À PROPOS DES TESTS DE L'HYPOTHÈSE SIMPLIFICATRICE POUR LES COPULES CONDITIONNELLES

Alexis Derumigny [1] & Jean-David Fermanian [2]

[1] *CREST-ENSAE, 3 avenue Pierre-Larousse, 92245 Malakoff cedex, France.*
alexis.derumigny@ensae.fr
[2] *CREST-ENSAE, J120, 3 avenue Pierre-Larousse, 92245 Malakoff cedex, France.*
jean-david.fermanian@ensae.fr. *This research has been supported by the Labex Ecodec.*

Résumé. Nous étudions "l'hypothèse simplificatrice" portant sur les copules conditionnelles dans un cadre général. Nous introduisons plusieurs tests de cette hypothèse pour des modèles de copules semi- et non-paramétriques. Nous proposons aussi des procédures de test proches basées sur des conditionnements par des ensembles plutôt que des conditionnements ponctuels. La distribution limite de telles statistiques de test sous l'hypothèse nulle est approchée par plusieurs schémas de ré-échantillonnage, dont la plupart sont nouveaux. Nous démontrons la validité d'un schéma particulier de ré-échantillonnage semi-paramétrique. Des simulations illustrent l'intérêt de nos résultats.

Mots-clés. Copule conditionnelle, hypothèse simplificatrice, ré-échantillonnage.

Abstract. We discuss the so-called "simplifying assumption" of conditional copulas in a general framework. We introduce several tests of the latter assumption for non- and semiparametric copula models. Some related test procedures based on conditioning subsets instead of point-wise events are proposed. The limiting distribution of such test statistics under the null are approximated by several bootstrap schemes, most of them being new. We prove the validity of a particular semiparametric bootstrap scheme. Some simulations illustrate the relevance of our results.

Keywords. Conditional copula, simplifying assumption, bootstrap.

1 Introduction

En statistique, il est très commun de distinguer deux sous-ensembles de variables : un vecteur aléatoire d'intérêt (aussi appelé variables expliquées), et un vecteur de covariables (variables explicatives). L'objectif est de prédire la loi du premier vecteur sachant le second. Formellement, soit $\mathbf{X}$ un vecteur aléatoire d-dimensionnel. On peut le décomposer en deux sous-vecteurs $\mathbf{X}_I$ et $\mathbf{X}_J$, tels que $\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_J)$, $I \cup J = \{1, \dots, d\}$, $I \cap J = \emptyset$, et nos modèles spécifient la loi conditionnelle de $\mathbf{X}_I$ sachant $\mathbf{X}_J = \mathbf{x}_J$ or sachant $\mathbf{X}_J \in A_J$ pour un ensemble mesurable $A_J \subset \mathbb{R}^{|J|}$. Nous utiliserons la notation standard pour les vecteurs : pour un ensemble d'indices I, $\mathbf{x}_I$ est le vecteur de dimension $|I|$ dont les composantes sont

les x_k, pour $k \in I$. Sans perdre de généralité, on suppose que I et J sont de la forme $I = \{1, \ldots, p\}$ et $J = \{p+1, \ldots, d\}$.

Par ailleurs, le problème de la dépendance entre les composantes d'un vecteur aléatoire de dimension d a été étudié de manière importantes dans de nombreux domaines. L'importance croissante des copules illustre le besoin en matière de modèles multivariés flexibles. Avec nos notations, lorsque des covariables sont présentes, il s'agit d'étudier la dépendance entre les composantes de $\mathbf{X}_I$ conditionnellement à $\mathbf{X}_J$. En conséquence, le concept de copule conditionnelle a été introduit par Patton (2006a, 2006b). Par définition, pour un ensemble borélien $A_J \subset \mathbb{R}^{d-p}$, la copule conditionnelle de $\mathbf{X}_I$ sachant $(\mathbf{X}_J \in A_J)$ est notée $C_{I|J}(\cdot|\mathbf{X}_J \in A_J)$. Il s'agit de la fonction de répartition du vecteur aléatoire $(F_{1|J}(X_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(X_p|\mathbf{X}_J \in A_J))$ conditionnellement à $(\mathbf{X}_J \in A_J)$. Ici, $F_{k|J}(\cdot|\mathbf{X}_J \in A_J)$ représente la loi conditionnelle de X_k sachant $\mathbf{X}_J \in A_J$, $k = 1, \ldots, p$. Dans ce qui suit, ces distributions conditionnelles seront supposées continues, impliquant l'existence et l'unicité de $C_{I|J}$ par le théorème de Sklar (1959). En d'autres termes, pour chaque $\mathbf{x}_I \in \mathbb{R}^p$, on a

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I|\mathbf{X}_J \in A_J\right) = C_{I|J}\left(F_{1|J}(x_1|\mathbf{X}_J \in A_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J \in A_J) \,|\, \mathbf{X}_J \in A_J\right).$$

En particulier, lorsque les évènements conditionnants sont des singletons, on obtient que la copule conditionnelle sachant $\mathbf{X}_J = \mathbf{x}_J$ est une fonction de répartition $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ sur $[0,1]^p$ telle que, pour tout $\mathbf{x}_I \in \mathbb{R}^p$,

$$\mathbb{P}\left(\mathbf{X}_I \leq \mathbf{x}_I|\mathbf{X}_J = \mathbf{x}_J\right) = C_{I|J}\left(F_{1|J}(x_1|\mathbf{X}_J = \mathbf{x}_J), \ldots, F_{p|J}(x_p|\mathbf{X}_J = \mathbf{x}_J) \,|\, \mathbf{X}_J = \mathbf{x}_J\right).$$

Le plus souvent, la dépendance de $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ par rapport à $\mathbf{x}_J$ est une source de complexité importante, en termes de spécification du modèle et d'inférence. Ainsi, la plupart des auteurs suppose l'hypothèse simplificatrice suivante :

Hypothèse $(\mathcal{SA})$: la copule conditionnelle $C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)$ ne dépend pas de $\mathbf{x}_J$, c'est-à-dire, pour chaque $\mathbf{u}_I \in [0,1]^p$, la fonction $\mathbf{x}_J \in \mathbb{R}^{d-p} \mapsto C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J)$ est une fonction constante (qui dépend de $\mathbf{u}_I$).

Sous l'hypothèse $(\mathcal{SA})$, on notera $C_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) =: C_{s,I|J}(\mathbf{u}_I)$. Cette égalité signifie que la dépendance en $\mathbf{X}_J$ parmi les composantes de $\mathbf{X}_I$ ne passe que par leurs marges conditionnelles. Nous remarquons que $C_{s,I|J}$ est différente a priori de la copule usuelle de $\mathbf{X}_I$.

En pratique, la spécification et l'estimation des modèles de copules conditionnelles sont loin d'être évidentes, en particulier lorsque les variables conditionnées et/ou conditionnantes sont nombreuses. L'hypothèse $(\mathcal{SA})$ est en particulier pertinente pour les modèles de "vines", étudiés par Acar et al. (2009), entres autres. En effet, pour construire les vines depuis un vecteur aléatoire $\mathbf{X}$ de dimension d, il est nécessaire de considérer des suites de copules bivariées conditionnelles $C_{I|J}$, où $I = \{i_1, i_2\}$ est un couple d'indice de $\{1, \ldots, d\}$, $J \subset \{1, \ldots, d\}$, $I \cap J = \emptyset$, et $(i_1, i_2|J)$ est un noeud du vine. En d'autres termes,

une copule bivariée conditionnelle est requise à chaque noeud du vine, et les tailles des ensembles conditionnants augmentent à chaque arbre. Sans hypothèse supplémentaire, la modélisation et l'estimation deviennent rapidement difficiles. C'est la raison pour laquelle la plupart des auteurs adoptent notre hypothèse simplificatrice ($\mathcal{SA}$) à chaque noeud du vine.

Néanmoins, l'hypothèse ($\mathcal{SA}$) peut sembler plutôt restrictive, même si elle peut être considérée comme acceptable pour des raisons pratiques. Le débat entre les défenseurs et les opposants de l'hypothèse simplificatrice reste encore ouvert. D'un côté, Hobæk-Haff et al. (2010) affirment que cette hypothèse est non seulement requise pour de l'inférence rapide, flexible , et robuste, mais qu'elle fournit aussi "une approximation plutôt bonne, même quand l'hypothèse simplificatrice est loin d'être vérifiée par le modèle réél". De l'autre côté, Acar et al. (2012) maintiennent que "cette vision est trop optimiste", et Spanhel et Kurz (2015) reconnaissent que "il est très improbable que le processus inconnu de génération des données satisfasse l'hypothèse simplificatrice dans un sens mathématique strict".

Ainsi, il y a un besoin pour des tests formels et universels de l'hypothèse simplificatrice. Il est probable que cette hypothèse soit acceptable dans certaines circonstances, alors qu'elle est trop brutale dans d'autres. Cela signifie que, pour des sous ensembles d'indices I et J donnés, on souhaite tester

$$\mathcal{H}_0 : C_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J) \text{ ne dépend pas de } \mathbf{x}_J,$$

contre l'hypothèse complémentaire. Dans la suite, nous proposerons plusieurs statistiques de test de $\mathcal{H}_0$, en supposant parfois que la copule conditionnelle appartient à une certaine famille paramétrique.

Tester $\mathcal{H}_0$ est étroitement lié au problème de copules à m échantillons, dans lequel nous avons m échantillons différents et indépendants d'une variable p-dimensionnelle $\mathbf{X}_I = (X_1, \ldots, X_p)$. Dans chaque échantillon, les observations sont i.i.d. avec leur propre lois marginales et leur propre copule $C_{I,k}$. Le problème de copules à m-échantillons consiste à tester si les m copules sont égales. On remarque qu'il est possible de fusionner tous les échantillons en un seul, et de créer les variables discrètes Y_i valant k lorsque i est dans le k-ième échantillon. Ainsi, le problème de copule à m échantillons est formellement équivalent au test de $\mathcal{H}_0$ avec la variables conditionnante $\mathbf{X}_J := Y$

Réciproquement, supposons que nous avons défini une partition $\{A_{1,J}, \ldots, A_{m,J}\}$ de $\mathbb{R}^{d-p}$ composée d'ensembles mesurables tels que $\mathbb{P}(\mathbf{X}_J \in A_{k,J}) > 0$ pour tout $k = 1, \ldots, m$, et que nous voulons tester

$$\bar{\mathcal{H}}_0 : k \in \{1, \ldots, m\} \mapsto C_{I|J}(\cdot\,|\mathbf{X}_J \in A_{k,J}) \text{ ne dépend pas de } k.$$

Alors on peut diviser l'échantillon en m sous-échantillons, chaque échantillon k rassemblant les observations telles que la variable conditionnante appartienne à $A_{k,J}$. Ainsi, $\bar{\mathcal{H}}_0$ est équivalente à un problème de copules à m échantillons.

Nous remarquons que $\bar{\mathcal{H}}_0$ ressemble à une conséquence de $\mathcal{H}_0$ alors que ce n'est pas le cas en général pour des $\mathbf{X}_J$ continus. Néanmoins, $\bar{\mathcal{H}}_0$ comporte la même intuition que $\mathcal{H}_0$. Puisqu'en pratique $\bar{\mathcal{H}}_0$ peut être réalisée beaucoup plus facilement (aucun lissage n'est alors nécessaire), certains chercheurs pourraient préférer cette hypothèse $\bar{\mathcal{H}}_0$ plutôt que l'hypothèse initiale $\mathcal{H}_0$.

2 Tests de $\mathcal{H}_0$

Une première idée naturelle est de construire un test de $\mathcal{H}_0$ basé sur une comparaison entre la copule conditionnelle $C_{I|J}$ estimée avec et sans l'hypothèse simplificatrice. Ces estimateurs seront appelés respectivement $\hat{C}_{s,I|J}$ et $\hat{C}_{I|J}$. Ainsi, en introduisant une certaine distance $\mathcal{D}$ entre copules conditionnelles, un test peut être basé sur la statistique $\mathcal{D}(\hat{C}_{I|J}, \hat{C}_{s,I|J})$. Nous pensons par exemple aux statistiques de test de type Kolmogorov-Smirnov

$$\mathcal{T}^0_{KS,n} := \|\hat{C}_{I|J} - \hat{C}_{s,I|J}\|_\infty = \sup_{\mathbf{u}_I \in [0,1]^p} \sup_{\mathbf{x}_J \in \mathbb{R}^{d-p}} |\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)|, \tag{1}$$

ou de type Cramer von-Mises

$$\mathcal{T}^0_{CvM,n} := \int \left(\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{x}_J) - \hat{C}_{s,I|J}(\mathbf{u}_I)\right)^2 w(d\mathbf{u}_I, d\mathbf{x}_J), \tag{2}$$

Pour évaluer $\hat{C}_{I|J}$, nous proposons d'utiliser l'estimateur non-paramétrique des copules conditionnelles introduit par Fermanian et Wegkamp (2012).

Ainsi, si l'on dispose d'un échantillon $(\mathbf{X}_i)_{i=1,\dots,n}$ i.i.d. de dimension d, notre estimateur de $C_{I|J}$ sera défini par

$$\hat{C}_{I|J}(\mathbf{u}_I|\mathbf{X}_J = \mathbf{x}_J) := \hat{F}_{I|J}\left(\hat{F}^-_{1|J}(u_1|\mathbf{X}_J = \mathbf{x}_J), \dots, \hat{F}^-_{p|J}(u_p|\mathbf{X}_J = \mathbf{x}_J) \,\middle|\, \mathbf{X}_J = \mathbf{x}_J\right),$$

où $\hat{F}^-_{i|J}$ est le pseudo-inverse de $\hat{F}_{i|J}$, pour $i = 1, \dots, p$; et $\hat{F}_{I|J}$ (respectivement $\hat{F}_{i|J}$) est l'estimateur par noyau de la fonction de répartition conditionnelle de $\mathbf{X}_I$ (respectivement X_i) sachant $\mathbf{X}_J$.

Pour l'estimateur $\hat{C}_{s,I|J}$, plusieurs choix sont possibles. On peut par exemple naïvement prendre un point $\mathbf{x}^*_J \in \mathbb{R}^{d-p}$ et poser $\hat{C}^{(1)}_{s,I|J}(\cdot) := \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}^*_J)$. Comme le choix de $\mathbf{x}^*_J$ est trop arbitraire, une alternative est de prendre l'estimateur

$$\hat{C}^{(2)}_{s,I|J}(\cdot) := \int \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)\, w(d\mathbf{x}_J),$$

pour une fonction w à variations bornées telle que $\int w(d\mathbf{x}_J) = 1$. Malheureusement, ce choix nécessite une procédure d'intégration sur un espace de dimension $d-p$, ce qui peut devenir un problème numérique important quand $d-p$ est supérieur à 3.

Pour éviter les intégrations multiples, on peut aussi rendre aléatoire la fonction de poids w. Par exemple, en choisissant la fonction de répartition empirique des $\mathbf{X}_J$, on obtient

$$\hat{C}^{(3)}_{s,I|J}(\cdot) := \int \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{x}_J)\, \hat{F}_J(d\mathbf{x}_J) = \frac{1}{n}\sum_{i=1}^{n} \hat{C}_{I|J}(\cdot|\mathbf{X}_J = \mathbf{X}_{i,J}). \tag{3}$$

En fait, des tests plus subtils sont possibles, en utilisant une autre définition : $\mathcal{H}_0$ est équivalente à l'indépendance des vecteurs aléatoires $\mathbf{Z}_{I|J} := (F_1(X_1|\mathbf{X}_J), \ldots, F_p(X_p|\mathbf{X}_J))$ et $\mathbf{X}_J$. On peut donc utiliser des statistiques de tests inspirées des tests classiques d'indépendance. Ainsi, on introduit les pseudo-observations

$$\Big(\hat{F}_{1|J}(X_{i,1}|\mathbf{X}_{i,J}), \ldots, \hat{F}_{p|J}(X_{i,p}|\mathbf{X}_{i,J})\Big)_{i=1,\ldots,n} := (\hat{\mathbf{Z}}_{i,I|J})_{i=1,\ldots,n},$$

ainsi qu'une estimation de la loi jointe du couple $(\mathbf{Z}_{I|J}, \mathbf{X}_J)$, qui peut être définie par

$$\begin{aligned} G_{I,J}(\mathbf{x}_I, \mathbf{x}_J) &:= \mathbb{P}(\mathbf{Z}_{I|J} \leq \mathbf{x}_I, \mathbf{X}_J \leq \mathbf{x}_J) \\ &\simeq \quad \hat{G}_{I,J}(\mathbf{x}) := n^{-1}\sum_{i=1}^{n} \mathbf{1}(\hat{\mathbf{Z}}_{i,I|J} \leq \mathbf{x}_I, \mathbf{X}_{i,J} \leq \mathbf{x}_J). \end{aligned}$$

À partir de là, on peut utiliser des statistiques de test d'indépendance comme celle du chi-deux. Soient $B_1, \ldots, B_N$ (resp. $A_1, \ldots, A_m$) des sous-ensembles de $\mathbb{R}^p$ (resp. $\mathbb{R}^{d-p}$). Nous définissons alors la statistique de test comme

$$\mathcal{I}_{\chi,n} = n\sum_{k=1}^{N}\sum_{l=1}^{m} \frac{\Big(\hat{G}_{I,J}(B_k \times A_l) - \hat{G}_{I,J}(B_k \times \mathbb{R}^{d-p})\hat{G}_{I,J}(\mathbb{R}^p \times A_l)\Big)^2}{\hat{G}_{I,J}(B_k \times \mathbb{R}^{d-p})\hat{G}_{I,J}(\mathbb{R}^p \times A_l)}.$$

Il est également possible de faire des tests de l'hypothèse simplificatrice dans un cadre semi-paramétrique. Supposons que le modèle est paramétré de la manière suivante : pour tout $\mathbf{x}_J$, il existe $\theta(\mathbf{x}_J)$ tel que $C_{I|J}(\cdot|\mathbf{x}_J) = C_{\theta(\mathbf{x}_J)}(\cdot)$, où $\mathcal{C} := \{C_\theta, \theta \in \Theta \subset \mathbb{R}^m\}$ est une famille paramétrique de copules. Notre problème est alors réduit au test du caractère constant de la fonction $\theta(\cdot)$. Autrement dit, on veut tester

$$\mathcal{H}_0^c : \text{la fonction } \mathbf{x}_J \mapsto \theta(\mathbf{x}_J) \text{ est une constante, notée } \theta_0.$$

On peut alors construire des statistiques de test basées sur la différence entre un estimateur $\hat{\theta}(\cdot)$ de $\theta(\cdot)$ et un estimateur $\hat{\theta}_0$ de θ_0 :

$$\mathcal{T}^c_\infty := \sup_{\mathbf{x}_J \in \mathbb{R}^{d-p}} \|\hat{\theta}(\mathbf{x}_J) - \hat{\theta}_0\|, \text{ ou } \mathcal{T}^c_2 := \int \|\hat{\theta}(\mathbf{x}_J) - \hat{\theta}_0\|^2 \omega(\mathbf{x}_J)\, d\mathbf{x}_J,$$

pour une fonction de poids ω.

Dans tous ces cas, les lois asymptotiques des statistiques de test sont difficiles à évaluer, et nous avons développé des procédures de ré-échantillonnage adaptées pour estimer des p-valeurs.

Néanmoins, le calcul de toutes ces statistiques de test dans les cadres semi- et nonparamétrique précédents nécessite une fenêtre de lissage h, dont le choix peut être difficile. C'est la raison pour laquelle il peut être intéressant de tester $\bar{\mathcal{H}}_0$ plutôt que $\mathcal{H}_0$ en introduisant des conditionnement par des ensembles. Dans ce cadre, toutes les statistiques de tests précédentes peuvent être adaptées, avec des schémas de ré-échantillonnage correspondants.

Bibliographie

[1] Aas, K., Czado, C., Frigessi, A., et Bakken, H. (2009), Pair-copula constructions of multiple dependence, *Insurance : Mathematics and Economics*, 44(2), 182-198.
[2] Abegaz, F., Gijbels, I. et Veraverbeke, N. (2012), Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, 110, $43 - 73$.
[3] Acar, E.F., Craiu, R.V. et Yao, F. (2011), Dependence calibration in conditional copulas : a nonparametric approach. *Biometrics*, 67, 445-453.
[4] Derumigny, A., et Fermanian, J.-D. (2016), *About tests of the "simplifying" assumption for conditional copulas*, Prépublication ArXiv :1612.07349.
[5] Fermanian, J.-D. et Wegkamp, M. (2012), Time-dependent copulas, *Journal of Multivariate Analysis*, 110, 19-29.
[6] Fermanian, J.-D. et Lopez, O. (2015), Single-index copulas, *Working paper Crest* 2015-12.
[7] Genest, C., Rémillard, B. et Beaudoin, A. (2009), Goodness-of-fit tests for copulas : A review and a power study, *Insurance : Mathematics and Economics*, 44, 199-213
[8] Gijbels, I., Omelka, M., et Veraverbeke, N. (2016), Nonparametric testing for no covariate effects in conditional copulas, *Statistics*, 1-35.
[9] Hobæk Haff, I., Aas, K. et Frigessi, A. (2010), On the simplified pair-copula construction–simply useful or too simplistic ? *Journal of Multivariate Analysis*, 101 1296–1310.
[10] Nagler, T. et Czado, C. (2016), *Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas*, Prépublication ArXiv :1503.03305
[11] Patton, A. (2006a), Modelling asymmetric exchange rate dependence, *International Economic Review*, 47, 527-556.
[12] Patton, A. (2006b), Estimation of multivariate models for time series of possibly different lengths, *Journal of Applied Econometrics*, 21, 147-173.
[13] Sklar, M. (1959), *Fonctions de répartition à n dimensions et leurs marges*, Université Paris 8,
[14] Spanhel, F., et Kurz, M.S. (2015), *Simplified vine copula models : Approximations based on the simplifying assumption*, Prépublication ArXiv :1510-06971.