REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohammed Seddik Ben Yahia-Jijel

Faculté des Sciences Exactes et Informatique

Département de Mathématiques

Mémoire

Présenté pour l'obtention du diplôme de

Master

Spécialité Mathématiques

Option Probabilités et statistique

Thème

Utilisation des copules pour l'estimation d'une distribution maltraitée discrète

Présenté par

Boussandel Salima

Gueham Mounira

Devant le jury composé de

Président	**Cheraitia Hassen**	M.C.B.	Université de Jijel
Encadreur	**Gherda Mebrouk**	M.A.A.	Université de Jijel
Examinatrice	**Abdi Zineb**	M.A.A.	Université de Jijel

Promotion 2019/2020

Remerciements

Avant, nous remercions Allah pour nous avoir donné la santé, la volonté, le courage et la détermination qui nous ont accompagnés tout au long de la préparation et l'élaboration de ce travail et qui nous ont permis d'achever ce modeste travail. Le présent travail est non seulement le résultat de notre courage,sacrifice, patience et endurance mais aussi une participation de plusieurs personnes qui nous sont chères.
Nous tenons d'abord à remercier très chaleureusement ***Mr.Mebrouk Gherda*** *qui nous a permis de bénéficier de son encadrement. Les conseils qu'il nous a prodigué, la patience, la confiance qu'il nous a témoignés ont été déterminants dans la réalisation de notre travail.*
Nos précieux remerciements vont à le président et les membres de jury :
Mr. Hassen Cheraitia *et* ***Mme. Zineb Abdi*** *pour l'honneur qu'ils nos ont fait en acceptant de juger nous travail et faire partie de ce jury.*
Nous n'oublions pas dans nos éloges l'ensemble du personnel du Département de mathématiques de l'Université Mohammed Seddik Ben Yahyia, et plus personnellement nos enseignants pour tous leurs efforts.
Nos remerciements vont également à tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicace

Au terme de cette étude un grand merci au Dieu, pour le courage et la force qui nous a offert pour terminer ce memoire.

A la lumière de mes jours, la source de mes efforts, la flamme de mon coeur, ma vie et mon bonheur ; mes parents que j'adore.

A mon encadreur Gherda Mebrouk

A mes chers fréres
Radwan,Taher,Jihad

A mes soeurs
Amel et Selma

A ma binôme Mounira

Finalement, à grâce aux personne les plus chères et proches.

Salima

Dédicace

Au terme de cette étude un grand merci au Dieu, pour le courage et la force qui nous a offert pour terminer ce memoire.

A la lumière de mes jours, la source de mes efforts, la flamme de mon coeur, ma vie et mon bonheur ; mes parents que j'adore.

A mon marie Youcef et sa famille

A mon encadreur Gherda Mebrouk

A mes chers fréres
Foad, Nassim, Mohammed

A mes soeurs
Widad et Ibtissem

A ma binôme Salima

Finalement, à grâce aux personne les plus chères et proches.

Mounira

Résumé

Dans ce mémoire, nous présentons une nouvelle méthode d'estimation des distributions discrètes décrit par **Victor Fassaluza**, **Luis Gustavo Esteves** et **Carlos Alberto de Braganca Pereira**. Cette méthode permet d'estimer la la fonction d'une distribution discrète en présence des données manquantes en utilisant les copules de Bernstein après avoir approximé cette distribution par des polynômes de Bernstein, et en fin, nous présentons une application sur des données réelles que nous allons comparer avec les différents estimateurs des probabilités théoriques pour chaqu'une des distances suivantes, Aitchison, euclidean, variation total, et Divergence symétrisée de Kullback – Leibler, afin d'évaluer la pertinence de cette méthode par rapport aux méthodes existantes.

Abstract

In this thesis, we present a new method of estimating distributionsdiscrete described by Victor Fassaluza, Luis Gustavo Esteves and Carlos Alberto de Braganca Pereira. This method makes it possible to estimate the function of a discrete distribution in the presence of missing data using copulas of Bernstein after having approximated this distribution by polynomials of Bernstein, and at the end we present an application on real data that we will compare with the different estimators of the theoretical probabilities for each of the following distances, Aitchison, euclidean, total variation, and Divergence Kullback - Leibler symmetrization, in order to assess the relevance of this method by compared to existing methods.

Table des matières

MOTIVATION

Les variables aléatoires discrètes est un outil indispensable pour la modélisation des durées de vie. Lorsque les mesures de ces durées de vie sont peu pratiques, voir insuffisantes, ou impossible à mesurer en échelle continue, il est raisonnable de considérer la durée de vie comme une variable aléatoire discrète. En pratique, nous rencontrons des situations où la durée de vie d'un équipement est considérée comme une variable aléatoire discrète, on peut considérer la durée de vie comme un nombre de succès, ...ect.

INTRODUCTION

Le concept de copule est l'un des moyens qui peuvent enrichir la notion de dépendance stochastique. Traiter cette notion à l'aide de corrélations linéaires seules est (en général) une très mauvaise idée. l'un des avantages du concept de copules c'est que tout modèle probabiliste multi-varié possède (au moins) une copule. Un modèle probabiliste est caractérisé par la donnée de la fonction de répartition F_X du vecteur X. En utilisant les copules, le théorème de Sklar nous assure l'existence d'une copule qui nous permet de séparer la modélisation des lois marginales (fonctions de répartition à une dimension) et de la structure de dépendance (copule). Le modèle probabiliste est alors obtenue par assemblage des lois marginales et de la copule. L'un des problèmes fondamentaux en statistique est l'estimation d'une fonction de répartition F à partir d'un échantillon de variables aléatoires réelles $X_1, X_2, ..., X_n$ indépendantes et de même loi inconnue.

Plusieurs estimateurs ont été proposés pour l'estimation de cette fonction, comme la fonction de répartition empiriques qui est un estimateur naturel, cette fonction ne fait appel à aucune structure algébrique ou topologique mais seulement à des notions ensemblistes et que les techniques d'estimation ne sont pas autre chose que des techniques de régularisation de cette référence empirique dont la donnée est équivalente à celle de l'échantillon. Estimateurs par lissage, estimateur à noyau, estimateurs splines et d'autres estimateurs.

Dans l'histoire de la théorie de l'approximation, les polynômes de Bernstein univariés et multivariés jouent un rôle central depuis le début du XX ème siècle. Les traitements des polynômes de Bernstein en une variable ou en plusieurs variables, non pas seulement été utilisés pour fournir une preuve constructive du célèbre théorème d'approximation de Weierstrass pour les fonctions continues sur des intervalles

compacts, y compris des estimations explicites du taux de convergence, mais aussi pour des applications plus avancées en analyse fonctionnelle et en conception assistée par ordinateur, comme les courbes et surfaces de Bézier. Les propriétés de conservation de forme et de lissage local des polynômes de Bernstein sont d'un intérêt central, en particulier les applications d'ingénierie. (Il peut être intéressant de noter ici que Donald Knuth a utilisé des courbes de Bézier pour la conception de polices TEX.) Les applications des polynômes de Bernstein pour modéliser la dépendance stochastique via des copules, en revanche, sont examinées beaucoup plus tard.

L'utilisation des copules à des fins de modélisation et de simulation, par exemple dans la gestion des risques, revêt une importance croissante. Rappelons qu'une copule C (d-dimensionnelle) est la fonction de distribution cumulée (cdf) d'un vecteur aléatoire $U = (U_1, \cdots, U_d)$ dont les distributions marginales unidimensionnelles sont uniformes sur l'intervalle $[0, 1]$. Le théorème bien connu de Sklar forme une base clé de la théorie des copules.

L'association entre variables aléatoires est un sujet d'intérêt dans de nombreux domaines scientifiques. La méthode la plus complète pour caractériser l'association entre les variables aléatoires consiste à déterminer la distribution conjointe de ces variables aléatoires en utilisant les fonctions de densité multivariée pour les variables absolument continues et les fonctions de masse de probabilité multivariée pour les variables discrètes. Actuelemnt les chercheurs s'intéressés à évaluer de telles associations (voir, par exemple, [4, 7]).

Le travail de notre mémoire est inspiré essentiellement d'un article de Victor Fossaluza publié en 2017, La motivation de cet article était une étude réalisée dans le cadre de l'Obsessive-Compulsive Programme sur les troubles du spectre de l'Institut de psychiatrie, faculté de médecine de l'Université de São Paulo.

L'objectif de cet article est d'introduire une méthode d'estimation des fonctions de masse de probabilité discrètes multivariées en présence de données manquantes (marginales). Pour cela,l'auteur a développé une méthode d'estimation qui utilise à la fois des fonctions de distribution empiriques et des polynômes de Bernstein.

La procédure consiste à estimer une fonction de distribution conjointe lisse, puis à appliquer une méthode qui transforme cette fonction en une fonction discrète, c'est-à-dire la fonction de masse de probabilité conjointe estimée. Les résultats de cette nouvelle méthode sont comparés à ceux des méthodes alternatives, en évaluant

les distances standard.

Ce mémoire se divise en trois chapitres.
Le premier chapitre est une introduction mathématique aux copules. On y retrouve les principaux théorèmes de base de la théorie des copules. Dans ce chapitre nous présentons quelques familles de copules paramétriques les plus utilisées en pratique, ainsi que les différentes méthodes d'estimation des copules (paramétriques, semi-paramétriques et non paramétriques). Le second chapitre est réparti en trois sections. Dans la première section nous allons présenté la définition et quelques propriétés de polynômes de Bernstein, ainsi que le théoreme de Weirstrass qui affirme la convergence uniforme d'une suite de polynômes. Dans la deuxiéme section nous présentons quelques propriétés des copules et densités de Bernstein. Dans la troisiéme section nous présentons comment passer de polynômes à copules de Bernstein. Dans le troisième chapitre, La première section décrit les méthodes existantes trouvées dans la litt érature qui seront considérées pour comparaison. La section 2 décrit l'estimateur proposé par Victor Fossaluza pour les fonctions de probabilité conjointes. La section 3 présente une discussion de la nouvelle méthode et comparaisons de cette méthode avec les méthodes alternatives utilisant à la fois des échantillons simulés et le vrai exemple OCD. Enfin, dans la section 4, nous présentons des commentaires et considérations pour les travaux futurs.

Chapitre 1

Rappel sur les copules

1.1 Quelques mesures de dépendance

En statistique, il existe un certain nombre de grandeurs proposées par des auteurs afin de mesurer la dépendance entre deux variables aléatoires.

1.1.1 Coefficient de corrélation de Pearson

Soit (X,Y) un vecteur aléatoire dans $\mathbb{R}^2$
alors $\rho(X,Y) = \frac{Cov(X,Y)}{\delta_X \delta_Y}$

où $cov(X,Y) = E(XY) - E(X)E(Y)$ est la covariance entre X et Y.

δ_X, δ_Y l'ecart type de X et Y respectivement. E(X) et E(Y) l'espérance de X et Y respectivement

Ce coefficient est la mesure la plus connue,est une mesure imparfaite de la dépendance. Il ne peut se mettre sous une forme ne faisant intervenir que la copule, les marginales restent présentent dans l'expression, ce coefficient n'est donc pas une mesure de dépendance, en plus c'est un coefficient de corrélation linéaire.

Les proprietés de coefficient de corrélation de Pearson sont

a) $\rho(aX, bY) = signe(ab)\rho(X,Y)$

b) $\rho(aX + b, cY + d) = signe(ac)\rho(X,Y)$

c) $\rho = 0$ n'implique pas que les variables X et Y sont indépendantes mais on ne peut pas dire qu'il n y'a pas une corrélation entre ces deux variables.

Définition 1.1. *Soient $(x_1, y_1), (x_2, y_2)$ deux observations d'un couple de variables aléatoires (X, Y), sont dit*

concordantes si

$$(x_1 - x_2)(y_1 - y_2) > 0 \Leftrightarrow ((x_1 < x_2) \text{ et } (y_1 < y_2) \quad ou \quad (x_1 > x_2 \text{ et } y_1 > y_2))$$

disconcordantes si

$$(x_1 - x_2)(y_1 - y_2) < 0 \Leftrightarrow ((x_1 < x_2 \text{ et } y_1 > Y_2) \quad ou \quad (x_1 > x_2 \text{ et } y_1 < Y_2))$$

Définition 1.2. *La fonction de concordance entre les deux variables aléatoires (X_1, Y_1) et (X_2, Y_2) est définie par*

$$Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{1.1}$$

1.1.2 Le tau de Kendall

Le τ de Kendall $\tau(X, Y) = E(F_X(X)F_Y(Y)) - 1$ ne dépend également pas des marginales, il est égale à la probabilité de concordance des rangs moins la probabilité de disconcordance des rangs, en effet :
soient $(x_1,y_1), (x_2,y_2)$ deux observations d'un couple de variables aléatoires (X, Y) avec (X, Y) sont i.i.d, alors

$$\tau = P[(x_1 - x_2)(y_1 - y_2) > 0] - P[(x_1 - x_2)(y_1 - y_2) < 0] \tag{1.2}$$

Dans le cas générale, on utilise l'estimateur

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{i=1}^{j-1} signe(x_j - x_i)(y_i - y_j) \tag{1.3}$$

on note par $z = (x_j - x_i)(y_i - y_j)$

$$signez = \begin{cases} 1 & \text{si } z \geq 0 \\ -1 & \text{si } z < 0 \end{cases}$$

1.1.3 Rho de Spearman

Le ρ_s de Spearman défini comme étant le coefficient de corrélation de Pearson calculé sur les rangs des deux variables éliminent par construction l'effet de dépendance aux lois marginales.

$$\rho_s(X,Y) = 3[P\left\{(X-X')(Y-Y')\right\} > 0 - [P\left\{(X-X')(Y-Y')\right\} < 0] \quad (1.4)$$

Le rho de Spearman peut s'écrir en fonction de coefficient de Pearson comme suit

$$\rho_s(X,Y) = \rho(F_X(x), F_Y(y)) \quad (1.5)$$

Remarque 1. *Le fait que $\tau = 0$ et $\rho_s = 0$ n'implique pas que les variables X et Y soient indépendantes. τ et ρ_s sont deux coefficients de corrélation non linéaire.*

1.2 La fonction de répartition

Définition 1.3. *soit (X,Y) un vecteur aléatoire à valeur dans $\mathbb{R}^2$, la loi de (X,Y) est caractérisée par $F(X,Y) = P(X \leq x, Y \leq y), \forall (x,y) \in \mathbb{R}^2$, F est la fonction de répartition bivariée de (X,Y).*

Définition 1.4. *(lois marginales de (X,Y)) lois de X et de Y prises séparament décrites par*

$$F_1(x) = P(X \leq x) = \lim_{y \to +\infty} F(x,y) = F(x,+\infty)$$
$$F_2(y) = P(Y \leq y) = \lim_{x \to +\infty} F(x,y) = F(+\infty,y).$$

Remarque 2. *Si X et Y sont independantes alors $F(x,y) = F_1(x)F_2(y)$*

Proposition 1.5. *soit F fonction de répartition bivariée alors*
1) F est continue à droite c-à-d $\lim_{x \to x_0 y \to y_0} F(x,y) = F(x_0,y_0)$.
2) $\lim_{x \to -\infty} F(x,y) = \lim_{y \to -\infty} F(x,y) = 0$.
3) $\lim_{\substack{x \to +\infty \\ y \to +\infty}} F(x,y) = 1$.
4) F est 2 -croissante pour tout $(a_1,a_2),(b_1,b_2)$ avec $-\infty \leq a_1 \leq b_1 \leq +\infty$ et $-\infty \leq a_2 \leq b_2 \leq +\infty$
$F(b_1,b_2) - F(\ a_1,b_2) - F(b_1,a_2) + F(a_1,a_2) \geq 0.$

Une fonction de deux variables vérifiant les quatre propriétés précédentes est une fonction de répartition bivariée .

1.3 Les copules

Les copules permettent d'étudier la dépendance entre plusieurs variables aléatoires, avec l'idée que cette dépendance ne doit pas contenir d'information provenant des lois marginales des variables elles-mêmes. Pour ce faire, on les « uniformise », c'est à dire qu'on se prémuni de « l'effet d'optique » dû au fait que ces variables peuvent avoir des lois marginales très différentes. En particulier, les copules permettent d'imposer une structure de dépendance à des lois marginales (ou des variables aléatoires) données séparément.
Dans toute la suite on prend $I = [0, 1]$.

Définition 1.6. *on appelle copule (de dimention 2 ou 2-dimentionnelle) toute fonction de répartition bivariée C ayant pour marginales la loi uniforme sur I. Autrement dit C vérifié les quatre proprietés d'une fonction de répartition bivariée avec de plus*

1) $\forall u, v \in I, C(u, 0) = C(0, v) = 0.$
2) $\forall u, v \in I, C(u, 1) = u, C(1, v) = v.$
3) $\forall u_1, v_1, u_2, v_2 \in I$, *avec* $u_1 < v_1$ *et* $u_2 < v_2$,
on a

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

telle que $C(v_1, v_2) = P(X \leq v_1, Y \leq v_2)$.

Exemple 1.7. $\forall u, v \in I$, $M(u, v) = \min(u, v)$ *définit une copule, en effet :*

$$\min(u, 0) = \min(0, v) = 0,\ \min(u, 1) = u, \min(1, v) = v.$$

$$\min(u_2, v_2) \leq \min(u_2, v_1),\ \min(u_1, v_2) \leq \min(u_1, v_1).$$

Théorème 1.8. *soit C une copule bivariée, pour tout u_1, v_1, u_2, v_2 avec $u_1 \leq u_2$ et $v_{1\leq}v_2$, on a $|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1|$, toute copule est continue.*

Théorème 1.9. *soit C une copule bivariée alors*

1) Les dérivées partielles existent p.s et $\frac{\partial C(u_1,u_2)}{\partial u_j} \leq 1, j = 1, 2.$
2) $u \longmapsto \frac{\partial C(u,v)}{\partial u}$ *et* $v \longmapsto \frac{\partial C(u,v)}{\partial v}$ *sont bien définie et décroissente.*

Remarque 3. *(la densité) Les copules admettrent des densités de probabilité.*
Si la densité c associée à la copule C existe, alors elle est définie par $c(u, v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$

Si la fonction de répartition conjointe H est absolument continue, en utilisant le théoreme de Sklar nous pouvons exprimer la densité d'un vecteur aléatoire (X,Y) en fonction de densité de sa copule et des fonctions de répartition marginales F et G , ainsi que les fonctions de densités f et g par $h(x,y) = c(F(x),G(y))f(x)g(y)$.

Théorème 1.10. *(Théoreme de Sklar)*

1) Si C une copule, si F et G deux fonctions de répartition alors $H(x,y)=C(F(x),G(y))$ est la fonction de répartition bivariée, ayant F et G pour marginales.
2) Soit H une fonction de répartition bivariée de marginales F et G il existe une copule C telle que pour tout $(x,y)\in \mathbb{R}^2$ $H(x,y) = C(F(x),G(y))$.
si de plus F et G sont continues alors C est unique.

Remarque 4. *Lorsque les marginales ne sont pas continues, il est toujours possible de définir une copule, mais celle-ci n'est plus unique et de ce fait perd beaucoup de son intérêt.*

1.3.1 Proprietés des copules

1) Toute copule C satisfait
$\forall u,v \in I, w(u,v) = \max(u+v-1,0) \leq C(u,v) \leq \min(u,v) = M(u,v)$
2) M et w sont deux copules "extrémes"
M : croissante en deux ces arguments
w : décroissante en deux ces arguments
3) $C_\perp(u,v) = uv$: composantes indépendantes
4) Si u_1, v_1, u_2, v_2 sont dans I
$|C(u_1,v_1) - C(u_2,v_2)| \leq |u_2 - u_1| + |v_2 - v_1|$.

Proposition 1.11. *Si $C_{x,y}$ est la copule de vecteur (X,Y), toute transformation croissante de (X,Y) a la méme copule.*

Corollaire 1.12. *Si F est croissante et G croissante, $C_{F(x),G(y)}(u,v) = C_{X,Y}(u,v)$.*
Si F est croissante et G décroissante, $C_{F(x),G(y)}(u,v) = u - C_{X,Y}(u,1-v)$.
Si F décroissante et G décroissante, $C_{F(x),G(y)}(u,v) = u+v-1+C_{X,Y}(1-u,1-v)$.
Si F est décroissante et G croissante, $C_{F(x),G(y)}(u,v) = v - C_{X,Y}(1-u,v)$.

Théorème 1.13. *(Forme de Frechet)*

Soit H une fonction de distribution conjointe d'un couple aléatoire (X,Y) de fonction de répartition marginales F et G. Pour tout copule bivariée C associée à H

et $\forall(u,v) \in I^2$,

on a

$$max(u+v-1,0) \leq C(u,v) \leq \min(u,v) \tag{1.6}$$

Proposition 1.14. *Soit* (X,Y) *un couple de variable aléatoire continue de copule* C*, l'expression de Tau de Kendall en terme de copule est la suivante*

$$\tau(X,Y) = 4\int_0^1\int_0^1 C(u,v)d(u,v) - 1 = 4\int_0^1\int_0^1 \partial_u C(u,v)\partial_v C(u,v)dudv - 1 = 4E(C(u,v)) - 1 \tag{1.7}$$

Proposition 1.15. *Soit* (X,Y) *un couple de variables aleatoires continues de copule* C *alors*

$$\rho_s(X,Y) = 12\int_0^1\int_0^1 uv\partial C(u,v) - 3 = 12\int_0^1\int_0^1 C(u,v)dudv - 3 \tag{1.8}$$

Il existe un grand nombre de copules adaptées à différentes situations, toute répartition associée à un vecteur dont les marginales sont uniformes sur [0,1] définie une copule.

1.3.2 Les copules paramétriques

Les copules archimédiennes

La notion de copule archimédienne, définie par GENEST et MACKAY [1986], regroupe un certain nombre de copules (Clayton, Gumbel, Franck), l'idée d'une copule archimédienne de générateur ϕ est que la tranformation $\omega(u) = \exp(-\phi(u))$ appliquée aux marginales « rend les composantes indépendantes »

$$\omega(C(u_1,\cdots,u_n)) = \prod_{i=1}^{n}\omega(u_i) \tag{1.9}$$

Définition 1.16. *(La copule archimédienne) La copule archimédienne de générateur* ϕ *est définie par*

$$C(u,v) = \begin{cases} \phi^{-1}(\phi(u_1) + \cdots + \phi(u_n)) & si \ \sum_{i=1}^{n}\phi(u_i) \leq \phi(0) \\ 0 & sinon \end{cases} \tag{1.10}$$

Le générateur doit être choisi de classe C^2 de sorte que

$$\phi(1) = 0, \quad \phi^{'}(u) \leq 0, \quad \phi^{''}(u) > 0.$$

Toutefois, quelques formes particulières sont souvent utilisé es en pratique du fait de leur simplicité de mise en oeuvre. On peut notamment citer les exemples ci-après

la copule	sa forme	le générateur
indépendance	uv	$-ln(v)$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-\frac{1}{\theta}}$	$(t^{-\theta} - 1), \theta \geq 0$
Frank	$\frac{-1}{\theta} \ln(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1})$	$-ln(\frac{e^{\theta t}-1}{e^{\theta}-1})$
Gumbel	$exp(-((-ln(u))^{\theta} + (-\ln(v))^{\theta})^{\frac{1}{\theta}}$	$(-ln(t))^{\theta}$

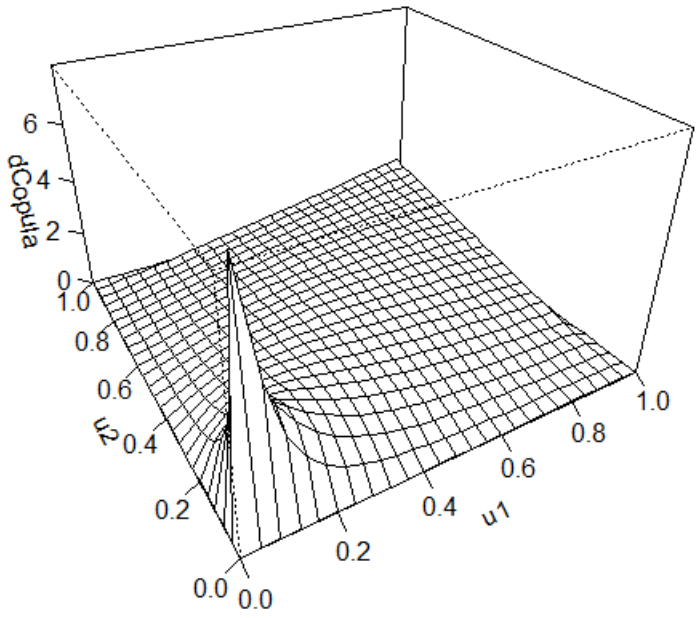

Figure 1.1- density de copule de clyton

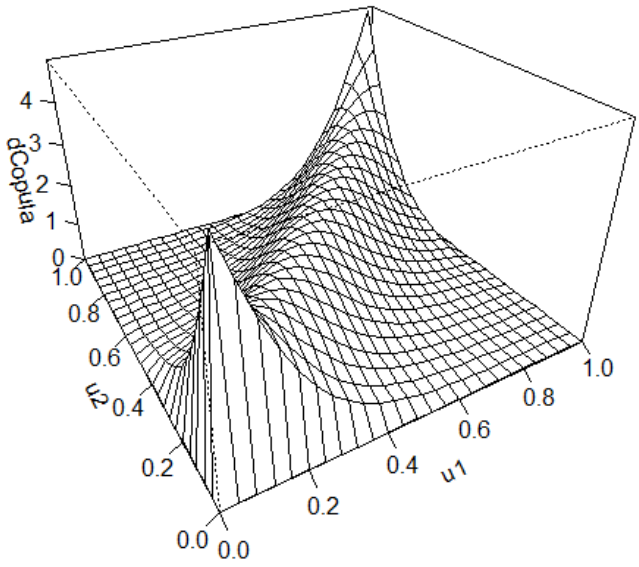

Figure 1.2- density de copule de frank

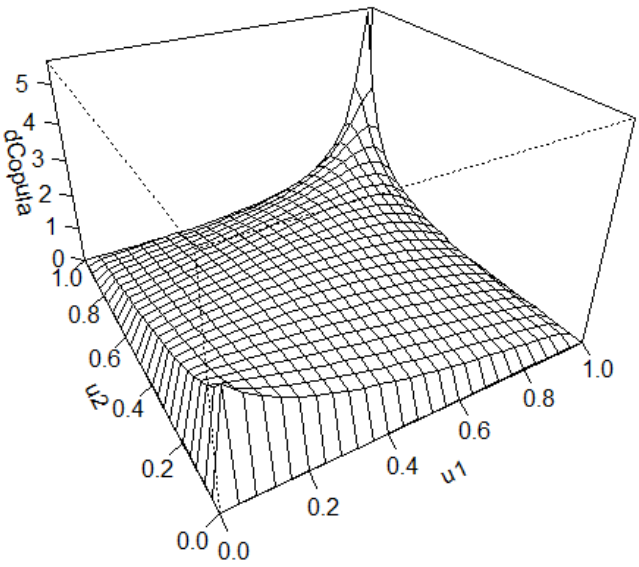

Figure 1.3- de copule de frank

Les copules elliptiques

Les copules elliptiques sont définies à partir des familles des lois elliptiques. Une copule est dite elliptique si elle est la copule d'une loi elliptique. On en considère ici deux cas particuliers, la copule gaussienne et la copule de Student.

Les copules gaussiennes La copule gaussienne ne présente pas de dépendance de queue et n'est donc pas adaptée à des valeurs extrêmes. L'importance de cette copule réside dans le fait qu'elle est sous-jacente à la distribution normale multivariée. En effet, modéliser la structure de dépendance d'un échantillon par une copule gaussienne est cohérent avec la mesure de cette dépendance par le coefficient de corrélation linéaire. Elle est définit par $C_\rho(u,v) = \phi_\rho(\phi^{-1}(u), \phi^{-1}(v)), \rho \in]-1,1[$

telles que ϕ : la fonction de répartition de la loi normale $N(0,1)$.

ϕ_ρ : distribution du vecteur gaussien (X,Y) centré de matrice de covariance

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

on a

$$\phi_\rho(X,Y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp[-\frac{(s^2+t^2-2\rho st)}{2(1-\rho^2)}]dsdt \qquad (1.11)$$

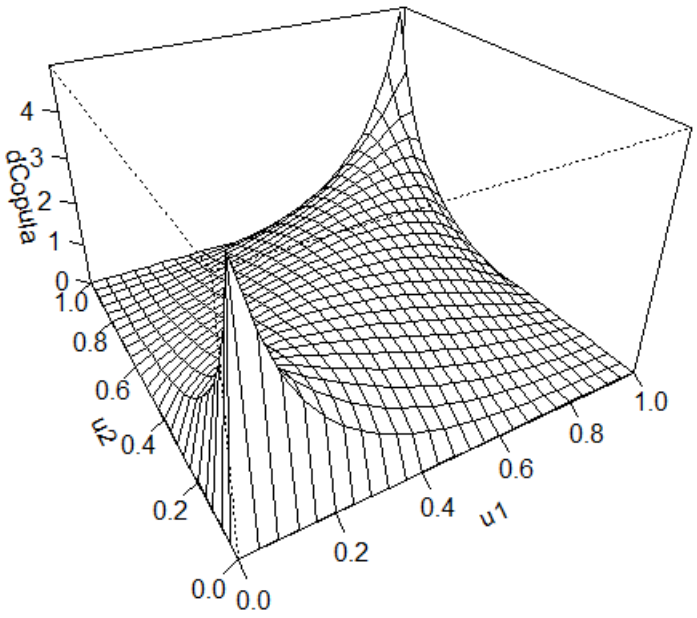

Figure 1.4- density de copule normale

La copule de Student La copule de Student (t copula) est la copule sous-jacente à une distribution multi-variée de Student. Cette structure de dépendance capte les dépendances extrêmes positives et négatives.

Elle est construite de la même manière que la copule gaussienne mais à partir de la distribution de Student centrée réduite

Définition 1.17. *Soit $\rho \in I$, la fonction de répartition de Student à υ degré de liberté est définie par*

$$t_{\rho,\upsilon}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho}}[1+\frac{(s^2+t^2-2\rho st)}{\upsilon(1-\rho^2)}]^{-\frac{\upsilon+2}{2}}dsdt \tag{1.12}$$

Définition 1.18. *La copule de Student est une copule paramétrique, paramétré e par le coefficient de corrélation linéaire ρ et de degré de liberté υ, cette copule est définie par*

$$C_{\rho,\upsilon}(u,v) = \ t_{\rho,\upsilon}(t_\upsilon^{-1}(u), t_\upsilon^{-1}(v)) = \int_{-\infty}^{t_\upsilon^{-1}(u)}\int_{-\infty}^{t_\upsilon^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho}}[1+\frac{(s^2+t^2-2\rho st)}{\upsilon(1-\rho^2)}]^{-\frac{\upsilon+2}{2}}dsdt \tag{1.13}$$

Si $\upsilon \to \infty$, alors la copule de Student converge vers la copule gaussienne et dans ce cas, il est trés difficile de différencier entre ces deux copules.

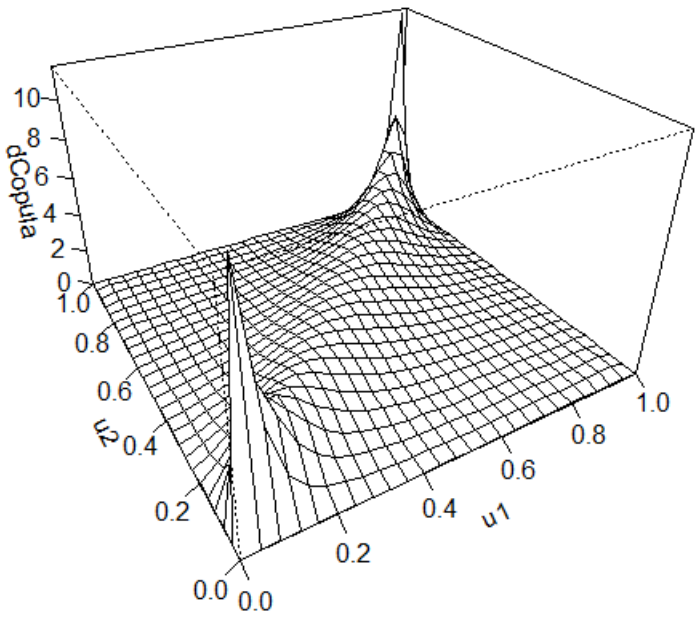

Figure 1.5.- density de copule de student

Copule des valeurs extrêmes

Une autre classe particuliére des copules et celle des valeurs extrêmes. Le non "extreme value copula" suggére un lien entre la théorie des valeurs extrêmes et ses copules.

Définition 1.19. *On appelle copule des valeurs extrêmes, toute copule C vé rifiant la propriété suivante*

$$C(u^t, v^t) = C^t(u, v), \forall u, v \in I, \forall t > 0 \tag{1.14}$$

où

$$C^{\frac{1}{t}}(u^t, v^t) = C(u, v)$$

Copule MinMax

Elle est définit par

$$C_{(m,M)}(u,v) = \begin{cases} v - [v^{\frac{1}{n}} + (1-u)^{\frac{1}{n}} - 1]^n, & \text{Si} 1 - (1-u)^{\frac{1}{n}} < v^{\frac{1}{n}} \\ v, & \text{Sinon} \end{cases} \tag{1.15}$$

1.3.3 La dépendance de queue

Définition 1.20. *Soit X et Y deux variables aleatoires continues de fonctions marginales F et G de copule C.*
Le coefficients de dépendance de queue supérieur de X et Y est définie par

$$\lambda_u = \lim_{u \to 1^-} P(Y > G^{-1}(u) / X > F^{-1}(u))$$

$$= \lim_{u \to 1^-} \frac{1 - 2u + C(u,v)}{1-u}$$

$$= \lim_{u \to 1^-} \frac{\overline{C}(u,v)}{1-u}. \tag{1.16}$$

La copule $\overline{C}(u,v)$ s'appelle la copule de survie.
Interprétation *Si cette limite existe et si $\lambda_u \in]0,1[$, X et Y sont dites asymptotiquement dépendantes au niveau supérieur de la queue.*

$\lambda_u = 0$, on dit que X et Y sont indépendantes au niveau supérieur de la queue.

Définition 1.21. $\lambda_L = \lim_{u \to 0^+} P(Y \leq G^{-1}(u) / X \leq F^{-1}(u)) = \lim_{u \to 0^+} \frac{C(u,v)}{u}$.

Si $\lambda_L = 0$, alors elle n'a pas dépendance de queue du niveau inferieur.

1.3.4 Copules associées à une copule

a) Copule de survie

Définition 1.22. *Soit $\widetilde{C}(u_1, \cdots, u_n)$ la fonction définie par*

$$\widetilde{C}(u_1, \cdots, u_n) = C(1-u_1, \cdots, 1-u_n) = P(u_1 > u, \cdots, u_n > u) \tag{1.17}$$

alors $\widetilde{C}$ est appelé la copule de survie de C.

Remarque 5. *Il ne faut pas comprendre que* $\widetilde{C}$ *est la distribution de survie* $\overline{C}$ *associée à C.*

Corollaire 1.23. $\forall u, v \in [0,1]$

$$\widetilde{C}(u,v) = 1 - (u+v) + C(u,v) \tag{1.18}$$

$\overline{C}$ *et* $\widetilde{C}$ *sont reliées par la relation suivante*

$$\forall u, v \in I, \overline{C}(u,v) = \widetilde{C}(1-u, 1-v) \tag{1.19}$$

b) La copule duale

Elle est définit par

$$P(X \leq x \ \ ou \ \ Y \leq Y) = P(X \leq x) + P(Y \leq Y) - P(X \leq x, Y \leq Y)$$

$$= F(x) + G(y) - H(x,y)$$

$$= F(x) + G(y) - C(F(x), G(y) \tag{1.20}$$

c) La co-copule

Elle est définit par

$$P(X \geq x \ \ ou \ \ Y \leq Y) = P(X > x) + P(Y \leq Y) - P(X > x, Y \leq Y)$$

$$= F(x) - \overline{C}(F(x), G(y))$$

$$= 1 - G(y) - C(1 - F(x), 1 - G(y)) \tag{1.21}$$

d) La copule mixte

Elle est définit par

$$P(X > x \ \ ou \ \ Y \leq Y) = P(X > x) + P(Y \leq Y) - P(X > x, Y \leq Y)$$

$$= F(x) - \overline{C}(F(x), G(y))$$

$$= 1 - G(y) - C(\overline{F(x)}, \overline{G(y)}) \tag{1.22}$$

notation La copule mixte est noté C"

$$C"(u,v) = 1 - v - C(1-u, 1-v) \tag{1.23}$$

1.4 Estimation statistique

Définition 1.24. *Un estimateur est une statistique permettant d'évaluer un paramètre inconnu relatif à une loi de probabilité (comme son espérance ou sa variance). Il peut par exemple servir à estimer certaines caractéristiques d'une population totale à partir de données obtenues sur un échantillon comme lors d'un sondage. La définition et l'utilisation de tels estimateurs constitue la statistique inférentielle.*

La qualité des estimateurs s'exprime par leur convergence, leur biais, leur efficacité et leur robustesse. Diverses méthodes permettent d'obtenir des estimateurs de qualités différentes.

1.4.1 Estimation non-paramétrique

Rappel sur la fonction de répartition empirique

Le cadre général de l'estimation non paramétrique d'une loi marginale s'appuie sur la fonction de répartition empirique, définie par

$$F_k(x) = \frac{1}{K} \sum_{i=1}^{k} 1_{\{x_i \leq x\}}. \tag{1.24}$$

Pour un n-échantillon $(x_1, \cdots, x_K)$ de loi F, en dimension n, si on se donne $(x_1^k, \cdots, x_n^k)_{1 \leq k \leq K}$ un K-échantillon du vecteur (de dimension n) X, on peut généraliser l'expression de la fonction de répartition empirique en posant :

$$F_K(x_1, \cdots, x_n) = \frac{1}{K} \sum_{i=1}^{K} 1_{\left\{x_1^k \leq x_1, \cdots, x_n^k \leq x_n\right\}} \tag{1.25}$$

Cet estimateur conduit à un estimateur non paramétrique naturel d'une copule.

1.4.2 Estimation paramétrique et semi-paramétrique

On se place ici dans le cas où la distribution conjointe dépend d'un paramètre, que l'on cherche à estimer.

Méthode des moments

Cette méthode est notamment utilisée pour les mesures de dépendance, l'estimateur des moments de la mesure de dépendance considérée est alors simplement obtenu en égalant l'expression paramétrique (analytique) de la mesure avec un estimateur non paramétrique de cette même mesure.

Par exemple, pour le tau de Kendall, on a $\widehat{\tau} = \frac{c-d}{c+d}$ avec c (resp. d) le nombre de paires concordantes (resp discordantes) dans l'échantillon. Donc pour une copule de Gumbel l'expression $\tau = 1 - \frac{1}{\theta}$ conduit à l'estimateur des moments $\widehat{\theta} = \frac{1}{1-\widehat{\tau}}$.

Maximum de vraisemblance

De l'égalité $F(x_1, \cdots, x_n) = C(F_1(x_1), \cdots, F_n(x_n))$ on déduit par dérivation l'expression de la densité du vecteur $(X_1, \cdots, X_n)$:

$$f(x_1, \cdots, x_n) = c(F_1(x_1), \cdots, F_n(x_n)) \prod_{i=1}^{n} f_i(x_i) \tag{1.26}$$

où $c(u_1, \cdots, u_n) = \frac{\partial^n C(u_1, \cdots, u_n)}{\partial u_1, \cdots, \partial u_n}$ désigne la densité de la copule. L'expression de log-vraissemblance de l'échantillon $(x_1^k, \cdots, x_n^k)_{1 \leq k \leq K}$ s'en déduit immédiatement

$$l(\theta) = \sum_{k=1}^{K} \ln(c(F_1(x_1^k; \theta), \cdots, F_n(x_n^k; \theta); \theta)) + \sum_{k=1}^{K} \sum_{i=1}^{n} f_i(x_i^k; \theta) \tag{1.27}$$

Il reste maximiser cette expression en θ, ce qui peut s'avérer en pratique fastidieux.

Méthode « IFM»

Cette méthode, proposée par Shih et Louis [1995] consiste à « découper » le problème d'estimation en deux étapes successives

l'estimation des paramètres $\theta_1, \cdots, \theta_n$ des marginales.

l'estimation du paramètre θ_c de la copule.

Le paramètre θ est donc décomposé sous la forme $\theta = (\ \theta_1, \cdots, \theta_n, \theta_c)$ et on commence par déterminer les estimateurs EMV des paramètres des marginales, soit

$$\widehat{\theta_i} = \arg\max_{\theta i} \sum_{k=1}^{K} f_i(x_i^k; \theta_i) \tag{1.28}$$

On « injecte » alors ces estimateurs dans la partie « copule » de la log-vraisemblance, ce qui conduit à $\widehat{\theta_c} = \arg\max_{\theta_c} \sum_{k=1}^{K} \ln(c(F_1(x_1^k;\widehat{\theta_1}),\cdots,F_n(x_n^k;\widehat{\theta_n});\theta_c))$. D'autres procédures peuvent être imaginées, comme par exemple l'estimation non paramétrique des marginales suivi d'un maximum de vraisemblance pour le paramètre de la copule (procédure « omnibus » de Genest, [1995] ou Shih et Louis, [1995]).

Chapitre 2

Polynômes et Copules de Bernstein

2.1 Les polynômes de Bernstein

Introduction

Les polynômes de Bernstein, nommés ainsi en l'honneur du mathématicien russe Sergeï Bernstein (1880-1968), permettent de donner une démonstration constructive et probabiliste du théorème d'approximation de Weierstrass. Ils sont également utilisés dans la formulation générale des courbes de Bézier.

Définition 2.1. *Soit f une fonction définie et continue sur [0, 1] à valeurs dans* $\mathbb{C}$. *Pour n entier naturel non nul donné, le n-ième polynôme de Bernstein associé à f est*

$$B_n(f) = \sum_{k=0}^{n} C_n^k f\left(\frac{k}{n}\right) X^k (1-X)^{n-k}. \tag{2.1}$$

Exemple 2.2. $\sum_{k=0}^{n} C_n^k (k-nX)^2 X^k (1-X)^{n-k} = nX(1-X)$, *en effet :*

a) on suppose que $\forall x \in [0,1]$, $f(x) = 1$.

Si, pour tout x de [0, 1], $f(x) = 1$, *alors pour n entier naturel non nul donné*

$$B_n(f) = \sum_{k=0}^{n} C_n^k X^k (1-X)^{n-k} = (X + (1-X))^n = 1.$$

Convergence uniforme de la suite des polynômes de Bernstein.

Rappelons d'abord le théoreme de Heine.

Théorème 2.3. *(théorème de Heine) Toute fonction f continue sur un intervalle fermé borné [a,b] est uniformément continue. Soit f une fonction continue sur* $[0,1]$ *à*

valeurs dans $\mathbb{R}$ *ou* $\mathbb{C}$. *On va montrer que la suite* $(B_n(f))_{n\in\mathbb{N}^*}$ *converge uniformément vers* f *sur* $[0,1]$.

a) Une majoration de $|f(x) - B_n(f)(x)|$:

Soit x un réel de $[0,1]$ et n un entier naturel non nul.

$$|f(x) - B_n(f)(x)| = \left| f(x) - \sum_{k=0}^{n} C_n^k f\left(\tfrac{k}{n}\right) x^k (1-x)^{n-k} \right|$$

$$= \left| f(x) \sum_{k=0}^{n} C_n^k x^k (1-x)^{n-k} - \sum_{k=0}^{n} C_n^k f\left(\tfrac{k}{n}\right) x^k (1-x)^{n-k} \right| \text{ (d'aprés (a))}$$

$$= \left| \sum_{k=0}^{n} C_n^k (f(x) - f\left(\tfrac{k}{n}\right)) x^k (1-x)^{n-k} \right| \quad \leq \sum_{k=0}^{n} C_n^k \left| f(x) - f\left(\tfrac{k}{n}\right) \right| x^k (1-x)^{n-k}$$

b) Pourquoi l'expression précédente est-elle petite ?

Tout d'abord, $\sum_{k=0}^{n} C_n^k x^k (1-x)^{n-k} = 1$ est une expression bornée uniformément en x.

Ensuite, pour x donnée et pour des k tels que $\left|x - \frac{k}{n}\right|$ est petit, $\left|f(x) - f\left(\frac{k}{n}\right)\right|$ est petit. Pour k tels que $\frac{k}{n}$ assez éloigné de x et décrivant donc un sous-enssemble J de [0,n], $\left|f(x) - f\left(\frac{k}{n}\right)\right|$ est bornée uniformément en x et il n'y a qu'à espérer que $\sum_{k\in J} C_n^k x^k (1-x)^{n-k}$ soit petit. Mais là, on dispose de

$$\sum_{k\in J} C_n^k (\tfrac{k}{n} - x)^2 x^k (1-x)^{n-k} \leq \sum_{k=0}^{n} C_n^k (\tfrac{k}{n} - x)^2 x^k (1-x)^{n-k} = \tfrac{x(1-x)}{n} \leq \tfrac{1}{4n}$$

c) Soit ε un réel strictement positif.

f est continue sur le segment [0, 1] et est donc d'une part bornée sur ce segment, et d'autre part uniformément continue sur ce segment d'après le théorème de Heine. Par suite, il existe un réel M tel que pour tout x de [0, 1], $|f(x)| \leq M$ et il existe un réel $\alpha > 0$ tel que $\forall (x,y) \in [0,1]^2, |x-y| < \alpha \implies |f(y) - f(x)| < \frac{\varepsilon}{2}$.

d) Soient n un entier naturel non nul et x un réel de [0, 1].

$$|f(x) - B_n(f)(x)| \leq \sum_{k=0}^{n} C_n^k \left| f(x) - f\left(\tfrac{k}{n}\right) \right| x^k (1-x)^{n-k}$$

$$= \sum_{\substack{k=0 \\ \left|x-\frac{k}{n}\right|<\alpha}}^{n} C_n^k \left| f(x) - f\left(\tfrac{k}{n}\right) \right| x^k (1-x)^{n-k} + \sum_{\substack{k=0 \\ \left|x-\frac{k}{n}\right|\geq\alpha}}^{n} C_n^k \left| f(x) - f\left(\tfrac{k}{n}\right) \right| x^k (1-x)^{n-k}$$

$$\leq \tfrac{\varepsilon}{2} \sum_{\substack{k=0 \\ \left|x-\frac{k}{n}\right|<\alpha}}^{n} C_n^k x^k (1-x)^{n-k} + 2M \sum_{\substack{k=0 \\ \left|x-\frac{k}{n}\right|\geq\alpha}}^{n} C_n^k x^k (1-x)^{n-k}$$

$$\leq \frac{\varepsilon}{2}\sum_{k=0}^{n} C_n^k x^k(1-x)^{n-k} + \frac{2M}{\alpha^2}\sum_{\substack{k=0\\ |x-\frac{k}{n}|\geq\alpha}}^{n} C_n^k(x-\tfrac{k}{n})^2 x^k(1-x)^{n-k}$$

$$\leq \frac{\varepsilon}{2} + \frac{2M}{\alpha^2}\sum_{k=0}^{n} C_n^k(x-\tfrac{k}{n})^2 x^k(1-x)^{n-k}$$

$$= \frac{\varepsilon}{2} + \frac{2M}{\alpha^2}\frac{x(1-x)}{n}.$$

(dans la deuxième somme, l'inégalité $\left|x-\frac{k}{n}\right| \geq \alpha$ s'écrit encore $1 \leq \frac{(x-\frac{k}{n})^2}{\alpha^2}$ et donc

$|f(x) - B_n(f)(x)| \leq \frac{\varepsilon}{2}\frac{2M}{\alpha^2}\frac{1/4}{n} = \frac{\varepsilon}{2} + \frac{2M}{2n\alpha^2}$,

(si $x \in [0,1], x(1-x) = \frac{1}{4} - (x-\frac{1}{2})^2 \leq \frac{1}{4}$ En résumé, ε strictement positif ayant été donné, on a montré que $\forall x \in [0,1], \forall n \in \mathbb{N}^*, |f(x) - B_n(f)(x)| \leq \frac{\varepsilon}{2} + \frac{2M}{2n\alpha^2}$.

or $\frac{\varepsilon}{2} + \frac{2M}{2n\alpha^2}$ tend vers $\frac{\varepsilon}{2}$ quand n tend vers $+\infty$. par suite , il existe un entier naturel non nul n_0 tel que tout entier natural n$\geq n_0$, $\frac{\varepsilon}{2} + \frac{2M}{2n\alpha^2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ et donc pour tout réel x $\in [0,1], |f(x) - B_n(f)(x)| < \varepsilon$.

on a montré que $\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}^* / \forall x \in [0,1], \forall n \in \mathbb{N}$,
$(n \geq n_0 \implies |f(x) - B_n(f)(x)| < \varepsilon)$,et donc que La suite $B_n(f))_{n\in\mathbb{N}^*}$ des polynomes de Bernstein converge uniformément vers f sur $[0,1]$.

Théorème 2.4. *(théoreme de Weirstrass) Toute fonction continue de* $[0,1]$ *dans* $\mathbb{R}$ *ou* $\mathbb{C}$ *est limite uniforme sur* $[0,1]$ *d'une suite de polynômes.*

Plus généralement, toute fonction continue sur un segment [a,b] de $\mathbb{R}$ *à valeur dans* $\mathbb{R}$ *ou* $\mathbb{C}$*, est limite uniforme sur ce segment d'une suite de polynômes.*

Rappel sur la loi Binomiale

Une loi binomiale est une loi de probabilité discrète décrite par deux paramètres n le nombre d'expériences réalisées, et p la probabilité de succès. Pour chaque expérience appelée épreuve de Bernoulli, on utilise une variable aléatoire qui prend la valeur 1 lors d'un succès et la valeur 0 sinon. La variable aléatoire, somme de toutes ces variables aléatoires, compte le nombre de succès suit une loi binomiale. Il est alors possible d'obtenir la probabilité de k succès dans une répétition de n expériences .

Définition 2.5. *La loi binomiale, de paramètres n et p, est la loi de probabilité discrète d'une variable aléatoire X dont la fonction de masse est donnée par*

$$P(X = k) = C_n^k P^k (1-P)^{n-k}, \forall k = 1, \cdots, n. \tag{2.2}$$

Propriétés

On rappelle que deux variables aléatoires Y_1 et Y_2 de loi discrète sont indépendantes si

$$P(Y_1 = \mathbf{k}, Y_2 = h) = P(Y_1 = k)P(Y_2 = h), E(X) = np; Var(X) = np(1-p). \quad (2.3)$$

Lien avec la loi Binomiale

D'un point de vue probabiliste, pour tout $p \in [0;1], B_i^m(p)$ est la probabilité $P(X = i)$, où X est une variable aléatoire suivant une loi binomiale de paramètre (m,p). C'est d'ailleurs l'interprétation qu'en fait Bernstein dans sa démonstration du théorème d'approximation de Weierstrass.

Voici une explication probabiliste

Soit $X_1, \cdots, X_n$ une suite de variables aléatoires réelles indépendantes telles que

$$\begin{cases} P(X_i = 1) = x \\ P(X_i = 0) = 1 - x \end{cases}$$

Autrement dit , on tire n fois 0 ou 1 avec à chaque fois la probabilité x d'obtenir 1 et $1 - x$ d'obtenir 0, on pose ensuite $S_n = \frac{X_1 + \cdots + X_n}{n}$

$X_1 + \cdots + X_n$ compte le nombre de fois ou on a tiré 1 par la loi des grands nombres, S_n converge en probabilité vers x.
Maintenant des calcules classiques montrent que $S_n$suit une loi binomiale , avec $P(S_n = \frac{k}{n}) = C_n^k x^k (1-x)^{n-k}$ on en déduit

$$E(f(S_n)) = \sum_{k=0}^{n} C_n^k f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} \quad (2.4)$$

En autre, puisque S_n converge en probabilité vers x , il est logique de pense que $p(B_n(f(x))) = E(f(S_n)) \to E(f(x)) = f(x)$.

Corollaire 2.6. *Si $(P_n)_{n\in\mathbb{N}}$ est une suite de polynômes convergeant uniformément sur $\mathbb{R}$ vers une fonction f, f est nécessairement un polynôme.*

Remarque 6. *Ce resultat montre que les séries entiéres usuelles de rayon infini (de somme e^x ou $cosx$) ne sont pas uniformément convergentes sur $\mathbb{R}$.*

Les propriétés des polynomes de Bernstein

Les polynomes de Bernstein possédent les propriétés suivantes Le degré de B_k^n est égale à n.

a) partition de l'unité

$$\sum_{i=0}^{n} B_i^n(x) = 1. \tag{2.5}$$

b) positivité

$$\forall i \in \{0, \cdots, n\}, B_i^n(x) \geq 0. \tag{2.6}$$

c) symétrie

$$\forall i \in \{0, \cdots, n\}, B_i^n(x) = B_{n-i}^n(1-x). \tag{2.7}$$

d) valeurs aux bords

$$\forall i \in \{0, \cdots, n\}, B_i^n(0) = \delta_{i,0}, B_i^n(1) = \delta_{i,n} \tag{2.8}$$

avec δ le symbole de Kronecher.

e) multiplicité des racines

pour $B_i^n, 0$ est une racine d'ordre de multiplicité i et 1 une racine de multiplicité $n-i$.

f) formules de récurrence, pour $n > 0$

$$B_i^n(x) = \begin{cases} (1-x)B_i^{n-1}(x) & \text{Si } \ i = 0 \\ (1-x)B_i^{n-1}(x) + xB_{i-1}^{n-1}(x) & \forall i \in \{0, \cdots, n\} \\ xB_{i-1}^{n-1}(x) & \text{Si } i = n \end{cases} \tag{2.9}$$

$$B_i^{n\prime}(x) = n(B_{i-1}^{n-1}(x) - B_i^{n-1}(x)).$$

g) décomposition sur la base canonique

$$B_i^n(x) = C_i^n \sum_{k=0}^{n-i} C_k^{n-i}(-1)^k x^{i+k} = \sum_{l=i}^{n} C_l^n C_i^l (-1)^{l-i} x^l. \tag{2.10}$$

h) inversement

$$x^p = \sum_{k=0}^{n-p} C_k^{n-p} \frac{1}{C_k^n} B_{n-k}^n(x) = \frac{1}{C_i^n} \sum_{s=p}^{n} C_p^s B_s^n(x). \tag{2.11}$$

B_k^n atteint sont maximum sur $[0,1]$ en $\frac{k}{n}$.

Théorème 2.7. *Pour tout $n \in \mathbb{N}$, les polynômes $(B_k^n)_{0<k<\infty}$ forment une base de $\mathbb{R}_n[X]$.*

2.2 Les copules de Bernstein

Nous rappelons la représentation de Sklar pour les distributions multivariées, Soit H une fonction de distribution k-dimontional avec des marges 1-dimensionnelles $F_1, \cdots, F_k$, alors il existe une fonction C de l'unité k-cube à l'unité cube telle que

$$H(x_1, \cdots, x_k) = C(F_1(x_1), \cdots, F_k(x_k)) \tag{2.12}$$

C est appelé $k-$copule, si chaque F_j est continue, la copule est unique.

Soit $\alpha(\frac{\upsilon_1}{m_1}, \cdots, \frac{\upsilon_k}{m_k})$ une constante de valeur réelle indexée par $(\upsilon_1, \cdots, \upsilon_k)$, $\upsilon_j \in \mathbb{N}_+$, telle que $0 \leq \upsilon_j \leq m_j$ bien que nous puissions simplement utiliser $\alpha_{\upsilon_1, \cdots, \upsilon_k}$, $0 \leq \upsilon_j \leq m_j, \forall j$, pour commodité conceptuelle nous ne le faisons pas, maintenant nous définissons l'objet à étudier.

Définition 2.8. *Soit* $P_{\upsilon_j,\ m_j}(u_j) = C_{\upsilon_j}^{m_j} u_j^{m_j - \upsilon_j} (1 - u_j)^{m_j - \upsilon_j}$
Si $C_B : [0,1]^k \to [0,1]$, *où*

$$C_B(u_1, \cdots, u_k) = \sum_{\upsilon_1} \cdots \sum_{\upsilon_k} \alpha(\tfrac{\upsilon_1}{m_1}, \cdots, \tfrac{\upsilon_k}{m_k})\ P_{\upsilon_1,\ m_j}(u_1) \cdots\ P_{\upsilon_k,\ m_j}(u_k), \cdots (*)$$

Satisfait aux propriétés de la fonction copule, alors C_B *est une copule de Bernstein pour tout* $m_j \geq 1$.

La copule de Bernstein généralise les familles de copules de polynômes, les polynômes sont des cas spéciaux de copules avec une section de polynômes dans une ou plusieurs variables, pour le cas 2-dimensionnels,

$$C(u_1, u_2) = u_1^3 a(u_2) + u_1^2 b(u_2) + u_1 c(u_2) + d(u_2) \tag{2.13}$$

est une copule appropriée pour un choix approprié de fonctions $a(\cdots)$, $b(\cdots)$, $c(\cdots)$ et $d(\cdots)$.

Cette copule a des sections cubiques et les fonctions mentionnées ci-dessus peuvent être des polynômes, puis elles peuvent être exactement écrites comme une copule de Berstein.

Théorème 2.9. *Soit* $C_{[0,1]^k}$ *l'espace des fonctions continues bornées dans l'hypercube k-dimensionnel* $[0,1]^k$. *Ensuite, l'ensemble des polynômes de Bernstein définis en* $(*)$ *est dense dans* $C_{[0,1]^k}$.

Propriétés des copules de Bernstein

Nous listons quelques propriétés de C_B en commun avec toutes les autres copules

(1) C_B est croissante dans tous ses arguments.

(2) C_B satisfait aux bornes de Frechet, c'est-à-dire

$$\min(0, u_1 + \cdots + u_k - (k-1)) \leq \ C_B(u_1, \cdots, u_k) \ \leq \min(u_1, \cdots, u_k)$$

ce qui signifie que C_B est minimun : c'est-à-dire : $C_B(u_1, \cdots, u_k) = 0$ Si $u_j = 0$ pour au moins un$j$, et $C_B(1, \cdots, 1, u_j, 1, \cdots, 1) = u_j$,$\forall j$.

(3) $\prod_{j=1}^{k} u_j$ est une copule pour des variables aléatoires indépendantes, c.-à-d. La copule de produit.

(4) C_B est lipschitzienne, c'est-à-dire

$$|C_B(x_1, \cdots, x_k) - C_B(y_1, \cdots, y_k)| \ \leq \sum_{j=1}^{k} |x_j - y_j| .$$

À la lumière de ces propriétés, le résultat suivant montre les propriétés associées spécifiques à la copule de Bernstein

Théorème 2.10. *$C_B(u_1, \cdots, u_k)$ est une copule de Bernstein si et seulement si*

$$\sum_{l_1=0}^{1} \cdots \sum_{l_k=0}^{1} (-1)^{l_1+\cdots+l_k} \alpha(\frac{\upsilon_1 + l_1}{m_1}, \cdots, \frac{\upsilon_k + l_k}{m_k}) \geq 0$$

et

$$\min(0, \frac{\upsilon_1}{m} + \cdots + \frac{\upsilon_k}{m} - (k-1)) \leq \alpha(\frac{\upsilon_1}{m}, \cdots, \frac{\upsilon_k}{m}) \leq \min(\frac{\upsilon_1}{m}, \cdots, \frac{\upsilon_k}{m}),$$

en particulier

$$\lim_{\upsilon_j \to 0} \alpha(\frac{\upsilon_1}{m}, \cdots, \frac{\upsilon_k}{m}) = 0, \forall j = 1, \cdots, k$$

et

$$\alpha(1, \cdots, 1, \frac{\upsilon_j}{m}, 1, \cdots, 1) = \frac{\upsilon_j}{m}, \forall j = 1, \cdots, k$$

Théorème 2.11. *toute copule $C(u_1, \cdots, u_k)$ peut être écrite comme $u_1 \cdots u_k + G(u_1, \cdots, u_k)$ où $u_1 \cdots u_k$ est la copule produit et $G(u_1, \cdots, u_k)$ est un terme de pertubation contenant toutes les informations sur la dépendance de $(u_1, \cdots, u_k)$.*

La densité de Bernstein Toute copule de Bernstein a une densité de copules, c'est parce que la copule de Bernstein est absolument continue, Définissez $\triangle_{1,\cdots,k}$ comme opérateur de différence avant k -dimensionel c-á-d

$$\triangle_{1,\cdots,k}\alpha(\frac{\upsilon_1}{m},\cdots,\frac{\upsilon_k}{m}) \equiv \sum_{l_1=0}^{1}\cdots\sum_{l_k=0}^{1}(-1)^{k+l_1+\cdots+l_k}\alpha(\frac{\upsilon_1+l_1}{m},\cdots,\frac{\upsilon_k+l_k}{m}) \quad (2.14)$$

En raison des propriétés préservant la convexité des polynômes de Bernstein, la densité de copules de Bernstein a la structure suivante

$$c_B(u_1,...,u_k) = m^k \sum_{\upsilon_1=0}^{m-1}\cdots\sum_{\upsilon_k=0}^{m-1}\triangle_{1,\cdots,k}\alpha(\frac{\upsilon_1}{m},\cdots,\frac{\upsilon_k}{m}) \times P_{\upsilon_1,\ m-1}(u_1)\cdots P_{\upsilon_k,\ m-1}(u_k), \quad (2.15)$$

telle que $c_B = \frac{\partial^k C_B}{\partial u_1\cdots\partial u_k}$

Pour plus de commodité, nous utilisons la définition suivante de la densité de copules de Bernstein

$$c_B(u_1,\cdots,u_k) = \sum_{\upsilon_1=0}^{m}\cdots\sum_{\upsilon_k=0}^{m}\beta(\frac{\upsilon_1}{m},\cdots,\frac{\upsilon_k}{m}) \times \prod_{j=1}^{k} C_{\upsilon_j}^{\ m} u_j^{\upsilon_j}(1-u_j)^{m-\upsilon_j} \quad (2.16)$$

où $\beta(\frac{\upsilon_1}{m},\cdots,\frac{\upsilon_k}{m})$ est défini par

$$\beta(\frac{\upsilon_1}{m},\cdots,\frac{\upsilon_k}{m}) \equiv (m+1)^k\triangle_{1,\cdots,k}\alpha(\frac{\upsilon_1}{m+1},\cdots,\frac{\upsilon_k}{m+1}) \quad (2.17)$$

Corollaire 2.12. *La densité de copule est toujours positive.*

2.3 De polynômes de Bernstein aux copules de Bernstein

Théorème 2.13. *Soit F la cdf d'un vecteur aléatoire $X = (X_1,...,X_d)$, c-á-d $F(x_1,\cdots,x_d) = P(X_1 \le x_1,\cdots,X_d \le x_d)$ avec cdfs marginales $F_1,\cdots,F_d$. Il existe alors une copule $C : [0,1]^d \to [0,1]$ tel que $F(x_1,\cdots,x_d) = C(F_1(x_1),\cdots,F_d(x_d))$ pour tous $x_1,\cdots,x_d \in \mathbb{R}$.*

Si $F_1,\cdots,F_d$ sont continus alors C est unique.

***Vice versa** pour une copule C et des cdfs univariés $F_1,\cdots,F_d$, la relation $F(x_1,\cdots,x_d) = C(F_1(x_1),\cdots,F_d(x_d))$ définit le cdf F de certains d -variable vecteur aléatoire avec cdfs marginaux $F_1,\cdots,F_d$.*

Ainsi, le théorème de Sklar assure que le cdf F de tout vecteur aléatoire à d-variable peut être écrit en termes de ses fonctions de distribution marginale $F_1, \cdots, F_d$ et une copule C appropriée qui décrit ainsi la structure de dépendance des composantes du vecteur. Une telle décomposition est souvent très utile en pratique ; pour une application exemplaire dans le contexte des copules de Bernstein. La définition de ce type de copule spécifique, construite au moyen de polynômes de Bernstein. Jusqu'à présent, la discussion des modèles potentiels de copules s'est principalement concentrée sur d'autres types, à savoir le cas elliptique (par exemple, la gaussienne et la copule t) ou le cas d'Archimède (par exemple les copules de Gumbel, Clayton et Frank). Il semble que le véritable impact des polynômes de Bernstein sur les modèles de copules n'ait été découvert que plus récemment, d'abord dans le cadre de la théorie de l'approximation et plus tard en particulier en rapport avec applications en finance . Les copules de Bernstein présentent plusieurs avantages par rapport aux approches traditionnelles. Les copules de Bernstein permettent une description très flexible, non paramétrique et essentiellement non symétrique des structures de dépendance également dans des dimensions plus élevées Les copules de Bernstein se rapprochent arbitrairement de toute autre copule donnée les densités de copules de Bernstein sont données sous une forme explicite et peuvent donc être facilement utilisées pour des études de simulation Monte Carlo. Dans cet article, nous passons en revue la construction des copules de Bernstein à travers des vecteurs aléatoires discrets avec des marges uniformes (appelées squelettes discrets), et soulignons leur connexion aux copules, et aux opérateurs de produits tensoriels de Bernstein. La représentation explicite des copules de Bernstein en termes d'opérateurs de Bernstein à produit tensoriel avec un squelette discret, à notre connaissance, n'a pas été mentionnée auparavant dans la littérature connexe. Cette approchLemme, entre autres, ouvre une approche pragmatique et d'économie de stockage pour adapter la structure de dépendance des données observées aux copules de Bernstein via des copules de tour, un sous-cas spécial de copules en damier basé sur la distribution empirique multivariée. La représentation du produit tensoriel pourrait également être utile dans d'autres études sur la préservation de la douceur globale pour l'approximation de la copule, car elle permet un transfert direct des résultats de la théorie de l'approximation multivariée dans le contexte de la copule.

Quelques faits mathématiques simples sur les polynômes de Bernstein et les copules de Bernstein

Les affirmations du lemme suivant sont bien connues dans la littérature, mais pour plus de commodité et une meilleure compréhension dans le contexte de la copule.

Lemme 2.14. *Soit* $B(m,k,z) = \begin{pmatrix} m \\ k \end{pmatrix} z^k(1-z)m-k, 0 \leq z \leq 1,$
$k = 0, \cdots, m \in \mathbb{N}$, *alors nous avons*

$$\int_0^1 mB(m-1,k,z)dz = 1 \tag{2.18}$$

pour $k = 0, \cdots, m-1$.

Lemme 2.15. *De plus,* $\frac{d}{dz}B(m,k,z) = m[B(m-1,k-1,z) - B(m-1,k,z)]$ *pour* $k = 0, \cdots, m$, *avec la convention* $B(m-1,-1,z) = B(m-1,m,z) = 0$. *Pour l'opé rateur de Bernstein défini par* $B_m f : z \mapsto \sum_{k=0}^{m} f\left(\frac{k}{m}\right) B(m,k,z)$ *pour les fonctions à valeurs réelles sur* $[0,1]$*et* $z \in [0,1]$, *cela donne*

$$\frac{d}{dz}B_m f(z) = m\sum_{k=0}^{m-1} \Delta_m f\left(\frac{k}{m}\right) B(m-1,k,z) \tag{2.19}$$

où $\Delta_m f(z) := f\left(z+\frac{1}{m}\right) - f(z)$*pour* $z \in [0,1]$ *désigne l'opérateur de différence directe.*

Théorème 2.16. *Pour* $d \in \mathbb{N}$ *soit* $U = (U_1, \cdots, U_d)$ *et soit un vecteur aléatoire dont la composante marginale* U_i *suit une distribution uniforme discrète sur*
$T_i := \{0, 1, \cdots, m_i - 1\}$ *avec* $m_i \in \mathbb{N}, i = 1, \cdots, d$, *soit plus loin* $p(k_1, \cdots, k_d) := p\left(\bigcap_{i=1}^{d}\{U_i = k_i\}\right)$ *pour tout* $(k_1, \cdots, k_d) \in \prod_{i=1}^{d} T_i$. *alors*

$$c_B^U(u_1, \cdots, u_d) := \sum_{k_1=0}^{m_1-1} \cdots \sum_{k_d=0}^{m_d-1} p(k_1, \cdots, k_d) \prod_{i=1}^{d} m_i B(m_i - 1, k_i, u_i), (u_1, \cdots, u_d) \in [0,1]^d \tag{2.20}$$

définit la densité d'une copule C_B^U *d-dimensionnelle, appelée copule de Bernstein. Nous appelons* c_B^U *la densité de copules de Bernstein induite par* U. *Le vecteur* U *est également appelé le squelette discret de la copule de Bernstein.*

Remarque 7. *La derniére equation montre que si* $U = (U_1, \cdots, U_d)$ *est un vecteur aléatoire avec une densité* c_B^U *de copules de Bernstein conjointe comme ci-dessus, alors tout vecteur aléatoire partiel* $V = (U_{i_1}, \cdots, U_{i_n})$ *avec* $n < d$ *et*

1≤ $i_1 < \cdots < i_n \leq d$ possède une densité de copules de Bernstein c_B^V donnée par

$$c_B^V(u_{i_1}, \cdots, u_{i_n}) = \sum_{k_{i_1}=0}^{m_{i_1}-1} \cdots \sum_{k_{i_n}=0}^{m_{i_n}-1} p\left(\bigcap_{l=1}^{n}\{U_{i_l} = k_{i_l}\}\right) \prod_{l=1}^{n} m_{i_l} B(m_{i_l}-1, k_{i_l}, u_{i_l}) \tag{2.21}$$

$(u_{i_1}, \cdots, u_{i_n}) \in [0,1]^n$

Théorème 2.17. *Dans les conditions du théorème précédent, la copule de Bernstein C_B^U induite par U est explicitement donnée par*

$$\begin{aligned} C_B^U(x_1, ..., x_d) &= \int_0^{x_d} ... \int_0^{x_1} c_B^U(u_1, ..., u_d) du_1 ... du_d \\ &= \sum_{k_1=0}^{m_1} ... \sum_{k_d=0}^{m_d} p\left(\bigcap_{i=1}^{d}\{U_i < k_i\}\right) \prod_{i=1}^{d} B(m_i, k_i, x_i), \text{ pour } (x_1, \cdots, x_d) \in [0,1]^d \end{aligned} \tag{2.22}$$

Remarque 8. *D'un point de vue probabiliste, à la lumière du lemme précédent, les densités de copules de Bernstein $c_B^U(u_1, \cdots, u_d\ )$ peuvent également être considérées comme des mélanges de densités de vecteurs aléatoires $Y(k_1, m_1, \cdots, k_d, m_d\ ) = (Y_{(k_1,m_1)}, \cdots, Y_{(k_d,m_d)})$ avec des composants indépendants qui suivent les distributions bêta avec paramètres $k_j + 1 et\ m_j + 1$ et de densité*

$$f_{Y_{(k_j,m_j)}}(z) = m_j(\begin{cases} m_j - 1 \\ k_j z^{k_j}(1-z)^{m_j-1-k_j} \end{cases}$$

$$= \frac{1}{B(k_j+1, m_j-k_j)} z^{k_j}(1-z)^{m_j-1-k_j} \tag{2.23}$$

pour $j = 1, \cdots, d$ et $z \in [0,1]$. Ici U est le vecteur aléatoire de mélange. D'un point de vue algorithmique, cette représentation est particulièrement utile pour les simulations de Monte Carlo avec des copules de Bernstein.

Chapitre 3

Ulistration par l'article de " Victor Fossaluza, Luís Gustavo Esteves and Carlos Alberto de Bragança Pereira" sous le titre"*Estimating Multivariate Discrete Distributions Using Bernstein Copulas*"

3.1 Introduction

Mesurer la dépendance entre les variables aléatoires est l'un des problèmes les plus fondamentaux en statistique, et par conséquent, elle est de grand intérêt. Les copules sont récemment devenues un outil important pour inférer correctement la distribution conjointe des variables d'intérêt. Bien que de nombreuses études aient traité le cas des variables continues, peu les études se sont concentrées sur le traitement des variables discrètes. Ce chapitre présente une approche non paramétrique de l'estimation des distributions discrètes conjointes avec support borné à l'aide des copules et des polynômes de Bernstein. Nous présentons une application sur des données réelles sur les troubles obsessionnels compulsifs (obsessive-compulsive disorder).

3.2 Solutions existantes

Tout d'abord, nous introduisons le cadre mathématique de ce problème. Soit F la fonction de distribution inconnue d'un vecteur aléatoire X qui prend des valeurs dans un sous-ensemble de $\mathbb{R}^p$. Un échantillon de taille n de X est représenté par $X_1, ..., X_n$, où $X_i = (X_{i1}, ..., X_{ip})$ et $i = 1, ..., n$. En d'autres termes, les X_i sont des variables aléatoires conditionnellement indépendantes et identiquement distribuées les réalisations deX_i sont notées x_i. En supposant que la distribution F est tirée d'une famille connue de distributions, nous représentons le modèle statistique par (X, F, P_θ), où X est l'espace d'échantillonnage, F est une sigma-algèbre de ses sous-ensembles et $P = \{p(\cdot \mid \theta) : \theta \in \Theta, \Theta \subset \mathbb{R}^p\}$est une famille de distributions indexées par le paramètre θ qui appartient à l'espace paramétrique Θ. L'estimation de F est alors réduite à celle du paramètre θ, et la structure de dépendance est limité à celui pris en charge par le modèle statistique sous-jacent. Depuis de nombreuses années, la distribution normale multivariée a été utilisée pour la plupart des analyses multivariées. Récemment, pour de nombreux phénomènes aléatoires dont les distributions sont asymétriques et possèdent des queues plus lourdes que celles de la distribution normale, des distributions alternatives, telles que les distributions asymétriques multivariées.

Dans les approches récentes, les copules sont devenues un outil populaire pour modéliser les structures de dépendance multivariées et pour obtenir de nouvelles distributions multivariées avec des marginaux donnés. En bref, une copule est une distribution multivariée dont les marginaux sont uniformes sur l'interval [0, 1]. Il existe de nombreuses familles paramétriques de copules, permettant la modé lisation de nombreuses structures de dépendance différentes.
Soit F une fonction de distribution p-dimensionnelle avec les marges $F_1, ..., F_p$. Sklar a d'abord montré qu'il existe une copule C de dimension p telle que :

$$F(x_1, ..., x_p) = C(F_1(x_1), ..., F_p(x_p)) \tag{3.1}$$

pour tout $x = (x_1, ..., x_p)$ dans le domaine de F. Si les variables $X_1, ..., X_p$ sont absolument continus, alors la copule C est unique ; sinon, C est uniquement déterminé sur Ran(F_1)... Ran (F_p). Ainsi, la copule peut être utilisée pour modéliser séparé ment les marges et la structure de dépendance. La représentation non unique d'une copule pour des distributions discrètes est un problème théorique qui doit être considéré dans le contexte d'une preuve analytique, mais cela ne limite pas ses

applications empiriques. Cependant, le théorème de Sklar ne nous dit pas comment trouver la copule C. Ce problème est largement discuté dans la littérature, et plusieurs solutions à ajuster plusieurs familles de copules (paramétriques) et choisissez l'une d'entre elles à l'aide de certains critères de sélection ou d'un test d'adéquation.

Des techniques non paramétriques peuvent également être appliquées pour estimer une distribution multivariée. Une solution populaire utilisant cette approche est l'application de la fonction de distribution empirique $F^{(n)} : \mathbb{R}^p \to [0,1]$, qui est défini, pour $(t_1, ..., t_p) \in R^p$, comme

$$F^{(n)}(t_1, ..., t_p) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_{1i} \leq t_1, ..., x_{pi} \leq t_p\}} \tag{3.2}$$

où $1_{\{A\}}$ est l'indicatrice de l'ensemble A. Cette approche équivaut à utiliser les fréquences relatives pour estimer la fonction de masse de probabilité conjointe. Les fréquences relatives coïncident avec l'estimation du maximum de vraisemblance sous l'hypothèse que les données sont tirées d'une distribution multinomiale. Un inconvénient de cette approche est que la probabilité de toute cellule non observée sera estimée à zéro. Une autre approche possible consiste à utiliser une fonction pour le lissage la distribution empirique. Nous pouvons considérer les polynômes de Bernstein à cet effet en raison de leur simplicité et de leurs bonnes propriétés mathématiques. Soit$h : [0,1] \longrightarrow \mathbb{R}$ une fonction continue. Le polynôme de Bernstein de degré m (multivarié) pour la fonction h, $B_h^m : [0,1]^p \to \mathbb{R}$, est défini par

$$B_h^m(x_1, ..., x_p) = \sum_{j1=0}^{m} ... \sum_{jp=0}^{m} h(\frac{j1}{m}, ..., \frac{jp}{m}) \prod_{i=1}^{p} \begin{pmatrix} m \\ ji \end{pmatrix} x_i^{ji} (1-x)^{m-ji} \tag{3.3}$$

Les polynômes de Bernstein multivariés pour la fonction h convergent uniformément vers la fonction h comme $m \to \infty$, et ses dérivés sont simples à obtenir. La fonction h doit être définie dans $[0,1]^p$, et donc, pour des raisons pratiques, les données qui ne prennent pas de valeurs dans $[0,1]^p$ doivent d'abord être transformées. Pour appliquer cette mé thode aux données OCD, par exemple, nous considérons la transformation $Y = X/20$. Des polynômes DE Bernstein ont été utilisés pour approximer une copule C en remplaçant simplement la fonction h par la copule. Le polynôme de Bernstein résultant,B_C^m, qui est aussi une copule qui converge fortement vers C, est appelé copule de Bernstein. Lorsque la vraie copule est inconnue, la copule empirique peut être utilisée à la place, et la fonction résultante est appelée la copule empirique de Bernstein. La copule empirique est définie par

$$C_n(u_1, ..., u_p) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{F_1(x_{1i}) \leq u_1, ..., F_p(x_{pi}) \leq u_p\}} \tag{3.4}$$

Notez que même lorsque $F_i, i \in 1, ..., n$, est inconnu, nous pouvons utiliser la distribution marginale empirique$F_i^{(n)}$ comme estimateur cohérent de F_i, selon le théorème de Glivenko Cantelli. D'autres estimateurs des distributions marginales pourraient être envisagés à la place, comme dans la procédure propos ée dans la section suivante. Nous avons jusqu'à présent obtenu une fonction continue comme estimation alors que notre objectif est clairement d'estimer une fonction de masse de probabilité (discrè te). Par conséquent, cette fonction doit être discrétisée pour obtenir une estimation adéquate. Ceci peut être réalisé comme suit : supposons, sans perte de généralité, que $X = (X_1, ..., X_p)$ est un vecteur aléatoire tel que toutes ses composantes $X_i, i = 1, 2, ..., p$, ont des valeurs dans l'ensemble $\Omega = \{0, 1, ..., k\}$ avec probabilité un. De plus, il existe toujours un vecteur aléatoire continu $Z = (Z_1, ..., Z_p)$ avec une fonction de distribution F telle que
$P(0 \leq Z_i \leq k) = 1$et $X_i = \sum_{j=0}^{k} j1_{\{j-0.5 \leq Z_i \leq j+0.5\}}$.
Il s'ensuit que

$$P(X_1 = x_1, ..., X_p = x_p) = P(x_1 - 0.5 \leq Z_1 \leq x_1 + 0.5, ..., x_p - 0.5 \leq Z_p \leq x_p + 0.5) \tag{3.5}$$

Soit F (ou une estimation $\widehat{F}$) la fonction de distribution conjointe continue du vecteur aléatoire Z, soit $B = [a, b] = [a_1, b1]...[a_1, b_p]$ et soit un curbe de dimension p avec tous ses sommets dans Ω. Le volume F de B est alors donné par

$$V_F(B) = \sum_c sgn(c)F(c) \tag{3.6}$$

où la somme est prise sur tous les sommets $c = (c_1, ..c_p)$ de B, et sgn (c) est donnée par

$$\text{sign(c)}= \begin{cases} 1 \text{ si} & c_j = a_j \text{ pour un nombre paire de } j \\ \text{-1 } si & \text{c}_j = a_j \text{ pour un nombre impaire de } j \end{cases}$$

En particulier, supposons que $b = (b_1, ..., b_p) \in \{0, 1, ..., k\}^p$, avec
$B = [b_1 - \frac{1}{2}, b_1 + \frac{1}{2}] = [b_1 - 0.5, b_1 + 0.5] \times [b_2 - 0.5, b_2 + 0.5] \times ... \times [b_p - 0.5, b_p + 0.5]$
avec $b_i \in \Omega, \forall i = 1, ..., p$ alors la probabilité de l'événement
$\{X = b\} = \{X_1 = b_1, ..., X_p = b_p\}$ peut être calculé (estimé) comme suit

$$P(X = b) = P(b_1 - 0,5 < Z_1 \leq b_1 + 0,5, ..., b_p - 0,5 < Z_p \leq b_p + 0,5]) = V_F(B) \tag{3.7}$$

Du fait qu'une faible convergence se produit aux points de discontinuité de la fonction de distribution limite F et que notre objectif est d'estimer une fonction de masse

de probabilité discrète, nous considérons des ensembles de la forme $B = [b_1 - \frac{1}{2},\ b_1 + \frac{1}{2}]$ avec $b_i \in \Omega,\ i = 1, ..., p$ de sorte que les sommets du rectangle p -dimensionnel B sont toujours des points de continuité de la fonction de distribution du vecteur aléatoire discret X. Ainsi, une telle discrétisation donne des estimations satisfaisantes pour la fonction de masse de probabilité de X.

3.3 Solutions proposées

La méthode proposée pour estimer la distribution conjointe d'un vecteur aléatoire discret consiste à utiliser des polynômes de Bernstein pour estimer les marginales et la copule. L'avantage de cette méthode est qu'elle permet d'utiliser toutes les observations, même en cas de valeurs manquantes de certaines variables. De plus, cette méthode est une approche non paramétrique et il existe peu de restrictions sur la structure de dépendance.

Premièrement, pour chaque variable aléatoire X_i, nous estimons les distributions marginales en utilisant la distribution marginale empirique avec n_i observations,

$F_i^{(n_i)}(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} 1_{\{x_{ij} \leq x\}}, i = 1, ..., p$, le polynôme de Bernstein de degré $m_i = \frac{n_i}{\log(n_i)}$ pour lisser cette fonction est

$$B_i^{m_i}(x) = \sum_{j=1}^{m_i} F_i(n_i)(\frac{j}{m_j}) \begin{pmatrix} m_i \\ j \end{pmatrix} x^j (1-x)^{(m_i - j)} \tag{3.8}$$

Comme cet estimateur converge vers la distribution marginale, nous estimons la copule en utilisant une version alternative de la copule empirique basée sur les n observations complètes et les estimations

$$B_i^{m_i}, i = 1, ..., p \quad C_n(u_1, ..., u_p) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{B_1^{m_1}(x_{1j}) \leq u_1, ..., B_p^{m_p}(x_{pj}) \leq u_p\}} \tag{3.9}$$

et lisser cette fonction pour obtenir la copule de Bernstein empirique correspondante

$$B_{C_n}^m(u_1, ..., u_p) = \sum_{j1=0}^{m} ... \sum_{jp=0}^{m} C_n(\frac{j1}{m}, ..., \frac{jp}{m}) \prod_{i=1}^{p} \begin{pmatrix} m \\ ji \end{pmatrix} x_i^{ji} (1-x)^{m-ji} \tag{3.10}$$

Notez que la construction de la copule C_n en utilisant des polynômes de Bernstein plutôt que des fonctions de distribution empiriques (marginales) donne, au moins dans les exemples présentés dans la section 4, des estimations non nulles pour les cellules non observées. Cette caractéristique justifie le choix de cette version alternative de la copule empirique.

L'estimation de la fonction de distribution conjointe est une discrétisation de la fonction suivante

$$\widehat{F}_{m,n}(x_1, ..., x_p) = B^m_{C_n}(B^m_1(x_1), ..., B^m_p(x_p)) \tag{3.11}$$

L'algorithme utilisé pour obtenir la solution proposée est assez simple et est résumé ci-dessous

1. Pour toutes les observations n_i de chaque variable X_i, estimer la fonction de distribution empirique marginale $F(n_i)$.
2. Lisser chaque fonction $F_i^{n_i}$ en utilisant un polynôme de Bernstein $B_i^{m_i}$ i de degré m_i.
3. Pour toutes les observations complètes du vecteur aléatoire X, estimer la copule empirique C_n.
4. Estimer la copule de Bernstein en lissant la copule empirique C_n à l'aide du polynôme de Bernstein multivarié de degré $B^m_{C_n}$.
5. Obtenir une estimation continue de la fonction de distribution multivariée $\widehat{F}_{m,n}$ donnée par l'equation (3.11).
6. Discrétiser $\widehat{F}_{m,n}$ pour obtenir une estimation de la fonction de masse de probabilité discrète multivariée.

3.4 Distances utilisées

Pour les exemples décrits ci-dessus, nous présentons les estimations des fonctions de masse de après ces exemples, nous présentons et comparons les probabilités obtenus par la méthode proposée et de comparer ses performances avec certaines solutions existantes.

a) La distribution empirique présentée dans qui est obtenue en utilisant uniquement le paires et la fonction de probabilité résultante, coïncide avec les fréquences relatives des points observés dans l'échantillon.

b) l'approximation t asymétrique multivariée obtenue par discrétisation d'un paramètre distribution continue multivariée, estimée par la méthode du maximum de vraisemblance en utilisant uniquement les paires complètes.

c) La discrétisation de la copule normale avec l'approximation marginale nor-

male de la distribution fonction obtenue en utilisant toutes les observations pour l'estimation de la distribution marginale et en utilisant seules les paires complètes pour l'estimation de la copule. Cette méthode est assez similaire à celle décrite en (b) considérant la distribution multivariée normale plutôt que la distribution t asymétrique, mais ici, il est possible d'estimer les distributions marginales en utilisant toutes les données disponibles, pas seulement les paires complètes.

d) La discrétisation de l'approximation polynomiale de Bernstein empirique présentée dans l'équation (3.3), en remplaçant la fonction h par la distribution empirique $F(n)$ obtenue en (a) et en utilisant uniquement les paires complètes.

e) La solution proposée décrite dans la section précédente, qui est obtenue en utilisant le Bernstein polynôme pour approximer les marges en utilisant toutes les observations et la copule approchée en utilisant les paires complètes. Pour tous les exemples, nous illustrons les estimations des distributions de probabilité par l'évolution de plusieurs distances entre les distributions estimée et théorique. Dans ce but, une notation doit être introduite.

Soit $\theta = (\theta_1, \cdots, \theta_k)$) les probabilit és théoriques, et soit $\widehat{\theta} = (\widehat{\theta}_1, \cdots, \widehat{\theta}_k)$ les probabilités estimées. Nous considérons les distances suivantes pour comparer les estimations

3.4.1 Distance d'Aitchison

$$\triangle(\widehat{\theta}, \theta) = \sqrt{\sum_{i=1}^{k} [\ln(\frac{\widehat{\theta}_i}{\theta_i}) - L]^2}, \text{ oú } L = \frac{1}{k} \sum_{i=1}^{k} \ln(\frac{\widehat{\theta}_i}{\theta_i})$$

3.4.2 Distance euclidienne

$$\delta(\widehat{\theta}, \theta) = \sqrt{\sum_{i=1}^{k} [\widehat{\theta}_i - \theta_i]^2}$$

3.4.3 Distance de variation totale

$$\tau(\widehat{\theta}, \theta) = \frac{1}{2} \sum_{i=1}^{k} \left| \widehat{\theta}_i - \theta_i \right|$$

3.4.4 Divergence symétrisée de Kullback – Leibler

$$D(\widehat{\theta},\theta) = \frac{1}{2}[\sum_{i=1}^{k} \theta_i \ln(\frac{\theta_i}{\widehat{\theta}_i}) + \sum_{i=1}^{k} \widehat{\theta}_i \ln(\frac{\widehat{\theta}_i}{\theta_i})]$$

Aitchison et Pawlowsky ont présenté de nombreux arguments en faveur de l'utilisation d'Aitchison distance pour les vecteurs de composition, c'est-à-dire lorsque la somme des composantes du vecteur est constante (dans notre cas, la somme des probabilités est égale à un).

À la fin de cette section, nous présentons les estimations de la distribution des données réelles décrites dans l'introduction. Dans ce cas, nous ne connaissons pas la distribution théorique, nous ne présentons que le les estimations et les distances calculées à partir de la distribution empirique.

3.5 Distribution en forme elliptique simulée

Dans cette section, nous simulons les données d'une distribution de forme elliptique avec des marginaux X_1 bêta-binômiale ($N_x = 20, a = 5, b = 5$) et binômiale $Y_1(N_y = 20, p = 0,5)$ et une copule normale avec paramètre $r = 0,7$.

Nous pouvons voir dans les tableaux 1, 2, et 3 que dans ces exemples, les solutions basées sur les distributions elliptiques, à savoir les distributions asymétriques τ et normales, donnent de meilleures estimations. Cette une estimation supérieure se produit parce que la fonction de masse de probabilité théorique est de forme elliptique. Cependant, dans des situations pratiques, nous n'avons aucune connaissance de la forme réelle de la distribution. Dans ce cas, la distribution empirique peut être une bonne base pour évaluer les estimations, malgré l'existence de nombreux points non observés qui sont estimés à zéro. Lorsque les estimations sont comparées à la distribution empirique, la solution que nous proposons semble produire de bons résultats, présence de données censurées.

Tableau 1. Distances entre les estimations et les probabilités théoriques pour 600 paires complètes d' observations. Les valeurs en gras mettent en évidence les petites distances.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	4.98521	0.02116	0.09988	0.04154
T-copula	1.44499	0.00629	0.02915	0.00345
Normal copula	**1.28402**	**0.00476**	**0.02418**	**0.00236**
Bernstein polynomial	3.45943	0.01388	0.07159	0.01870
Bernstein copula	3.23712	0.01217	0.06360	0.01578

Tableau 2. Distances entre les estimations et les probabilités théoriques pour le cas des données censurées dans un seul marginal, avec 1000 observations dans un marginal et 200 dans l'autre. Les valeurs mettre en évidence les plus petites distances

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	9.87909	0.03454	0.17083	0.12490
T-copula	1.28441	**0.00493**	**0.02291**	**0.00239**
Normal copula	**1.28040**	0.00554	0.02738	0.00284
Bernstein polynomial	4.84901	0.02049	0.11110	0.03982
Bernstein copula	3.28689	0.01171	0.06340	0.01530

Tableau 3. Distances entre les estimations et les probabilités théoriques pour le cas des données censurées pour les deux variables, avec 600 observations pour chaque variable, dont 300 forment des paires complètes. Les valeurs mettent en évidence les plus petites distances.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	7.32955	003035	0.13826	0.09162
T-copula	1.12383	0.00419	0.02221	0.00185
Normal copula	**1.06365**	**0.00375**	**0.01891**	**0.00146**
Bernstein polynomial	4.54073	0.01934	0.10051	0.03531
Bernstein copula	3.54526	0.01377	0.06743	0.01963

3.5.1 Distribution asymétrique simulée

Dans cette section, nous présentons les données simulées pour une distribution asymétrique avec marges Binômiales X_2 bêta($N_x = 20, a = 0,85, b = 1,1$) et binômiale $Y_2(N_y = 15,\ \pi = 0,6)$ et un Gumbel copule avec le paramètre $\theta = 0,7$.

Dans le cas d'une distribution asymétrique, notre solution proposée donne une meilleure estimation dans les trois cas considérés , le cas des données censurées, le cas avec données manquantes dans une variable et le cas avec des données manquantes dans Tableau 4.

Tableau 4. Distances entre les estimations et les probabilités théoriques pour 600 paires complètes d'observations. Les valeurs en gras mettent en évidence les petites distances.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	6.17032	0.02767	0.12761	0.06377
T-copula	5.47625	0.03287	0.14429	0.07724
Normal copula	5.76598	0.03293	0.14785	0.08020
Bernstein polynomial	5.41325	**0.02436**	0.11969	0.05380
Bernstein copula	**5.07634**	0.02519	**0.11842**	**0.05068**

Tableau 5. Distances entre les estimations et les probabilités théoriques pour le cas des données censurées dans un seul marginal, avec 1000 observations dans un marginal et 200 dans l'autre. Les valeurs audacieuses mettre en évidence les plus petites distances.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	8.77534	0.03906	0.19104	0.14363
T-copula	5.23773	0.03130	0.13494	0.07356
Normal copula	5.07437	0.02892	0.12727	0.06549
Bernstein polynomial	5.65580	0.02626	0.13135	0.06562
Bernstein copula	**4.86027**	**0.02558**	**0.11172**	**0.05379**

Tableau 6. Distances entre les estimations et les probabilités théoriques pour le cas des données censurées dans les deux variables, avec 600 observations pour chaque variable, dont 300 forment des paires complètes. L'audacieux les valeurs mettent en évidence les plus petites distances.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
empirical	7.32005	0.03284	0.15576	0.09786
T-copula	4.79233	0.02760	0.12321	0.05917
Normal copula	5.06253	0.02819	0.12855	0.06325
Bernstein polynomial	5.09522	0.02253	0.11486	0.05028
Bernstein copula	**4.35547**	**0.01957**	**0.09863**	**0.03595**

La nouvelle méthode a donné de meilleurs résultats que les autres solutions présentées. Il convient de souligner que la méthode a été élaborée pour les cas avec une grande proportion de données exemple de la sous section 3.5.1 considérant les données censurées dans un seul marginal. Tailles d'échantillon n_2 f_1,. . . , 2000g avec la même proportion de données censurées de l'exemple ont été considérées. Dans ce cas, le soutien de la distribution compte 336 points. A noter que pour les petits échantillons ($n < 750$), la méthode proposée présente les plus petites distances. Bien que la méthode ait été développée pour de petits échantillons, il semble que les estimations convergent vers la distribution théorique lorsque n augmente. Cependant, une enquête détaillée sur les propriétés asymptotiques de la nouvelle méthode sont nécessaires et reste un probléme ouvert.

3.5.2 Etude des données réelles

Dans cette section, nous présentons les estimations des données réelles décrites dans l'introduction. le YBOCS est l'une des mesures des résultats les plus largement utilisées dans les études de traitement de l'obsessionnel compulsif trouble (TOC). Les scores YBOCS totaux comprennent un nombre entier variant de 0 à 40 et visent pour évaluer la gravité des symptômes obsessionnels compulsifs. Le score total YBOCS est la somme de deux sous-échelles, allant de 0 à 20, dont l'une mesure la gravité de la contrainte et l'autre d'obsession. Les travaux proposent qu'au lieu de la somme, il serait préférable de considérer le maximum de ces deux sous-échelles, appelées M-YBOCS. Ainsi, les psychiatres se sont intéressés à comprendre les propriétés de cette nouvelle échelle, comme la distribution de probabilité des scores M-YBOCS avant et après que les patients aient reçu un traitement pour le TOC. Comme déjà mentionné dans l'introduction, l'ensemble de données contient 1001 observations des scores au temps initial et seulement 213 au moment final. Cela se produit parce que

de nombreux patients abandonnent le traitement. Les causes du décrochage peuvent être extrêmement différentes, telles qu'une réduction des symptômes le patient estime qu'il / elle n'a pas besoin de traitement, ou même aggrave les symptômes, provoquant le patient pour discréditer le traitement. Le petit nombre de paires complètes dans la base de données rend difficile d'estimer la distribution conjointe. Dans les problèmes réels, il y a peu de cas où la loi de probabilité qui génère les données est révélé. Dans de tels cas, une méthode assez courante pour évaluer si les méthodes proposées sont adéquates est comparer les estimations avec les données observées. Dans les modèles prédictifs, par exemple, il est courant de vérifier une certaine distance entre les valeurs prévues et observées. De cette façon, nous comparons la distance entre les estimations et la fonction de probabilité empirique (qui est la fréquence relative de chaque observation). La solution proposée donne des distances plus petites que les approches existantes (tableau 7).

Tableau 7. Distances entre les estimations et les probabilités empiriques pour les données réelles.

	Aitchison	**Euclidean**	**Variation total**	**Kullback – Leibler**
T-copula	11.56219	0.03941	0.22684	0.19899
Normal copula	12.07092	0.04143	0.24701	0.21760
Bernstein polynomial	11.35361	0.03933	0.22910	0.19125
Bernstein copula	**10.81475**	**0.03703**	**0.21020**	**0.17184**

L'estimation par la distribution empirique présente de nombreux zéros en raison du petit nombre d'observations. Les chercheurs pensent que la proportion de points non observés diminuerait si l'échantillon avait moins d'abandons. En outre, ils estiment que les hypothèses communes de normalité ou même des hypothèses de symétrie n'ont aucun sens dans ce cas. La méthode proposée attribue des valeurs positifs aux probabilités aux cellules non observées et capture la nature asymétrique des données.

Conclusions

Dans ce travail, une nouvelle approche du problème de l'estimation des distributions bivariées discrètes est présenté. La procédure, qui consiste essentiellement à estimer à la fois les marginaux et la copule en utilisant des polynômes de Bernstein, vise à résoudre trois problèmes importants : le traitement des données bivariées en présence de valeurs manquantes marginales (en utilisant toutes les informations disponibles, y compris les paires d'observations incomplètes), la possibilité d'obtenir des estimations positives pour les cellules non observées, donnant ainsi des distributions discrètes estimées «plus lisses» ; et la considération d'une grande variété des structures de dépendance entre les variables aléatoires pertinentes. La nouvelle approche convient aux ces cas en raison de sa nature assez non restrictive et non paramétrique. L'utilisation des polynômes de Bernstein donnent de meilleurs résultats que la distribution empirique pour estimer la distribution marginale. C'est important de noter que la copule empirique de Bernstein ne produit une copule qu'asymptotiquement, et d'autres méthodes pourraient être utilisées pour estimer la copule de Bernstein. En tous cas, la méthode proposée a montré des estimations raisonnables pour les fonctions de masse de probabilité étudiées. le nouvelle méthode peut également être appliquée aux variables aléatoires à p-dimensions, $p > 2$.

La nouvelle méthode a été appliquée à plusieurs exemples de données simulées, et selon quelques mesures typiques de la distance (entre les distributions estimée et théorique), il est mieux adopté que certaines des solutions existantes dans les cas où le nombre de points à estimer est supérieur à la taille de l'échantillon . Bien que la méthode ait été développée pour de petits échantillons, il semble que les estimations proposées convergent vers la distribution théorique, mais des études plus détaillées sont encore nécessaires. La nouvelle méthode a également été appliquée aux données échantillonnées auprès d'adultes diagnostiqués avec une trouble obsessionnel compulsif. L'estimation obtenue par la méthode a été appréciée par les chercheurs en psychiatrie.

Bien que la nouvelle méthode présente des avantages pratiques par rapport aux alternatives existantes présentées, certains aspects n'ont pas été abordés ici, à savoir : il faudra encore développer la nouvelle procédure sous plusieurs aspects qui n'ont pas été abordés ici : (i) l'étude des propriétés asymptotiques pour les grands taille d'échantillon n et / ou pour un degré polynomial supérieur m ; ii) une justification officielle de la nouvelle procédure sous une approche décisionnelle théorique.

Bibliographie

[1] C. Cottin and D. Pfeifer : *From Bernstein polynomials to Bernstein copulas*, May 2014.

[2] Michel Denuit and Philippe Lambert : *Constraints on concordance measures in bivariate discrete data*, Belguin, 4 april 2002.

[3] Victor Fossaluza, Luís Gustavo Esteves and Carlos Alberto de Bragança Pereira :*Estimating Multivariate Discrete Distributions Using Bernstein Copulas*, 2017.

[4] H.Joe, *Dependence Modeling with Copulas, CRC Press : Boca Raton, FL*, USA, 2014.

[5] G. Jogesh Babu and Yogendra P. Chaubey : *Smooth estimation of a distribution and density function on a hypercube using Bernstein polynomials for dependent random vectors*, 21 november 2005.

[6] Alexandre Leblanc : *On estimating distribution functions using Bernstein polynomials*, 20 November 2011.

[7] R. B. Nelsen, *An Introduction to Copulas, 2nd ed, Springer Verlag*, New York, NY, USA, 2006.

[8] Frédéric Planchet : *Dépendance stochastique, Introduction à la théorie des copules*, Décembre 2010.

[9] Alexandre Popier : *copules, Université du Maine, Le Mans*, Septembre 2010.

[10] Jean-Louis Rouget : *Polynômes de Bernstein*, 2007.

[11] A. Sancetta and S. E. Satchell : *The Bernstein copula and its applications to modeling and approximations of multivariate distributions*, June 2004.

[12] *Théorie des Valeurs Extr^emes*, Groupe de Recherche Op´erationnelle Cr´edit Lyonnais Bercy-Expo – Immeuble Bercy Sud – 4‘eme etage 90, Quai de Bercy — 75613 Paris Cedex 12, France, Janvier 2002.

[13] Pravin K. Trivedi and David M. Zimmer : *Copula Modeling : An Introduction for Practitioners*, 2005.