MTCopula: Génération de données synthétiques et complexes basées sur les Copules

Fodil Benali, Damien Bodénès, Nicolas Labroche, Cyril de Runz

MTCopula: Génération de données synthétiques et complexes basées sur les Copules

Fodil Benali*,**, Damien Bodénès*, Nicolas Labroche **, Cyril de Runz**

*Adwanted Group, Paris
{fbenali,dbodenes}@adwanted.com,
**BDTLN - LIFAT, University of Tours, Blois, France
{nicolas.labroche, cyril.derunz}@univ-tours.fr

Résumé. Cet article est une version courte de Benali et al. (2021) [1]. La plupart des techniques existantes de génération de données ne fonctionnent bien que pour de faibles dimensions et échouent à capturer les dépendances complexes entre les dimensions des données. L'identification de la bonne combinaison de modèles et de leurs paramètres respectifs reste un problème ouvert. Nous présentons MTCopula, une nouvelle approche de génération de données synthétiques complexes, flexible et extensible, qui choisit automatiquement le meilleur modèle de copules et les marginales les mieux ajustées pour capturer la complexité des données en se reposant sur le critère d'information d'Akaike.

1 Introduction

De nos jours, il peut être difficile d'obtenir des données de valeur et de qualité en quantité, du fait des moyens de collecte et des possibles problèmes de confidentialité, comme c'est le cas de la planification publicitaire, notre contexte industriel. Dans ce contexte, seuls de petits volumes données complexes et de haute qualité (multidimensionnelles, multivariées, catégorielles/continues, etc.), représentatifs de l'ensemble des données, sont disponibles pour générer un jeu de données synthétiques large et réaliste. Par conséquent, il existe un réel besoin pour un générateur de données complexes réalistes.

Notre objectif est de générer de nouvelles données qui conservent les mêmes caractéristiques que les données originales, telles que la distribution de leurs attributs et leur interdépendance, afin que tout travail effectué sur les données d'origine puisse être réalisé en utilisant les données synthétiques (Petricioli et al., 2020). Ceci ne peut pas être fait en utilisant la méthode habituelle de génération de données synthétiques unidimensionnelles car, lorsqu'elle est appliquée dans un contexte hautement dimensionnel, elle ne permet pas de modéliser la dépendance entre les variables. Pour résoudre ces problèmes, plusieurs travaux récents se sont concentrés sur des approches d'apprentissage profond comme le Generative Adversarial Network (GAN), mais ces approches nécessitent une grande quantité de données pour l'étape d'apprentissage et ne peuvent donc pas être utilisées pour notre problème.

1. http://ceur-ws.org/Vol-2840/paper8.pdf

Les modèles basés sur les copules suscitent un intérêt croissant pour l'estimation (Salinas et al., 2019) et l'échantillonnage (Li et al., 2020) à partir d'une fonction de distribution multivariée. Les copules (Nelsen, 2007) sont des distributions de probabilités conjointes dans lesquelles toute probabilité univariée continue est représentée par une fonction de distribution et dans lesquelles toute distribution de probabilité continue univariée peut être insérée en tant que marge. La copule capture le comportement conjoint des variables et modélise la structure de dépendance, tandis que chaque marge modélise le comportement individuel de la variable correspondante. Ainsi, notre problème se résume à construire une distribution de probabilité conjointe qui s'ajuste au mieux à la distribution marginale de chaque variable et permet de capturer les différentes dépendances entre ces variables. Ce problème est souvent vu comme une tâche d'apprentissage de structure qui peut être résolu en maximisant la vraisemblance ou un critère issu de la théorie de l'information (Portet, 2020).

La copule est un outil mathématique flexible qui peut supporter différentes configurations en termes d'ajustement de distributions marginales et de modèles de copules. Le choix de la meilleure configuration n'est pas simple. Par exemple, dans la littérature, les générateurs de données basés sur les copules utilisent généralement le modèle de copule gaussienne, alors même que ce modèle a des difficultés à capturer les dépendances de queue, ce qui peut affecter la qualité des données générées.

Dans ce travail, nos contributions sont les suivantes : (1) nous formalisons le problème de la génération de données complexes et synthétiques ; (2) nous proposons MTCopula pour apprendre les copules et choisir automatiquement les marginales et le modèle de copule qui s'ajuste le mieux aux données que nous cherchons à générer ; et (3) nous présentons des expériences montrant la manière dont MTCopula préserve les relations implicites entre les variables dans les ensembles de données synthétiques sur un cas d'utilisation réel et des ensembles de données de la littérature.

Ce document présente la formulation de notre problème (section 2), notre solution pour modéliser et générer des données avec leurs structures de dépendance (section 3), les expériences réalisées (section 4) et finalement conclut et ouvre une discussion sur nos travaux futurs (section 5).

2 Formulation du problème

Notre objectif est, étant donné un ensemble d'observations complexes et représentatives L_o, de générer un ensemble de données synthétiques L_s qui est similaire à l'ensemble de données original L_o avec les propriétés suivantes :

— pour chaque variable de l'ensemble de données, les valeurs générées doivent être cohérentes avec la distribution de la variable, et
— les dépendances entre les variables doivent rester les mêmes dans le nouvel ensemble de données.

Cet objectif peut être reformulé comme suit : trouver automatiquement le modèle statistique reposant sur des copules qui permet la meilleure génération de données possibles. Dans notre cas, cela peut être fait, premièrement, en estimant les paramètres des marginales et, deuxièmement, en estimant les paramètres des copules qui représentent les relations de dépendances entre marginales. Cependant, cet ajustement ne sera presque

jamais exact. Ainsi, le problème consiste à déterminer les paramètres du modèle qui minimisent la quantité relative d'information perdue.

Dans la littérature, le critère d'information d'Akaike (AIC) (Akaike, 1974) est souvent utilisé, car il offre un compromis entre la qualité de l'ajustement et la simplicité du modèle en le pénalisant proportionnellement à son nombre de paramètres. Cela permet de diminuer le risque de sur- et sous-adaptation en même temps. Toutefois, ce critère n'a pas été utilisé dans le contexte de la détermination automatique des meilleures marginales ou des meilleurs modèles de copules pour la génération de données.

Dans ce qui suit, nous formulons donc notre problème sur la base de l'AIC sans perte de généralité car tout autre test pourrait être utilisé, tel que le test de *Kolmogorov-Smirnov*, qui ne pénalise pas les modèles ayant le plus de paramètres. Sur la base de l'AIC, notre problème de génération de données synthétiques devient le problème d'optimisation en deux étapes suivant :

1. L'échantillonnage des valeurs cohérentes avec le comportement de chaque variable $x_{j\in[\![1;d]\!]}$ (où d est le nombre de variables) consiste à trouver la fonction de densité de distribution marginale correspondante (f_j, γ_j) telle que :

$$minimiser\ AIC = 2k - 2\ln(\hat{\mathcal{L}}(\hat{\gamma_j}|x_j)),\ j = 1..d \tag{1}$$

   où $\hat{\mathcal{L}}(\hat{\gamma_j}|x_j) = \prod_{i=1}^{n} f_j(x_{ij}|\hat{\gamma_j})$ représente l'estimateur du maximum de vraisemblance (MLE ou Maximum Likelihood Estimator) d'une densité marginale candidate f_j avec un vecteur de paramètres $\hat{\gamma_j}$ de dimension k obtenu ainsi :

$$\hat{\gamma_j} = arg\max_{\gamma_j} \prod_{i=1}^{n} f_j(x_{ij}|\gamma_j) \tag{2}$$

   où n est le nombre d'observations.

2. Caractériser le comportement d'interdépendance des variables consiste à trouver la densité de distribution conjointe, ici les paramètres de la copule, (h, θ) que :

$$minimiser\ AIC = 2k - 2\ln(\hat{\mathcal{L}}(\hat{\theta}|x_1, .., x_j, .., x_d)) \tag{3}$$

   où $\hat{\mathcal{L}}(\theta|x_1, .., x_j, .., x_d) = \prod_{i=1}^{n} h(x_1, .., x_j, .., x_d|\hat{\theta})$ est le MLE du modèle h avec pour paramètre θ. $\hat{\theta}$ est déterminé par :

$$\hat{\theta} = arg\max_{\theta} \prod_{i=1}^{n} h(x_{i1}, ..., x_{ij}, ..., x_{id}; \theta). \tag{4}$$

3 Solution proposée

Notre système, appelé MTCopula, se décompose en trois étapes : (1) préparation des données, (2) apprentissage du modèle de copule, et (3) génération de données synthétiques. Il est à noter que seule l'étape (1) est spécifique aux problèmes de données ayant besoin d'un encodage numérique, tandis que les étapes (2) et (3) sont entièrement génériques à tout scénario de génération de données synthétiques.

Nous présenterons ici uniquement les étapes 2 et 3. (Benali et al., 2021) fournit plus d'information sur l'étape 1, notamment pour le traitement des séries temporelles ou des données catégorielles.

3.1 Apprentissage des copules

Le processus d'apprentissage des copules se fait en deux étapes : l'ajustement des distributions marginales et l'ajustement des modèles de copules.

Notre système propose deux méthodes pour estimer les distributions marginales. La première est non-paramétrique, via la distribution empirique, et la seconde est paramétrique et utilise l'estimation du maximum de vraisemblance (MLE). L'AIC est utilisé pour automatiser le choix de la meilleure distribution marginale parmi un ensemble de distributions présélectionnées. Actuellement, nous choisissons, sans perte de généralité, parmi les distributions bornées suivantes : Gaussienne tronquée, GaussienneKDE (Kernel Density Estimator), Bêta, Exponentielle tronquée et Uniforme.

Les distributions marginales estimées sont utilisées pour construire des observations de pseudo-copules. Un critère de sélection de modèle, tel que l'AIC, est utilisé pour sélectionner la copule C qui s'adapte le mieux aux données de la pseudo-copule et qui caractérise la dépendance entre les marginaux.

La plupart des travaux, réalisés dans la génération de données synthétiques basée sur les copules, utilisent une copule gaussienne utilisant le facteur de corrélation de Pearson (Li et al., 2020) avec une approche MLE pour estimer les marginales. Notre système offre une certaine flexibilité en termes de choix de modèle de copule basé sur l'AIC, qui, à son tour, permet d'apprendre différents modèles de copule et de choisir le modèle qui correspond le mieux aux données d'entrée. Pour le moment, nous ajustons deux modèles, la copule gaussienne et la copule basée sur T-Student (que nous appelons T-copule), car ils sont capables de capturer différentes structures de dépendance : linéaire comme la corrélation en utilisant la copule gaussienne et l'indice de Pearson, et un comportement non linéaire comme la dépendance des queues de distributions en utilisant la T-copule. Par ailleurs, le facteur de corrélation de Pearson (utilisé généralement dans les copules gaussiennes) n'est pas invariant sous une transformation non linéaire strictement monotone, ce qui peut avoir un impact sur le processus d'estimation lors de la normalisation avec des fonctions de distributions marginales. Notre contribution MTCopula utilise l'inversion du τ de Kendall, qui est basée sur la relation entre le paramètre de corrélation des copules elliptiques (T-copule ou copule gaussienne) et le τ de Kendall de deux variables aléatoires. Pour la T-copule, une autre étape est nécessaire pour estimer les degrés de liberté, qui est basée sur le MLE avec la matrice de corrélation maintenue fixe.

3.2 Génération et reconstruction de données

Pour la génération de données synthétiques, les échantillons de copules sont générés par échantillonnage à partir de la fonction de densité de copules c qui correspond à la fonction de distribution conjointe estimée de copules C. Ensuite, la transformation de probabilité inverse est appliquée pour ramener les échantillons de copules à la distribution naturelle des données.

Pour générer des variables aléatoires corrélées, notre méthode utilise la factorisation de Cholesky, qui est couramment utilisée dans la simulation de Monte Carlo pour produire des estimations efficaces des valeurs simulées (Zhu et al., 2017). Une fois le processus de génération de données synthétiques terminé, dans le cas de génération de données contenant des variables catégorielles ou des series temporelles, une opération de reconstruction est effectuée afin de reconvertir dans les formats adéquats (Benali et al., 2021).

4 Expériences

Dans cette section, nous décrivons une partie des expériences qui ont été menées pour valider la capacité de MTCopula à générer des données synthétiques [2] détaillées dans (Benali et al., 2021). Nous cherchons dans cet article à répondre à la question suivante : « Comment s'affranchir du principal goulot d'étranglement des méthodes basées sur les copules qui est (*i*) de pouvoir choisir parmi les modèles de marginales, et (*ii*) de choisir parmi les modèles copules qui peuvent avoir des propriétés différentes pour capturer la dépendance ? » MTCopula automatise le processus en utilisant le critère AIC comme mesure pour déterminer automatiquement le meilleur modèle soit pour les marginaux soit pour les copules (Copules Gaussiennes, T-Copules). Nous montrons dans quelle mesure ce choix est efficace dans notre contexte.

Pour les expériences présentées ci-après, nous utilisons les 2 jeux de données. Le jeu de données XYZ, contenant 1000 instances de 3 variables continues, a été généré en utilisant un mélange de distributions bêta et gaussiennes avec une corrélation entre Y et Z uniquement, afin de simuler des distributions marginales complexes. Le jeu de données Abalone (4177 instances, 8 variables continues, discrètes et catégorielles) provient de la plateforme de jeux de données de l'UCI [3].

4.1 AIC pour le choix des marginales

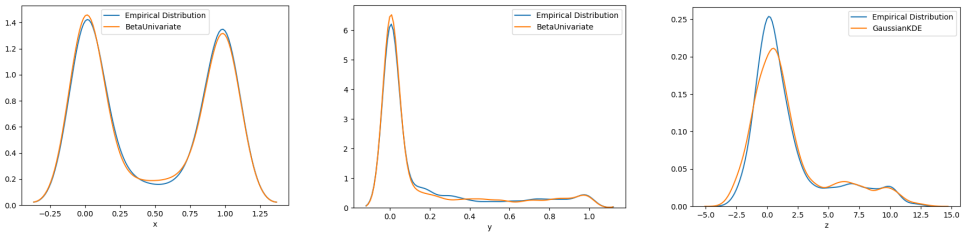
Pour évaluer l'importance de l'AIC dans la sélection de la distribution marginale correspondant le mieux au comportement des variables marginales, nous ajustons une liste de distributions bornées : distribution bêta, distribution uniforme, exponentielle tronquée, gaussienne tronquée et estimation de la densité du noyau, en utilisant la méthode MLE pour chaque variable. Pour chacune de ces distributions, nous évaluons l'AIC en utilisant les paramètres ajustés. La distribution présentant la valeur minimale de l'AIC est sélectionnée pour modéliser le comportement de la variable. Notez que nous utilisons une liste de distributions bornées afin d'éviter de générer des valeurs aberrantes. Le tableau 1 illustre l'évaluation de l'AIC pour l'ajustement des distributions marginales des variables de l'ensemble de données XYZ.

D'après le tableau 1, nous pouvons observer que, pour les variables X et Y, La distribution Bêta a une très petite valeur d'AIC (-11718.86 et -11001.61 respectivement). Par conséquent, nous remarquons que la distribution des données réelles (couleur bleue sur la figure 1) et la distribution ajustée (couleur orange sur la figure 1) sont presque identiques (voir les figures 1a et 1b). En revanche, pour la variable Z, la valeur de l'AIC minimum n'est pas aussi faible (4435.44) par rapport aux autres variables. Par conséquent, nous observons une différence significative entre la distribution ajustée et la distribution réelle des données dans la figure 1c. Cela s'explique par le fait que l'AIC estime la quantité relative d'informations perdues par un modèle donné : moins un modèle perd d'informations, plus la qualité de ce modèle est élevée.

2. Les codes sources sont disponibles sur `https://github.com/cderunz/MTCopula`.
3. `https://archive.ics.uci.edu/ml/datasets.php`

Variable	Bêta	KDE	Uniform	Truncated Exponential	Truncated Gaussian
X	-11718.86	-281.47	4.0	98.99	133.62
Y	-11001.61	-1116.15	3.96	-1497.21	-690.05
Z	240273.73	4435.44	5480.97	5040.59	4896.43

TAB. 1 – AIC sur les marginales du jeu de données XYZ.



(a) Dist. bêta ajustée pour X (b) Dist. bêta ajustée pour Y (c) Dist. KDE ajustée pour Z

FIG. 1 – Distributions marginales obtenues après ajustement.

4.2 AIC pour le choix du modèle de copule

Dans cette expérience, nous étudions l'impact du choix du modèle de copule sur la qualité de la génération de données, et nous démontrons l'importance de l'AIC pour choisir le meilleur modèle de copule. À cette fin, nous ajustons deux modèles de copules, le gaussien et le T-Copula, sur deux ensembles de données différents, XYZ et Abalone. Pour les deux modèles, nous utilisons la méthode de Kendall pour estimer la matrice de corrélation P. Le degré de liberté ν de la T-Copula est estimé par la méthode CMLE (Genest et al., 1995) avec la matrice de corrélation P maintenue fixe. Les résultats sont moyennés après **10 exécutions**. La figure 2 illustre l'évaluation de la RMSE pour estimer la préservation des dépendances à l'aide des deux copules.

(a) Variation de la RMSE pour XYZ (b) Variation de la RMSE pour Abalone

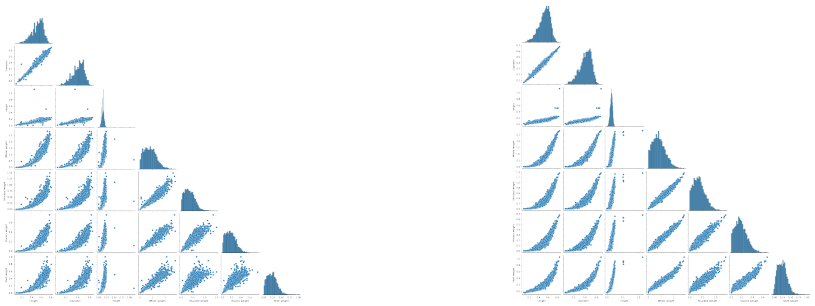
FIG. 2 – Evaluation de la préservation des structures de dépendances pour différents modèles de copules.

D'après la figure 2a), nous pouvons observer que, pour le jeu de données XYZ, la copule gaussienne est plus performante que la T-copule. D'autre part, comme le montre la figure 2b, la T-copule surpasse la copule gaussienne sur le jeu de données Abalone. Cela est dû au fait que le jeu de données XYZ ne présente pas de dépendance de queue. Par conséquent, l'utilisation de la T-copule aura un impact sur la matrice de corrélation en prenant en compte les dépendances dans les queues qui n'apparaissent pas dans les données originales. À l'inverse, l'ensemble de données Abalone présente une structure de

dépendance de queue, comme l'illustre la figure 3a. Par conséquent, l'utilisation d'une T-copule pour la génération de données corrigera les dépendances dans les queues, alors que ce n'est pas le cas avec la copule gaussienne. Pour l'instant, nous utilisons la T-copule uniquement pour la modélisation de dépendance de queue symétrique. C'est la raison pour laquelle, nous ne contrôlons pas la structure de la queue supérieure dans les données synthétiques générées comme le montre la figure 3b. Les résultats du tableau 2 confirment ces conclusions. Pour l'ensemble de données XYZ, le AIC minimal qui s'ajuste le mieux aux données correspond à la copule gaussienne $(3993,73)$. D'autre part, la T-copule a la valeur minimale de AIC qui s'ajuste le mieux au jeu de données Abalone (9507.26). Cela confirme l'intérêt de l'AIC pour choisir le meilleur modèle de copule qui s'adapte le mieux au processus de génération des données.

Database	**Copula Model**	**AIC Value**
XYZ	Gaussian	3993.73
	T-Student	3998.18
Abalone	Gaussian	12388.88
	T-Student	9507.26

Tab. 2 – AIC des modèles de copule gaussien et T-copule.



(a) Données réelles. (b) Données générées à l'aide d'une T-copule.

Fig. 3 – Visualisation du jeu de données Abalone et des données générées.

Dans cette section, nous avons démontré l'efficacité de MTCopula pour sélectionner, parmi différentes combinaisons d'ajustements de marginales et de modèles de copules, les modèles les plus appropriés qui représentent le mieux le processus de génération de données, et nous avons montré l'importance et la pertinence du critère AIC dans ce processus.

5 Conclusion

Dans cet article, nous avons proposé MTCopula, une solution flexible, extensible et générique pour la génération de données synthétiques complexes. Elle incorpore différents modèles de copules (pour l'instant les copules gaussiennes et les T-copules) afin de capturer différentes structures de dépendance, y compris les dépendances de queue. Pour garantir la préservation des dépendances pendant le processus d'apprentissage des copules, MTCopula fait appel au τ de Kendall, qui est robuste aux valeurs aberrantes

et invariant sous des transformations strictement monotones. Contrairement à la littérature qui utilise uniquement la distribution gaussienne pour modéliser les marginales, notre solution incorpore une variété de distributions bornées afin de s'adapter au mieux au comportement des variables et de ne pas générer de valeurs aberrantes. De plus, MTCopula est moins restrictif en termes de quantité de données d'entrée et est plus explicable que les GANs. MTCopula est capable de sélectionner automatiquement les distributions marginales univariées et le modèle de copule les mieux adaptés aux données d'entrée. Pour cela, il utilise la MLE pour ajuster le modèle des distributions marginales possibles, puis l'AIC pour choisir à la fois la meilleure distribution et le meilleur modèle de copule entre la T-copule et le modèle gaussien. MTCopula gère plusieurs types de données, notamment des ensembles de données tabulaires complexes et des séries chronologiques multiples/multivariées. Les expériences proposées montrent l'intérêt et l'efficacité de MTCopula par rapport aux méthodes existantes.

Références

Akaike, H. (1974). A new look at the statistical model identification. *IEEE trans. on automatic control 19*(6), 716–723.

Benali, F., D. Bodenes, N. Labroche, et C. de Runz (2021). Mtcopula : Synthetic complex data generation using copula. In *DOLAP 2021*, Volume 2840, pp. 51–60. CEUR-WS.org.

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

Li, Z., Y. Zhao, et J. Fu (2020). Sync : A copula based framework for generating synthetic data from aggregated sources.

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

Petricioli, L., L. Humski, M. Vranić, et D. Pintar (2020). Data set synthesis based on known correlations and distributions for expanded social graph generation. *IEEE Access 8*, 33013–33022.

Portet, S. (2020). A primer on model selection using the akaike information criterion. *Infectious Disease Modelling 5*, 111–128.

Salinas, D., M. Bohlke-Schneider, L. Callot, R. Medico, et J. Gasthaus (2019). High-dimensional multivariate forecasting with low-rank gaussian copula processes. In *Advances in NIPS*, pp. 6827–6837.

Zhu, H., L. Zhang, T. Xiao, et X. Li (2017). Generation of multivariate cross-correlated geotechnical random fields. *Computers and Geotechnics 86*, 95–107.

Summary

This paper is a short version of Benali et al. (2021). Most of the existing techniques work well for low-dimensional data and fail to capture complex dependencies between data dimensions. Moreover, identifying the right combination of models and their respective parameters is still an open problem. We present MTCopula, a novel flexible and extendable synthetic complex data generation approach that automatically chooses the best Copula model and the best-fitted marginals to catch the data complexity relying on Akaike Information Criterion.