[-\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\big],$$

et la matrice de variabilité

$$\boldsymbol{J}(\boldsymbol{\beta}) = E[U(\boldsymbol{\beta})U(\boldsymbol{\beta})^\top].$$

Alors, la matrice de Godambe est donnée par l'expression

$$\mathcal{G}(\beta_0) = \boldsymbol{H}(\boldsymbol{\beta})^{-1}\boldsymbol{J}(\boldsymbol{\beta})(\boldsymbol{H}(\boldsymbol{\beta})^{-1})^\top.$$

La matrice d'information de Godambe est difficile à déterminer en raison du calcul des dérivées multiples. Mais elle est avantageuse en terme de calculs comparée à la méthode du maximum de vraisemblance.

La méthode IFM, tout comme la précédente est sensible à une éventuelle mauvaise spécification des marges.

2.7.2 Estimation semi-paramétrique

L'approche semi-paramétrique suppose la copule paramétrique et les marges non paramétriques. Cette méthode est aussi connue sous le nom de pseudo-maximum de vraisemblance ou encore maximum de vraisemblance canonique (CML pour « Canonical Maximum Likelihood »).

Contrairement à la méthode IFM dans laquelle l'estimation du paramètre de la copule dépend de celle des marges, la méthode CML telle que proposée par (Genest *et al.*, 1995) ne fait aucune hypothèse sur les distributions marginales. Elle utilise les *cdf* empiriques pour les marges

$$\hat{F}_j(x_{j,i}) = \frac{1}{n+1}\sum_{l=1}^{n} \mathbb{1}_{\{x_{j,l}\leq x_{j,i}\}}, \quad j \in \{1,\ldots,d\}. \tag{2.9}$$

Dans le système d'équations (2.9), on utilise $n+1$ au lieu de n pour éviter le problème d'existence des limites de la fonction de densité de la copule.

Après avoir estimé les marges de façon non paramétrique, on estime le paramètre de la copule en maximisant la fonction de pseudo log-vraisemblance :

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log \left(c(\hat{F}_1(x_{1,i}), \ldots, \hat{F}_d(x_{d,i}); \theta) \right).$$

Le paramètre de la copule comme on peut le constater est estimé indépendamment de la spécification des lois des marges. Les temps de calcul sont acceptables, du moins par rapport aux méthodes précédentes.

(Kim *et al.*, 2007) ont mené une étude comparative des méthodes d'estimations paramétriques et semi-paramétriques. Il en ressort que :

1. Les méthodes IFM et FML ne sont pas robustes en raison d'éventuelles mauvaises spécifications des distributions des marges de la copule.
2. La méthode CML a presque le même algorithme que l'IFM, mais le fait que le paramètre de la copule est estimé indépendamment de la spécification des lois marginales la rend robuste.
3. Un autre avantage de la CML sur l'IFM est que la CML estime la distribution des marges de façon non paramétrique.
4. Leurs résultats de simulation montrent qu'en général, dans la plupart des situations pratiques, l'estimateur de CML est meilleur que les estimateurs IFM et FML .

2.7.3 Estimation non-paramétrique

Lorsque le modèle paramétrique de la copule est mal spécifié, l'utilisation des méthodes d'estimation paramétrique et semi-paramétrique conduit à un mauvais

ajustement du modèle aux données, d'où l'utilité de la méthode d'estimation non paramétrique. Sans trop aller en profondeur, il est bon de souligner qu'en général, puisque la copule est une *cdf*, son estimation non paramétrique s'appuie sur la *cdf* empirique définie par

$$F_n(x) = \frac{1}{n+1} \sum_{k=1}^{n} \mathbb{1}_{\{x_k \leq x\}}, \tag{2.10}$$

pour un échantillon de n observations $(x_1, \ldots, x_n)$ de loi F.

En dimension d, on obtient un estimateur naturel non empirique de la copule en généralisant la formule (2.10). Ainsi, du théorème de Sklar, il vient que

$$\begin{aligned} C(u_1, \ldots, u_d) &= C(F_1(x_1), \ldots, F_d(x_d)) \\ &= F(x_1, \ldots, x_d) \\ &= P(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ &= \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{\{x_{1,i} \leq x_1, \ldots, x_{d,i} \leq x_d\}}. \end{aligned}$$

La section 2.7 est d'une grande importance pour la suite. En effet, dans le chapitre qui suit, nous utiliserons en grande partie certaines notions - en l'occurrence celles relatives à l'estimation du paramètre de la copule et en particulier la méthode d'estimation semi-paramétrique CML - pour étudier notre modèle proposé.

CHAPITRE III

LE MODÈLE DE RÉGRESSION QUANTILE BINAIRE À BASE D'UNE COPULE

3.1 Introduction

Dans la littérature, plusieurs auteurs ont analysé le modèle de régression quantile binaire linéaire. (Kordas, 2006) a lissé la fonction indicatrice avec un noyau, puis pour résoudre la question, a transformé le problème d'optimisation résultant en un programme linéaire. (Hashem *et al.*, 2016) quant à eux ont ajouté au modèle de Kordas une pénalité, le groupe Lasso. En procédant ainsi, ils ont pu simultanément estimer les paramètres du modèle et faire de la classification à l'aide d'une approche bayésienne. (Aristodemou *et al.*, 2019) ont présenté une nouvelle façon d'estimer les paramètres du même modèle sans transformer le problème d'optimisation initial en un programme linéaire. Après avoir lissé la fonction indicatrice à l'aide d'un noyau comme dans (Kordas, 2006), ils remarquent que le problème d'optimisation ressemble à un problème de régression non linéaire pondéré. Ils appliquent l'algorithme IRLS non linéaire (non linear iteratively reweighted least square) et selon eux, obtiennent de façon efficace de bons résultats.

Tous ces auteurs et de nombreux autres ont exploré le cas unique où la variable latente et les covariables sont liées par une fonction linéaire. Étant donné qu'en pratique cette relation n'est pas toujours linéaire, nous avons entrepris d'étudier

le cas où ce lien est plus général [1]. Dans notre travail, nous modélisons la relation entre la variable d'intérêt et les covariables avec les copules. De façon prosaïque, comme nous l'avions vu au chapitre précédent, la copule s'apparente à une boite noire qui prend en entrée plusieurs variables (au moins deux) et présente en sortie la relation qui lie ces variables entre elles.

L'un des résultats que nous trouvons intéressants qui est démontré dans notre travail est que, dans le cas d'une copule normale de marges normales standard, notre modèle est équivalent au modèle de Kordas. Notre modèle se veut donc une généralisation du modèle de (Kordas, 2006).

Les travaux de (Noh *et al.*, 2015) - où la variable réponse est continue - nous ont été d'une très grande utilité. En effet, explorant l'estimation semi-paramétrique des quantiles conditionnels de modèles multivariés en fonction des copules, il nous a permis de résoudre le problème de régression quantile binaire à l'aide d'une copule.

3.2 La régression quantile binaire à base d'une copule

Dans cette partie, nous allons développer la théorie derrière la régression quantile binaire avec la dépendance entre les variables mesurées à l'aide des copules.

3.2.1 Le modèle

Soit X un vecteur $p \times 1$, la variable explicative observable et Y^* une variable latente continue (non observable et de distribution supposée normale) liés par une copule C_θ de paramètre $\theta \in \Theta \subseteq \mathbb{R}$. Le théorème de Sklar garantit l'existence

1. Le lien entre les covariables et la variable réponse quel qu'il soit peut être capté par la copule.

d'une telle copule. Considérons le modèle

$$\begin{aligned} Y^* &= m(X,\theta) + U \\ Y &= \mathbb{1}\{Y^* \geq y^*\}, \end{aligned} \tag{3.1}$$

où m est une fonction quelconque, U est un bruit non observable, Y une variable binaire observable et $y^* \in \mathbb{R}$ une constante.

Pour simplifier, nous allons considérer $p = 1$, c'est-à-dire X un vecteur de $\mathbb{R}$. La méthode se généralise bien au cas $p \geq 2$.

3.2.2 Expression d'un quantile en fonction d'une copule

Dans ce qui suit, nous donnons l'expression d'un quantile en fonction d'une copule selon que la variable d'intérêt est ou non binaire.

Proposition 3.2.1. *Soit Y, Y^*, X vérifiant le modèle latent (3.1). Soit C_θ la copule de paramètre θ qui modélise la dépendance entre Y^* et X. Pour $\tau \in (0,1)$, on a la relation*

$$Q_{Y^*|X=x}(\tau) := F^{-1}_{Y^*|X=x}(\tau) = F^{-1}_{Y^*}\big(D^{-1}(\tau;\theta,u)\big), \quad u \in (0,1),$$

où D^{-1} est l'inverse de la fonction

$$v \mapsto D(v;\theta,u) = \left.\frac{\partial C_\theta(u,v)}{\partial u}\right|_{u=F_X(x)} \tag{3.2}$$

par rapport à sa composante $v \in (0,1)$ et F_X la cdf de la variable aléatoire X.

La preuve de cette proposition est donnée en appendice à la section A.3.1.

3.2.3 La relation entre le quantile binaire et la copule

La proposition qui suit nous donne la relation qui existe entre la fonction de quantiles et la copule lorsque la variable d'intérêt est binaire.

Proposition 3.2.2. *Soit Y, Y^*, X vérifiant le modèle latent (3.1). Soit C_θ la copule de paramètre θ qui modélise la dépendance entre Y^* et X. On utilise la propriété d'équivariance par transformation monotone et la Proposition 3.2.1 pour aboutir à la proposition suivante :*

$$Q_{Y|X=x}(\tau) = \mathbb{1}\big\{F_{Y^*}^{-1}\big(D^{-1}(\tau;\theta,u)\big) \geq y^*\big\}, \quad \tau \in (0,1), \tag{3.3}$$

avec $u = F_X(x)$. La preuve de cette proposition est donnée en appendice à la section A.3.2.

Dans la section qui suit, on applique cette proposition à la copule gaussienne.

Cas de la copule gaussienne

Supposons que la copule C_ρ qui lie les variables Y^* et X est une copule gaussienne de paramètre $\rho \in (-1,1)$. On montre en appendice à la section A.3.3, en supposant sans nuire à la généralité que les variables Y^* et X ont une distribution normale standard [2], que pour tout $\tau \in (0,1)$

$$D^{-1}(\tau;\rho,u) = \Phi\big(\Phi^{-1}(\tau)\sqrt{1-\rho^2} + \rho\Phi^{-1}(u)\big),$$

où Φ est la *cdf* de la distribution normale standard et $u = \Phi(x)$. De l'équation (3.3), il découle que

$$\begin{aligned} Q_{Y|X=x}(\tau) &= \mathbb{1}\left\{\Phi^{-1}\left(\Phi\left(\Phi^{-1}(\tau)\sqrt{1-\rho^2} + \rho\Phi^{-1}(u)\right)\right) \geq y^*\right\} \\ &= \mathbb{1}\big\{\Phi^{-1}(\tau)\sqrt{1-\rho^2} + \rho x \geq y^*\big\}. \end{aligned}$$

Donc

$$Q_{Y|X}(\tau) = \mathbb{1}\big\{X^\top \boldsymbol{\beta}_\tau \geq y^*\big\},$$

2. Pour des variables normales et non standard, les normaliser et obtenir de ce fait des variables normales centrées réduites.

avec

$$(X, \boldsymbol{\beta}_\tau) = \left(\begin{pmatrix} 1 \\ \\ x \end{pmatrix}, \begin{pmatrix} \Phi^{-1}(\tau)\sqrt{1-\rho^2} \\ \\ \rho \end{pmatrix} \right).$$

Il s'ensuit que, lorsque la copule est gaussienne et ses marges aussi, on retrouve le modèle de (Kordas, 2006) défini à l'équation (1.6) avec $y^* = 0$. Notre modèle proposé est donc une généralisation du modèle de Kordas.

3.2.4 La régression quantile binaire vue comme un problème d'optimisation

Comme au Chapitre 1, le quantile défini à l'équation (3.3) est la solution au problème d'optimisation

$$\min_\theta E_{Y|X=x}\Big[\rho_\tau\left(Y - \mathbb{1}\{m_\tau(x,\theta) \geq y^*\}\right)\Big],$$

dont la version empirique est donnée par

$$\min_\theta \sum_{i=1}^{n} \rho_\tau\big(y_i - \mathbb{1}\{m_\tau(x_i,\theta) \geq y^*\}\big), \tag{3.4}$$

où la fonction $m_\tau(x,\theta)$ est définie par l'équation

$$m_\tau(x,\theta) = F^{-1}_{Y^*}(D^{-1}(\tau;\theta,u)), \tag{3.5}$$

avec $u = F_X(x)$.

Une alternative au problème (3.4) est d'y remplacer la fonction indicatrice par un noyau pour obtenir le problème

$$\min_\theta \sum_{i=1}^{n} \rho_\tau(y_i - \tilde{m}_\tau(x_i,\theta)), \tag{3.6}$$

où

$$\tilde{m}_\tau(x_i,\theta) = K\left(\frac{m_\tau(x_i,\theta) - y^*}{h_n}\right),$$

avec K un noyau utilisé pour lisser la fonction indicatrice, h_n un paramètre de lissage et la fonction $m_\tau(.,.)$ définie à l'équation (3.5).

3.3 Estimation des paramètres

Deux alternatives pour estimer la fonction quantile s'offrent à nous :

— la méthode de Kordas : on procède comme dans les travaux de (Kordas, 2006) en résolvant le problème d'optimisation (3.6) ;

— la méthode « two stages » : on résout le problème en deux étapes. On estime d'abord le paramètre de la copule indépendamment du problème d'optimisation, ensuite, on effectue un « plug-in » de ce paramètre estimé dans l'expression de la fonction quantile. Cette approche constitue l'épine dorsale de notre méthode et nous la développons suffisamment dans la suite.

3.3.1 Estimation des paramètres par la méthode de Kordas

La méthode de (Kordas, 2006) et ses variantes se sont avérées infructueuses pour estimer les paramètres de notre modèle proposé. La fonction objective du problème d'optimisation (3.6) - comme on peut le voir aux Figures 3.1, 3.2, 3.3 et 3.4 - n'étant ni convexe, ni convave, aucune méthode d'optimisation n'a fourni de résultats appropriés. Les différents cas examinés et les expressions de la fonction D^{-1} correspondante (et donc implicitement de la fonction objective) sont répertoriés dans le Tableau 3.1. Le lecteur peut les retrouver dans (Bernard et Czado, 2015)

La copule de Gumbel comme on peut le constater n'admet pas d'expression analytique pour D^{-1}. Mais, on la calcule bien numériquement.

Les Figures 3.1, 3.2, 3.3 et 3.4 illustrent la courbe de la fonction objective en fonction du paramètre de la copule. À gauche, on peut voir la fonction originale (c'est-à-dire la fonction objective du problème (3.4)) et à droite la même fonc-

Copules	$D^{-1}(\tau;\theta,u)$
Normale	$\Phi(\Phi^{-1}(\tau)\sqrt{1-\theta^2}+\theta\Phi^{-1}(u))$ $\theta \in (-1,1)$
Frank	$-\frac{1}{\theta}\ln\left(1-\frac{\tau(1-e^{-\theta})}{e^{-\theta u}+\tau(1-e^{-\theta u})}\right)$ $\theta \in \mathbb{R}-\{0\}$
Clayton	$\left(\left(\tau^{-\frac{\theta}{1+\theta}}-1\right)u^{-\theta}+1\right)^{-1/\theta}$ $\theta > 0$
Gumbel	-* $1 \leq \theta < \infty$

Tableau 3.1: Expression analytique de la fonction D^{-1} aussi appelée « h-inverse function » selon la copule de paramètre θ. * Gumbel n'admet pas d'expression analytique pour D^{-1}.

tion lissée (l'indicatrice est remplacée par un noyau d'Épachnikov [3] comme dans l'équation (3.6)).

Pour ce qui est des simulations, les marges de la copule sont générées selon une distribution normale centrée et réduite de taille $n = 1000$ avec une dépendance de tau de Kendall égale à 0.6.

3. Le noyau d'Épachnikov a pour équation $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\{|u| \leq 1\}$.

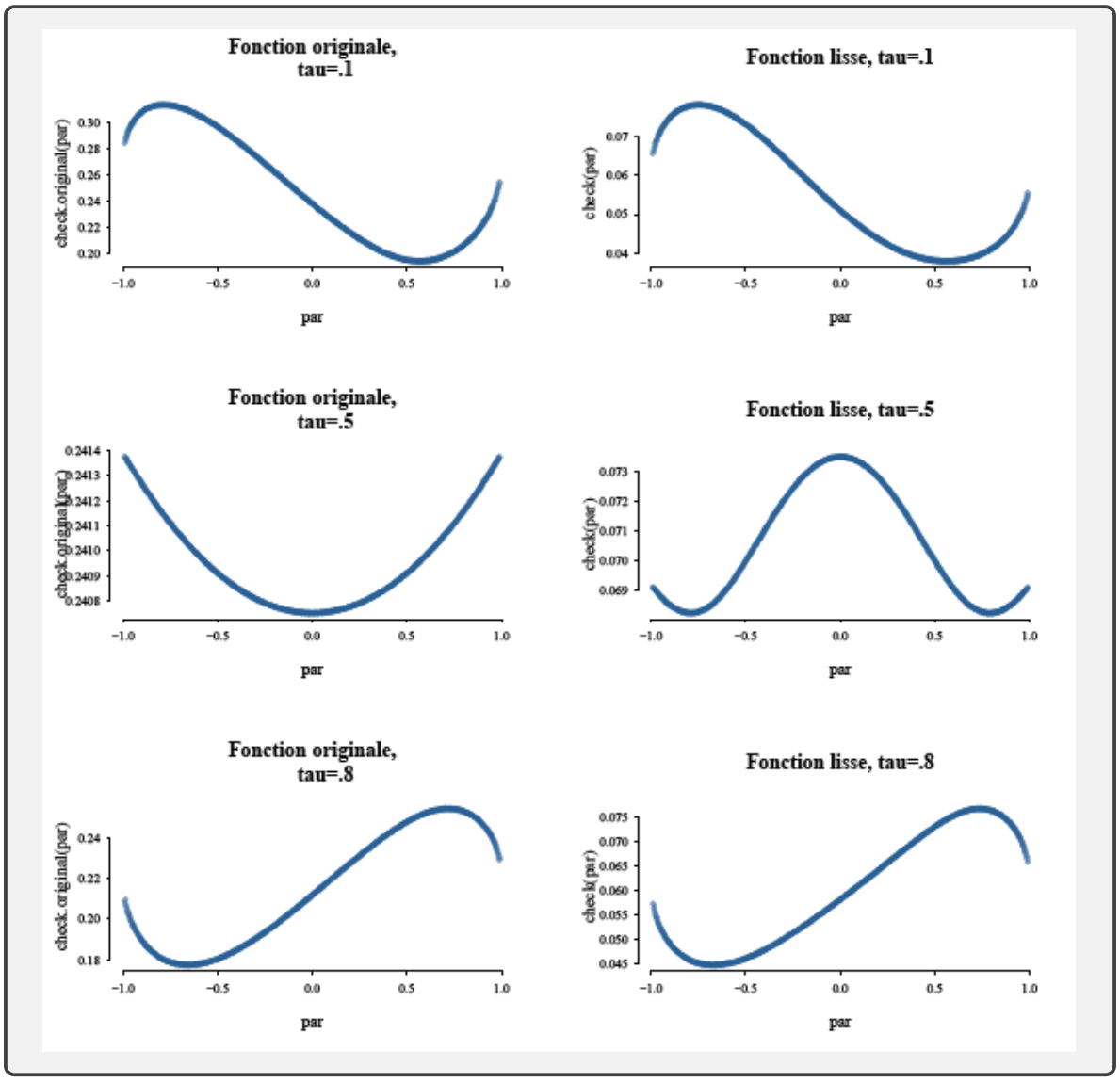

Figure 3.1: Courbes de la fonction objective lorsque la copule est normale bivariée aux quantiles $\tau = (.1, .5, .8)$. `tau` est le quantile et `par` le paramètre de la copule.

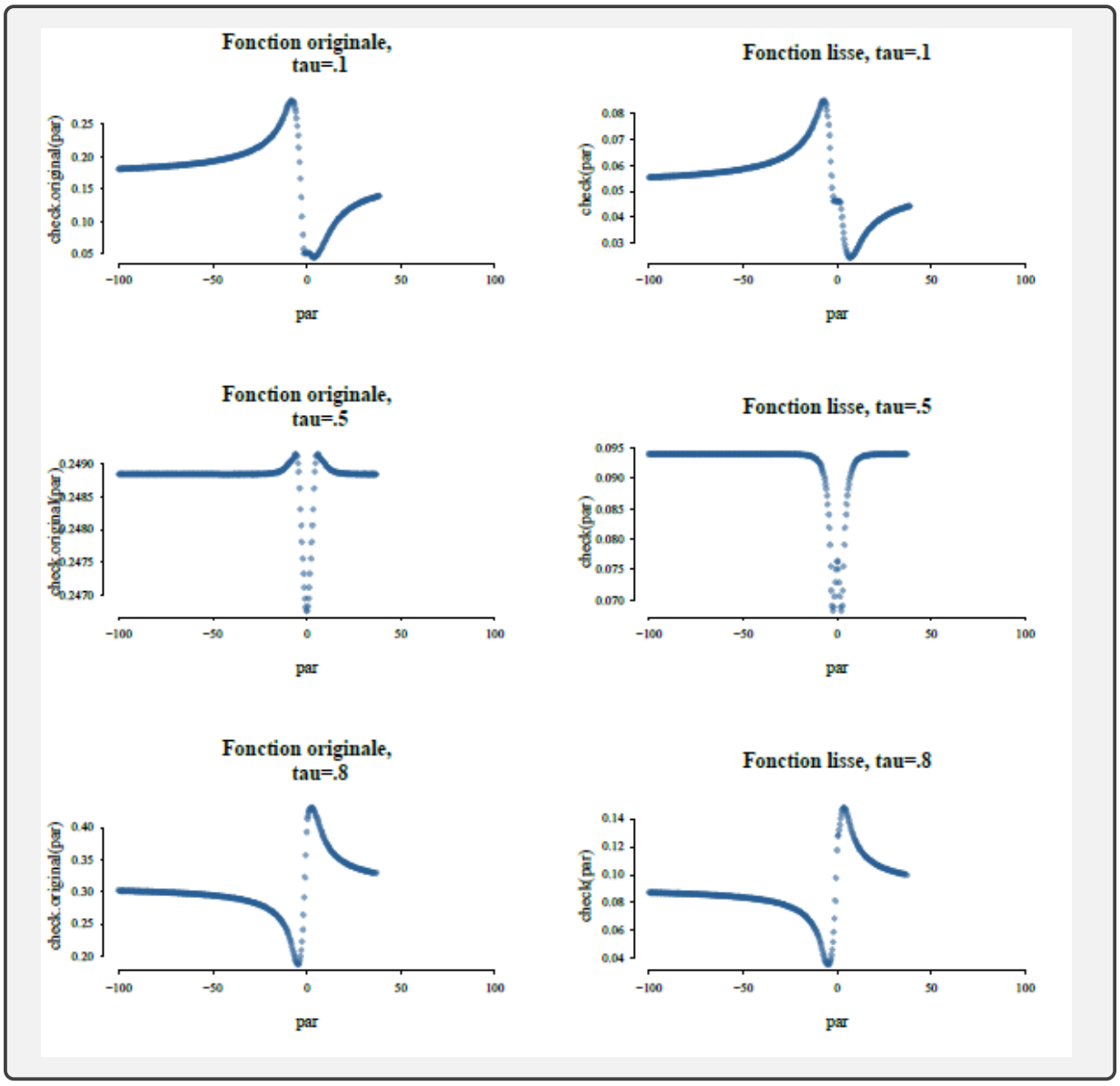

Figure 3.2: Courbe de la fonction objective fonction de la copule de Frank bivariée aux quantiles $\tau = (.1, .5, .8)$. `tau` est le quantile et `par` le paramètre de la copule.

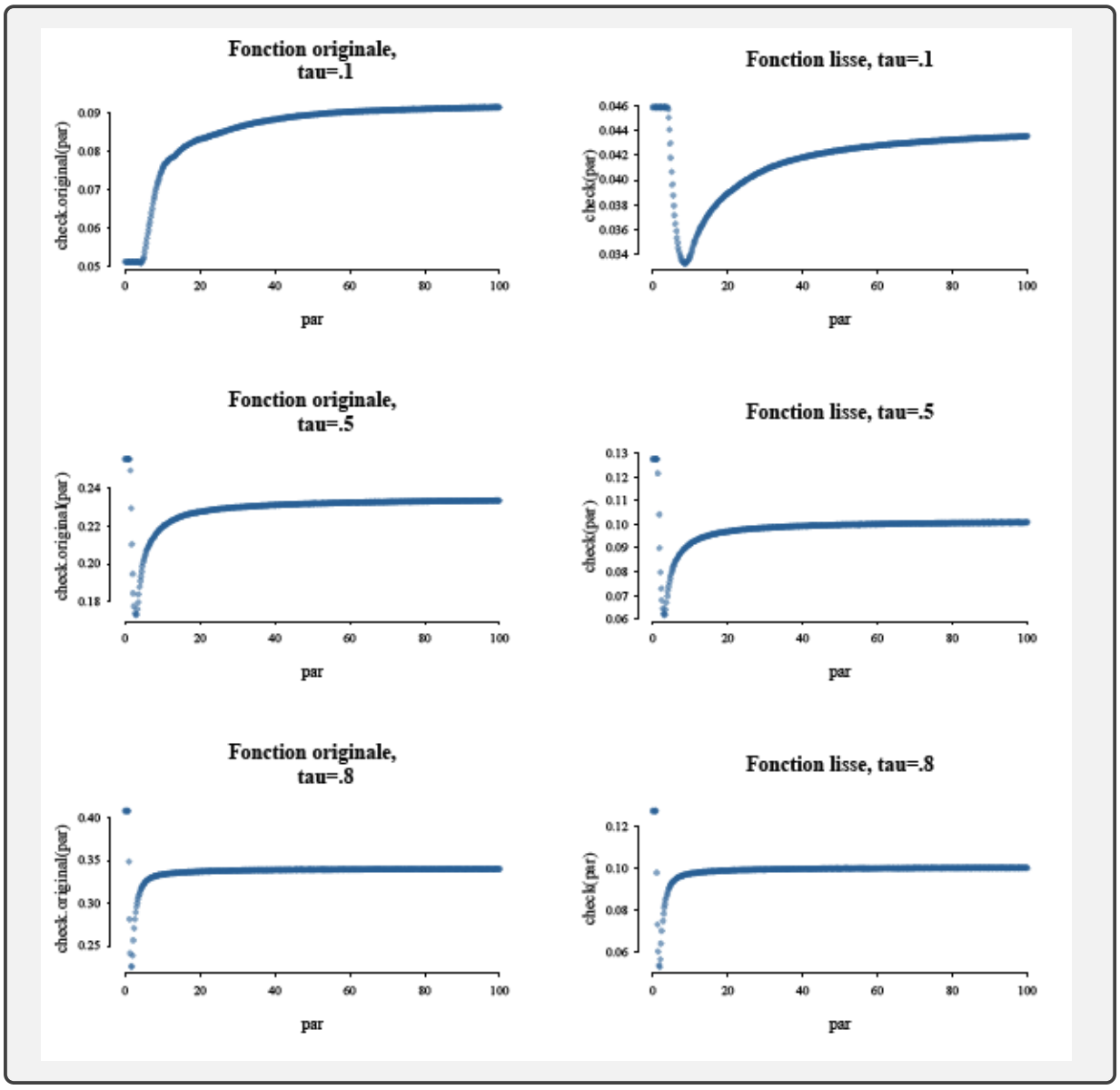

Figure 3.3: Courbe de la fonction objective fonction de la copule de Clayton bivariée aux quantiles $\tau = (.1, .5, .8)$. `tau` est le quantile et `par` le paramètre de la copule.

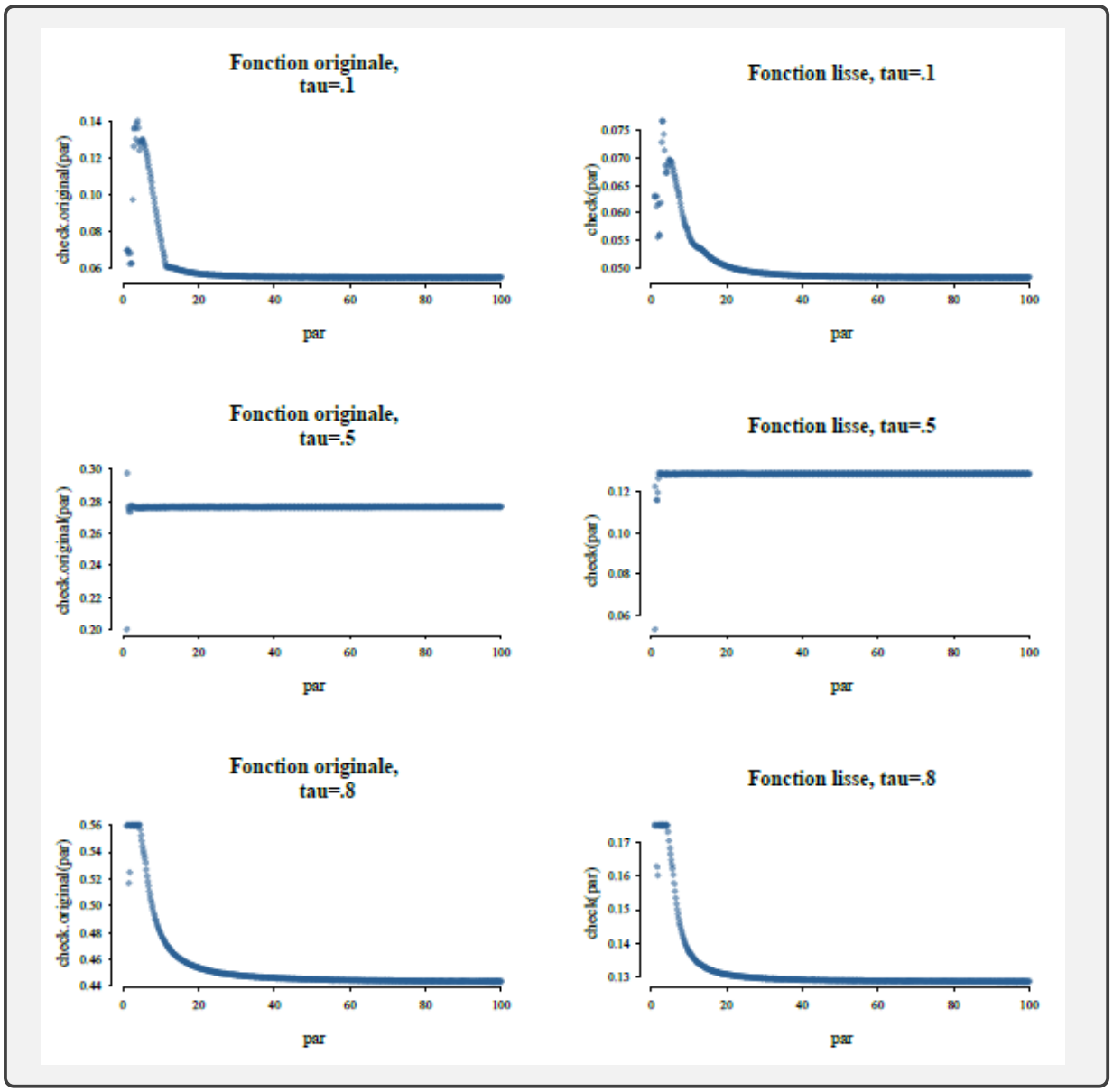

Figure 3.4: Courbe de la fonction objective fonction de la copule de Gumbel bivariée aux quantiles $\tau = (.1, .5, .8)$. `tau` est le quantile et `par` le paramètre de la copule.

Heureusement pour nous, la fonction objective du problème (3.4) ne dépend que du paramètre de la copule. Le résoudre se résume donc tout naturellement à estimer la copule. Une fois que le paramètre de la copule est obtenu, on est capable de calculer trivialement le quantile conditionnel estimé en se servant de l'équation (3.5). Le problème qui initialement consistait à estimer le paramètre de la copule en résolvant un problème d'optimisation difficile devient un tout autre problème dont la résolution nécessite deux étapes : la première, qui consiste à estimer la copule (le paramètre de la copule et ses marges) et la seconde à estimer le quantile conditionnel par « plug-in » de l'estimateur de la copule dans l'expression analytique de la fonction de quantile conditionnel. Une approche similaire a été utilisée dans les travaux de (Noh *et al.*, 2015) dans le cas de la variable réponse continue.

3.3.2 Estimation du quantile conditionnel

Estimer le paramètre de la copule en résolvant un problème d'optimisation s'est avéré fastidieux. Pour pallier ces difficultés, on procède autrement. D'abord on estime la copule indépendamment de la fonction du quantile conditionnel et ensuite on estime cette fonction de quantile conditionnel en se servant de la copule estimée.

Parmi toutes les méthodes d'estimation de la copule énumérées au Chapitre 2, nous choisissons d'utiliser la méthode semi-paramétrique CML pour des raisons précédemment expliquées. Cette méthode estime les marges de la copule de façon non paramétrique et le paramètre de la copule de façon paramétrique par la méthode du maximum de vraisemblance. Il faut par conséquent déterminer la fonction de vraisemblance à maximiser.

Fonction de vraisemblance d'un couple de variables aléatoires dont l'une est continue et l'autre discrète.

Soit (X, Y) un couple de variables aléatoires avec Y la variable binaire issue de la variable latente Y^* et X une variable continue. Soit F_X et F_{Y^*} les *cdf* respectives des variables X et Y^*, et soit C_θ la copule qui lie la distribution conjointe de (X, Y^*), $F^*(.,.)$, aux fonctions marginales.

Proposition 3.3.1. *La densité du couple (X, Y) est de la forme*

$$f(x,y) = f_X(x) \times \Big[C_\theta^{01}\big(\pi_0, F_X(x)\big)\Big]^{\mathbb{1}_{\{y=0\}}} \times \Big[1 - C_\theta^{01}\big(\pi_0, F_X(x)\big)\Big]^{\mathbb{1}_{\{y=1\}}},$$

avec $\pi_0 = F_{Y^*}(y^*) = P(Y^* < y^*) = P(Y = 0)$ et

$$C_\theta^{01}(\pi_0, F_X(x)) = \frac{\partial}{\partial F_X(x)} C_\theta(\pi_0, F_X(x)). \tag{3.7}$$

La preuve de cette proposition est détaillée en appendice à la section A.3.4.

Ainsi, la fonction de vraisemblance est définie par

$$\begin{aligned} L(\theta, \pi_0; \boldsymbol{x}, \boldsymbol{y}) &= \prod_{i=1}^{n} f(x_i, y_i) \\ &= \prod_{i=1}^{n} \Bigg\{ f_X(x_i) \times \Big[C_\theta^{01}\big(\pi_0, F_X(x_i)\big)\Big]^{\mathbb{1}_{\{y_i=0\}}} \times \\ &\qquad \Big[1 - C_\theta^{01}\big(\pi_0, F_X(x_i)\big)\Big]^{\mathbb{1}_{\{y_i=1\}}} \Bigg\}. \end{aligned} \tag{3.8}$$

Estimation du paramètre de la copule

Nous comptons estimer le paramètre de la copule de façon paramétrique par la méthode du maximum de vraisemblance. Sauf que, la fonction de vraisemblance

telle que définie ci-dessus dépend aussi des marges $\pi_0 = F_{Y^*}(y^*) = P(Y = 0)$ et $F_X(x)$, probabilités à estimer.

L'estimation du paramètre de la copule dépend de celle des marges. La méthode CML telle que proposée par (Genest *et al.*, 1995) ne fait aucune hypothèse sur les distributions marginales. Elle utilise les fonctions de distribution empirique des marges comme suit :

$$F_n(x) = \frac{1}{n+1} \sum_{k=1}^{n} \mathbb{1}_{\{x_k \leq x\}}. \tag{3.9}$$

On utilise $n + 1$ au lieu de n afin d'éviter des problèmes lors des simulations [4] π_0 peut être estimée soit par l'équation (3.9), auquel cas son estimateur est de la forme $\hat{\pi}_0 = \dfrac{\#\{y_i = 0\}}{n+1}$, soit en maximisant la fonction de vraisemblance (3.8) - méthode que nous privilegions dans notre démarche. La notation $\#\Omega$ désigne le cardinal de l'ensemble Ω.

Une fois les marges de la copule estimée, on estime le paramètre de la copule (et éventuellement π_0) en résolvant le problème d'optimisation

$$\arg\max_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}; (\boldsymbol{x}, \boldsymbol{y})),$$

où l est la fonction de log-vraisemblance définie par

$$\begin{aligned} l(\boldsymbol{\alpha}; (\boldsymbol{x}, \boldsymbol{y})) = \sum_{i=1}^{n} \Bigg\{ & \log \hat{f}_X(x_i) + \mathbb{1}_{\{y_i=0\}} \times \log C_\theta^{01}\big(\pi_0, F_n(x_i)\big) + \\ & \mathbb{1}_{\{y_i=1\}} \times \log \Big[1 - C_\theta^{01}\big(\pi_0, F_n(x_i)\big)\Big] \Bigg\}, \end{aligned} \tag{3.10}$$

avec $\boldsymbol{\alpha} = (\theta, \pi_0)$, $F_n(x)$ la *cdf* empirique de la variable aléatoire X et $\hat{f}_X$ l'estimateur empirique de la densité de la variable aléatoire X. Une façon d'obtenir $\hat{f}_X$

4. En effet, on relève le problème d'inexistence de limites de la fonction de densité de la variable aléatoire X lorsque sa *cdf* vaut 0 ou 1.

serait par exemple d'effectuer une dérivation numérique de la fonction F_n définie par l'équation (3.9).

Ainsi, une estimation du quantile conditionnel est donnée par

$$\hat{Q}_{Y|X=x}(\tau) = \mathbb{1}\big\{\widehat{m_\tau(x,\theta)} \geq \hat{y}^*\big\},$$

où

$$\widehat{m_\tau(x,\theta)} = \hat{F}^{-1}_{Y^*}\big(D^{-1}(\tau;\hat{\theta},F_n(x))\big), \tag{3.11}$$

$\hat{y}^*$ satisfait $\hat{\pi}_0 = P(Y^* < \hat{y}^*)$, $\hat{F}_{Y^*}(.) := F_n(.)$ et $D := C^{01}$ définie à l'équation (3.7).

3.3.3 Application à quelques copules

Dans cette section, nous appliquons notre modèle proposé aux familles de copules elliptiques et archimédiennes. Les formules analytiques des estimateurs $\widehat{m_\tau(.,.)}$ (et donc implicitement des estimateurs des quantiles conditionnels) sont repertoriés dans le Tableau 3.2.

La preuve est détaillée en appendice à la section A.3.5 seulement pour la copule normale.

3.4 Quid de la probabilité de succès conditionnelle ?

On montre en appendice à la section A.3.1 que

$$P(Y^* < y^*|X = x) = \frac{\partial C_\theta(F_{Y^*}(y^*), F_X(x))}{\partial F_X(x)}.$$

Comme la somme des probabilités sur le support de la variable Y est égale à 1, c'est-à-dire

$$P(Y = 0|X = x) + P(Y = 1|X = x) = 1,$$

Copules	$\widehat{m_\tau(x,\theta)}$
Normale	$\hat{F}_{Y^*}^{-1}\left[\Phi\left(\Phi^{-1}(\tau)\sqrt{1-\hat{\theta}^2}+\hat{\theta}\Phi^{-1}(F_n(x))\right)\right]$ $\hat{\theta}\in(-1,1)$
Frank	$\hat{F}_{Y^*}^{-1}\left[-\frac{1}{\hat{\theta}}\ln\left(1-\frac{\tau(1-e^{-\hat{\theta}})}{e^{-\hat{\theta}F_n(x)}+\tau(1-e^{-\hat{\theta}F_n(x)})}\right)\right]$ $\hat{\theta}\in\mathbb{R}-\{0\}$
Clayton	$\hat{F}_{Y^*}^{-1}\left[\left(\left(\tau^{-\frac{\hat{\theta}}{1+\hat{\theta}}}-1\right)(F_n(x))^{-\hat{\theta}}+1\right)^{-1/\hat{\theta}}\right]$ $\hat{\theta}\geq 0$
Gumbel	-* $1\leq\hat{\theta}<\infty$

Tableau 3.2: Expression analytique de l'estimateur $\widehat{m_\tau(.,.)}$ selon la copule de paramètre θ. * Gumbel n'admet pas d'expression analytique.

et que

$$P(Y^*<y^*|X=x)=P(Y=0|X=x),$$

alors, nous avons

$$\begin{aligned}P(Y=1|X=x) &= 1-P(Y=0|X=x)\\ &= 1-\frac{\partial C_\theta(F_{Y^*}(y^*),F_X(x))}{\partial F_X(x)}.\end{aligned}$$

D'où

$$P(Y=1|X=x)=1-\frac{\partial C_\theta(\pi_0,F_X(x))}{\partial F_X(x)}, \tag{3.12}$$

avec $\pi_0=F_{Y^*}(y^*)=P(Y=0)$.

Il s'ensuit qu'un estimateur de la probabilité de succès conditionnelle est donné par l'expression

$$\widehat{P(Y=1|X=x)} = 1-\left.\frac{\partial C_\theta(v,u)}{\partial u}\right|_{v=\hat{\pi}_0,u=F_n(x)}. \tag{3.13}$$

Le Tableau 3.3 nous donne l'expression de l'estimateur de la probabilité de succès conditionnelle dépendamment de quelques copules de paramètre θ.

Copules	$\widehat{P(Y = 1\|X = x)}$
Normale	$\Phi\left(\dfrac{\hat{\theta}\Phi^{-1}(F_n(x)) - \Phi^{-1}(\hat{\pi}_0)}{\sqrt{1-\hat{\theta}^2}}\right)$ $\hat{\theta} \in (-1, 1)$
Frank	$1 - \dfrac{e^{-\hat{\theta}F_n(x)}(1 - e^{-\hat{\theta}\hat{\pi}_0})}{e^{-\hat{\theta}F_n(x)} + e^{-\hat{\theta}\hat{\pi}_0} - e^{-\hat{\theta}} - e^{-\hat{\theta}(F_n(x)+\hat{\pi}_0)}}$ $\hat{\theta} \in \mathbb{R} - \{0\}$
Clayton	$1 - (F_n(x)^{-\hat{\theta}} + \hat{\pi}_0^{-\hat{\theta}} - 1)^{\frac{-1}{\hat{\theta}}-1} F_n(x)^{-\hat{\theta}-1}$ $\hat{\theta} \geq 0$
Gumbel	$1 - \dfrac{1}{F_n(x)}\big[-\ln F_n(x)\big]^{\hat{\theta}-1}\big[(-\ln F_n(x))^{\hat{\theta}} +$ $(-\ln \hat{\pi}_0)^{\hat{\theta}}\big]^{\frac{1-\hat{\theta}}{\hat{\theta}}} \hat{C}_G(F_n(x), \hat{\pi}_0)$ où $\quad \hat{C}_G(u, v) = \exp\Big\{-\big[(-\ln u)^{\hat{\theta}} + (-\ln v)^{\hat{\theta}}\big]^{\frac{1}{\hat{\theta}}}\Big\}$ et $\quad 1 \leq \hat{\theta} < \infty$

Tableau 3.3: Expression analytique de l'estimateur de la probabilité de succès conditionnelle selon la copule de paramètre θ.

3.4.1 Intervalle de la probabilité de succès conditionnelle de chaque individu

Dans cette partie, on s'intéresse à l'intervalle dans lequel se trouve la probabilité de succès conditionnelle de chaque individu. Pour le calculer, on procède comme dans les travaux de (Kordas, 2006).

Définissons comme au Chapitre 1, $P_{1|X} = P(Y = 1|X = x)$. La caractérisation

principale du quantile [5] entraine que

$$m_\tau(x,\theta) \quad \begin{matrix}\leqslant\\ \geqslant\end{matrix} \quad y^* \iff P_{1|X} \quad \begin{matrix}\leqslant\\ \geqslant\end{matrix} \quad 1-\tau,$$

relation qui est équivalente à

$$P\left(Y=1\middle| m_\tau(x,\theta) \quad \begin{matrix}\leqslant\\ \geqslant\end{matrix} \quad y^*\right) \quad \begin{matrix}\leqslant\\ \geqslant\end{matrix} \quad 1-\tau. \tag{3.14}$$

L'équation (3.14) est équivalente au système d'équations

$$P(Y=1|X=x)\begin{cases} > 1-\tau & \text{si } m_\tau(x,\theta) > y^* \ , \\ = 1-\tau & \text{si } m_\tau(x,\theta) = y^* \ , \\ < 1-\tau & \text{si } m_\tau(x,\theta) < y^* \ . \end{cases} \tag{3.15}$$

Il en résulte qu'un quantile, disons la médiane, sépare l'échantillon en deux groupes : ceux qui sont en dessous de la médiane conditionnelle et ceux qui y sont au dessus. Un changement de variables $U \to F^{-1}_{U|X}(\tau|X)$ conduit à

$$\begin{aligned} P_{1|X} &= \int_{\mathbb{R}} \mathbb{1}\big\{m(x,\theta)+U \geq y^*\big\} dF_{U|X}(U|X) \\ &= \int_0^1 \mathbb{1}\big\{m(x,\theta)+F^{-1}_{U|X}(\tau|X) \geq y^*\big\} dF_{U|X}(F^{-1}_{U|X}(\tau|X)) \\ &= \int_0^1 \mathbb{1}\big\{m(x,\theta)+Q_{U|X}(\tau) \geq y^*\big\} d\tau, \\ &= \int_0^1 \mathbb{1}\big\{m_\tau(x,\theta) \geq y^*\big\} d\tau, \end{aligned}$$

qui désigne la probabilité de succès conditionnelle en fonction d'une grille de quantiles $\tau \in (0,1)$ avec $m_\tau(.,.)$ définie à l'équation (3.5). On approxime cette probabilité conditionnelle sur une grille de quantiles estimés en procédant comme suit :

— Estimer le paramètre de la copule $\hat{\theta}$;

5. C'est une valeur qui divise la distribution en deux parties complémentaires.

— Calculer $\widehat{m_\tau(x,\theta)}$ en se servant de l'équation (3.11) sur la grille de quantiles

$$\eta = \{\tau_1, \ldots, \tau_m : \tau_1 < \cdots < \tau_m\};$$

— Déterminer enfin les quantités

$$\hat{\tau}_i = \arg\min_{\tau \in \eta} \{\tau : \widehat{m_\tau(x_i,\theta)} \geq \hat{y}^*\}, \quad i = 1, \ldots, n.$$

Une estimation de l'intervalle de la probabilité conditionnelle de succès de l'individu i est donnée par l'équation

$$\hat{P}_{i,1|X_i=x_i}(\eta) = [1 - \hat{\tau}_i, 1 - \hat{\tau}_{i,-1}),$$

où $\hat{\tau}_{i,-1}$ est le quantile qui précède $\hat{\tau}_i$.

Par exemple, admettons qu'on effectue des estimations sur la grille de quantiles $\eta = \{0.05, 0.1, \ldots, 0.95\}$ et que pour l'individu i, $\widehat{m_\tau(x_i,\theta)} - \hat{y}^*$ est positif si $\tau \geq 0.4$ et négatif si $\tau < 0.4$; alors, $\hat{\tau}_i = 0.40$ et $\hat{P}_{i,1|X_i}(\theta) = [0.60, 0.65)$. Si pour tous les quantiles de l'ensemble η , $\widehat{m_\tau(x_i,\theta)} - \hat{y}^*$ est positif, alors la probabilité de succès conditionnelle de l'individu i appartient à l'intervalle $[.95, 1)$. Si par contre $\widehat{m_\tau(x_i,\theta)} - \hat{y}^*$ est négatif, cette probabilité est dans l'intervalle $(0, 0.05]$.

Cette partie est suffisamment illustrée au chapitre 4 portant sur les simulations.

CHAPITRE IV

ÉTUDES DE SIMULATION ET ANALYSE DE DONNÉES RÉELLES

Dans ce chapitre, nous conduisons trois études de simulation afin d'évaluer la performance de notre méthodologie proposée en opposition aux méthodes existantes, en l'occurrence les modèles linéaires généralisés. Dans le premier scénario, on évalue le biais et l'erreur quadratique moyenne (MSE) des estimateurs des marges et du paramètre de la copule. Dans le deuxième, on étudie le choix de la copule et on effectue une analyse de sensibilité à la mauvaise spécification de la copule. Dans le troisième scénario, on compare notre approche à nos compétiteurs en terme de l'erreur de classification.On suppose en amont que la variable latente a une distibution normale.

4.1 Génération des données

Avant de procéder à l'évaluation du biais et MSE des estimateurs de la copule, il est de bon ton de savoir générer les données mixtes issues d'une copule. Nous proposons deux façons (**algorithmes 3 et 4**) de simuler un jeu de données (X, Y) de taille n tels que X, Y^* et Y vérifient le modèle (3.1) et ses hypothèses.

En supposant que $X \sim F_X$ et $Y^* \sim F_{Y^*}$, nous avons :

Algorithme 3 Simulation de données mixtes liées par une copule

a. Calculer le paramètre de la copule θ à partir d'un tau de Kendall fixé ;

b. Générer un couple $(U, V^*) \sim C_\theta$ de taille n ;

c. Définir $X = F_X^{-1}(U)$;

d. Définir $Y^* = F_{Y^*}^{-1}(V^*)$;

e. Retourner $(X, Y^*) \sim C_\theta$;

f. Définir $y^* = F_{Y^*}^{-1}(\pi_0)$, $\pi_0 \in (0, 1)$ fixé ;

g. Déterminer les indices j pour lesquels $Y_j^* \geq y_j^*$ et tirer les X_j correspondants ;

h. Déterminer les indices k pour lesquels $Y_k^* < y_k^*$ et tirer les X_k correspondants ;

i. Retourner $X = (X_j, X_k)$;

j. Construire la variable binaire Y qui prend 1 pour tous les indices j et 0 pour tous les indices k ;

k. Retourner $(X, Y) \sim C_\theta$ de taille n tel que $P(Y = 0) = \pi_0$.

Une autre façon de le faire serait de procéder comme avec la logit en générant $Y \sim \mathcal{B}(n, \mu)$, où

$$\mu = P(Y = 1|X = x) = 1 - \frac{\partial C(\pi_0, F_X(x))}{\partial F_X(x)}, \tag{4.1}$$

avec $\pi_0 = F_{Y^*}(y^*) = P(Y = 0)$ fixé et n la taille de l'échantillon.

Algorithme 4 Simulation de données mixtes liées par une copule

a. Générer un échantillon $X \sim F_X$ de taille n ;

b. Définir $U := F_X(X) \sim U[0,1]$, uniforme sur $[0,1]$;

c. Calculer le paramètre de la copule θ à partir d'un tau de Kendall fixé ;

d. Définir $\mu = 1 - D(\pi_0; \theta, U)$ pour $\pi_0 \in (0,1)$ et C_θ fixés. D définie à l'équation (3.2) ;

e. Générer $Y \sim \mathcal{B}(n, \mu)$, loi binomiale ;

f. Retourner $(X, Y) \sim C_\theta$ de taille n tel que $P(Y = 0) = \pi_0$.

Les données simulées, on peut estimer les paramètres puis évaluer leurs biais et MSE.

Dans la suite, nous fixons $B = 1000$ réplications et utilisons la fonction `optim` de `R` pour résoudre les problèmes d'optimisation.

4.2 Évaluation du biais et MSE des estimateurs de la copule

Il s'agit ici de vérifier que les estimateurs du paramètre de la copule et/ou de π_0 sont de « bons » estimateurs dans le sens où ils sont sans biais et de variances minimales. Ceci suppose d'estimer au préalable la copule en se servant de l'**Algorithme 5** ci-dessous. Une fois les paramètres de la copule estimés, on se sert de l'**Algorithme 6** pour calculer leurs biais et leurs MSE. On applique ces algorithmes à un jeu de données simulées selon la copule de tau de Kendall égal à 0.41 et on estime ensuite le paramètre de la même copule. Par exemple, si nous simulons les données selon la copule gaussienne, alors on estimera les paramètres de la copule gaussienne. Les résultats concernant les copules gaussienne, de Frank, de Clayton et de Gumbel sont mentionnés dans le Tableau 4.1 et illustrés par les

Figures 4.1, 4.2, 4.3 et 4.4.

Algorithme 5 ESTIMATION DES PARAMÈTRES DE LA COPULE

a. Générer $(X, Y) \sim C_\theta$ de taille n tel que $P(Y = 0) = \pi_0$ selon **Algorithme 4** ;

b. Définir $\hat{U} := \hat{F}_X(x) = \dfrac{1}{n+1} \sum_{k=1}^{n} \mathbb{1}_{\{x_k \leq x\}}$;

c. Définir $\hat{f}_X(x)$ - dérivée numérique de $\hat{F}_X(x)$;

d. Définir

$$L(\theta, \pi_0; \boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} \left\{ \hat{f}_{X_i}(x_i) \times \left[C_\theta^{01}\big(\pi_0, \hat{U}_i\big) \right]^{\mathbb{1}_{\{y_i=0\}}} \times \left[1 - C_\theta^{01}\big(\pi_0, \hat{U}_i\big) \right]^{\mathbb{1}_{\{y_i=1\}}} \right\};$$

e. Définir $l(\theta, \pi_0; \boldsymbol{x}, \boldsymbol{y}) = log\big(L(\theta, \pi_0; \boldsymbol{x}, \boldsymbol{y})\big)$;

f. Calculer $\arg\max_{(\theta, \pi_0)} l(\theta, \pi_0; \boldsymbol{x}, \boldsymbol{y})$;

g. Retourner $(\hat{\theta}, \hat{\pi}_0)$.

Algorithme 6 BIAIS ET MSE DES ESTIMATEURS

a. Pour b allant de 1 à B réplications,

a.1. Générer $(X^{(b)}, Y^{(b)}) \sim C_\theta$ de taille n tel que $P(Y^{(b)} = 0) = \pi_{0b}$ selon **Algorithme 4** ;

a.2. Calculer $(\hat{\theta}_b, \hat{\pi}_{0b})$ selon **Algorithme 5** ;

b. Calculer

$$Biais_{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \theta), \quad Biais_{\hat{\pi}_0} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\pi}_{0b} - \pi_0);$$

c. Calculer

$$MSE_{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2, \quad MSE_{\hat{\pi}_0} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\pi}_{0b} - \pi_0)^2;$$

d. Retourner $Biais_{\hat{\theta}}, Biais_{\hat{\pi}_0}, MSE_{\hat{\theta}}, MSE_{\hat{\pi}_0}$.

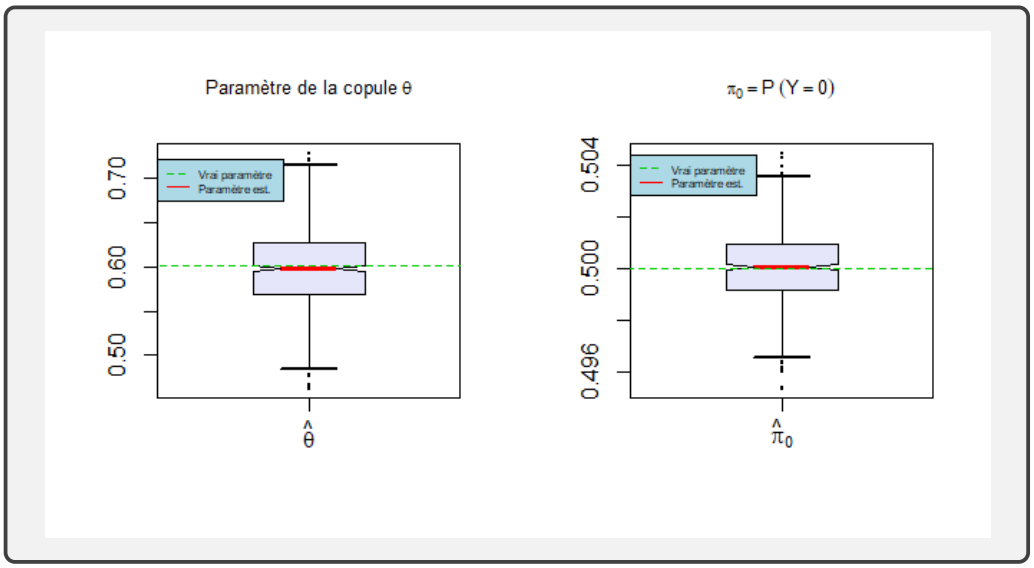

Figure 4.1: Diagramme à moustache des paramètres estimés du modèle de régression quantile à base de la copule normale de tau de Kendall égal à 0.41.

Tableau 4.1: Biais et MSE des estimateurs des paramètres θ et de π_0. Les données sont simulées selon la copule considérée de tau de Kendall 0.41.

	Copule gaussienne		Copule de Frank		Copule de Clayton		Copule de Gumbel	
	θ	π_0	θ	π_0	θ	π_0	θ	π_0
Biais	-3.46e-03	3.94e-05	-1.28e-03	1.04e-08	1.07e-02	4.21e-02	3.48e-03	1.98e-05
MSE	2.00e-03	1.90e-06	2.43e-01	2.83e-16	2.54e-06	1.71e-07	1.13e-02	4.04e-07

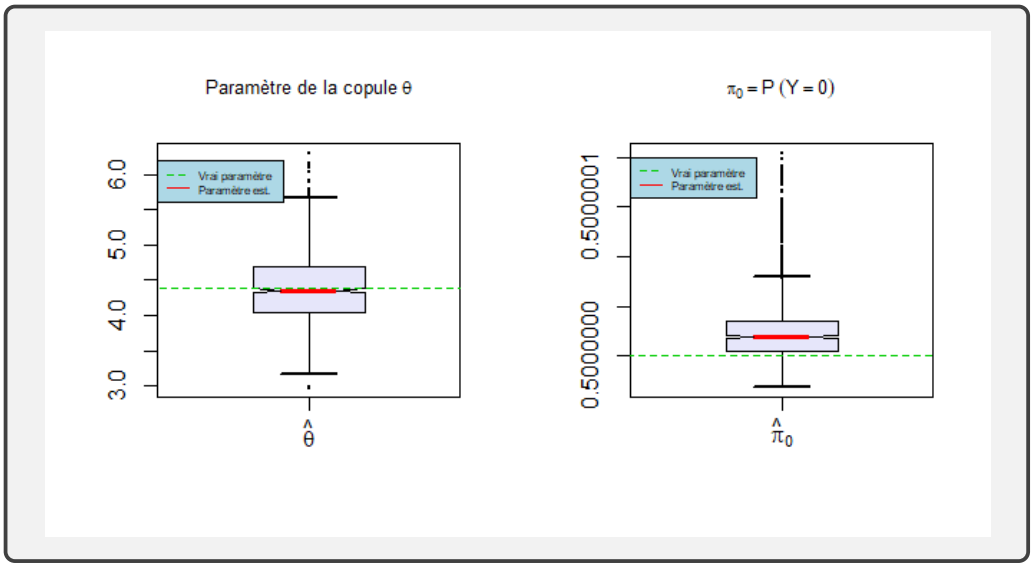

Figure 4.2: Diagramme à moustache des paramètres estimés du modèle de régression quantile à base de la copule de Frank de tau de Kendall égal à 0.41.

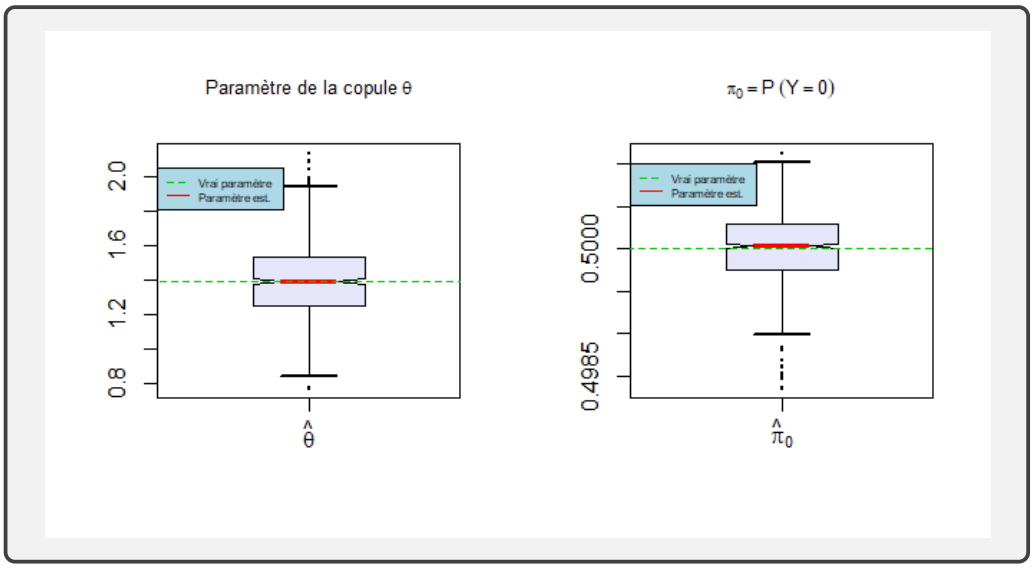

Figure 4.3: Diagramme à moustache des paramètres estimés du modèle de régression quantile à base de la copule de Clayton de tau de Kendall égal à 0.41.

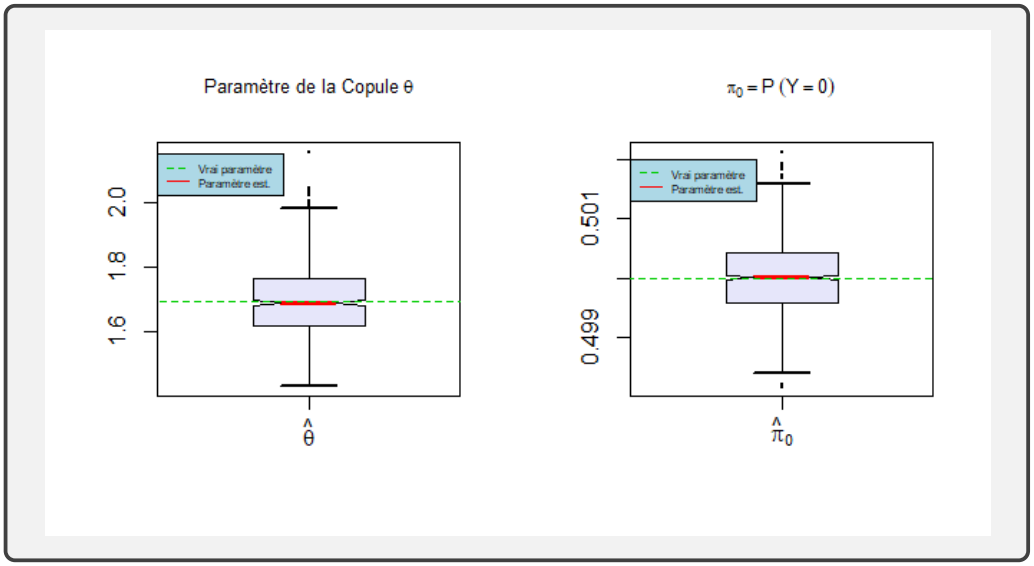

Figure 4.4: Diagramme à moustache des paramètres estimés du modèle de régression quantile à base de la copule de Gumbel de tau de Kendall égal à 0.41.

On peut observer dans le Tableau 4.1, que, quelque soit la copule, le biais et MSE des paramètres estimés tendent vers 0. Par conséquent, les estimateurs sont sans biais et de MSE presque nuls. Ils sont donc de bons estimateurs.

Les Figures 4.1, 4.2, 4.3 et 4.4 représentent respectivement les diagrammes à moustaches des paramètres estimés du modèle de régression quantile binaire à base des copule normale, de Frank, de Clayton et de Gumbel. La ligne interrompue bleue représente la vraie valeur du paramètre et la ligne rouge au milieu du box, la valeur estimée du même paramètre. De prime abord, on observe que les données formées par l'ensemble des paramètres estimés sont symétriques. De plus, on note des valeurs en dehors des moustaches : ce sont des valeurs aberrantes. On peut aussi voir que la ligne bleue et le segment rouge (la médiane) semblent pratiquement superposées : ceci indique que l'estimateur est non biaisé en considérant que la médiane et la moyenne coincident par la symétrie de la loi de l'estimateur, et de ce fait, corrobore l'interprétation du Tableau 4.1 donnée ci-dessus.

4.3 Choix de la copule et étude de la sensibilité

4.3.1 Choix de la copule

En pratique, on ne connait pas la distribution jointe des données observées et il est donc impossible de connaitre *a priori* la copule sous-jacente aux données recueillies. Pourtant, notre modèle a la copule pour épine dorsale. Il est donc impératif d'élaborer une méthode capable d'estimer la copule qui s'ajuste le mieux aux données. La fonction `fitCopula` du module `copula` par exemple permet de sélectionner la copule en utilisant le critère d'information d'Aïkake, AIC [1], lorsque

1. Le critère AIC sélectionne la copule de plus petit AIC (ou de façon équivalente la copule de plus grande valeur de log-vraisemblance).

les variables sont continues. Nous les avons testées [2] mais sans succès sur notre modèle où les variables sont mixtes (la covariable est continue et la variable d'intérêt binaire). Nous utilisons donc la fonction de vraisemblance de notre modèle pour calculer l'AIC. L'étude est effectuée avec la copule normale, la copule de Frank, de Clayton et de Gumbel, et implémentée dans l'**Algorithme 7**.

Algorithme 7 Sélection de la copule

a. Pour b allant de 1 à B réplications,

a.1. Générer $\left(X^{(b)}, Y^{(b)}\right) \sim C_\theta$ de taille n tel que $P\left(Y^{(b)} = 0\right) = \pi_0^{(b)}$. C_θ est la copule gaussienne par exemple ;

a.2. Pour chaque copule C fixée parmi les copules de Clayton, de Frank, de Gumbel et la copule gaussienne,

a.2.i. calculer $\left(\hat{\theta}^{c(b)}, \hat{\pi}_0^{c(b)}\right)$;

a.2.ii. calculer $AIC^{c(b)} = 2k - l\left(\hat{\theta}^{c(b)}, \hat{\pi}_0^{c(b)}; \boldsymbol{x}^{(b)}, \boldsymbol{y}^{(b)}\right)$,
l est la fonction de log-vraisemblance,
$k = 2$ est le nombre de paramètres du modèle ;

a.3. Sélectionner la copule C de AIC minimal ;

b. Retourner la copule la plus sélectionnée au terme des B réplications.

Dans le Tableau 4.2, nous comparons la fonction `fitCopula` de `R` - généralement utilisée pour calculer l'AIC dans le cas des variables continues - avec notre algorithme. Les données sont simulées selon une copule gaussienne de tau de Kendall égal à 0.41.

2. `fitCopula`, `BicopSelect` et `xvCopula`.

`Modèle ajusté`	Normal	Frank	Clayton	Gumbel	Total
`fitCopula`	0	0	0.9991	0.0009	1
`Loglikelihood`	0.648	0.107	0.062	0.183	1

Tableau 4.2: Probabilité de sélection de la copule avec `fitCopula` versus la log-vraisemblance de notre modèle. Les données sont simulées selon la copule normale de tau de Kendall égal à 0.41.

La fonction `fitCopula` face à un jeu de données mixtes ne sélectionne pas la bonne copule. Pour la suite des études, nous utiliserons par conséquent notre propre critère AIC.

Nous procédons à la sélection du modèle de copule à partir de plusieurs jeux de données simulées selon la copule normale, la copule de Frank, celle de Clayton et enfin celle de Gumbel. Les résultats sont mentionnés dans le Tableau 4.3.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel	Total
Normal	0.643	0.111	0.057	0.189	1
Frank	0.131	0.675	0.044	0.15	1
Clayton	0.057	0.038	0.905	0	1
Gumbel	0.192	0.109	0.001	0.698	1

Tableau 4.3: Probabilité de sélection de la copule via le critère AIC.

La bonne copule, comme nous pouvons le constater, n'est pas sélectionnée à 100%. Notre approche pourrait par conséquent sélectionner une mauvaise copule avec une certaine probabilité, d'où la question : que se passerait-il si une mauvaise copule

était sélectionnée [3] ? La mauvaise spécification de la copule aurait-elle un impact sur l'estimation des paramètres ? On se servira du tau de Kendall - mesure d'association commune à toutes les copules - pour étudier ce phénomène de mauvaise spécification.

4.3.2 Analyse de sensibilité à la mauvaise spécification de la copule

Dans cette section, on mesure la sensibilité au choix de la mauvaise copule sur le tau de Kendall, en se servant du biais et de l'erreur quadratique moyenne. On procède comme dans l'**Algorithme 8**.

Algorithme 8 ÉTUDE DE LA SENSIBILITÉ AU CHOIX DE LA MAUVAISE COPULE

a. Fixer un tau de Kendal κ et calculer le paramètre θ de la copule gaussienne par exemple ;

b. Pour b allant de 1 à B réplications,

a.1. Générer $(X^{(b)}, Y^{(b)}) \sim C_\theta^{Gauss}$ de taille n tel que $P(Y^{(b)} = 0) = \pi_0^{(b)}$;

a.2. Estimer $\theta^{c(b)}$ le paramètre d'une copule C considérée ;

a.3. Calculer $\hat{\kappa}^c$ à partir de la relation existante entre le tau de Kendall et le paramètre de la copule ;

c. Retourner $\hat{\boldsymbol{\kappa}}^{\boldsymbol{c}} = (\hat{\kappa}^{c(1)}, \ldots, \hat{\kappa}^{c(B)})$;

d. Calculer

$$Biais_{\kappa^c} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\kappa}^{c(b)} - \kappa), \quad MSE_{\kappa^c} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\kappa}^{c(b)} - \kappa)^2;$$

e. Retourner $Biais_{\kappa^c},\ MSE_{\kappa^c}$.

---

3. Il se pose ici le problème de mauvaise spécification de la copule.

Pour illustrer cet algorithme, on simule les données selon une copule gaussienne de tau de Kendall égal à 0.41.

Modèle ajusté	Normal	Frank	Clayton	Gumbel
Biais	-4.36e-05	-3.02e-02	-2.68e-02	3.40e-03
MSE	0.0013	0.0021	0.0021	0.0014

Tableau 4.4: Biais et MSE du tau de Kendall.

Il ne semble pas y avoir de différences importantes entres les paramètres estimés des copules incorrectement specifiées et ceux estimés de la copule correcte. Ainsi, le choix de la copule importe peu. La mauvaise spécification de la copule n'altère pas les estimations. Par conséquent, on s'attend à des résultats semblables si on choisit une copule plutôt qu'une autre.

4.4 Erreur de classification

Dans cette section, on mesure la performance de notre modèle en comparaison aux modèles SVM (Support Vector Machine) et logit, c'est-à-dire leur habilité à prédire la variable binaire d'intérêt. On utilise l'approche de validation croisée qui utilise une partie du jeu de données - jeu d'entrainement - pour entrainer le modèle et une autre - jeu de test - pour prédire. La méthodologie à suivre pour l'implémenter à notre modèle est indiquée à l'**Algorithme 9**. Les modèles de génération des données explorés sont les suivants :

M1. (X, Y) généré selon une copule parmi la copule normale, de Frank, de Clayton et de Gumbel de tau de Kendall égal à 0.5, selon l'**Algorithme 3** ;

M2. $Y|X$ suit une logit tel que $X \sim \mathcal{N}(0, 1)$ et $(\beta_0, \beta_1) = (\beta_0, 1)$, où β_0 vérifie la

relation

$$\begin{aligned}\pi_0 &= 1-p \\ &= 1-P(Y=1) \\ &= 1-\int_{\mathbb{R}} P\left(Y=1|X=x\right)\phi(x)dx \\ &= 1-\int_{\mathbb{R}} \left(1+\exp(-\beta_0-\beta_1 x)\right)^{-1}\phi(x)dx,\end{aligned}$$

ϕ étant la disribution de la loi normale standard ;

M3. $Y^* = X + \varepsilon$, avec $X \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \mathcal{N}(0,1)$, un modèle linéaire ;

M4. Un autre modèle homoscédastique symétrique mais avec les queues épaisses pour la variable latente Y^* : $Y^* = X + \varepsilon$, $X \sim \mathcal{N}(0,1)$ et $\varepsilon \sim$ *student* à 3 degrés de liberté ;

M5. Un modèle homoscédastique asymétrique : $Y^* = X + \varepsilon$, avec $X \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \chi^2_{(1)}$;

M6. Un modèle hétéroscédastique : $Y^* = X + (2+X)\varepsilon$ avec $X \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \mathcal{N}(0,1)$;

M7. Un dernier modèle qui est simulé comme suit :

— simuler

$$Y \sim \mathcal{B}(n, 1-\pi_0)$$

pour un certain paramètre $\pi_0 \in (0,1)$; n est la taille de l'échantillon ;

— pour tout i, si $Y_i = 1$, alors piger X_i suivant une distribution de Khi deux à 3 degrés de liberté, sinon piger X_i suivant une distribution de Cauchy standard.

Pour les modèles M_i, $i \in \{1,\ldots,6\} - \{2\}$, une fois la variable latente Y^* générée, on discrétise en utilisant la règle de classification $Y = \mathbb{1}\{Y^* \geq y^*\}$, où $y^* = F^{-1}_{Y^*}(\pi_0)$ avec $\pi_0 \in (0,1)$ fixé. Nous considérons $\pi_0 = 0.5$ dans le cas balancé et

$\pi_0 = 0.9$ dans le cas non balancé. Dans les estimations, nous utilisons l'hypothèse de normalité $Y^* \sim \Phi$, où Φ est la *cdf* de la loi normale.

Algorithme 9 ERREUR DE CLASSIFICATION

a. Pour b allant de 1 à B réplications,

a.1. Générer $(X^{(b)}, Y^{(b)}) \sim C_\theta^{Gauss}$ de taille n tel que $P(Y^{(b)} = 0) = \pi_0^{(b)}$;

a.2. Effectuer le « 2-Folds cross-validation » comme suit :

a.2.1. Diviser le jeu de données en 2 : $(X_{train}^{(b)}, Y_{train}^{(b)})$ et $(X_{test}^{(b)}, Y_{test}^{(b)})$;

a.2.2. Calculer $(\hat{\theta}_{train}^{(b)}, \hat{\pi}_{0\,train}^{(b)})$ selon **Algorithme 5** ;

a.2.3. Calculer le seuil de classification

$$\hat{y}^{(b)*} = \Phi^{-1}(\hat{\pi}_{0\,train}^{(b)});$$

a.2.4. Calculer

$$y_{pred}^{(b)} = \mathbb{1}\left\{\Phi^{-1}\big(D^{-1}(\hat{\theta}_{train}^{(b)}; \tau, F_n(x_{test}^{(b)}))\big) \geq \hat{y}^{(b)*}\right\};$$

a.2.5. Calculer à chaque itération l'erreur de classification

$$Err_b = \mathbb{1}\{y_{pred}^{(b)} \neq y_{test}^{(b)}\};$$

b. Retourner enfin l'erreur de classification moyenne de toutes les réplications

$$Err = \frac{1}{B}\sum_{b=1}^{B} Err_b.$$

(i) **Cas des données balancées**

Les données sont balancées lorsque $\pi_0 = .5$, ce qui mène dans la collecte des données à des échantillons avec la même proportion de 1 que de 0. Naturellement, il n'est pas utile dans ce cas de calculer les erreurs de classification à un quantile autre que la médiane. On fixe par conséquent le quantile à 0.5. Le résultat des simulations est consigné dans le Tableau 4.5.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel	Logit	SVM
M1 (Normal)	0.255	0.254	0.258	0.256	**0.253**	**0.253**
M1 (Frank)	0.228	0.227	0.232	0.228	**0.225**	**0.225**
M1 (Clayton)	0.251	0.248	**0.245**	0.253	0.247	**0.245**
M1 (Gumbel)	0.256	0.254	0.265	0.254	0.254	**0.253**
M2 (Logit)	0.330	0.330	0.331	0.333	**0.328**	0.329
M3 (student 3 ddl)	0.255	0.254	0.258	0.258	**0.252**	0.253
M4 (Chi 1 ddl)	0.274	0.269	0.300	0.296	0.243	**0.237**
M5 (lm)	0.254	0.253	0.258	0.258	**0.253**	**0.253**
M6 (Hétéroscédastique)	0.351	0.349	0.340	**0.339**	0.349	0.344
M7 (Mélange)	0.240	0.234	0.225	**0.224**	0.326	0.333

Tableau 4.5: Erreur de classification au quantile 0.5 avec $\pi_0 = 0.5$.

(ii) **Cas des données non balancées**

Les données sont non balancées lorsque $\pi_0 \neq .5$, ce qui mène dans la collecte des données à des échantillons avec les proportions de 1 et de 0 différentes. Dans ce cas, nous fixons[4] $\pi_0 = .9$. On calcule les erreurs de classification aux quantiles $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$, la prédiction pour Logit et SVM étant faites avec un seuil standard de 0.5. Les résultats sont consignés dans

4. La variable binaire est constituée d'environ 90% de 0 et 10% de 1

les Tableaux 4.6, 4.7, 4.8, 4.9 et 4.10.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel
M1 (Normal)	0.0993	0.1003	0.1003	**0.0979**
M1 (Frank)	0.0997	**0.0997**	0.0997	0.0999
M1 (Clayton)	0.0997	0.0997	**0.0997**	0.0997
M1 (Gumbel)	0.0940	0.1001	0.1002	**0.0899**
M2 (Logit)	**0.0994**	**0.0994**	**0.0994**	**0.0994**
M3 (student 3ddl)	**0.0987**	0.0999	0.0999	0.0999
M4 (Chi2 1ddl)	**0.0995**	0.1003	0.1003	0.1003
M5 (lm)	**0.0989**	0.0999	0.0999	0.0999
M6 (Hétéroscédastique)	**0.0999**	**0.0999**	**0.0999**	**0.0999**
M7 (Mélange)	0.0992	**0.0991**	**0.0991**	**0.0991**

Tableau 4.6: Erreur de classification au quantile 0.1 avec $\pi_0 = 0.9$.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel
M1 (Normal)	0.0966	0.1003	0.1003	**0.0947**
M1 (Frank)	0.1003	**0.0997**	0.0997	0.1011
M1 (Clayton)	0.0998	0.0997	**0.0997**	0.1005
M1 (Gumbel)	0.0860	0.0969	0.0982	**0.0813**
M2 (Logit)	**0.0994**	**0.0994**	**0.0994**	**0.0994**
M3 (student 3ddl)	**0.0949**	0.0997	0.0998	0.0998
M4 (Chi2 1ddl)	**0.0965**	0.1002	0.1003	0.1003
M5 (lm)	**0.0963**	0.0998	0.0998	0.0998
M6 (Hétéroscédastic)	0.1002	**0.0999**	**0.0999**	**0.0999**
M7 (Mélange)	0.1003	**0.0991**	**0.0991**	**0.0991**

Tableau 4.7: Erreur de classification au quantile 0.25 avec $\pi_0 = 0.9$.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel	Logit	SVM
M1 (Normal)	0.0939	0.0979	0.0991	0.0937	**0.0927**	0.0991
M1 (Frank)	0.1036	0.1018	0.1006	0.1060	0.1044	**0.0998**
M1 (Clayton)	0.1030	0.0999	**0.0997**	0.1068	0.1039	**0.0997**
M1 (Gumbel)	0.0762	0.0810	0.0855	0.0748	**0.0728**	0.0818
M2 (Logit)	0.0996	0.0995	**0.0994**	0.0995	0.0998	**0.0994**
M3 (student 3ddl)	0.0853	0.0955	0.0981	0.0982	**0.0786**	0.0901
M4 (Chi2 1ddl)	0.0853	0.0964	0.0990	0.0990	**0.0697**	0.0794
M5 (lm)	0.0933	0.0980	0.0987	0.0987	**0.0928**	0.0984
M6 (Hétéroscédastique)	0.1042	0.1020	0.1009	0.1012	0.1046	**0.1000**
M7 (Mélange)	0.1122	0.1050	0.1025	0.1037	0.1046	**0.0991**

Tableau 4.8: Erreur de classification au quantile 0.5 avec $\pi_0 = 0.9$.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel
M1 (Normal)	0.1207	0.1365	0.1404	**0.1142**
M1 (Frank)	0.1322	0.1478	0.1475	**0.1278**
M1 (Clayton)	0.1396	0.1501	0.1399	**0.1362**
M1 (Gumbel)	0.0986	0.1124	0.1175	**0.0906**
M2 (Logit)	0.1248	0.1318	**0.1182**	0.1197
M3 (student 3ddl)	**0.1128**	0.1293	0.1268	0.1256
M4 (Chi2 1ddl)	**0.1041**	0.1210	0.1122	0.1105
M5 (lm)	**0.1208**	0.1367	0.1410	0.1427
M6 (Hétéroscédastique)	**0.1381**	0.1551	0.1576	0.1620
M7 (Mélange)	**0.1447**	0.1573	0.1595	0.1664

Tableau 4.9: Erreur de classification au quantile 0.75 avec $\pi_0 = 0.9$.

Modèle ajusté / Vrai modèle	Normal	Frank	Clayton	Gumbel
M1 (Normal)	0.2339	0.2423	0.2669	**0.1966**
M1 (Frank)	0.2676	0.2741	0.3075	**0.2256**
M1 (Clayton)	0.3279	0.3349	0.3761	**0.2807**
M1 (Gumbel)	0.1888	0.1931	0.2171	**0.1541**
M2 (Logit)	**0.3283**	0.3381	0.3964	0.3995
M3 (student 3ddl)	**0.2600**	0.2719	0.3211	0.3185
M4 (Chi2 1ddl)	**0.2881**	0.2982	0.3758	0.3642
M5 (lm)	**0.2333**	0.2418	0.2665	0.2685
M6 (Hétéroscédastique)	**0.2707**	0.2781	0.3035	0.3090
M7 (Mélange)	**0.2653**	0.2662	0.2855	0.2949

Tableau 4.10: Erreur de classification au quantile 0.9 avec $\pi_0 = 0.9$.

Commentaires :

1. En général, les erreurs de classification de tous les modèles - le modèle SVM, le modèle de régression logistique et le nôtre - sont plus petites dans le cas des données non balancées (celles-ci sont en effet plus grandes quand les données sont balancées) ;
2. Quand les données sont balancées et issues du modèle 7, notre modèle performe mieux que le modèle de régression logistique et le modèle SVM ;
3. Quand les données sont non balancées, notre modèle est moins performant au fur et à mesure que le quantile s'éloigne de la probabilité de succès marginale. En outre, nous avons l'avantage d'effectuer de la classification avec notre modèle aux quantiles autres que la médiane de la variable latente.

Conclusion : Pour $p = 1 - \pi_0$ fixé, notre modèle proposé performe mieux lorsque l'écart $|\tau - p|$ tend vers 0, τ étant le quantile en lequel on évalue la performance de notre modèle et p la probabilité de succès marginale.

4.5 Calcul de la probabilité de succès conditionnelle

Nous avons calculé au Chapitre 3 la probabilité de succès à base de notre modèle. Dans cette section, on illustre la courbe de probabilité de succès conditionnelle de plusieurs modèles : le modèle logit, le modèle probit et le nôtre. Comme on peut voir à la Figure 4.5, à la médiane, quand la copule sous-jacente aux données est normale de tau de Kendall égal à 0.41, les courbes de probabilité de succès sont confondues. Ce n'est pas le cas, Figure 4.6, pour la copule de Clayton de même tau de Kendall. Ainsi, lorsque la copule est normale et ses marges aussi, les probabilités de succès conditionnelles des modèles probit, logit et du nôtre semblent égales.

Dans la suite, en implémentant l'**Algorithme 10** obtenu en s'inspirant de (Kordas, 2006), on illustre graphiquement à la Figure 4.7 les intervalles de probabilité de succès conditionnelle de chaque individu pour un jeu de données construit à base de la copule normale de tau de Kendall égal à 0.41 de marges normales standard. On y observe que, pour notre modèle, lorsque la copule est normale et ses marges aussi, la région formée par ces intervalles de probabilité de succès s'apparente à une région de confiance de la probabilité de succès conditionnelle.

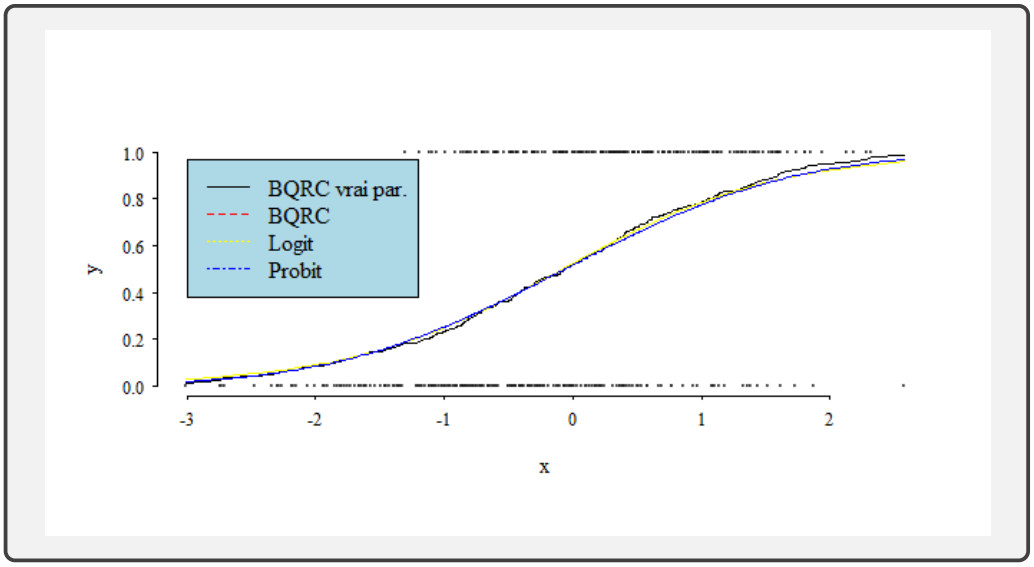

Figure 4.5: Courbe de probabilités de succès conditionnelles, copule normale de tau de Kendall égal à 0.41 versus autres modèles. `BQRC vrai par.` est le modèle de régression médiane à base de la copule normale où le paramètre de la copule est le vrai paramètre ; `BQRC` est le modèle de régression médiane à base de la copule normale où le paramètre de la copule est estimé.

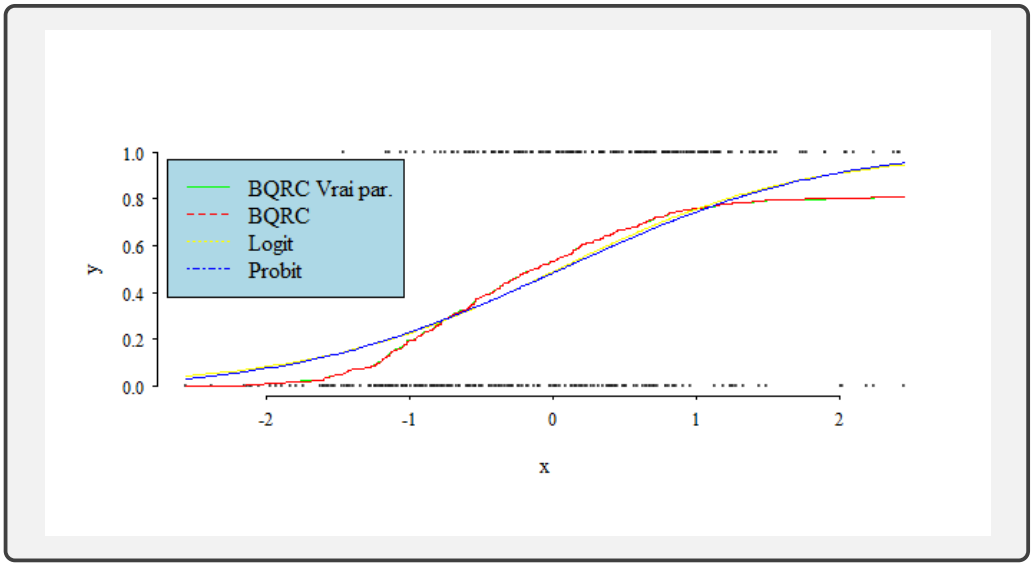

Figure 4.6: Courbe de probabilités de succès conditionnelles, copule de Clayton de tau de Kendall égal à 0.41 versus autres modèles. `BQRC vrai par.` est le modèle de régression médiane à base de la copule de Clayton où le paramètre de la copule est le vrai paramètre ; `BQRC` est le modèle de régression médiane à base de la copule de Clayton où le paramètre de la copule est estimé.

Algorithme 10 Estimateur de l'intervalle de probabilité de succès conditionnelle de chaque individu

a. Estimer le paramètre de la copule en se servant de l'**Algorithme 5**, obtenir $\hat{\theta}$;

b. Calculer pour chaque individu i la quantité $\widehat{m_\tau(x_i, \theta)}$ (en se servant de l'équation (3.11)) sur une grille ordonnée de quantiles

$$\eta = \{\tau_1, \ldots, \tau_m : \tau_1 < \cdots < \tau_m\};$$

c. Déterminer ensuite la solution

$$\hat{\tau}_i = \arg\min_{\tau \in \eta} \{\tau : \widehat{m_\tau(x_i, \theta)} \geq y^*\}, \quad i = 1, \ldots, n,$$

où y^* vérifie la relation $\pi_0 = P(Y^* < y^*)$, $\pi_0 \in (0, 1)$ fixé ;

d. Un intervalle contenant l'estimateur de la probabilité de succès conditionnelle de l'individu i est

$$[1 - \hat{\tau}_i, 1 - \hat{\tau}_{i,-1}),$$

où $\hat{\tau}_{i,-1}$ est le quantile qui précède directement $\hat{\tau}_i$;

e. Si $\widehat{m_\tau(x_i, \theta)} \geq y^*$ pour tout i, alors l'intervalle recherché est

$$[\tau_m, 1);$$

Si par contre $\widehat{m_\tau(x_i, \theta)} < y^*$ pour tout i, l'intervalle recherché est

$$(0, \tau_1].$$

On constate sur la Figure 4.7 que la région formée par les intervalles - qui contiennent les probabilités de succès conditionnelles de chaque individu - s'apparente à une région de confiance de la probabilité de succès conditionnelle.

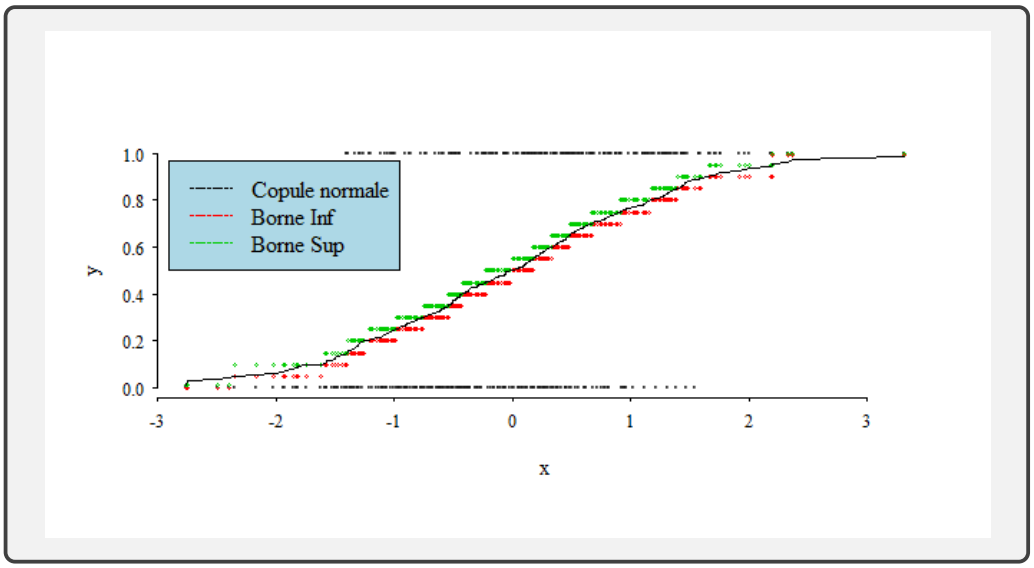

Figure 4.7: Intervalle de probabilité de succès conditionnelle de chaque individu : cas de la copule normale de tau de Kendall égal à 0.41.Borne Inf et Borne Sup signifient respectivement borne inférieure et borne supérieure ; Copule normale est la courbe de la probabilité de succès conditionnelle résultant de la copule normale de tau de Kendall 0.41.

4.6 Analyse de données réelles

Dans cette section, nous analysons des données réelles de méthylation de l'ADN avec notre modèle - le modèle de régression quantile binaire à base d'une copule.

La méthylation de l'acide désoxyribonucléique (ADN) est un processus biologique qui consiste en l'ajout des groupes méthyles (CH3) à la place d'un atome d'hydrogène (H) sur la molécule d'ADN. Elle est connue pour être impliquée dans le développement des tumeurs, en l'occurence du cancer de sein (des changements dans le profil de méthylation des cellules tumorales mammaires par exemple ont déjà été observés). Les objectifs des chercheurs sont donc entre autre d'étudier les changements de la méthylation globale de l'ADN dans des tissus normaux ainsi

qu'identifier les sites spécifiques différentiellement méthylés qui sont associés à une augmentation du risque de tumeur.

Le jeu de données interne à `R` dans le module `pcev` est constitué des variables

1. `methylation` : des niveaux de méthylation de 5 986 sites CpG, mesurés sur 40 échantillons en utilisant le séquençage bisulfite. Chaque échantillon correspond à l'un des trois types de cellules : Cellules B (8 échantillons), lymphocytes T (19 échantillons) et monocytes (13 échantillons). Les sites CpG sont situés autour du gène BLK connu pour être différentiellement méthylé chez les cellules de type T et les monocytes versus les cellules de type B.
2. `pheno` : le phénotype d'intérêt, variable binaire, avec $Y = 1$ si l'échantillon provient d'une cellule B et $Y = 0$ sinon.
3. `position` : des informations sur la position génomique de chaque site CpG.

La variable binaire `pheno` est la variable d'intérêt et la variable `methylation` la covariable qui est une matrice de 5 986 variables. Il nous faut, pour pouvoir appliquer notre modèle à ce jeu de données réelles, une seule covariable représentant la variable `methylation`. L'analyse en composante principale, l'ACP, est l'instrument qui va remédier à ce problème. En effet, c'est un outil intéressant de compression et de synthèse de l'information, très utile lorsqu'on est en présence d'une somme importante de données quantitatives à traiter et qui n'identifie pas au préalable la variable dépendante ou indépendante. À partir de p variables, elle définit k nouvelles variables qui sont des combinaisons linéaires des variables initiales. Ces nouvelles variables contiennent une bonne proportion de l'information totale sur toutes les variables de départ.

Nous allons donc dans un premier temps effectuer une analyse ACP sur le vecteur de covariables `methylation` et ne retenir que la première composante principale

comme la covariable de notre modèle. Notre jeu de données constitué, nous sélectionnons la copule qui se l'ajuste le mieux. Passée cette étape, nous calculons les erreurs de classification de notre modèle à base non seulement de la copule sélectionnée mais également des autres copules (mauvaise spécification de la copule). Le calcul de ces erreurs se fera aussi pour les modèles `logit` et `SVM`, modèles contre lesquels on se compare. On calcule ensuite les intervalles de la probabilité de succès conditionnelle de chaque individu et enfin, on en fait une représentation graphique.

4.6.1 Sélection de la copule et estimation des paramètres

La sélection de la copule qui s'ajuste le mieux aux données se fait par le critère AIC. Le critère sélectionne le modèle de plus petit AIC. Les résultats sont consignés dans le Tableau 4.11.

Modèles ajustés	Normal	Frank	Clayton	Gumbel
AIC	201.94	201.91	201.25	201.98

Tableau 4.11: Valeurs de l'AIC pour la sélection de la copule issue des données réelles.

Le critère AIC comme nous pouvons le constater sélectionne la copule de Clayton de paramètre estimé à 0.58. Cette copule est donc celle qui s'ajuste le mieux à ce jeu de données.

4.6.2 Erreur de classification

Avant de calculer les erreurs de classification, on souhaite déterminer le quantile optimal, c'est-à-dire celui en lequel l'erreur de classification de notre modèle est

minimal. On procède par validation croisée avec la règle 50-50 pour le faire. La Figure 4.8 illustre les chemins des erreurs de classification en fonction d'une grille de quantiles. On y observe que l'erreur de classification augmente à mesure qu'on s'éloigne du quantile $\tau = 0.4$.

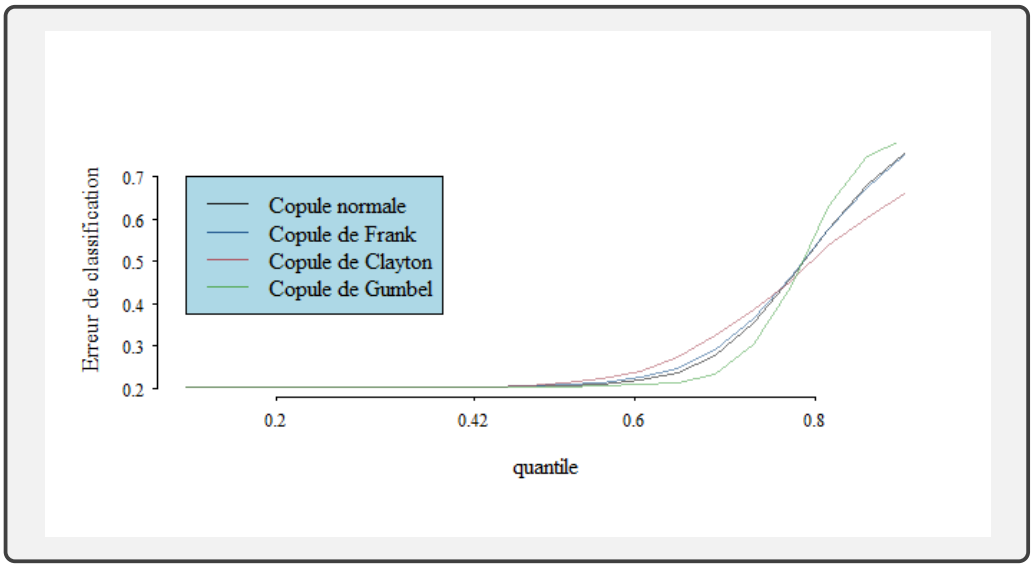

Figure 4.8: Chemins de l'erreur de classification en fonction d'une grille de quantiles. Copule normale, Copule de Frank, Copule de Clayton et Copule de Gumbel représentent respectivement ces chemins pour les copules normale, de Frank, de Clayton et de Gumbel obtenus des données réelles de méthylation de l'ADN.

On calcule enfin les erreurs de classification de notre modèle au quantile 0.5 et au quantile optimal, 0.1, en comparaison avec le modèle de régression logistique et le modèle SVM. Les résultats sont consignés dans les Tableaux 4.12 et 4.13.

Modèles ajustés	Gaussienne	Frank	Clayton	Gumbel	Logit	SVM
Erreurs de classification	0.2052	0.2122	0.2066	0.2048	0.2108	**0.2027**

Tableau 4.12: Erreurs de classification au quantile 0.5.

Modèles ajustés	Gaussienne	Frank	Clayton	Gumbel
Erreurs de classification	0.2027	0.2027	0.2027	0.2027

Tableau 4.13: Erreurs de classification au quantile 0.1.

Au quantile 0.5, le modèle de régression logistique a relativement la plus grande erreur de classification (hormis la copule de Frank) et le modèle SVM la plus petite. Notre modèle proposé performerait en général mieux que le modèle de régression logistique pour ce jeu de données réelles de méthylation de l'ADN. De plus, nous avons l'avantage de pouvoir effectuer de la classification aux quantiles autres que la médiane.

4.6.3 Probabilité de succès conditionnelle

Dans cette section, on illustre à la Figure 4.9 la courbe de probabilité de succès conditionnelle de notre modèle [5] versus celle des modèles logit et probit. Ces probabilités sont calculées à partir du jeu de données réelles de méthylation de l'ADN. Nous y illustrons également les intervalles de probabilités conditionnelles de chaque individu.

5. La copule utilisée est celle sélectionnée par le critère AIC

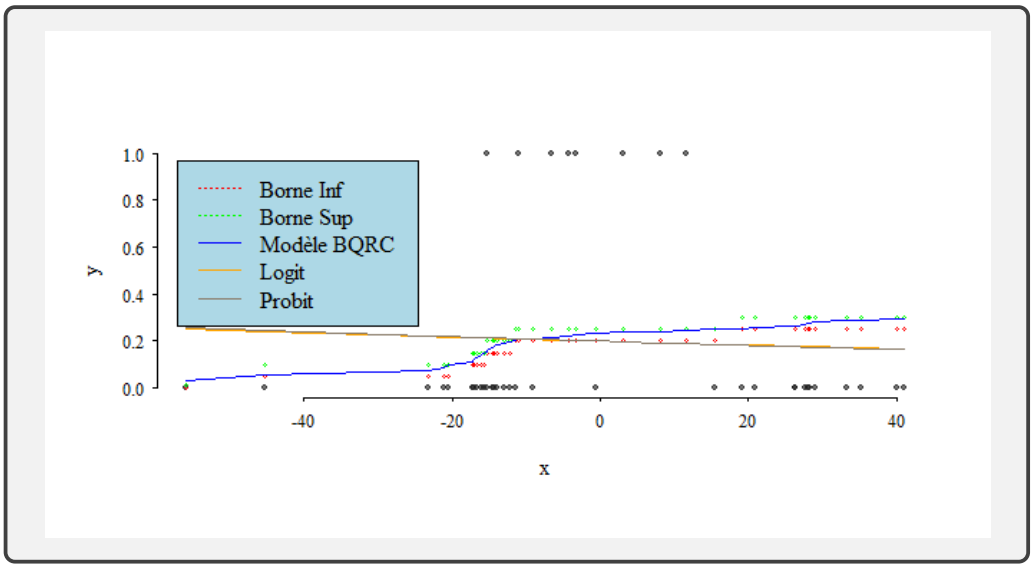

Figure 4.9: Courbe de probabilités de succès conditionnelles - notre modèle versus les modèles logit et probit. Borne Inf et Borne Sup signifient respectivement bornes inférieures et supérieures ; Modèle BQRC, Logit et Probit sont respectivement les courbes de probabilité de succès conditionnelle de la copule de Clayton, du modèle de régression logistique et du modèle probit obtenues des données de méthylation de l'ADN.

Comme on peut le voir à la Figure 4.9, les modèles de régression logit et probit (sur la première composante principale) semblent ne pas être adaptés à ce jeu de données, contrairement au modèle de régression quantile binaire à base de la copule de Clayton.

CONCLUSION

Nous avons tout au long de notre travail présenté une nouvelle méthode de régression, la régression quantile binaire à base d'une copule. Cette méthode peut être assimilée à un mélange de deux méthodes : la méthode de Kordas pour résoudre le modèle de régression quantile binaire linéaire (Kordas, 2006) et la méthode de Noh (Noh *et al.*, 2015) pour le problème de régression quantile basée sur les copules, où toutes les variables sont continues. Il nous a fallu passer par trois principaux chapitres pour la déployer.

Le chapitre 1, qui traite exclusivement de la méthode de Kordas pour le modèle de régression quantile binaire linéaire, présente des subtilités dans l'estimation des paramètres et le calcul de la probabilité de succès conditionnelle. L'estimation des paramètres n'est pas chose aisée puisque, pour le faire, Kordas est obligé de transformer le problème d'optimisation initial en un programme linéaire. Et nous savons tous ô combien les méthodes de programmation linéaire, bien qu'efficaces pour certains problèmes - en occurrence pour celui de Kordas, sont rudimentaires et fastidieuses pour ne pas dire complexes. C'est d'ailleurs pour cette raison que plusieurs auteurs se sont investis dans la recherche de nouvelles méthodes de résolution du problème d'optimisation de (Kordas, 2006). Parmi eux, nous avons mentionné (Aristodemou *et al.*, 2019) qui ont tranformé le problème d'optimisation initial en un problème résoluble par la méthode qu'ils ont nommée « Non Linear Least Asymmetric Weighted squares ». Bien que nous ne l'ayons pas mentionné dans notre travail, il y a également (Benoit *et al.*, 2017) qui y sont allés avec la méthode bayésienne. Ce qu'il faudrait enfin retenir de la méthode de Kor-

das c'est cette façon ingénieuse d'estimer la probabilité de succès conditionnelle étant donné que le calcul de cette quantité n'est possible que numériquement.

Au chapitre 2, nous avons fait un développement non exhaustif de la théorie des copules. Ce chapitre s'est avéré importantissime pour la construction de notre modèle ainsi que tous les développements théoriques y afférents. Nous y avons principalement présenté les caractéristiques des deux principales familles de copules ainsi que les méthodes d'estimation de celles-ci. De l'estimation paramétrique à l'estimation non parmétrique en passant par l'estimation semi-paramétrique, une méthode proposée par (Genest *et al.*, 1995) a été retenue pour être appliquée à notre modèle proposé : il s'agit de la méthode d'estimation semi-paramétrique de la copule par maximum de vraisemblance canonique (CML pour Canonical Maximum Likelihood). Elle consiste à estimer les marges de la copule de façon non paramétrique (*cdf* empiriques pour les marges) et le paramètre de la copule de façon paramétrique en maximisant la fonction de pseudo-vraisemblance du modèle.

Le chapitre 3 quant à lui a été entièrement consacré à la construction de notre modèle, l'estimation du quantile conditionnel et le calcul de la probabilité de succès conditionnelle. Il est bon de souligner que notre modèle a plusieurs avantages sur celui de Kordas. En effet, il est exempt du problème de translation-échelle contrairement aux modèles de régression quantile. Ce gros avantage nous a permis d'estimer le quantile conditionnel en deux étapes : nous avons en premier lieu estimé le paramètre de la copule par la méthode CML et en second effectué un « plug-in » du paramètre estimé de la copule dans l'expression analytique du quantile conditionnel. Il nous a également permis d'avoir une expression analytique de la probabilité de succès conditionnelle estimée, ce qui n'était pas possible avec le modèle de (Kordas, 2006). Pour Kordas, la difficulté dans le calcul de la probabilité de succès conditionnelle l'a poussé à estimer le plus « petit » intervalle qui la

contient. Lorsqu'on aplique la même méthode à notre modèle, on obtient aussi des intervalles ; sauf que la région formée par ceux-ci s'apparente à une région de confiance de la probabilité de succès conditionnelle.

Un des résultats cruciaux que nous avons obtenu pour notre modèle proposé est le suivant : « *lorsque la copule est normale et ses marges aussi, notre modèle est équivalent au modèle de Kordas* ». Cette proposition implique implicitement que nous pouvons, à partir de notre modèle, calculer l'expression analytique de la probabilité de succès conditionnelle du modèle de Kordas. Rappelons qu'il n'est pas possible de calculer analytiquement cette quantité à partir du modèle de Kordas. De plus, toujours sous l'hypothèse de la copule normale de marges normales, on observe que les courbes des probabilités de succès conditionnelle de notre modèle, de ceux de la logit et de la probit sont presque confondues. Ceci implique que lorsque la copule est gaussienne et ses marges aussi, la probabilité de succès de notre modèle est très proche de celle de la logit et de la probit.

Dans la partie simulation, nous avons comparé notre modèle avec le modèle de régression logistique, que nous savons tous difficilement imbattable sur le pan de la classification. Nous avons pu construire un scénario dans lequel notre modèle performe mieux que le modèle de régression logistique. Nous avons même appliqué notre modèle à un jeu de données réelles qu'on peut retrouver dans le module `pcev` du logiciel `R`. On a pu, après analyse de ce jeu spécifique de méthylation de l'ADN, conclure que notre modèle proposé comparativement aux modèles de régression logit et probit, serait le plus approprié pour expliquer la relation entre la variable d'intérêt et la covariables.

Bien qu'ayant obtenu des résultats satisfaisants, il se trouve que l'estimation paramétrique de la copule, l'épine dorsale de notre modèle, présente des inconvénients. En effet, quand la fonction de régression n'est pas monotone, pire encore pour un

prédicteur de grande dimension, aucune copule paramétrique ne s'ajuste aux données[6] (Dette *et al.*, 2014). Le lecteur peut observer à la Figure 4.10 qu'avec le jeu de données simulé selon la distribution $(Y_i, X_i),\ i = 1, \ldots, n$, i.i.d. telle que $Y_i = (X_i - .5)^2 + \sigma\varepsilon_i,\ X_i \sim U[0,1],\ \sigma = .25$ et ε_i de loi normale standard pour tout i, la copule paramétrique n'est pas adaptée pour modéliser la distribution jointe de (X, Y). Par contre, la copule non paramétrique s'ajuste parfaitement au nuage de points.

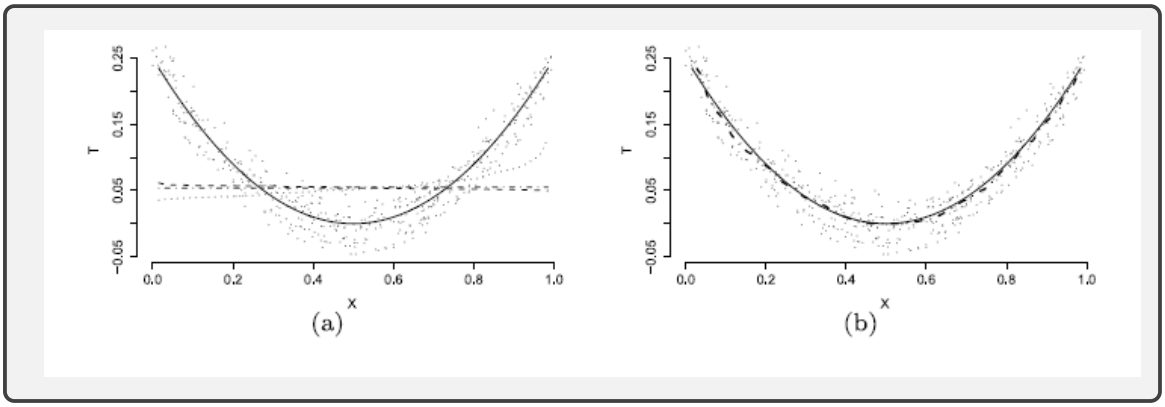

Figure 4.10: Estimations de régression quantile basées sur des copules de l'exemple minimaliste unidimensionnel. Les copules de Gauss, de Gumbel et de Frank sont utilisées pour l'estimation paramétrique des copules dans (a), tandis que (b) représente l'ajustement de régression résultant d'une estimation non paramétrique de la densité de la copule. Figure 1 de (De Backer *et al.*, 2017).

Comme perspective, nous envisageons donc d'explorer notre modèle en dimension supérieure à 2 où la distribution jointe, modélisée par la copule, sera estimée de façon non paramétrique.

6. Les courbes représentatives des copules paramétriques sont monotones

RÉFÉRENCES

Aravkin, A. Y., Kambadur, A., Lozano, A. C. et Luss, R. (2014). Sparse quantile huber regression for efficient and robust estimation. *arXiv preprint arXiv :1402.4624*.

Aristodemou, K., He, J. et Yu, K. (2019). Binary quantile regression and variable selection : A new approach. *Econometric Reviews*, *38*(6), 679–694.

Benoit, D. F., Van den Poel, D. *et al.* (2017). bayesqr : A bayesian approach to quantile regression. *Journal of Statistical Software*, *76*(7), 1–32.

Bernard, C. et Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, *138*, 104–126.

Cherubini, U., Luciano, E. et Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

De Backer, M., El Ghouch, A., Van Keilegom, I. *et al.* (2017). Semiparametric copula quantile regression for complete or censored data. *Electronic Journal of Statistics*, *11*(1), 1660–1698.

Dette, H., Van Hecke, R. et Volgushev, S. (2014). Some comments on copula-based regression. *Journal of the American Statistical Association*, *109*(507), 1319–1324.

d'Haultfoeuille, X. et Givord, P. (2014). La régression quantile en pratique.

Embrechts, P. (2009). Copulas : A personal view. *Journal of Risk and Insurance*, *76*(3), 639–650.

Fathia, K. (2018). Sur les copules bivariées.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon, 3ˆ e serie, Sciences, Sect. A*, *14*, 53–77.

Genest, C., Ghoudi, K. et Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, *82*(3), 543–552.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Hashem, H., Vinciotti, V., Alhamzawi, R. et Yu, K. (2016). Quantile regression with group lasso for classification. *Advances in Data Analysis and Classification*, *10*(3), 375–390.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica : journal of the Econometric Society*, 505–531.

Jennings, L., Wong, K. et Teo, K. (1996). Optimal control computation to account for eccentric movement. *The ANZIAM Journal*, *38*(2), 182–193.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Joe, H. et Xu, J. J. (1996). The estimation method of inference functions for margins for multivariate models.

Kim, G., Silvapulle, M. J. et Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, *51*(6), 2836–2850.

Koenker, R. (2005). Quantile regression cambridge univ.

Koenker, R. et Bassett Jr, G. (1978). Regression quantiles. *Econometrica : journal of the Econometric Society*, 33–50.

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, *21*(3), 387–407.

Manski, C. F. (1985). Semiparametric analysis of discrete response : Asymptotic properties of the maximum score estimator. *Journal of econometrics*, *27*(3), 313–333.

Mkhadri, A., Ouhourane, M. et Oualkacha, K. (2017). A coordinate descent algorithm for computing penalized smooth quantile regression. *Statistics and Computing*, *27*(4), 865–883.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Nie, N. H., Bent, D. H. et Hull, C. H. (1975). *SPSS : Statistical package for the social sciences*, volume 227. McGraw-Hill New York.

Noh, H., Ghouch, A. E. et Van Keilegom, I. (2015). Semiparametric conditional quantile estimation through copula-based multivariate models.

Journal of Business & Economic Statistics, *33*(2), 167–178.

Schweizer, B. (1991). Thirty years of copulas. In *Advances in probability distributions with given marginals* 13–50. Springer.

Shih, J. H. et Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384–1399.

Sklar, A., SKLAR, A. et Sklar, C. (1959). Fonctions de répartition à n dimensions et leurs marges.

Tibiletti, L. (1995). Beneficial changes in random variables via copulas : An application to insurance. *The Geneva Papers on Risk and Insurance Theory*, *20*(2), 191–202.

APPENDICE A

PREUVES DE PROPOSITIONS

A.1 Preuves du chapitre 1

A.1.1 Preuve de la Proposition 1.2.1

On aimerait trouver la valeur de θ qui minimise la fonction de perte espérée

$$\begin{aligned} E[\rho_\tau(Y-\theta)] &= \int_{\mathbb{R}} \rho_\tau(Y-\theta)d\theta \\ &= \tau\int_{y\geq\theta}(y-\theta)f_Y(y)dy + (\tau-1)\int_{y<\theta}(y-\theta)f_Y(y)dy \\ &= \tau\int_{y\geq\theta}(y-\theta)dF_Y(y) + (\tau-1)\int_{y<\theta}(y-\theta)dF_Y(y). \end{aligned}$$

Désignons par $\hat{\theta}$ la solution à ce problème de minimisation. Alors, elle vérifie les conditions de KKT, Karush Kuhn Tucker :
Condition de premier ordre : $\dfrac{d}{d\theta}E[\rho_\tau(Y-\theta)] = 0$

En utilisant la **règle d'intégration de Leibniz** pour intervertir la dérivée et

l'intégrale, il vient que :

$$\begin{aligned}
0 &= \frac{d}{d\theta}E[\rho_\tau(Y-\theta)] \\
&= \frac{d}{d\theta}\left[\tau\int_{y\geq\theta}(y-\theta)dF_Y(y)+(\tau-1)\int_{y<\theta}(y-\theta)dF_Y(y)\right] \\
&= -\tau\int_{y\geq\theta}dF_Y(y)+(1-\tau)\int_{y<\theta}dF_Y(y) \\
&= -\tau F_Y(y)\Big|_\theta^\infty+(1-\tau)F_Y(y)\Big|_{-\infty}^\theta \\
&= -\tau(1-F_Y(\theta))+(1-\tau)(F_Y(\theta)-0) \\
&= -\tau+F_Y(\theta).
\end{aligned}$$

Il s'ensuit que

$$\hat{\theta}=F_Y^{-1}(\tau),$$

qui est par définition le quantile d'ordre τ de la distribution de la *v.a.* Y.

La fonction $E[\rho_\tau(Y-\theta)]$ étant convexe, cette solution est un minimum global.

A.1.2 Preuve de la propriété d'équivariance par transformation monotone

Comme g est croissante, alors

$$P\Big(Y\leq Q_\tau(Y)\Big)=P\Big(g(Y)\leq g(Q_\tau(Y))\Big),$$

et par la définition de $Q_\tau(Y)$, on a clairement $P\Big(Y\leq Q_\tau(Y)\Big)\geq\tau$. Par ailleurs, $Q_\tau\left(g(Y)\right)=\inf\{y \text{ tel que } F_{g(Y)}(y)\geq\tau\}$. Il s'ensuit que $g\big(Q_\tau(Y)\big)\geq Q_\tau\big(g(Y)\big)$.

Réciproquement, en définissant $g^-(v)=sup\{x \text{ tel que } g(x)\leq v\}$, on a pour $u=Q_\tau\big(g(Y)\big)$

$$\tau\leq P\big(g(Y)\leq u\big)\leq P\big(Y\leq g^-(u)\big).$$

Par définition de $u=Q_\tau\big(g(Y)\big)$ et de $Q_\tau(Y)$, on a la relation $g^-(u)\geq Q_\tau(Y)$. Et comme g est continue à gauche, $g\big(g^-(u)\big)\leq u$. D'où $Q_\tau\big(g(Y)\big)=u\geq g\big(g^-(u)\big)\geq g\big(Q_\tau(Y)\big)$, ce qui termine la preuve.

A.1.3 Preuve de la Proposition 1.4.1

Soit $i \in \{1, 2, \ldots, n\}$. Posons

$$L_i = \rho_\tau\big(y_i - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\}\big)$$

et

$$S_i = \Big[y_i - (1-\tau)\Big]\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\}.$$

Alors,

$$L_i = \begin{cases} \rho_\tau(y_i - 1) & \text{si } \boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\ , \\ \rho_\tau(y_i) & \text{si } \boldsymbol{x}_i^\top \boldsymbol{\beta} < 0\ . \end{cases}$$

Comme

$$\begin{aligned} \rho_\tau(y_i) &= \tau|y_i|\mathbb{1}(y_i \geq 0) + (1-\tau)|y_i|\mathbb{1}(y_i \leq 0) \\ &= \tau|1|\mathbb{1}(y_i = 1) + (1-\tau)|0|\mathbb{1}(y_i \leq 0) \\ &= \tau\mathbb{1}\{y_i = 1\}, \end{aligned}$$

et

$$\begin{aligned} \rho_\tau(y_i - 1) &= \tau|y_i - 1|\mathbb{1}(y_i - 1 \geq 0) + (1-\tau)|y_i - 1|\mathbb{1}(y_i - 1 \leq 0) \\ &= \tau|1 - 1|\mathbb{1}(y_i = 1) + (1-\tau)|0 - 1|\mathbb{1}(y_i = 0) \\ &= (1-\tau)\mathbb{1}\{y_i = 0\}, \end{aligned}$$

alors,

$$\begin{aligned} L_i &= \rho_\tau(y_i - 1)\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} + \rho_\tau(y_i)\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} < 0\} \\ &= (1-\tau)\mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} + \tau\mathbb{1}\{y_i = 1\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} < 0\} \\ &= (1-\tau)\mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} + \tau\mathbb{1}\{y_i = 1\}(1 - \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\}) \\ &= (1-\tau)\mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} + \tau\mathbb{1}\{y_i = 1\} - \tau\mathbb{1}\{y_i = 1\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} \\ &= \tau\mathbb{1}\{y_i = 1\} - \Big[\tau\mathbb{1}\{y_i = 1\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} - (1-\tau)\mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\}\Big]. \end{aligned}$$

Si $y_i = 1$,

$$S_i = \Big[1 - (1-\tau)\Big]\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} = \tau \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\};$$

et si $y_i = 0$,

$$S_i = \Big[0 - (1-\tau)\Big]\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} = -(1-\tau)\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\}.$$

Il en découle que

$$\tau \mathbb{1}\{y_i = 1\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} - (1-\tau)\mathbb{1}\{y_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{\beta} \geq 0\} \quad = \quad S_i.$$

D'où

$$L_i = -S_i + constante, \quad \forall i \in \{1, 2, \ldots, n\}.$$

Donc

$$\arg\min_{\boldsymbol{\beta}} L_{n\tau}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} -S_{n\tau}(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} S_{n\tau}(\boldsymbol{\beta}).$$

A.2 Preuves du chapitre 2

A.2.1 Preuve de la Proposition 2.3.1

Soit α et β deux fonctions strictement croissantes, X et Y deux variables aléatoires de marges F_1 et G_1 et de copule $C_{X,Y}(u,v) = H_1(x,y)$. Désignons par F_2 et G_2 les marges des variables $\alpha(X)$ et $\beta(Y)$ et par H_2 leur loi conjointe. Alors, pour tous $u, v \in I$ on a :

$$\begin{aligned}
C_{\alpha(X),\beta(Y)}(u,v) &= C_{\alpha(X),\beta(Y)}(F_2(x), G_2(y)) \\
&= H_2(x,y) \\
&= P\left(\alpha(X) \leq x, \beta(Y) \leq y\right) \\
&= P\left(X \leq \alpha^{-1}(x), Y \leq \beta^{-1}(y)\right) \\
&= H_1(\alpha^{-1}(x), \beta^{-1}(y)) \\
&= C_{X,Y}(F_1(\alpha^{-1}(x)), G_1(\beta^{-1}(y))) \\
&= C_{X,Y}(F_2(x), G_2(y)).
\end{aligned}$$

A.2.2 Preuve de la Proposition 2.3.3 : théorème de Sklar

Les copules sont par définition des distributions conjointes de variables uniformes. Ainsi,

$$C(u_1, u_2) = P[U_1 \leq u_1, U_2 \leq u_2],$$

avec $u_i = F_i(x_i)$ pour tout i. Comme les F_i sont continues alors les $F_i(X_i)$ sont uniformes sur $I = [0, 1]$. Soit F la *cdf* jointe des variables aléatoires (X_1, X_2). Alors,

$$\begin{aligned} C(u_1, u_2) &= P(U_1 \leq u_1, U_2 \leq u_2) \\ &= P(F_1(X_1) \leq u_1, F_2(X_2) \leq u_2) \\ &= P(X_1 \leq F_1^{-1}(u_1), X_2 \leq F_2^{-1}(u_2)) \\ &= F(F_1^{-1}(u_1), F_2^{-1}(u_2)) \\ &= F(x_1, x_2). \end{aligned}$$

A.2.3 Preuve de la Proposition 2.3.5

Pour tout $t \in I$, on a

$$\begin{cases} C^-(t, 0) = C^-(0, t) = \max(t - 1, 0) = 0, \\ \\ C^-(t, 1) = C^-(1, t) = \max(t, 0) = 0. \end{cases}$$

Soit $u_1, u_2, v_1, v_2 \in I$ tels que $u_1 \geq u_2$ et $v_1 \geq v_2$. Alors, nous avons

$$\max(u_1 + v_2 - 1, 0) - \max(u_2 + v_2 - 1, 0) \geq \max(u_1 + v_1 - 1, 0) - \max(u_2 + v_1 - 1, 0).$$

Donc W est une copule. On montre de même que M est une copule.

Il reste à montrer que

$$W \leq C(u, v) \leq M.$$

Soit $u, v \in I$. On a

$$\begin{cases} C(u,v) \leq C(u,1) = u, \\ \\ C(u,v) \leq C(1,v) = v. \end{cases}$$

Il s'ensuit que

$$C(u,v) \leq \min(u,v) = M.$$

Par ailleurs, comme toute copule est 2-croissante, alors elle vérifie l'inégalité

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0,$$

pour tous $u_1, u_2, v_1, v_2 \in I$ tels que $u_1 \geq u_2$ et $v_1 \geq v_2$. En particulier pour $u_1 = u, v_1 = v, u_2 = v_2 = 1$, on a en plus de la relation $C(u,v) \geq 0$ que :

$$\begin{aligned} C(1,1) - C(1,v) - C(u,1) + C(u,v) \geq 0 \quad &\Longrightarrow \quad 1 - v - u + C(u,v) \geq 0 \\ &\Longrightarrow \quad C(u,v) \geq u + v - 1 \\ &\Longrightarrow \quad C(u,v) \geq \max(u + v - 1, 0) \\ &\Longleftrightarrow \quad C(u,v) \geq W. \end{aligned}$$

A.2.4 Preuve de la Proposition 2.4.2

Le rhô de Spearman est défini comme étant proportionnel à la différence entre la probabilité de concordance et celle de discordance des couples aléatoires (X_1, Y_1) et (X_2, Y_3) :

$$\rho_{X,Y} = 3P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) > 0\Big] - 3P\Big[\big((X_1 - X_2) - (Y_1 - Y_3)\big) < 0\Big].$$

Ainsi, on a

$$\begin{aligned} P\left[((X_1 - X_2) - (Y_1 - Y_3)) > 0\right] &= \iint_{I^2} P\left[((X_1 - X_2) - (Y_1 - Y_3)) > 0 \middle| X_2 = u, Y_3 = v\right] dudv \\ &= \iint_{I^2} \left\{P\left(X_1 > u, Y_1 > v\right) + P\left(X_1 < u, Y_1 < v\right)\right\} dudv \\ &= \iint_{I^2} (1 - u - v + 2C(u,v)) dudv \\ &= 2\iint_{I^2} C(u,v) dudv. \end{aligned}$$

De la même façon,

$$\begin{aligned} P\left[\left((X_1 - X_2) - (Y_1 - Y_3)\right) < 0\right] &= 1 - P\left[((X_1 - X_2) - (Y_1 - Y_3)) > 0\right] \\ &= 1 - 2\iint_{I^2} C(u,v) dudv. \end{aligned}$$

On en déduit que

$$\rho_{X,Y} = 12\iint_{I^2} C(u,v) dudv - 3.$$

A.2.5 Preuve de la Proposition 2.4.3

On le démontre de façon similaire au cas du rhô de Spearman.

A.3 Preuves du chapitre 3

Dans cette partie, toutes les preuves sont effectuées uniquement pour la copule gaussienne. Le cas des autres copules se fait de façon analogue.

A.3.1 Preuve de la Proposition 3.2.1

Par définition, la copule de paramètre θ a pour expression :

$$\begin{aligned} C_\theta(u, v^*) &= F_{XY^*}(x, y^*) \\ &= P(X \leq x, Y^* \leq y^*) \\ &= \int_{-\infty}^{x} P(Y^* \leq y^* | X = t) f_X(t) dt, \end{aligned}$$

où $u = F_X(x)$, $v^* = F_{Y^*}(y^*)$ avec F_X et F_{Y^*} les *cdf* respectives des variables X et Y^*.

Par définition, nous avons

$$\begin{aligned} D(v^*;\theta,u) := \frac{\partial C_\theta(F_X(x),F_{Y^*}(y^*))}{\partial F_X(x)} &= \frac{\partial x}{\partial F_X(x)}\frac{\partial C_\theta(F_X(x),F_{Y^*}(y^*))}{\partial x} \quad \text{(règle de la chaîne)} \\ &= \frac{1}{f_X(x)}\frac{\partial}{\partial x}\int_{-\infty}^{x} P(Y^* \le y^*|X=t)f_X(t)dt \\ &= P(Y^* \le y^*|X=x) \\ &= F_{Y^*|X=x}(y^*). \end{aligned}$$

Donc

$$\tau := F_{Y^*|X=x}(y^*) = D(v^*;\theta,u) \in (0,1).$$

Il s'ensuit, en plus de la propriété de l'inverse de la composée de deux fonctions [1], que le quantile conditionnel de Y^* sachant X a pour expression

$$Q_{Y^*|X=x}(\tau) := F^{-1}_{Y^*|X=x}(\tau) = F^{-1}_{Y^*}\big(D^{-1}(\tau;\theta,u)\big)$$

où D est définie à l'équation (3.2).

A.3.2 Preuve de la Proposition 3.2.2

Supposons que

$$Y = \mathbb{1}\{Y^* \ge y^*\};$$

Alors, comme la fonction $g : t \mapsto g(t) = \mathbb{1}(t \ge y^*)$ est croissante, on a d'après la **propriété d'équivariance par tranformation monotone** des quantiles, pour $\tau \in (0,1)$, que

$$\begin{aligned} Q_{g(Y^*)|X=x}(\tau) &= g(Q_{Y^*|X=x}(\tau)) \\ &= \mathbb{1}\{Q_{Y^*|X=x}(\tau) \ge y^*\} \\ &= \mathbb{1}\big\{F^{-1}_{Y^*}\big(D^{-1}(\tau;\theta,u)\big) \ge y^*\big\}, \end{aligned}$$

---

1. Si f et g sont inversibles, alors $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$; ce qui équivaut à $\big[g(f(x))\big]^{-1} = f^{-1}(g^{-1}(x))$ pour tout x tel que les fonctions $f(x)$ et $g^{-1}(x)$ sont définies.

où $u = F_X(x)$ et θ le paramètre de la copule qui lie X et Y^*. Donc, pour x quelconque dans le support de X, nous avons

$$Q_{Y|X}(\tau) = \mathbb{1}\left\{F_{Y^*}^{-1}\big(D^{-1}(\tau;\theta,u)\big) \geq y^*\right\}. \tag{A.1}$$

A.3.3 Preuve des formules du Tableau 3.1

La preuve se fait en deux étapes : on calcule d'abord la dérivée partielle de la copule, puis on détermine l'inverse de l'expression obtenue.

(i) Déterminons $\frac{\partial C_\theta(u,v)}{\partial u}$, où C_θ est la copule gaussienne bivariée de marges gaussiennes [2] et de paramètre $\theta \in (-1,1)$.

Par définition,

$$\begin{aligned} C_\theta(u,v) &= P(X \leq x, Y \leq y) \\ &= P\big(X \leq \Phi^{-1}(u), Y \leq \Phi^{-1}(v)\big) \\ &= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left\{ - \frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)} \right\} dxdy. \end{aligned}$$

Posons

$$g(x,y) = \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left\{ - \frac{x^2 - 2\theta xy + y^2}{2(1-\theta^2)} \right\},$$
$$b = \Phi^{-1}(u) \quad \text{et} \quad d \;=\; \Phi^{-1}(v).$$

Alors,

$$\begin{aligned} \frac{\partial C_\theta(u,v)}{\partial u} &= \frac{\partial}{\partial u} \int_{-\infty}^{b} \int_{-\infty}^{d} g(x,y)dxdy \\ &= \frac{\partial b}{\partial u} \frac{\partial}{\partial b} \int_{-\infty}^{b} \int_{-\infty}^{d} g(x,y)dxdy \quad \text{(Règle de la chaîne)} \\ &= \frac{1}{\Phi'(\Phi^{-1}(u))} \int_{-\infty}^{d} \left(\frac{\partial}{\partial b} \int_{-\infty}^{b} g(x,y)dx \right) dy \quad \text{(Théorème de Leibniz)} \\ &= \frac{1}{\phi(b)} \int_{-\infty}^{d} g(b,y)dy. \end{aligned}$$

---

2. C'est-à-dire $u = \Phi(x)$ et $v = \Phi(y)$, où Φ est la *cdf* d'une loi normale centrée réduite.

Mais

$$\begin{aligned} g(b,y) &= \frac{1}{2\pi\sqrt{1-\theta^2}}\exp\left\{-\frac{b^2-2\theta by+y^2}{2(1-\theta^2)}\right\} \\ &= \frac{1}{2\pi\sqrt{1-\theta^2}}\exp\left\{-\frac{(y-\theta b)^2-\theta^2b^2+b^2}{2(1-\theta^2)}\right\} \\ &= \frac{1}{2\pi\sqrt{1-\theta^2}}\exp\left\{-\frac{(y-\theta b)^2}{2(1-\theta^2)}-\frac{b^2(1-\theta^2)}{2(1-\theta^2)}\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{b^2}{2}\Big\}\frac{1}{\sqrt{2\pi(1-\theta^2)}}\exp\left\{-\frac{(y-\theta b)^2}{2(1-\theta^2)}\right\} \\ &= \phi(b)\phi\Big(\frac{y-\theta b}{\sqrt{1-\theta^2}}\Big). \end{aligned}$$

Il s'ensuit que

$$\begin{aligned} D(v;\theta,u) &:= \frac{\partial C_\theta(u,v)}{\partial u} \\ &= \frac{1}{\phi(b)}\int_{-\infty}^{d}\phi(b)\phi\Big(\frac{y-\theta b}{\sqrt{1-\theta^2}}\Big)dy \\ &= \int_{-\infty}^{d}\phi\Big(\frac{y-\theta b}{\sqrt{1-\theta^2}}\Big)dy \\ &= \Phi\Big(\frac{d-\theta b}{\sqrt{1-\theta^2}}\Big). \end{aligned}$$

Donc

$$D(v;\theta,u)=\Phi\Big(\frac{d-\theta b}{\sqrt{1-\theta^2}}\Big), \tag{A.2}$$

avec $b=\Phi^{-1}(u)$ et $d = \Phi^{-1}(v)$.

(ii) Déterminons l'inverse partiel de $D(v,\theta)$ par rapport à sa composante v.

Pour tout $\tau\in(0,1)$,

$$\begin{aligned} D(v;\theta,u)=\tau &\iff \Phi\Big(\frac{d-\theta b}{\sqrt{1-\theta^2}}\Big)=\tau \\ &\iff \frac{d-\theta b}{\sqrt{1-\theta^2}}=\Phi^{-1}(\tau) \\ &\iff d=\Phi^{-1}(\tau)\sqrt{1-\theta^2}+\theta b \\ &\iff \Phi^{-1}(v)=\Phi^{-1}(\tau)\sqrt{1-\theta^2}+\theta\Phi^{-1}(u) \\ &\iff v=\Phi\big(\Phi^{-1}(\tau)\sqrt{1-\theta^2}+\theta\Phi^{-1}(u)\big). \end{aligned}$$

L'inverse partiel de $D(v;\theta,u)$ par rapport à sa composante v est donc

$$D^{-1}(v;\theta,u)=\Phi\big(\Phi^{-1}(\tau)\sqrt{1-\theta^2}+\theta\Phi^{-1}(u)\big) \tag{A.3}$$

A.3.4 Preuve de la Proposition 3.3.1

De la relation

$$P(X < x) = P(Y = 0, X < x) + P(Y = 1, X < x),$$

nous avons

$$P(Y = y, X < x) = \begin{cases} F^*(y^*, x) & \text{si } y = 0 \ , \\ F_X(x) - F^*(y^*, x) & \text{si } y = 1 \ . \end{cases}$$

Ainsi, en posant

$$F^*(y^*, x) = C_\theta(F_{Y^*}(y^*), F_X(x)),$$

la densité de (X, Y), $f(x, y)$, a pour expression

$$\begin{aligned} f(x,y) &= \frac{\partial}{\partial x} P(Y = y, X < x) \\ &= \left[\frac{\partial}{\partial x} P(Y = 0, X \leq x)\right]^{\mathbb{1}_{\{y=0\}}} \times \left[\frac{\partial}{\partial x} P(Y = 1, X \leq x)\right]^{\mathbb{1}_{\{y=1\}}} \\ &= \left[\frac{\partial}{\partial x} P(Y^* < y^*, X \leq x)\right]^{\mathbb{1}_{\{y=0\}}} \times \left[\frac{\partial}{\partial x} P(Y^* \geq y^*, X \leq x)\right]^{\mathbb{1}_{\{y=1\}}} \\ &= \left[\frac{\partial}{\partial x} F^*(y^*, x)\right]^{\mathbb{1}_{\{y=0\}}} \times \left[\frac{\partial}{\partial x}\Big(F_X(x) - F^*(y^*, x)\Big)\right]^{\mathbb{1}_{\{y=1\}}} \\ &= \left[\frac{\partial}{\partial x} C_\theta(F_{Y^*}(y^*), F_X(x))\right]^{\mathbb{1}_{\{y=0\}}} \times \left[\frac{\partial}{\partial x}\Big(F_X(x) - C_\theta(F_{Y^*}(y^*), F_X(x))\Big)\right]^{\mathbb{1}_{\{y=1\}}}. \end{aligned}$$

Comme

$$\begin{aligned} \frac{\partial}{\partial x} C_\theta(F_{Y^*}(y^*), F_X(x)) &= \frac{\partial F_X(x)}{\partial x} \times \frac{\partial}{\partial F_X(x)} C_\theta(F_{Y^*}(y^*), F_X(x)) \\ &= f(x) \times C_\theta^{01}(F_{Y^*}(y^*), F_X(x)), \end{aligned}$$

alors, nous avons

$$f(x, y) = f_X(x) \times \left[C_\theta^{01}\big(\pi_0, F_X(x)\big)\right]^{\mathbb{1}_{\{y=0\}}} \times \left[1 - C_\theta^{01}\big(\pi_0, F_X(x)\big)\right]^{\mathbb{1}_{\{y=1\}}},$$

avec $\pi_0 = F_{Y^*}(y^*) = P(Y^* < y^*) = P(Y = 0)$ et

$$C_\theta^{01}(F_{Y^*}(y^*), F_X(x)) = \frac{\partial}{\partial F_X(x)} C(F_{Y^*}(y^*), F_X(x)).$$

A.3.5 Preuve des formules du Tableau 3.2 : cas de la copule gaussienne

On procède en deux étapes pour calculer cet estimateur. On calcule d'abord le quantile conditionnel en fonction de la copule, ensuite, on effectue un « plug-in » de la copule estimée dans l'expression calculée à l'étape 1.

Étape 1 : Calcul du quantile conditionnel à base d'une copule normale de paramètre θ.

L'expression du quantile conditionnel à base d'une copule quelconque de paramètre θ est donné par l'équation (3.3) de la Proposition 3.2.2. Si la copule est normale de paramètre θ, alors

$$Q_{Y|X}(\tau) = \mathbb{1}\big\{F_{Y^*}^{-1}\big(D^{-1}(\tau;\theta,u)\big) \geq 0\big\}. \tag{A.4}$$

Étape 2 : Après avoir estimé la copule (ses marges de façon non paramétrique et son paramètre de façon paramétrique), on se sert des équations (A.3) et (A.4) pour obtenir

$$\hat{Q}_{Y|X}(\tau) = \mathbb{1}\left\{\hat{F}_{Y^*}^{-1}\left[\Phi\left(\Phi^{-1}(\tau)\sqrt{1-\hat{\theta}^2} + \hat{\theta}\Phi^{-1}\left(F_n(x)\right)\right)\right] \geq 0\right\},$$

avec $\hat{F}_{Y^*}(y^*) = \Phi(y^*)$, puisqu'on suppose en pratique que la variable latente est normale.

A.3.6 Preuve des formules du Tableau 3.3 : cas de la copule gaussienne

L'expression de l'estimateur de la probabilité de succès conditionnelle dont il est question ici est donnée par l'équation (3.13). Elle est définie par

$$\begin{aligned} P(\widehat{Y=1|X}=x) &= 1 - \left.\frac{\partial C(u,v)}{\partial u}\right|_{v=\hat{\pi}_0, u=F_n(x)} \\ &= 1 - \left. D(v;\theta,u)\right|_{v=\hat{\pi}_0, u=F_n(x)}. \end{aligned}$$

Lorsque la copule est gaussienne de paramètre $\hat{\theta}$, il découle de l'équation (A.2) que

$$\begin{aligned} P(Y = \widehat{1|X} = x) &= 1 - \Phi\left(\frac{\Phi^{-1}(\hat{\pi}_0) - \hat{\theta}\Phi^{-1}(F_n(x))}{\sqrt{1-\hat{\theta}^2}}\right) \\ &= \Phi\left(\frac{\hat{\theta}\Phi^{-1}(F_n(x)) - \Phi^{-1}(\hat{\pi}_0)}{\sqrt{1-\hat{\theta}^2}}\right). \end{aligned}$$

Note : Le lecteur peut trouver les expressions analytiques de D et D^{-1} dans les travaux de (Bernard et Czado, 2015).

APPENDICE B

QUELQUES CODES UTILISÉS POUR LES SIMULATIONS

B.1 Étude des simulations

B.1.1 Simulation des données mixtes comme dans l'algorithme 2

Voici le code `R` qui nous a permis de simuler les données mixtes issues d'une copule.

```
#####################################
#Simulate mixed data from a copula #
#####################################
simCopG <- function(pool.size, p, kendall, cop){
  #input:
  # @pool.size: sample size
  # @kendall: Kendall's tau
  # @p: success probability to balance or unbalance data
  # @cop: copula family [Gaussian=1, Frank=5, Clayton=3, Gumbel=4]
  # output: list of 2
  # data = matrix n*2 and copula parameter calculated
  #from Kendall's tau

  if(cop=="Gaussian") {
  Family = 1
  par = VineCopula:::BiCopTau2Par(family = Family, tau = kendall)
```

```
    }
    if(cop=="Clayton") {
    Family = 3
    par = VineCopula:::BiCopTau2Par(family = Family, tau = kendall)
       }
       if(cop=="Frank") {
       Family = 5
       par = VineCopula:::BiCopTau2Par(family = Family, tau = kendall)
    }
    if(cop=="Gumbel") {
    Family = 4
    par = VineCopula:::BiCopTau2Par(family = Family, tau = kendall)
 }
  # Generate x
  x = rnorm(pool.size)
  u = pnorm(x)
  p.cop = VineCopula:::BiCop(Family, par)

  #conditionnal success probability
  pp <-  1 - VineCopula:::BiCopHfunc1(u1=u, u2=rep(1-p, pool.size),
                                      obj = p.cop)
  #Generate y
  y = rbinom(pool.size, 1, pp)
  #data
  data <- data.frame(y=y,x=x)
  t <- list(data, par)
  names(t) <- c("data", "copula_parameter")
  t
}
```

```
###################
#      Example      #
###################
#Simulate mixed balanced data from a Gaussian copula with
#Kendall's tau = .41.

data=simCopG(pool.size=400, p=.5, kendall=.41, cop="Gaussian")$data
```

B.1.2 Fonction de log-vraisemblance : cas de la copule gaussienne

Dans cette section, nous donnons le code `R` qui calcule la fonction de $-$ log-vraisemblance de notre modèle dans le cas de la copule gaussienne.

```
lik.gauss <- function(x, y, s){
# Input
# @x: Feature or covariate
# @y: Binary response variable
# @s: A vector of copula parameter and success probability
# Output: - Loglikelihood
 n <- length(y)
 par <- s[1]
 p <- s[2]

 # x estimated density
 h0 <- ks:::hns(x, deriv.order=0)
 f0 <- ks:::kdde(x, h=h0, deriv.order=0, eval.points=x)$estimate

 # Empirical c.d.f. of x
 u <- (rank(x) - 0.5)/n
```

```
  # Gaussian copula derivative w.r.t.its first component
  cop <- VineCopula:::BiCop(1, par)
  D <- VineCopula:::BiCopHfunc1(u, rep(1-p, n), cop)

  # - Log-likelihood
  t <- f0*ifelse(y==0, D, 1-D)
  - sum(log(t))
}
```

La fonction de log-vraisemblance nous permet de calculer l'AIC, mais aussi d'estimer les paramètres. Nous utilisons la fonction `optim` de `R` pour résoudre nos problèmes d'optimisation.

B.1.3 Estimation des paramètres : Cas de la copule gaussienne

Étant donné un jeu de données mixtes liées par une copule de tau de Kendall fixé, on estime les paramètres de la copule et la probabilité de succès marginale comme suit.

```
##########################################
# Maximization of Log-likelihood function #
##########################################
par <- optim(par0=c(.4,.2), fn=lik.gauss, x = data[,-1],
        y = data[,1], lower = c(-.99, 0),
        upper = c(.99,1), method = "L-BFGS-B")$par

#Estimated copula parameter
cop.par = par[1]

#Estimated marginal succes propability
p = par[2]
```

B.1.4 Erreur de classification

Cas des copules

Dans cette section, nous calculons l'erreur de classification de notre modèle - par validation croisée - lorsque les données sont issues d'une copule parmi celle de Gauss, de Frank, de Clayton et de Gumbel. L'hypothèse de normalité de la variable latente est primordiale dans le calcul du seuil pour la classification.

```
classErrCop <- function(pool.size, p, kendall, cop, d, tau){
  #input:
  #pool.size: pool size
  #@Kendall: Kendall's tau
  #@p: marginal success probability
  #@cop: copula family ['Gaussian', 'Frank', 'Clayton', 'Gumbel']
  #@d: copula dimension
  #@tau: quantile where we wish to compute classification error
  # output: classification error Matrix

  if(cop=="Gaussian") {
    par = VineCopula:::BiCopTau2Par(family = 1, tau=kendall)
    p.cop = copula:::normalCopula(param=par, dim=d)
    u.alpha = copula:::rCopula(pool.size,p.cop)
    y.star=qnorm(u.alpha[,1])
    y.c=qnorm(u.alpha[,2])
    dat.mat.pool=data.frame(y.star=y.star,y.c=y.c)
    yub <- quantile(y.star, 1 - p)
    #-----------------------------------#
    ## case
    ind.case <- which(y.star>=yub)
    case.pool <- dat.mat.pool[ind.case,]
```

```
#-----------------------------------#
## controls
ind.con<-which(y.star<yub)
con.pool<-dat.mat.pool[ind.con,]
#-----------------------------------#
cc.pool<-rbind(case.pool,con.pool)

# CC sample
y=c(rep(1,length(ind.case)), rep(0,length(ind.con)))
x=cc.pool[,2]
}
else if(cop=="Clayton") {
par = VineCopula:::BiCopTau2Par(family = 3, tau = kendall)
p.cop = copula:::claytonCopula(param=par, dim=d)
u.alpha = copula:::rCopula(pool.size,p.cop)
y.star=qnorm(u.alpha[,1])
y.c=qnorm(u.alpha[,2])
dat.mat.pool=as.data.frame(cbind(y.star=y.star,y.c=y.c))
yub<- quantile(y.star,1-p)
#-----------------------------------#
## case
ind.case<-which(y.star>=yub)
case.pool<-dat.mat.pool[ind.case,]
#-----------------------------------#
## controls
ind.con<-which(y.star<yub)
con.pool<-dat.mat.pool[ind.con,]
#-----------------------------------#
cc.pool<-rbind(case.pool,con.pool)
```

```
 # CC sample
 y=c(rep(1,length(ind.case)),rep(0,length(ind.con)) )
 x=cc.pool[,2]
 }
 else if(cop=="Frank") {
 par = VineCopula:::BiCopTau2Par(family = 5, tau = kendall)
 p.cop = copula:::frankCopula(param=par, dim=d)
 u.alpha = copula:::rCopula(pool.size,p.cop)
 y.star=qnorm(u.alpha[,1])
 y.c=qnorm(u.alpha[,2])
 dat.mat.pool=data.frame(y.star=y.star,y.c=y.c)
 yub<- quantile(y.star,1-p)
 #----------------------------------#
 ## case
 ind.case<-which(y.star>=yub)
 case.pool<-dat.mat.pool[ind.case,]
 #----------------------------------#
 ## controls
 ind.con<-which(y.star<yub)
 con.pool<-dat.mat.pool[ind.con,]
 #----------------------------------#
 cc.pool<-rbind(case.pool,con.pool)

 # CC sample
 y=c(rep(1,length(ind.case)),rep(0,length(ind.con)) )
 x=cc.pool[,2]
}
else if(cop=="Gumbel") {
```

```
    par = VineCopula:::BiCopTau2Par(family = 4, tau = kendall)
    p.cop = copula:::gumbelCopula(param=par, dim=d)
    u.alpha = copula:::rCopula(pool.size,p.cop)
    y.star=qnorm(u.alpha[,1])
    y.c=qnorm(u.alpha[,2])
    dat.mat.pool=data.frame(y.star=y.star,y.c=y.c)
    yub<-quantile(y.star,1-p)
    #----------------------------------#
    ## case
    ind.case<-which(y.star>=yub)
    case.pool<-dat.mat.pool[ind.case,]
    #----------------------------------#
    ## controls
    ind.con<-which(y.star<yub)
    con.pool<-dat.mat.pool[ind.con,]
    #----------------------------------#
    cc.pool<-rbind(case.pool,con.pool)

    # CC sample
    y=c(rep(1,length(ind.case)),rep(0,length(ind.con)) )
    x=cc.pool[,2]
  }
  else {try(print("Unrecognized copula... Enter a copula between
  [Gaussian, Frank, Clayton, Gumbel]"), silent = TRUE)
  }

  data <- data.frame(cbind(y,x))

  #---2-fold cross-validation: 50-50 ---#
```

```
I <- sort(sample(1:pool.size, floor(.5*pool.size)))
appren <- data[I,]
test <- data[-I,]

# Parameters estimation
par1 <- optim(c(.4,.2),  lik.gauss, x = appren[,-1],
              y = appren[,1], lower = c(-.99, 0),
              upper = c(.99,1), method = "L-BFGS-B")$par

par2 <- optim(c(1,.2), lik.frank, x = appren[,-1],
              y = appren[,1], lower = c(-100,0),
              upper = c(100,1), method = "L-BFGS-B")$par

par3 <- optim(c(1.5,.2),  lik.clayton, x = appren[,-1],
              y = appren[,1], lower = c(.01, 0),
              upper = c(100, 1), method = "L-BFGS-B")$par

par4 <- optim(c(1.3,0.2),  lik.gumbel, x = appren[,-1],
              y = appren[,1], lower = c(1,0),
              upper = c(100, 1), method = "L-BFGS-B")$par

u <- (rank(test[, 2]) - 0.5)/nrow(test)
p.cop = VineCopula:::BiCop(1, par1[1])
hat.v <- VineCopula:::BiCopHinv1(u1=u, u2=rep(tau,nrow(test)),
                                 obj = p.cop)

#----------------------------------------#
# Copula-based binary quantile regression models
hat.y.star <-  qnorm(hat.v)
```

```
yub<- qnorm(1-par1[2]) #treshold
pred.gauss <- as.numeric(hat.y.star >= yub)

p.cop <- VineCopula:::BiCop(5, par2[1])
hat.v <-  VineCopula:::BiCopHinv1(u1=u, u2=rep(tau,nrow(test)),
                                  obj = p.cop)
hat.y.star <-  qnorm(hat.v)
yub <- qnorm(1-par2[2])  #treshold
pred.frank <- as.numeric(hat.y.star >= yub)

p.cop <- VineCopula:::BiCop(3, par3[1])
hat.v <- VineCopula:::BiCopHinv1(u1=u, u2=rep(tau,nrow(test)),
                                  obj = p.cop)
hat.y.star <-  qnorm(hat.v)
yub<- qnorm(1-par3[2]) #treshold
pred.clay <- as.numeric(hat.y.star >= yub)

p.cop <- VineCopula:::BiCop(4, par4[1])
hat.v <- VineCopula:::BiCopHinv1(u1=u, u2=rep(tau,nrow(test)),
                                  obj = p.cop)
hat.y.star <-  qnorm(hat.v)
yub<- qnorm(1-par4[2]) #treshold
pred.gum <- as.numeric(hat.y.star >= yub)

#---------------------------------------#
# logistic regression model
model <- glm(y~x, data = appren, family = binomial(link = "logit"))
eta <- predict(model, newdata = test)
pred.logit <- as.numeric(1/(1+exp(-eta)) >= .5)
```

```
  #---------------------------------------#
  # SVM model
  mo <- e1071:::svm(formula = y~x, data = appren,
                    type = 'C-classification',
                    kernel = 'linear')
  pred.svm <- as.numeric(as.vector(predict(mo, newdata = test)))

  #---------------------------------------#
  # Classification errors' matrix
  y.test <- test[,1]
  cbind(mean((pred.gauss - y.test)^2),
        mean((pred.frank - y.test)^2),
        mean((pred.clay - y.test)^2),
        mean((pred.gum - y.test)^2),
        mean((pred.logit - y.test)^2),
        mean((pred.svm - y.test)^2))
}


##############################
#   Example: Frank Copula    #
##############################
# Data is generated from a Frank copula with p = .1
# We predict interest binary variable with different models
# and then calculate classification errors at the quantile  .5
#through 1000 replications
#-----------------------------------------#
#-- use n-1 processors to make parallelization
```

```
library(doParallel)
library(doSNOW)
library(foreach)
library(doRNG)
cores <- detectCores()
cl <- makeCluster(cores - 1)#not to overload your computer
registerDoSNOW(cl)
#-----------------------------------------------#
#-------------B replications-------------------#
B = 1000
pb <- txtProgressBar(max = B, style = 3)
progress <- function(n) setTxtProgressBar(pb, n) #progress bar
opts <- list(progress = progress)
set.seed(1)
df <- foreach(i=1:B, .options.snow = opts) %dorng% {
  classErrCop(pool.size=400, p=.1, kendall=.5, cop="Frank",
  d=2, tau=.5)
}
close(pb)
stopCluster(cl)
sf <- data.frame(matrix(unlist(df), nrow=length(df), byrow = TRUE))

ErClasFrank = colMeans(sf)
names(ErClasFrank) = c('normal', 'frank', 'clayton', 'gumbel',
'logit', 'SVM')
ErClasFrank
#  normal     frank   clayton    gumbel     logit       SVM
#0.103585 0.101830 0.100640 0.105965 0.104360 0.099840
```