Gulf Journal of Mathematics
Vol 20, Issue 1 (2025) 1-21

ISSN: 2309-4966
https://doi.org/10.56947/gjom.v20i.2857

ROBUST GOODNESS-OF-FIT TEST FOR COPULAS BY THE MAXIMUM MEAN DISCREPANCY ESTIMATOR

N'DRI HUBERT BIAN[1*], SERGE-HIPPOLYTE ARNAUD KANGA[1], GUEÏ CYRILLE OKOU[2] and OUAGNINA HILI[1]

Abstract. This paper discusses the robustness of suitability tests using the empirical copula process. It compares the empirical copula with a parametric estimate of the copula obtained under the null hypothesis. Fitting tests using Kendall inversion and maximum likelihood estimation methods may be less effective, especially at the sight of outliers. We recommend to establish a goodness-of-fit procedure with the Maximum Mean Discrepancy (MMD) estimator, a robust estimator, in order to assess the empirical size and power of the test under several families of alternative copulas. Then a comparative large-scale simulation study on contaminated and uncontaminated data samples is conducted to analyze the power and also the robustness of test. This approach using the MMD estimator may offer a valid alternative for goodness-of-fit tests. It is also robust in the presence of perturbations and outliers in the observations of joint distributions.

1. Introduction

Assume a random vector $\mathbf{Y} = (Y_1, ..., Y_d)$ with continuous marginal distribution functions $\underline{F}_1, ..., \underline{F}_d$. In his work, [27] proved that the joint distribution function $\bar{H}$ of $\mathbf{Y}$ can be defined by

$$\bar{H}(\mathbf{y}) = \underline{C}\left\{\underline{F}_1(y_1), ..., \underline{F}_d(y_d)\right\}, \quad \mathbf{y} = (y_1, ..., y_d) \in \mathbb{R}^d, \tag{1.1}$$

where the *copula* $\underline{C} : [0,1]^d \longrightarrow [0,1]$, is a multivariate distribution function with uniform margins. Copula functions contain all the necessary information about the dependence structure that exists between the components of the random vector. In addition, they represent a tool for modeling dependency between variables over time [28]. Representation (1.1) allows statisticians to model marginal behavior and the dependency pattern represented by the joint distribution function separately. Copulas have also become more and more widely used to model multivariate distributions in a wide diversity of applications : finance [10, 23], hydrology [16, 17], actuarial sciences [14].

Here, we assume that the unknown copula $\underline{C}$ belongs to a family of parametric copulas $\mathcal{B} = \{\underline{C}_\theta : \theta \in \mathcal{O} \subset \mathbb{R}^p\}$, p non null integer. The next step is to estimate

Date: Received: Apr 11, 2025; Accepted: May 2, 2025.

* Corresponding author.

2020 *Mathematics Subject Classification.* 62F35; 62H05; 62H10; 62H15; 62H30.

Key words and phrases. maximum mean discrepancy estimator, robust goodness-of-fit test, empirical copula process, pseudo-observation, parametric bootstrap.

the parameter vector $\theta = (\theta_1, \ldots, \theta_p)$ from a random sample. For bivariate copulas with one real-valued parameter, a popular approach is to apply the method of moments based on the inversion of association measurements, that is, Spearman's rho or Kendall's tau. Again, as studied in [18, 21, 22], the pseudo-likelihood approach is a natural estimation method in the multi-parameter and multivariate case. However, both of these estimation methods have their disadvantages. Indeed, they lack robustness from a statistical point of view. Specifically, consider a perturbed copula of the form $\mathcal{L} = (1-\zeta)\underline{C} + \zeta\mathcal{D}$, where $\zeta \in [0;1]$ represents the contamination level and $\mathcal{D} \neq \underline{C}$ is a disturbing copula. Therefore, the semi-parametric inference procedure for copulas requires an efficient estimator that is consistent, robust to outliers, and can also be applied in the case of high dimensions.

To address this concern, [2] developed a semi-parametric inference approach for parametric copulas through mixture copulas. They proved its robustness in the presence of outliers. The procedure may be used in the event of model misspecification. This method would then be independent of the particular choice of models. Thus, this minimum distance estimator is based on the distance of the maximum mean discrepancy from two arbitrary probability distributions. It should also be noted that other distances are well known in the literature to induce robustness, for example total variation distance [33], Hellinger distance [4]. However, the estimation methods proposed in these articles are not explicit for the calculations.

A crucial question that is presently the subject of active investigation is whether the hypothesis $\underline{C} \in \mathcal{B} = \{\underline{C}_\theta : \theta \in \mathcal{O} \subset \mathbb{R}^p\}$, on which the copula parameter estimation step discussed above is relied upon is really valid or not. More specifically, we want to test

$$\mathbb{H}_0 : \underline{C} \in \mathcal{B} \text{ versus } \mathbb{H}_1 : \underline{C} \notin \mathcal{B}$$

A fairly large number of goodness of fit tests have been developed in the literature. Among others, [20, 5, 7] have shown the performance of fit tests for several families of copulas, even copulas with singular components, through simulations. However, to test $\mathbb{H}_0$ the marginal distributions are handled as unknown nuisance parameters and are replaced by the pseudo-observations. This procedure illustrates that better fit tests can also be obtained from the empirical copula process

$$\underline{\mathbb{C}}_n(\mathbf{t}) \;=\; \sqrt{n}\big\{\underline{C}_n(\mathbf{t}) - \underline{C}_{\hat{\theta}_n}(\mathbf{t})\big\}, \quad \mathbf{t} = (t_1, ..., t_d) \in [0,1]^d, \tag{1.2}$$

where $\underline{C}_n$ is the empirical copula [11] of the data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ generally defined by

$$\underline{C}_n(\mathbf{t}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\mathbf{1}\big(\hat{T}_{ij} \leq t_j\big) \;\; \mathbf{t} \in [0;1]^d, \tag{1.3}$$

where $\mathbf{1}(.)$ is the indicator function, which is equal to 1 if the statement in parentheses is true and zero otherwise.

In addition, $\underline{C}_{\hat{\theta}_n}$ is an estimate of $\underline{C}_\theta$ achieved by supposing that $\mathbb{H}_0 : \underline{C} \in \mathcal{B}$ holds and the random vector $\big(\hat{T}_{i1}, ..., \hat{T}_{id}\big)$ are the pseudo-observations computed from $\mathbf{Y}_i$ by $\hat{T}_{ij} = R_{ij}/(n+1)$, with R_{ij} the rank of Y_{ij} among $Y_{1j}, ..., Y_{nj}$. This

estimate is derived from a $\hat{\theta}_n$ estimator of θ such as the MMD estimator described above. As mentioned in [20], the Cramér-von Mises statistic defined by

$$\bar{\mathcal{S}}_n = n \int_{[0,1]^d} \left\{ \underline{C}_n(\mathbf{t}) - \underline{C}_{\hat{\theta}_n}(\mathbf{t}) \right\}^2 d\underline{C}_n(\mathbf{t}) \tag{1.4}$$

yielded the best results on the whole. It is also an omnibus statistic, that is, one that can be applied to all copula functions and involves no strategic choice in its application. The convergence of this test statistic is proved by [19] under certain suitable regularity conditions on the prespecified parametric family $\underline{C}_\theta$ and the sequence of estimators $\hat{\theta}_n$. They also establish that the test based on $\bar{\mathcal{S}}_n$ is consistent, that is, if $\underline{C} \notin \mathcal{B}$, then as $n \to \infty$, the probability of rejecting $\mathbb{H}_0$ when $\mathbb{H}_0$ is false, tend to 1.
In reality, the limiting distribution of $\bar{\mathcal{S}}_n$ relies on the family of copulas under the null hypothesis, as well as on the value of the unknown parameter θ. Therefore, the asymptotic relationship of the test statistic can only be tabulated and estimated p-values for this statistic can be derived using a parametric bootstrap procedure, validated by [19].

The choice of dependence parameter estimation method in implementing of this parametric bootstrap procedure is a relevant factor for fit tests using the empirical copula process. [6] have shown that the estimation approach can significantly influence the power of the Cramér-von Mises statistic. Their ability to detect deviations from $\mathbb{H}_0$ is more interesting in the case of fixed alternatives than in the case of mixtures. Simulation studies also carried out by [20] and [5] proved that goodness-of-fit tests are well suited to selecting the best parametric copula. But, these studies are generally based on undisturbed data that are simulated using a pre-specified parametric copula model. Extending these results from simulated data to the case of real financial market data, a study by [30] indicates that the copula adequacy tests investigated no longer identify the best parametric copula. The author suggests that this lack of power may in fact be due to a failure in the robustness of fit tests in the face of data contamination, generally by outliers and measurement errors. It is therefore necessary to know, which goodness-of-fit test for copulas is robust to such data contamination ? To remedy this, [31, 32] has introduced goodness-of-fit tests for archimedean and elliptical copulas based on the exclusion of outliers after their identification within the framework of mixture copulas. Although this approach made it easier for the various goodness-of-fit tests to maintain their nominal level, the power of these tests is considerably reduced with addition of a second disturbing parametric copula, and also after the exclusion of outliers. A likely explanation for this loss of test power is that excluding outliers reduces the average sample size.
Testing the null hypothesis involves first estimating the copula parameters by a consistent and robust estimator. The objective in this field is to investigate the finite sample performance of the goodness-of-fit test for copulas with the consistent and robust maximum mean discrepancy estimator. This study also examines the robustness of test for absolutely continuous copulas to data contamination at outliers. Fitting test may be better suited to identifying a true underlying copula

that is perturbed by the combination of another parametric copula through a mixing copula.

The rest of the document is structured as follows. In the second section, we present the MMD estimator based on the Gaussian kernel and its asymptotic representation. The third section provides the theoretical tools to develop the matching test using the empirical process of the underlying copula and the parametric bootstrap procedure. Section 4 firstly introduces simulation results to analyze nominal level and power of the test for four exchangeable copula families. Next, it describes the influence of outliers on the efficiency of the goodness-of-fit test using mixture copulas. Section 5 concludes.

2. MAXIMUM MEAN DISCREPANCY ESTIMATOR

In this part, we present the maximum mean discrepancy estimator. We define the reproducing kernel Hilbert functional space related to the estimator. We also recall some theories on the minimum distance estimator using the maximum mean discrepancy. The maximum mean discrepancy is defined in terms of specific functional spaces attesting to the difference between the distributions [9, 13].

2.1. Maximum mean discrepancy.

Let $\phi : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ be a positive definite kernel function on $\mathcal{T}$, and assume the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\phi$ associated with the kernel ϕ, equipped with inner product $\langle ., . \rangle_{\mathcal{H}_\phi}$ and norm $\|.\|_{\mathcal{H}_\phi}$. Let $\mathcal{F}$ denote the set of Borel probability measures $\mathbb{P}$ such that $\int_{\mathcal{T}} \sqrt{\phi(y,y)}\mathbb{P}(dy) < \infty$. We now recall the concept of kernel mean embedding, a Hilbert space embedding that maps probability distributions into the RKHS $\mathcal{H}_\phi$. Consider $\Gamma_{\mathbb{P}} \in \mathcal{H}_\phi$, the kernel mean embedding defined by the formula

$$\Gamma_{\mathbb{P}} \quad = \quad \mathbb{E}_{\mathbb{P}}[\phi(Y, .)] = \int_{\mathcal{T}} \phi(y, .)\mathbb{P}(dy), \tag{2.1}$$

for a probability measure $\mathbb{P}$.
Applications and the theoretical features of those embedding have been extensively discussed in [24]. The mean embedding fulfils the formula
$\mathbb{E}_{\mathbb{P}}[f(Y)] = \langle f, \Gamma_{\mathbb{P}} \rangle_{\mathcal{H}_\phi}$ for all function $f \in \mathcal{H}_\phi$ and reduces the metric on $\mathcal{H}_\phi$ generated by the inner product to define a pseudo-metric on $\mathcal{F}$ called the maximum mean discrepancy $\mathcal{M} : \mathcal{F} \times \mathcal{F} \to [0, +\infty[$, that is,

$$\mathcal{M}(\mathbb{P}, \mathbb{Q}) \quad = \quad \|\Gamma_{\mathbb{P}} - \Gamma_{\mathbb{Q}}\|_{\mathcal{H}_\phi} \tag{2.2}$$

$$\mathcal{M}(\mathbb{P}, \mathbb{Q}) \quad = \quad \left\| \int_{\mathcal{T}} \phi(y, .)\mathbb{P}(dy) - \int_{\mathcal{T}} \phi(y, .)\mathbb{Q}(dy) \right\|_{\mathcal{H}_\phi} \tag{2.3}$$

The squared $\mathcal{M}$ has a particularly simple term that can be obtained by applying of the reproducing property $(f(y) = \langle f, \phi(y, .) \rangle_{\mathcal{H}_\phi})$:

$$\begin{aligned}\mathcal{M}^2(\mathbb{P},\mathbb{Q}) &:= \left\|\int_{\mathcal{T}}\phi(y,.)\mathbb{P}(dy)-\int_{\mathcal{T}}\phi(y,.)\mathbb{Q}(dy)\right\|^2_{\mathcal{H}_\phi}\\ &= \int_{\mathcal{T}}\int_{\mathcal{T}}\phi(y,w)\mathbb{P}(dy)\mathbb{P}(dw)+\int_{\mathcal{T}}\int_{\mathcal{T}}\phi(y,w)\mathbb{Q}(dy)\mathbb{Q}(dw)\\ &\quad -2\int_{\mathcal{T}}\int_{\mathcal{T}}\phi(y,w)\mathbb{P}(dy)\mathbb{Q}(dw),\end{aligned}$$

thus giving a closed form statement until the calculation of expectations. MMD is actually an integral probability pseudo-metric as it may be identified as :

$$\mathcal{M}(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}_\phi}\leq 1}\left|\int_{\mathcal{T}}f(y)\mathbb{P}(dy)-\int_{\mathcal{T}}f(y)\mathbb{Q}(dy)\right|.$$

If $P\mapsto\Gamma_{\mathbb{P}}$ is injective then the kernel ϕ is said to be characteristic. This allows MMD to be a metric, not just a pseudo-metric.

2.2. **Robust maximum mean discrepancy estimator.** Given a family of parametric copulas $\mathcal{B}$, we research the best fitting copula within this family. Assume $\mathbf{Y}_1,...,\mathbf{Y}_n$ is an i.i.d. random sample from distribution function $\underline{C}_\theta(\mathbf{T})$, $\mathbf{T}=(T_1,...,T_d)$ where $T_1=\underline{F}_1,...,T_d=\underline{F}_d$ are continuous margins and $\underline{C}_\theta$ is an absolutely continuous copula. Within the framework of copula inference, the marginal distribution function remains unknown. The copula are therefore not directly observable. It is convenient to specify a pseudo-observation $(\hat{\mathbf{T}}_i)_{i=1,...,n}$, where $\hat{\mathbf{T}}_i:=(\hat{T}_{i1},\ldots,\hat{T}_{id})$ and

$$\hat{T}_{ij}:=\frac{n}{n+1}\underline{F}_{nj}(Y_{ij}),\ \ \underline{F}_{nj}(t)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_{ij}\leq t),$$

for every $(i,j)\in\{1,....,n\}\times\{1,...,d\}$ and $t\in\mathbb{R}$. To compute an estimator de θ, the MMD criterion to be minimized is then $\mathcal{M}(\mathbb{P}_\theta,\mathbb{P}_n)=\|\Gamma_{\mathbb{P}_\theta}-\Gamma_{\mathbb{P}_n}\|_{\mathcal{H}_\phi}$, for some $\mathcal{H}$, Reproducing Kernel Hilbert Space to which a kernel ϕ is attached, where $\mathbb{P}_\theta$ the law driven by $\underline{C}_\theta$ on the hypercube $\mathcal{T}:=[0,1]^d$ and $\mathbb{P}_n=(1/n)\sum_{i=1}^{n}\delta_{\hat{\mathbf{T}}_i}$ represents the empirical measure related to pseudo-observations $(\hat{\mathbf{T}}_i)_{i=1,...,n}$. Then, our estimator $\hat{\theta}_n$ of the copula parameter θ can be expressed as follows

$$\hat{\theta}_n = \arg\min_{\theta\in\mathcal{O}}\mathcal{M}^2(\mathbb{P}_\theta,\mathbb{P}_n) \tag{2.4}$$

$$\hat{\theta}_n = \arg\min_{\theta\in\mathcal{O}}\int_{\mathcal{T}}\int_{\mathcal{T}}\phi(\mathbf{t},\mathbf{v})\mathbb{P}_\theta(d\mathbf{t})\mathbb{P}_\theta(d\mathbf{v})-\frac{2}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi(\mathbf{t},\hat{\mathbf{T}}_i)\mathbb{P}_\theta(d\mathbf{t}). \tag{2.5}$$

Denoting $\underline{c}_\theta$, the density of the copula $\underline{C}_\theta$ relative to the Lebesgue measure, the above expression can be established as follows :

$$\hat{\theta}_n=\arg\min_{\theta\in\mathcal{O}}\int_{\mathcal{T}}\int_{\mathcal{T}}\phi(\mathbf{t},\mathbf{v})\underline{c}_\theta(\mathbf{t})\underline{c}_\theta(\mathbf{v})d\mathbf{t}d\mathbf{v}-\frac{2}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi(\mathbf{t},\hat{\mathbf{T}}_i)\underline{c}_\theta(\mathbf{t})d\mathbf{t}. \tag{2.6}$$

Through this last expression, the estimator $\hat{\theta}_n$ depends on the kernel, so its choice is a crucial issue. The choice of our kernel is based on the Gaussian kernels $\phi(\mathbf{t},\mathbf{v}) := \exp(-\|g(\mathbf{t}) - g(\mathbf{v})\|_2^2/\gamma^2)$ providing excellent results (where g is the identity application or reciprocal bijection of the distribution function of a standard gaussian random variable).
The asymptotic normality of the resulting estimator $\hat{\theta}_n$ is established by [2]. Using the regularity conditions of these authors, we illustrate the limit representation Θ of $\Theta_n = \sqrt{n}\big(\hat{\theta}_n - \theta\big)$ (for more details, see [2]).
Consider φ_θ, measurable and $\mathbb{P}-$integrable functions for all $\theta \in \mathcal{O}$, defined as follows :

$$\varphi_\theta(\boldsymbol{z}) = \int \phi(\mathbf{t},\mathbf{v})\mathbb{P}_\theta(d\mathbf{t})\mathbb{P}_\theta(d\mathbf{v}) - 2\int \phi(\mathbf{t},\boldsymbol{z})\mathbb{P}_\theta(d\mathbf{t}),$$

Then the theoretical loss function and its empirical expression are respectively of the following form :

$$\psi(\theta) = \int_{[0,1]^d} \varphi_\theta(\boldsymbol{z})\mathbb{P}(d\boldsymbol{z}), \quad \psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\theta(\mathbf{T}_i) = \int_{[0,1]^d} \varphi_\theta(\boldsymbol{z})\mathbb{P}_n(d\boldsymbol{z}).$$

Let $\{\varphi_\theta(.); \theta \in \mathcal{O}\}$ be a class of measurable functions and for all $z \in [0,1]^d$, the application $\theta \mapsto \varphi_\theta(z)$ is continuous on the compact parameter space $\mathcal{O}$.

Assuming that the estimator $\hat{\theta}_n$ is a zero of the loss function $\psi_n(\theta)$ and expand $\nabla_\theta\psi_n$ using the Taylor expansion, we have :

$$0 = \nabla_\theta\psi_n(\hat{\theta}_n) = \nabla_\theta\psi_n(\theta) + \nabla_{\theta,\theta}\psi_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta) \tag{2.7}$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_n$ and θ. Hence, we have :

$$\hat{\theta}_n - \theta = -\Big[\nabla_{\theta,\theta}\psi_n(\tilde{\theta}_n)\Big]^{-1} \times \nabla_\theta\psi_n(\theta) \tag{2.8}$$

Let's denote $M_n = \nabla_{\theta,\theta}\psi_n(\tilde{\theta}_n)$ a d-dimensional Hessian matrix. By definition of M_n and $A = \mathbb{E}[\nabla^2_{\theta,\theta}\varphi_\theta(\mathbf{T})]$, we obtain $M_n \xrightarrow{\mathbb{P}-a.s} A$ as $n \to +\infty$.
The function $\nabla_\theta\varphi_\theta(.)$ being continuous on the right and in addition of bounded variations, with integration by part, we obtain :

$$\begin{aligned}
\sqrt{n}\left\{\nabla_\theta\psi_n(\theta) - \mathbb{E}[\nabla_\theta\varphi_\theta(\mathbf{T})]\right\} &= \sqrt{n}\left\{\int_{[0,1]^d}\nabla_\theta\varphi_\theta(\boldsymbol{z})\mathbb{P}_n(d\boldsymbol{z}) - \int_{[0,1]^d}\nabla_\theta\varphi_\theta(\boldsymbol{z})\mathbb{P}(d\boldsymbol{z})\right\} \\
&= \sqrt{n}\int_{[0,1]^d}\nabla_\theta\varphi_\theta(\boldsymbol{z})(\mathbb{P}_n - \mathbb{P})(d\boldsymbol{z}) \\
&= (-1)^d\int_{[0,1]^d}\sqrt{n}\nabla_\theta\varphi_\theta(d\boldsymbol{z})(\mathbb{P}_n - \mathbb{P})(\boldsymbol{z})
\end{aligned}$$

With condition 7 (see [2]), the continuous application theorem et the weak convergence of the empirical process $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ to $\eta(\boldsymbol{z}) - \sum_{k=1}^{d}\dot{C}_k(\boldsymbol{z})\eta_k(\boldsymbol{z})$ (see Proposition 3.1 in [26] and under condition 9 see [2]). We deduce that the weak

limit of $\sqrt{n}\nabla_\theta\psi_n(\theta)$ exists, is centered and Gaussian, namely;

$$\int \left\{ \eta(\boldsymbol{z}) - \sum_{k=1}^{d} \dot{\underline{C}}_k(\boldsymbol{z})\eta_k(\boldsymbol{z}) \right\} \nabla_\theta\varphi_\theta(d\boldsymbol{z}).$$

Applying the formula (2.8), we get :

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta) &= M_n^{-1}\sqrt{n}\nabla_\theta\psi_n(\theta) + o_p(1) \\ &= M_n^{-1}\int_{[0,1]^d} \sqrt{n}\nabla_\theta\varphi_\theta(d\boldsymbol{z})(\mathbb{P}_n - \mathbb{P})(\boldsymbol{z}) + o_p(1) \end{aligned}$$

Hence, using Slutsky lemma and the fact that the limit matrix A is invertible, the limit law Θ of Θ_n is :

$$\Theta = A^{-1}\int \left\{ \eta(\boldsymbol{z}) - \sum_{k=1}^{d} \dot{\underline{C}}_k(\boldsymbol{z})\eta_k(\boldsymbol{z}) \right\} \nabla_\theta\varphi_\theta(d\boldsymbol{z}).$$

3. GOODNESS-OF-FIT TEST PROCEDURE

3.1. **Asymptotic results.** This section introduces a test procedure based on a non-parametric consistent estimation of a copula by the empirical copula. The empirical copula is defined by [11] as a distribution function of the sample of normalised ranks. He proved the low consistency of the empirical process to a centered gaussian limit under the independence conditions, that is in the specific case as $\underline{C}_\theta(a,b) = ab$. This outcome has been applied in the context of a general distribution by [15], [12] and [29]. An idea emitted by [12] and explored by [25] and [19] is to base a goodness-of-fit test on a edited copy of $\underline{\mathbb{C}}_{n,\theta}$, namely $\underline{\mathbb{C}}_n = \sqrt{n}\big(\underline{C}_n - \underline{C}_{\hat{\theta}_n}\big)$, where $\hat{\theta}_n$ is the consistent estimator of θ. Under the three natural assumptions listed below, the asymptotic study of the empirical copula process $\underline{\mathbb{C}}_n$ has been carried out in [25, 6].

$\tilde{\mathcal{A}}_1$ **:** For any $\theta \in \mathcal{O}$, the first order partial derivatives of $\underline{C}_\theta$ exist and are continuous.

$\tilde{\mathcal{A}}_2$ **:** For any $\theta \in \mathcal{O}$, $\Theta_n = \sqrt{n}\big(\hat{\theta}_n - \theta\big)$ converge to Θ.

$\tilde{\mathcal{A}}_3$ **:** For any $\theta \in \mathcal{O}$,

$$\sup_{\|\theta^* - \theta\| < \varepsilon} \sup_{\mathbf{t}\in[0,1]^d} |\dot{\underline{C}}_{\theta^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \longrightarrow 0,$$

as $\varepsilon \to 0$, where

$$\dot{\underline{C}}_\theta(\mathbf{t}) = \nabla_\theta\underline{C}_\theta = \left(\frac{\partial\underline{C}_\theta(\mathbf{t})}{\partial\theta_1}, ..., \frac{\partial\underline{C}_\theta(\mathbf{t})}{\partial\theta_p} \right)^\top, \quad \mathbf{t} \in [0,1]^d$$

.

Proposition 3.1. *Under $\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2$ and $\tilde{\mathcal{A}}_3$, the goodness-of-fit process $\underline{\mathbb{C}}_n = \sqrt{n}(\underline{C}_n - \underline{C}_{\hat{\theta}_n})$ converges weakly in $D([0,1]^d)$ to the tight centered Gaussian process*

$$\underline{\mathbb{C}}(\mathbf{t}) = \underline{\mathbb{C}}_\theta(\mathbf{t}) - \dot{\underline{C}}_\theta^\top(\mathbf{t})\Theta, \quad \mathbf{t} \in [0,1]^d, \tag{3.1}$$

$$\underline{\mathbb{C}}_\theta(\mathbf{t}) = \eta_\theta(\mathbf{t}) - \sum_{j=1}^{d} \frac{\partial \underline{C}_\theta}{\partial t_j}(\mathbf{t})\eta_\theta(1,...,1,t_j,1,...,1),$$

and where η_θ is a tight centered gaussian process on $[0,1]^d$ with covariance function

$$\mathbb{E}[\eta_\theta(\mathbf{t})\eta_\theta(\mathbf{v})] = \underline{C}_\theta(\mathbf{t}\wedge\mathbf{v}) - \underline{C}_\theta(\mathbf{t})\underline{C}_\theta(\mathbf{v}), \quad \mathbf{t},\mathbf{v}\in[0,1]^d. \tag{3.2}$$

Proof. Let us consider the following decomposition of the empirical process

$$\underline{\mathbb{C}}_n = \sqrt{n}\big(\underline{C}_n - \underline{C}_{\hat{\theta}_n}\big) = \sqrt{n}\big(\underline{C}_n - \underline{C}_\theta\big) + \sqrt{n}\big(\underline{C}_\theta - \underline{C}_{\hat{\theta}_n}\big) = \sqrt{n}\big(\underline{C}_n - \underline{C}_\theta\big) - \sqrt{n}\big(\underline{C}_{\hat{\theta}_n} - \underline{C}_\theta\big).$$

By putting $\underline{\mathbb{C}}_n = \underline{\mathbb{C}}_{n,\theta} - \mathcal{Z}_{n,\theta}$, this means $\underline{\mathbb{C}}_{n,\theta} = \sqrt{n}\big(\underline{C}_n - \underline{C}_\theta\big)$ and $\mathcal{Z}_{n,\theta} = \sqrt{n}\big(\underline{C}_{\hat{\theta}_n} - \underline{C}_\theta\big)$.
As assumption $\tilde{\mathcal{A}}_1$ is satisfied, the empirical process $\underline{\mathbb{C}}_{n,\theta}$ converges to $\underline{\mathbb{C}}_\theta$.
Next, let us prove that $\mathcal{Z}_{n,\theta}$ is close to $\dot{\underline{C}}_\theta^\top\hat{\theta}_n$ for n sufficiently large. It is then necessary to prove that $|\underline{C}_{\hat{\theta}_n} - \underline{C}_\theta|$ converges in probability to 0. By applying the mean-value theorem there exists a θ^* such that $\|\theta_n^* - \theta\| \le \|\hat{\theta}_n - \theta\|$ and

$$\begin{aligned} |\underline{C}_{\hat{\theta}_n}(\mathbf{t}) - \underline{C}_\theta(\mathbf{t})| &= \|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t})| \\ &\le \|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})|| + \|\hat{\theta}_n - \theta\||\dot{\underline{C}}_\theta(\mathbf{t})| \end{aligned}$$

Since assumption $\tilde{\mathcal{A}}_2$ implies that $\Theta_n := \sqrt{n}\big(\hat{\theta}_n - \theta\big)$ converges in distribution to Θ, under Slutsky's theorem the secondary term converges of inequality to 0 in probability. For the first term, design for all $\beta > 0$ the existence of an $M = M_\beta$ so that $\mathbb{P}(\|\Theta_n\| \ge M) < \beta$ (for the sequence Θ_n is tight). In addition, we have

$$\begin{aligned} \mathbb{P}\left\{\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha\right\} &= \mathbb{P}\left\{\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha, \|\Theta_n\| < M\right\} \\ &\quad + \mathbb{P}\left\{\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha, \|\Theta_n\| \ge M\right\}. \end{aligned}$$

Or

$$\mathbb{P}\left\{\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha, \|\Theta_n\| \ge M\right\} \le \mathbb{P}\left\{\|\Theta_n\| \ge M\right\}.$$

Then

$$\begin{aligned} \mathbb{P}\left\{\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha\right\} &\le \mathbb{P}\left\{\sup_{\mathbf{t}\in[0,1]^d} |\|\hat{\theta}_n - \theta\||\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| \ge \alpha, \|\Theta_n\| < M\right\} \\ &\quad + \mathbb{P}(\|\Theta_n\| \ge M) \\ &\le \mathbb{P}\left\{\sup_{\|\theta_n^* - \theta\| \le N/\sqrt{n}} \sup_{\mathbf{t}\in[0,1]^d} |\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_\theta(\mathbf{t})| > \frac{\alpha}{M}\right\} + \beta, \end{aligned}$$

where the latter inequality is caused by the inclusion of the matching events and from the proper choice of the constant M depending on β.

Since $\beta > 0$ may be chosen arbitrarily small, and applying the assumption $\tilde{\mathcal{A}}_3$, we have :

$$\lim_{n\to+\infty} \mathbb{P}\left\{ \sup_{\|\theta_n^*-\theta\|\leq M/\sqrt{n}} \sup_{\mathbf{t}\in[0,1]^d} |\dot{\underline{C}}_{\theta_n^*}(\mathbf{t}) - \dot{\underline{C}}_{\theta}(\mathbf{t})| > \frac{\alpha}{M} \right\} = 0.$$

□

3.2. **Test statistic and parametric bootstrap.** As the shape of the statistic (1.4) is not an explicit expression in common use, through [19], it is feasible to extend it from a parametric bootstrap procedure. It can be presented as follows.

$$\bar{\mathcal{S}}_{n,m} = \int_{[0,1]^d} \underline{\mathbb{C}}^2_{n,m}(\mathbf{t}) d\mathbf{t}, \tag{3.3}$$

where $\underline{\mathbb{C}}_{n,m} = \sqrt{n}(\underline{C}_n - \tilde{\underline{C}}_m)$ and $\tilde{\underline{C}}_m$ is the empirical copula calculated by equation (1.3) from an artificial sample of $\underline{C}_{\hat{\theta}_n}$ in order to approximate it. The Cramér-von Mises test statistic was selected due to its full good performance in the large-scale simulation studies of [5] and [20].
In the bivariate case, we obtain the expression of the Cramér-von Mises statistic, the tool for setting up a parametric bootstrap procedure. It is characterized by the following relationship.

$$\begin{aligned} \bar{\mathcal{S}}_{n,m} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\prod_{l=1}^{2}\left[1-\hat{T}_{il}\vee\hat{T}_{kl}\right] + \frac{n}{m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}\prod_{l=1}^{2}\left[1-\hat{V}_{jl}\vee\hat{V}_{kl}\right] \\ &\quad -\frac{2}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\prod_{l=1}^{2}\left[1-\hat{T}_{il}\vee\hat{V}_{jl}\right], \end{aligned} \tag{3.4}$$

where $p \vee q = \max(p, q)$.

The approximate p-value of the test employing Cramér-von Mises statistic will be obtained by applying the parametric bootstrap method described below. This one or two level procedure for testing goodness of fit can be gleaned from the work of [19], in which the authors proved the algorithm's validity. It Thus is a computationally intensive technique that involves generating random numbers of the assumed copula and estimating the copula parameters. However, it is a generic re-sampling technique, very simple to implement, employed to perform goodness-of-fit tests based on (1.2) in a multi-variate framework [3], [19]. For some large integer N and with an estimator $\hat{\theta}_n$ of θ based on $\mathbf{Y}_1, ..., \mathbf{Y}_n$, the parametric bootstrap consist in iterating the steps for all $k \in \{1, ..., N\}$.

To carry out the above task for a chosen family family of copulas $\underline{C}_\theta$, the dependence parameter must be estimated from the available data. We then note the use of the MMD estimator, implemented in R package (MMDCopula [1]).

4. NUMERICAL EXPERIMENTS

In this section, we illustrate the performance of the simulation-based test. We consider one dataset with outliers and another without outliers.

Algorithm 1 Approximation of test p-value

(i) : Produce a random sample $\mathbf{Y}_1^{(k)}, ..., \mathbf{Y}_n^{(k)}$ from the null hypothesis copula $\underline{C}_{\hat{\theta}_n}$.

(ii) : Let $\underline{C}_{\hat{\theta}_n}^{(k)}$ and $\hat{\theta}_n^{(k)}$ denote the version of $\underline{C}_{\hat{\theta}_n}$ and $\hat{\theta}_n$ estimated from the random sample $\mathbf{Y}_1^{(k)}, ..., \mathbf{Y}_n^{(k)}$.

(iii) : Construct an approximate expression for the test statistic under $\mathbb{H}_0$ as $\bar{\mathcal{S}}_{n,m}^{(k)} = n \int_{[0,1]^d} \left\{\underline{C}_n^{(k)}(\mathbf{t}) - \tilde{\underline{C}}_m^{(k)}(\mathbf{t})\right\}^2 d\mathbf{t}$, where $\tilde{\underline{C}}_m^{(k)}$, the approximation of $\underline{C}_{\hat{\theta}_n}^{(k)}$.

(iv) : Assuming that large values of test statistic (3.3) lead to rejection of $\mathbb{H}_0$, an approximate p-value for the test is finally obtained by $N^{-1} \sum_{k=1}^{N} \mathbf{1}\big(\bar{\mathcal{S}}_{n,m}^{(k)} \geq \bar{\mathcal{S}}_{n,m}\big)$.

4.1. **Study of power.** The most fundamental issue concerns the probability of first order error : does the test maintain its nominal size ? One interesting property is the test's ability to approximate the nominal level or size, usually chosen at 5%. The aim of this section is to demonstrate the ability of our test to maintain its nominal level and power under several alternatives.

If a test has a rejection probability lower than or equal to the significance level $\alpha \in (0, 1)$, then such a test is called an α-level test. Common values for α are 0.05 and 0.01. As a test of level α is also a test of level $\alpha' > \alpha$, the smallest of these α is called the test size and stands for the maximum probability of first type error [8]. Power of a test is the probability of rejecting the hypothesis $\mathbb{H}_0$ with reason, that is, when the alternative hypothesis is "true". Next, we examine the power of the test in relation to various alternatives. In terms of distributions, the four families of 2-dimensional one-parameter copulas considered for the study are absolutely continuous cumulative distribution functions, namely the Gaussian, Clayton, Gumbel and Frank distributions. Our simulations are performed on the basis of the following elements :

- Assumed exchangeable bivariate copula model under $\mathbb{H}_0$ (4 choices : Gaussian, Clayton, Gumbel, Frank),
- 2-dimensional copula model from which the data were generated (4 choices : Gaussian, Clayton, Gumbel, Frank),
- Level of dependency, from Kendall's tau : $\tau \in \{0.25; 0.50; 0.75; 0.85\}$,
- Size of each sample drawn from copula : $n \in \{100; 250; 500\}$,
- Nominal level of the test through the study : $\alpha = 5\%$.

For every test combination, 10000 replicates have been carried out to estimate the empirical level or power of all the tests considered. In other terms, 10000 random samples are generated and then employed to estimate the rejection percentages of the four null hypotheses considered. The estimated rejection rates of the null hypothesis in the test for a family of Gaussian, Clayton, Frank and

Gumbel copulas are recorded respectively in tables 1, 2, 3 et 4 for sample sizes 100, 250 and 500. In addition, empirical levels are summarized in italics.

In the light of the values of the various tables concerning empirical levels, we can state that all tests are overall satisfactory for maintaining their prescribed nominal level of 5%. There is a good trade-off between the probability of empirical and theoretical error of the first type in the various cases. We notice this as the sample size and the value of τ increase. But, this is not entirely true for certain Kendall's tau, for example with $\tau = 0.85$, whose approach remains overly conservative in certain circumstances.
All the copula models under consideration throughout the investigation have the independence copula and the Fréchet-Hoeffding upper bound, $M(a, b) = \min(a, b)$ as their limiting copulas at $\tau = 0$ and $\tau = 1$ respectively. This is justified by the results of the various tables, namely, as the value of τ changes from 0 to 1, the power of the test grows, steadies and, in some instance, decreases.
For each case, when the sample size becomes significant, our test has reasonably good power to reveal departures from the null hypothesis when the true copula is not the same. In addition, the power is also acceptable in several test combinations, with a value of at least 80%. In a nutshell, we can see that the performance of our various tests improves with the sample size, as it should in order to ensure the consistency of the approach.

In table 1, when testing the Gaussian hypothesis against the alternatives, Clayton and Gumbel having non-zero lower and upper tail dependence coefficients respectively, the approach does not fare as well for the smallest size. For example, for $n = 100$, we have 29.78% at $\tau = 0.25$ and 36.08% at $\tau = 0.75$ for the Clayton alternative and for the Gumbel alternative 69.89% at $\tau = 0.25$. For example, in Table 2, we note that when the Clayton copula is tested on a random sample of size $n = 250$, the rejection rate of the null hypothesis is around 95.31%, when the data are generated from the Gaussian copula with $\tau = 0.25$. We also notice in this table that the powers are higher than for testing the Gaussian, Frank and Gumbel hypotheses. This illustrates that it is easier to test the Clayton hypothesis. table 3 provides the rates of rejection of the Gumbel hypothesis. We can also see that our approach achieves very good performance. However, we find low proportions, namely $09.70\%, 04.11\%, 22\%$ when the data are drawn from Gaussian and Frank copulas respectively. As dependency increases, the test remains powerful in detecting deviations from the Gumbel copula. The results in table 4 illustrate the usefulness of our estimator in terms the good performance of our approach with a significant rejection rate, that is, above 80%. But low power, for size $n = 100$, comes from Gaussian copulas with moderate dependency. This is explained by the fact that these copulas are similar when the dependencies are moderate.

In summary, the results presented in the various tables show that the method chosen to estimate the copula family parameter for the null hypothesis has a significant impact on the performance of the goodness-of-fit test procedure. Overall, the performance of our approach with the MMD estimator is satisfactory for moderate and large sample sizes that evolve with the level of dependence. The

empirical level and power of the test are also appreciable. So, this approach may provide an alternative for fit testing.

TABLE 1. Empirical level and power of test at significance threshold $\alpha = 5\%$: percentage of rejecting, as estimated from 10000 replications, of the null hypothesis *bivariate Gaussian copula* under four copula models and four levels of dependance.

alternative	**sample**	**Kendall tau** τ			
copula	**size** n	0.25	0.50	0.75	0.85
Gaussian	100	*0.08*	*2.48*	*5.00*	*0.48*
	250	*1.89*	*5.00*	*5.00*	*0.93*
	500	*4.92*	*5.00*	*5.00*	*0.11*
Clayton	100	29.78	85.42	36.08	99.00
	250	97.88	100.00	99.49	99.66
	500	100.00	100.00	99.97	99.96
Gumbel	100	69.89	96.91	85.90	99.96
	250	98.79	97.90	96.65	99.85
	500	80.76	97.97	99.86	99.94
Frank	100	72.20	83.53	95.12	99.97
	250	95.47	95.52	97.89	100.00
	500	99.10	95.84	97.74	99.99

4.2. **Robustness to perturbation of the data.** In this part, we provide a simulation study on the robustness of fit test using the empirical process of gaussian, clayton, gumbel and frank copulas. We are looking at the test's ability to approximate its nominal level, as well as its performance to identify alternatives in the presence of contaminated data.

In a nutshell, we wish to answer the question of whether the fitting test with the robust MMD estimator is likely to maintain its performance with a data sample where we note the presence of outliers. We then enjoy the disturbance effect caused by the mixture of observations coming of a parametric copula with the original structure. To evaluate the power and nominal level of test, we identify four kinds of outliers. They come from :

- Gaussian copula with a Kendall's tau equal to 0.2;
- Clayton copula with a Kendall's tau equal to 0.5;
- Frank copula with a Kendall's tau equal to -0.8;
- Frank copula with a Kendall's tau equal to 0.5.

TABLE 2. Empirical level and power of test at significance threshold $\alpha = 5\%$: percentage of rejecting, as estimated from 10000 replications, of the null hypothesis *bivariate Clayton copula* under four copula models and four levels of dependance.

alternative	**sample**	**Kendall tau** τ			
copula	**size** n	0.25	0.50	0.75	0.85
Gaussian	100	89.09	97.36	92.20	100.00
	250	95.31	100.00	99.66	99.53
	500	98.79	99.95	100.00	99.90
Clayton	100	*5.00*	*5.00*	*0.17*	*5.00*
	250	*5.00*	*5.00*	*5.00*	*0.04*
	500	*5.00*	*5.00*	*5.00*	*0.32*
Gumbel	100	95.76	100.00	99.56	99.96
	250	98.88	99.96	99.70	100.00
	500	100.00	100.00	100.00	100.00
Frank	100	98.41	99.27	97.98	100.00
	250	97.80	100.00	99.23	100.00
	500	99.97	100.00	100.00	100.00

TABLE 3. Empirical level and power of test at significance threshold $\alpha = 5\%$: percentage of rejecting, as estimated from 10000 replications, of the null hypothesis *bivariate Gumbel copula* under four copula models and four levels of dependance.

alternative	**sample**	**Kendall tau** τ			
copula	**size** n	0.25	0.50	0.75	0.85
Gaussian	100	09.70	53.04	100.00	98.80
	250	99.93	93.93	84.85	99.98
	500	86.84	98.95	98.69	100.00
Clayton	100	96.45	99.99	89.94	100.00
	250	99.98	100.00	99.75	100.00
	500	100.00	100.00	100.00	100.00
Gumbel	100	*5.00*	*0.04*	*5.00*	*0.00*
	250	*2.64*	*5.00*	*5.00*	*0.00*
	500	*5.00*	*5.00*	*5.00*	*5.00*
Frank	100	04.11	22.00	83.46	100.00
	250	79.75	79.07	97.94	100.00
	500	98.00	99.60	99.96	100.00

TABLE 4. Empirical level and power of test at significance threshold $\alpha = 5\%$: percentage of rejecting, as estimated from 10000 replications, of the null hypothesis *bivariate Frank copula* under four copula models and four levels of dependance.

alternative	**sample**	**Kendall tau τ**			
copula	**size n**	0.25	0.50	0.75	0.85
Gaussian	100	68.80	32.46	60.80	97.45
	250	82.88	84.91	83.63	99.84
	500	86.93	99.09	97.15	99.96
Clayton	100	96.03	99.97	100.00	99.58
	250	99.58	99.87	97.07	100.00
	500	99.82	100.00	97.24	100.00
Gumbel	100	81.69	99.71	99.88	99.99
	250	96.00	99.89	79.19	100.00
	500	89.81	96.06	93.06	100.00
Frank	100	*5.00*	*5.00*	*0.57*	*5.00*
	250	*5.00*	*5.00*	*5.00*	*0.67*
	500	*5.00*	*5.00*	*5.00*	*0.24*

We thus considered perturbed models in which a quantity of observations has been substituted by random observations generated from another parametric copula. We set the proportions equal to $\zeta = 5\%$ and $\zeta = 10\%$ outliers. As bivariate mixing copulas, we propose the following models :

- $\mathcal{L}_{GF} = (1-\zeta)\underline{C}_{Gaussian} + \zeta\mathcal{D}_{Frank}$;
- $\mathcal{L}_{CF} = (1-\zeta)\underline{C}_{Clayton} + \zeta\mathcal{D}_{Frank}$;
- $\mathcal{L}_{GG} = (1-\zeta)\underline{C}_{Gumbel} + \zeta\mathcal{D}_{Gaussian}$;
- $\mathcal{L}_{GC} = (1-\zeta)\underline{C}_{Gumbel} + \zeta\mathcal{D}_{Clayton}$.

The data is selected from the desired copula and sampled on $[0;1]^2$. In order to encompass a wide field of dependence structure, our four choices of mixing copulas are focused on mixing various types of tail dependence.
To highlight the effect of the mixture two parametric bivariate copulas, scatter plots of four mixing copulas are shown in Figures 1 and 2 (sample size $n = 500$, contamination proportions $\zeta = 5\%$ and $\zeta = 10\%$ respectively for the perturbing bivariate copula).

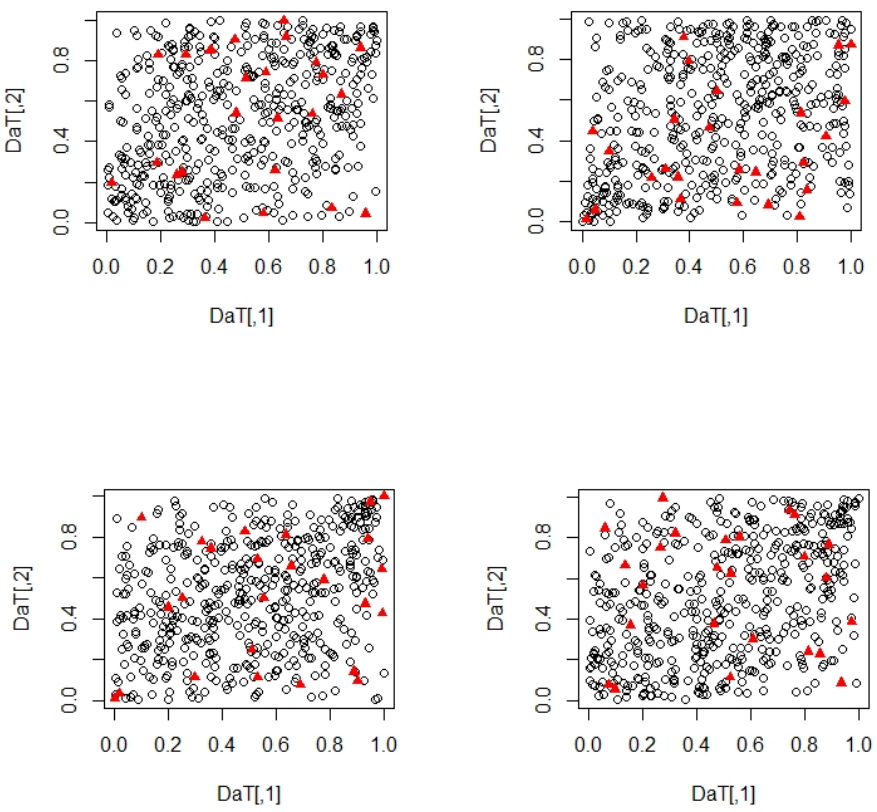

FIGURE 1. Scatter diagrams of random samples of size 500 from four bivariate mixture copulas. Data from the raw sample (the "true" copula, $\tau = 0.25$) are indicated in black (circle), while data samples drawn from the parametric copula perturbing the mixture (proportion $\zeta = 5\%$) are indicated in red color (triangles). The mixture distribution is composed of Gaussian copula perturbed by Frank copula ($\tau = -0.8$, upper left plot), Clayton copula perturbed by Frank copula ($\tau = 0.5$, upper right plot), Gumbel copula perturbed respectively by Gaussian distribution ($\tau = 0.2$, lower left plot) and Clayton copula ($\tau = 0.5$, lower right plot).

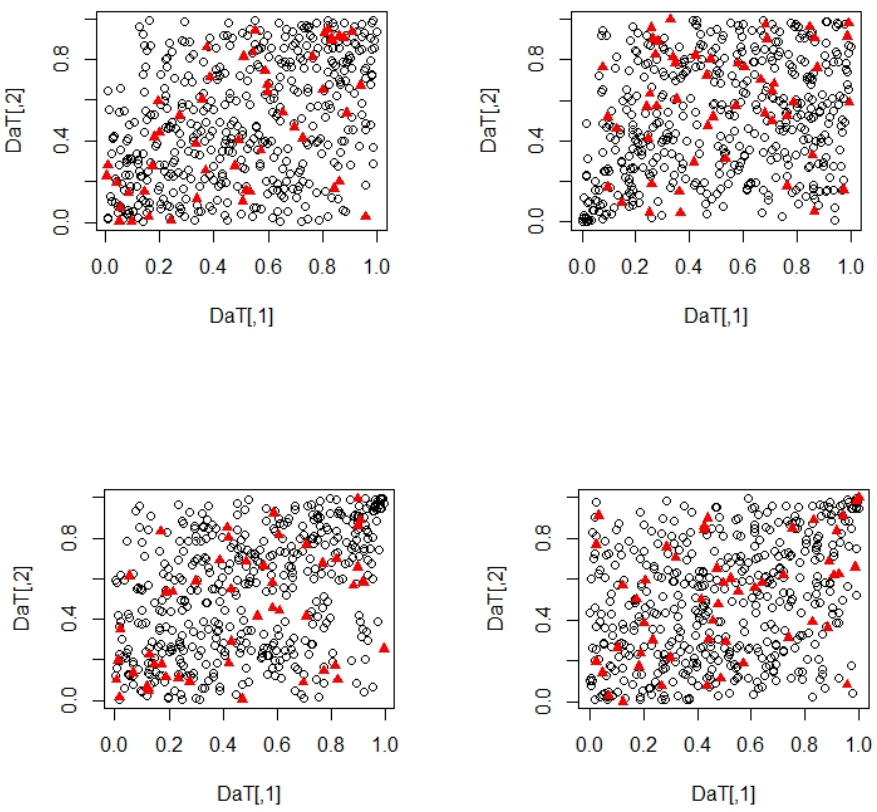

FIGURE 2. Scatter diagrams of random samples of size 500 from four bivariate mixture copulas. Data from the raw sample (the "true" copula, $\tau = 0.25$) are indicated in black (circle), while data samples drawn from the parametric copula perturbing the mixture (proportion $\zeta = 10\%$) are shown in red color (triangles). The mixture distribution is composed of a Gaussian distribution perturbed by Frank copula ($\tau = -0.8$, upper left plot), Clayton copula perturbed by Frank copula ($\tau = 0.5$, upper right plot), Gumbel copula perturbed respectively by Gaussian distribution ($\tau = 0.2$, lower left plot) and Clayton copula ($\tau = 0.5$, lower right plot).

In this simulation study, four families of absolutely continuous bivariate copulas are chosen for the null hypothesis, namely the Gaussian, Clayton, Gumbel and Frank copula, as well as under different alternative mixture copulas. Our procedure has been investigated for three sample sizes, $n = 100$, $n = 250$ and $n = 500$ in order to assess the improvement of the power of the goodness-of-fit test with increasing sample size. In addition, the approach is implemented with a choice of

copula parameters under the null hypothesis that corresponds to Kendall's tau, $\tau = 0.25$, and with various options of perturbing copula parameters according to $\tau \in \{-0.8; 0.2; 0.5\}$.

The values of the numerical experiment on the power of the adequacy test for parametric 2-dimensional mixture copulas and different contamination rates are summarized in Tables 5 and 6. In table 5, the results for copulas for the null hypothesis with the mixing parameter $\zeta = 5\%$ of the perturbation copula and the three different contaminated sample sizes are reported. Table 6 provides the percentages of the different copulas with the null hypothesis for an outlier proportion $\zeta = 10\%$. As we can observe in both tables, the test maintains its nominal level for the different sample sizes. Rejection rates remain mostly lower up to 0.01% for the principal component of the composition.

Concerning the influence of data perturbations, the tables 5 and 6 ($\zeta = 5\%$, $\zeta = 10\%$) indicate that the test performance is relatively significant for all choices of mixture copulas. A few exceptions occur in relation to the smaller sample size $n = 100$ (e.g. rejection rates of $8.80\%; 42.74\%$ and 37.01% respectively for

TABLE 5. Empirical level and power of test at significance threshold $\alpha = 5\%$: rate of rejecting as estimated from 10000 replications, of the null hypothesis for the contaminated data samples. The mixture distribution was parameterized with $\tau = 0.25$ while the parameters of the disruptive copula were chosen according to $\tau \in \{-0.8; 0.2; 0.5\}$ and $\zeta = 5\%$.

	Copula	$\zeta = 5\%$		
True copula	under $\mathcal{H}_0$	$n = 100$	$n = 250$	$n = 500$
Gaussian/Frank	Gaussian	**4.53**	**0.84**	**5.00**
	Clayton	66.41	99.53	99.68
	Gumbel	63.20	67.43	98.30
	Frank	86.73	88.50	60.15
Clayton/Frank	Gaussian	97.15	98.63	99.56
	Clayton	**5.00**	**5.00**	**5.00**
	Gumbel	89.93	99.68	99.97
	Frank	96.50	99.19	99.61
Gumbel/Gaussian	Gaussian	97.55	98.33	99.95
	Clayton	97.85	99.57	100
	Gumbel	**5.00**	**5.00**	**02.63**
	Frank	96.8	92.74	90.30
Gumbel/Clayton	Gaussian	88.21	98.29	100
	Clayton	84.26	97.91	99.80
	Gumbel	**04.19**	**01.09**	**5.00**
	Frank	08.80	60.93	95.71

TABLE 6. Empirical level and power of test at significance threshold $\alpha = 5\%$: rate of rejecting as estimated from 10000 replications, of the null hypothesis for the contaminated data samples. The mixture distribution was parameterized with $\tau = 0.25$ while the parameters of the disruptive distribution were chosen according to $\tau \in \{-0.8; 0.2; 0.5\}$ and $\zeta = 10\%$.

	Copula	$\zeta = 10\%$		
True copula	under $\mathcal{H}_0$	$n = 100$	$n = 250$	$n = 500$
Gaussian/Frank	Gaussian	**0.04**	**0.66**	**5.00**
	Clayton	86.01	86.70	99.36
	Gumbel	37.01	69.28	99.73
	Frank	61.81	98.64	86.16
Clayton/Frank	Gaussian	33.33	95.56	99.97
	Clayton	**5.00**	**5.00**	**4.96**
	Gumbel	23.76	98.68	100
	Frank	55.79	99.96	100
Gumbel/Gaussian	Gaussian	85.21	95.44	97.90
	Clayton	55.17	99.25	99.98
	Gumbel	**5.00**	**5.00**	**5.00**
	Frank	89.95	90.73	94.58
Gumbel/Clayton	Gaussian	89.88	97.09	98.91
	Clayton	98.62	99.91	99.97
	Gumbel	**0.01**	**5.00**	**5.00**
	Frank	42.74	84.04	99.62

the Gumbel/Clayton mixture and Frank's copula as copula in $\mathcal{H}_0$ and for the Gaussian/Frank mixture and in $\mathcal{H}_0$ the Gumbel copula).

The proportion of rejection increases as the sample size becomes larger, observed for the uncontaminated data and also with contamination rates from 5% to 10%. We can see that the rejection percentage for the majority of cases is greater than 80%, proving the power of the approach with the MMD estimator. Furthermore, compared with the mixture copula studies of [31, 32], the power of the test is comparatively high. This perfectly illustrates the robustness of the goodness-of-fit test with the robust MMD estimator in the presence of outliers. Finally, the presence of outliers in the distributions had no effect on the power or empirical level of the adequacy test based on the MMD estimator, a likely alternative in many situations.

5. CONCLUSION

In this study, we proposed the use of a robust minimum distance estimator, the MMD estimator, and the Cramér-von Mises statistic on the basis of the empirical

copula to adequately identify a parametric exchangeable copula. After stating the asymptotic results of the procedure, we illustrated with the second-level parametric bootstrap procedure the capacity of test to approximate its empirical level and power as sample size increases. The results on the comparative studies of the powers of the contaminated and non-contaminated samples (0%; 5% and 10%) showed the stability in the proportions of rejection of the null hypothesis. In a word, the robustness of the test of fitting the exchangeable parametric copulas to the contaminations of the data was achieved through the MMD estimator. In sum, this fitting approach could be a valid alternative for an adequate method for selecting copula models in the presence of outliers.

References

1. Alquier P., Chérief-Abdellatif B. E., Derumigny A. and Fermanian J.-D. (2020). Package r : 'mmdcopula'. In CM Cuadras, J Fortiana, and JA Rodríguez Lallena, editors, Distributions With Given Marginals and Statistical Modelling. https://github.com/AlexisDerumigny/MMDCopula.
2. Alquier P., Chérief-Abdellatif B. E., Derumigny A. and Fermanian J.-D. (2022). Estimation of copulas via maximum mean discrepancy. Journal of the American Statistical Association pp 1997–2012. https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2024836.
3. Babu G. J. and Rao C. R. (2004). Goodness-of-fit tests when parameters are estimated. The Indian Journal of Statistics pp 63–74 https://www.jstor.org/stable/25053332
4. Baraud Y., Birgé L. and Sart M. (2017). A new method for estimation and model selection: $\rho-$estimation. Invent math 207(2):425–517. http://link.springer.com/10.1007/s00222-016-0673-5.
5. Berg D. (2009). Copula goodness-of-fit testing: an overview and power comparison. The European Journal of Finance 15(7-8) : 675–701. http://www.tandfonline.com/doi/abs/10.1080/13518470802697428.
6. Berg D. and Quessy J. F. (2009). Local sensitivity analyses of goodness-of-fit tests for copulas. Scandinavian Journal of Statistics, 36 : 389 - 412. https://doi.org/10.1111/j.1467-9469.2009.00643.x.
7. Bian N. H., Hili O. and Okou G. C. (2021). A goodness-of-fit test based on kendall's process : Durante bivariate's copula models. Afrika Statistika 16 (3):2851 – 2882. http://dx.doi.org/10.16929/as/2021.2851.187
8. Bickel P. J. and Doksum K. A. (2007). Mathematical statistics: Basic ideas and selected topics (second ed). Pearson Prentice Hall, Upper Saddle River 1 https://doi.org/10.1201/9781315369266
9. Borgwardt K. M., Gretton A., Rasch M. J. and Smola A. (2006). Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22 : 49–57. https://doi/10.1093/bioinformatics/btl242
10. Cherubini U., Luciano E. and Vecchiator W. (2004). Copula methods in finance. Wiley finance series, John Wiley & Sons, Hoboken, NJ. https://DOI:10.1002/9781118673331
11. Deheuvels P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Bulletin de la Classe des sciences 65(1) : 274–292. https://www.persee.fr/doc/barb_0001-4141_1979_num_65_1_58521.
12. Fermanian J. D. (2005). Goodness-of-fit tests for copulas. Journal of Multivariate Analysis 95(1):119–152. https://doi.org/10.1016/j.jmva.2004.07.004.
13. Fortet R. and Mourier E. (1953). Convergence de la répartition empirique vers la répartition théorique. Ann Sci école Norm Sup 70(3):267–285. DOI:10.24033/asens.1013, http://www.numdam.org/item?id=ASENS_1953_3_70_3_267_0.

14. Frees E. W. and Valdez E. A. (1998). Understanding relationships using copulas. North American Actuarial Journal 2(1):1–25. http://www.tandfonline.com/doi/abs/10.1080/10920277.1998.10595667
15. Gaenssler P. and Stute W. (1987). Seminar on Empirical Processes, DMV Seminar, vol 9. Birkhäuser Basel, Basel http://link.springer.com/10.1007/978-3-0348-6269-1.
16. Genest C. and Favre A. C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. Journal of Hydrologic Engineering 12(4) : 347–368 https://ascelibrary.org/doi/10.1061/%28ASCE%291084-0699%282007%2912%3A4%28347%29
17. Genest C., Favre A. C., Béliveau J., and Jacques C. (2007). Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data: Metaelliptical copulas and frequency analysis. Water Resources Research 43(9). http://doi.wiley.com/10.1029/2006WR005275
18. Genest C., Ghoudi K. and Rivest L. P. (1995). A semiparametrie estimation procedure of dependence parameters in multivariate families of distributions. Biometrika p 10 https://doi.org/10.2307/2337532
19. Genest C. and Rémillard B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 44 1096-1127. https://doi/10.1214/07-AIHP148
20. Genest C., Rémillard B. and Beaudoin D. (2009). Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics 44(2):199–213. https://linkinghub.elsevier.com/retrieve/pii/S0167668707001205
21. Genest C. and Werker B. J. M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In CM Cuadras, J Fortiana, and JA Rodríguez Lallena, editors,Distributions With Given Marginals and Statistical Modelling pp 103–112. https://doi.org/10.1007/978-94-017-0061-0_12
22. Kim G., Silvapulle M. J. and Silvapulle P. (2007) Comparison of semiparametric and parametric methods for estimating copulas. Computational Statistics & Data Analysis 51(6):2836–2850. https://doi.org/10.1016/j.csda.2006.10.009.
23. McNeil A. J., Frey R. and Embrechts P. (2005). Quantitative risk management: concepts, techniques and tools. Princeton series in finance, Princeton University Press, Princeton, NJ, oCLC: ocm60796246.
24. Muandet K., Fukumizui K. and Sriperumbudur B. (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. Foundations and Trends®in Machine Learning 10(1-2):1–141. http://www.nowpublishers.com/article/Details/MAL-060.
25. Quessy J.-F. (2005). Méthodologie et application des copules: tests d'adéquation, tests d'indépendance et bornes pour la valeur-à-risque. Pro-Quest LLC, Ann Arbor, MI, thesis (Ph.D.)-Université Laval (Canada).
26. Segers J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. Bernoulli 18:764–782. https://doi.org/10.3150/11-BEJ387
27. Sklar A. (1959). Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris 8:229–231. https://books.google.ci/books?id=nreSmAEACAAJ
28. Tiemtoré H. and Bagré R. G. (2025). Dynamics of dependency : finite difference solutions to the black-scholes PDE with copulas. Gulf Journal of Mathematics 19(1) : 374-391. https://doi.org/10.56947/gjom.v19i1.2576
29. Tsukahara H. (2005). Semiparametric estimation in copula models. Canadian Journal of Statistics 33(3):357–375. https://onlinelibrary.wiley.com/doi/10.1002/cjs.5540330304.
30. Weiß G. (2011). Are Copula-GoF-tests of any practical use? Empirical evidence for stocks, commodities and FX futures. The Quarterly Review of Economics and Finance 51(2):173–188 https://linkinghub.elsevier.com/retrieve/pii/S1062976910000748

31. Weiß G. (2011). On the robustness of goodness-of-fit tests for copulas. SSRN Journal p 44 DOI:10.2139/ssrn.1927881
32. Weiß G. (2013). Identifying mixture copula components using outlier detection methods and goodness-of-fit test. Journal of Risk, Forthcoming, Available at SSRN p 48 http://dx.doi.org/10.2139/ssrn.1927881
33. Yatracos Y. G. (1985). Rates of convergence of minimum distance estimators and kolmogorov's entropy. The Annals of Statustics pp 768–774. https://doi.org/10.1214/aos/1176349553

[1] Laboratoire de Statistique, Probabilités et Recherche Opérationnelle (LASPRO), Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro, Côte d'Ivoire
Email address: hubertbn@yahoo.fr serge.kanga@inphb.ci o_hili@yahoo.fr

[2] Université Jean-Lorougnon-Guédé, Daloa, Côte d'Ivoire
Email address: okou.guei.cyrille@gmail.com