Estimation of copulas via Maximum Mean Discrepancy

Pierre Alquier [*], Badr-Eddine Chérief-Abdellatif [†], Alexis Derumigny [‡], and Jean-David Fermanian[§]

October 2, 2020

Abstract

This paper deals with robust inference for parametric copula models. Estimation using Canonical Maximum Likelihood might be unstable, especially in the presence of outliers. We propose to use a procedure based on the Maximum Mean Discrepancy (MMD) principle. We derive non-asymptotic oracle inequalities, consistency and asymptotic normality of this new estimator. In particular, the oracle inequality holds without any assumption on the copula family, and can be applied in the presence of outliers or under misspecification. Moreover, in our MMD framework, the statistical inference of copula models for which there exists no density with respect to the Lebesgue measure on $[0,1]^d$, as the Marshall-Olkin copula, becomes feasible. A simulation study shows the robustness of our new procedures, especially compared to pseudo-maximum likelihood estimation. An R package implementing the MMD estimator for copula models is available.

[*]RIKEN AIP, Nihonbashi 1-chome Mitsui Building (15th floor), 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan. Email: pierrealain.alquier@riken.jp

[†]Department of Statistics, University of Oxford. 24-29 St Giles' Oxford OX1 3LB, United Kingdom. Email: badr.eddine.cherief.abdellatif@gmail.com

[‡]University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science, Zilverling 4062, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: a.f.f.derumigny@utwente.nl

[§]CREST, ENSAE, Institut Polytechnique de Paris, 5 Avenue Le Chatelier, 91120 Palaiseau, France. Email: jean-david.fermanian@ensae.fr

1 Introduction

1.1 Context

Since the seminal work of Sklar [34], it is well known that every d-dimensional distribution F can be decomposed as $F(\boldsymbol{x}) = C\big(F_1(x_1), \ldots, F_d(x_d)\big)$, for all $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Here, $F_1, \ldots, F_d$ are the marginal distributions of F and C is a distribution on the unit cube $[0,1]^d$ with uniform margins, called a copula. This allows any statistician to split the complex problem of estimating a multivariate distribution into two simpler problems which are the estimation of the margins on one side, and of the copula on the other side. Copulas have become increasingly useful to model multivariate distributions in a wide variety of applications : finance, insurance, hydrology, engineering and so on. We refer to [26, 21] for a general introduction and a background on copula models.

In most cases, a copula of interest C belongs to a parametric family $\mathcal{C} := \{C_\theta, \theta \in \Theta \subset \mathbb{R}^p\}$ and one is interested in the estimation of the "true" value of the parameter θ. Typically, the goal is to evaluate the underlying copula only, without trying to specify the marginal distributions. In such a case, the most popular method for estimating parametric copula models is by Canonical Maximum Likelihood or CML, shorter ([17, 32]). This is a semi-parametric analog of Maximum Likelihood Estimation for copula models for which the margins are left unspecified and replaced by nonparametric counterparts. The method of moments is also a popular estimation technique, most often when $p = 1$, and is usually done by inversion of Kendall's tau or Spearman's rho. Both of these estimators have been implemented in the R package VineCopula [31] and attain the usual $\sqrt{n}$ rate of convergence as if the margins were known: see [35] for the asymptotic theory.

Nevertheless, both of these approaches suffer from drawbacks. In particular, they are not robust statistically speaking. For instance, assume that the true copula is slightly perturbed in the sense that $C := (1-\varepsilon)C_{\theta_0} + \varepsilon\tilde{C}$ for a small $\varepsilon > 0$ and a copula $\tilde{C} \neq C_{\theta_0}$. In general, there is no guarantee that the estimators obtained by CML or by the method of moments should be close to θ_0 when $\varepsilon \neq 0$, as pointed out in [20] for instance.

In the literature, there are very few attempts to build robust estimation methods for

semi-parametric copula models that would be "omnibus" (i.e. not dependent on some particular choices of models). Using Mahalanobis distances computed using robust estimates of covariance and location, [24] identified some points which seem not to follow the assumed dependence structure. Then, some copula parameters are obtained through the minimization of weighted goodness of fit statistics. In the semiparametric copula-based multivariate dynamic (SCOMDY) framework ([9]), [22] built a minimum density power divergence estimator which shows some resistance to some types of outliers. [14] proposed a parametric robust estimation method based on likelihood depth ([29]). Recently, [18] have considered robust and nonparametric estimation of the coefficient of tail dependence in presence of random covariates, that may be a way of estimating copulas for some particular models. Therefore, even if many estimators have been proposed for Huber contaminated models in general parametric cases, this has not been the case for semiparametric copula models yet. This paper is an attempt to fill this gap.

To this goal, we need to consider a relevant distance between distributions. The Maximum Mean Discrepancy (MMD) between two arbitrary probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathbb{D}(\mathbb{P},\mathbb{Q}) := \sup_{f\in\mathcal{F}} \left| \int f\, d\mathbb{P} - \int f\, d\mathbb{Q} \right|,$$

where $\mathcal{F}$ is the unit ball in an universal reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined on a compact metric space, with an associated kernel K and a norm $\|\cdot\|_{\mathcal{H}}$. It can be proved that $\mathbb{D}(\mathbb{P},\mathbb{Q})$ is the distance between the kernel mean embeddings of the two underlying probabilities, i.e. $\mathbb{D}(\mathbb{P},\mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$ (see [25], Section 3.5, that provides a state-of-the-art introduction to the theory of RKHS and MMD). When the kernel K is characteristic (i.e. when the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective), MMD becomes a distance between the two probabilities $\mathbb{P}$ and $\mathbb{Q}$. Such a distance can be easily empirically estimated and has been used many times in different areas of statistics and machine learning. See [13, 19] for the two-sample test problem, e.g. As a tool for parametric estimation, even though it was implicitly used in specific examples in machine learning [15], MMD has been studied as a general method for inference only recently [2, 4, 10, 11]. In the latter papers, it appeared that MMD criteria lead to consistent estimators that are unconditionally robust to model misspecification. Moreover, the flexibility offered by the choice of a kernel, which can be

used to build a trade-off between statistical efficiency and robustness, is another advantage of such estimators. Thus, it seems natural to apply such inference techniques to copulas, for which the risk of misspecification is significant most of the time. In this paper, we will study a general semi-parametric inference procedure for copulas that is robust w.r.t. corrupted data, and that can be applied in case of model misspecification. Note that other distances are known to induce robustness, like the total variation distance [40] or the Hellinger distance [3]. However, the estimation procedures proposed in these papers are not computable in practice. Also, we refer the reader to [3] for a thorough discussion on why the MLE, based on the Kullback divergence, cannot enjoy the same robustness properties.

The rest of the paper is organized as follows: the remaining of the introduction yields notations and the definition of our estimators. Section 2 contains our theoretical results: non-asymptotic oracle inequalities, consistency and asymptotic distributions of our estimators. Section 3 provides experimental results. A simulations study confirms the robustness of MMD. We also provide an R Package, called MMDCopula [1], which allows statisticians to apply our algorithms.

Note that our package computes the MMD estimator by a stochastic gradient algorithm, described in Section 3. From [4, 10], such an algorithm can be implemented to compute the MMD estimator as long as it is possible to sample from the model. Thus, our package has been built on the package VineCopula [31], which allows to sample from the most popular copula families. This package also provided us some helpful formulas for the densities of some copulas, and their differentials. More details about the implementation can be found in Section 3.

1.2 Notations

Let $(\boldsymbol{X}_i)_{i=1,\dots,n}$ be an i.i.d. sample of d-dimensional random vectors, whose underlying copula is denoted by C_0 and whose margins are denoted by $F_1, \dots, F_d$. The latter ones will be left unspecified and, to simplify, we assume they are continuous. Let us define the unfeasible random variables $U_k := F_k(X_k)$, $k \in \{1, \dots, d\}$, and $\boldsymbol{U} := (U_1, \dots, U_d)$, for a

given random vector $\boldsymbol{X} := (X_1, \ldots, X_d)$ whose underlying copula is C_0 and underlying margins are $F_1, \ldots, F_d$. Obviously, the cdf of $\boldsymbol{U}$ is C_0 and its law is denoted by $\mathbb{P}_0$. The empirical measure associated to $(\boldsymbol{X}_i)_{i=1,\ldots,n}$ is denoted as $\mathbb{P}_n$.

We consider a particular parametric family of copulas $\mathcal{C} := \{C_\theta, \theta \in \Theta \subset \mathbb{R}^p\}$ (the family "of interest") and we search the best-suited copula inside the latter family. When the model is correctly specified, there exists a "true" parameter $\theta_0 \in \Theta$ i.e. $C_0 = C_{\theta_0}$. More generally, possibly in case of misspecification, we focus on a "pseudo-true" parameter $\theta_0^* \in \Theta$ so that a particular distance between C_0 and C_θ is minimized over $\theta \in \Theta$. In our case, this chosen distance will be the MMD. Denoting by $\mathbb{P}_\theta^U$ the law induced by C_θ on the hypercube $\mathcal{U} := [0,1]^d$, the pseudo-true value is formally defined as

$$\theta_0^* := \arg\min_{\theta \in \Theta} \mathbb{D}(\mathbb{P}_\theta^U, \mathbb{P}_0).$$

In the copula-related literature with unknown margins, it is common to define a pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\ldots,n}$, where $\hat{\boldsymbol{U}}_i := (\hat{U}_{i,1}, \ldots, \hat{U}_{i,d})$ and

$$\hat{U}_{i,k} := F_{n,k}(X_{i,k}), \quad F_{n,k}(t) := n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_{i,k} \leq t),$$

for every $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, d\}$ and every real number t. Our goal will be to evaluate the pseudo-true parameter θ_0^* with MMD techniques, from the initial sample $(\boldsymbol{X}_i)_{i=1,\ldots,n}$ or from the pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\ldots,n}$.

A relevant idea will be to work on the hypercube $\mathcal{U} := [0,1]^d$ instead of $\mathbb{R}^d$. To be specific, imagine we observe n i.i.d. realizations of $\boldsymbol{U}$, called $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$, and let $\mathbb{P}_n^U$ be the associated empirical measure on $\mathcal{U}$. To obtain an estimator of θ, the MMD criterion to be minimized is then $\mathbb{D}(\mathbb{P}_\theta^U, \mathbb{P}_n^U) := \|\mu_{\mathbb{P}_\theta^U} - \mu_{\mathbb{P}_n^U}\|_{\mathcal{H}_U}$, for some RKHS $\mathcal{H}_U$, that is associated with a kernel $K_U : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$. As in [4], we have

$$\begin{aligned}\mathbb{D}^2(\mathbb{P}_\theta^U, \mathbb{P}_n^U) = &\int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_\theta^U(d\boldsymbol{u})\, \mathbb{P}_\theta^U(d\boldsymbol{v}) - 2 \int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_\theta^U(d\boldsymbol{u})\, \mathbb{P}_n^U(d\boldsymbol{v}) \\ &+ \int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_n^U(d\boldsymbol{u})\, \mathbb{P}_n^U(d\boldsymbol{v}).\end{aligned}$$

Since we do not observe some realizations of $\boldsymbol{U}$, we have to replace them by pseudo-

observations in the latter criterion. This yields the approximate criterion

$$\mathbb{D}^2(\mathbb{P}_\theta^U, \hat{\mathbb{P}}_n^U) = \int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_\theta^U(d\boldsymbol{u})\, \mathbb{P}_\theta^U(d\boldsymbol{v}) - 2 \int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_\theta^U(d\boldsymbol{u})\, \hat{\mathbb{P}}_n^U(d\boldsymbol{v}) + \int K_U(\boldsymbol{u}, \boldsymbol{v}) \hat{\mathbb{P}}_n^U(d\boldsymbol{u})\, \hat{\mathbb{P}}_n^U(d\boldsymbol{v}),$$

where $\hat{\mathbb{P}}_n^U$ denotes the empirical measure associated with the pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\dots,n}$. Then, our estimator of θ_0^* is defined as

$$\begin{aligned}\hat{\theta}_n &:= \arg\min_{\theta \in \Theta} \mathbb{D}(\mathbb{P}_\theta^U, \hat{\mathbb{P}}_n^U) \\ &= \arg\min_{\theta \in \Theta} \int K_U(\boldsymbol{u}, \boldsymbol{v}) \mathbb{P}_\theta^U(d\boldsymbol{u})\, \mathbb{P}_\theta^U(d\boldsymbol{v}) - \frac{2}{n} \sum_{i=1}^n \int K_U(\boldsymbol{u}, \hat{\boldsymbol{U}}_i) \mathbb{P}_\theta^U(d\boldsymbol{u}). \end{aligned} \tag{1}$$

If C_θ has a density c_θ w.r.t. the Lebesgue measure on $[0,1]^d$, this criterion may be rewritten as

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} \int K_U(\boldsymbol{u}, \boldsymbol{v}) c_\theta(\boldsymbol{u}) c_\theta(\boldsymbol{v})\, d\boldsymbol{u}\, d\boldsymbol{v} - \frac{2}{n} \sum_{i=1}^n \int K_U(\boldsymbol{u}, \hat{\boldsymbol{U}}_i) c_\theta(\boldsymbol{u})\, d\boldsymbol{u}. \tag{2}$$

It is clear from the definition that $\hat{\theta}_n$ depends on the kernel K_U. Thus, the choice of the latter kernel is a very important question. The experimental study in Section 3 shows that, for the most common parametric copulas, Gaussian kernels $K_G(\boldsymbol{u}, \boldsymbol{v}) := \exp(-\|h(\boldsymbol{u}) - h(\boldsymbol{v})\|^2/\gamma^2)$ lead to very good results (h being the identity map or the inverse of the c.d.f of a standard Gaussian random variable, applied coordinatewise) . Interestingly, it empirically seems that the optimal value of γ depends only on the model, and not on the sample size. Actually, this fact was rigorously proven in [10] for the Gaussian mean model, and we conjecture that it holds more generally. In our case, this allows to calibrate γ once and for all through a preliminary set of simulations. Note that [15] proposed a median heuristic to calibrate γ that yields good results in practice. Alternatively, [4] proposed to minimize the asymptotic variance of the estimated parameter, which we could do thanks to our Theorem 4. A more complete discussion on the choice of the kernel can be found page 14 in [4].

Remark 1. *An alternative approach would be to directly work with the initial observations $\boldsymbol{X}_i$, instead of the pseudo-observations $\hat{\boldsymbol{U}}_i$. In this case, we apply the same strategy, but*

with the initial sample. The "feasible" law of $\boldsymbol{X}_i$ *will be semi-parametric, because its margins are non-parametrically estimated. To obtain an estimator of* θ*, the criterion to be minimized would now be* $\mathbb{D}(\mathbb{P}_\theta^X, \mathbb{P}_n^X) := \|\mu_{\mathbb{P}_\theta^X} - \mu_{\mathbb{P}_n^X}\|_{\mathcal{H}_X}$*, for some RKHS* $\mathcal{H}_X$*, that is associated with a kernel* $K_X : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$*. Here,* $\mathbb{P}_\theta^X$ *denotes the law of* $\boldsymbol{X}$ *given by* $F_1, \ldots, F_d$ *and* C_θ*. Applying Sklar's theorem, note that, for every* $\boldsymbol{x} = (x_1, \ldots, x_d)$*,* $\mathbb{P}_\theta^X(\boldsymbol{X} \leq \boldsymbol{x}) = C_\theta\big(F_1(x_1), \ldots, F_d(x_d)\big)$*. As above,*

$$\mathbb{D}^2(\mathbb{P}_\theta^X, \mathbb{P}_n^X) = \int K_X(\boldsymbol{x}, \boldsymbol{y}) \mathbb{P}_\theta^X(d\boldsymbol{x})\, \mathbb{P}_\theta^X(d\boldsymbol{y}) - 2 \int K_X(\boldsymbol{x}, \boldsymbol{y}) \mathbb{P}_\theta^X(d\boldsymbol{x})\, \mathbb{P}_n^X(d\boldsymbol{y}) + \int K_X(\boldsymbol{x}, \boldsymbol{y}) \mathbb{P}_n^X(d\boldsymbol{x})\, \mathbb{P}_n^X(d\boldsymbol{y}).$$

Since we do not know the margins of $\boldsymbol{X}$*, this yields the approximate criterion*

$$\mathbb{D}^2(\hat{\mathbb{P}}_\theta^X, \mathbb{P}_n^X) = \int K_X(\boldsymbol{x}, \boldsymbol{y}) \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x})\, \hat{\mathbb{P}}_\theta^X(d\boldsymbol{y}) - 2 \int K_X(\boldsymbol{x}, \boldsymbol{y}) \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x})\, \mathbb{P}_n^X(d\boldsymbol{y}) + \int K_X(\boldsymbol{x}, \boldsymbol{y}) \mathbb{P}_n^X(d\boldsymbol{x})\, \mathbb{P}_n^X(d\boldsymbol{y}),$$

where, for every $\boldsymbol{x} = (x_1, \ldots, x_d)$*, we define* $\hat{\mathbb{P}}_\theta^X(\boldsymbol{X} \leq \boldsymbol{x}) = C_\theta\big(F_{n,1}(x_1), \ldots, F_{n,d}(x_d)\big)$*. Then, this provides another estimator*

$$\hat{\theta}_n^X := \arg\min_{\theta \in \Theta} \mathbb{D}(\hat{\mathbb{P}}_\theta^X, \mathbb{P}_n^X) = \arg\min_{\theta \in \Theta} \int K(\boldsymbol{x}, \boldsymbol{y}) \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x})\, \hat{\mathbb{P}}_\theta^X(d\boldsymbol{y}) - \frac{2}{n} \sum_{i=1}^n \int K(\boldsymbol{x}, \boldsymbol{X}_i) \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x}).$$

Unfortunately, the evaluation of any integral of the type $\int \psi(\boldsymbol{x})\, \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x})$ *is costly in general. Indeed,*

$$\int \psi(\boldsymbol{x})\, \hat{\mathbb{P}}_\theta^X(d\boldsymbol{x}) \simeq n^{-d} \sum_{i_1, \ldots, i_d = 1}^n \psi(X_{i_1,1}, \ldots, X_{i_d,d}) c_\theta\big(F_{n,1}(X_{i_1,1}), \ldots, F_{n,d}(X_{i_d,d})\big).$$

Therefore, it is more convenient to deal with the first method, especially if d *is large. This is our choice in this paper.*

2 Theoretical results

We now study the theoretical properties of the estimator defined by (1). Since we will work with pseudo-observations from now on, we forget the upper index "U" to lighten

notations. Thus, the law induced by the pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\dots,n}$, previously denoted $\hat{\mathbb{P}}_n^U$, simply becomes $\hat{\mathbb{P}}_n$. Moreover, $\mathbb{P}_n^U$, the law of the unobservable sample $(\boldsymbol{U}_i)_{i=1,\dots,n}$ becomes $\mathbb{P}_n$. Recall that the true underlying law is $\mathbb{P}_0$, and $\mathbb{P}_0 = \mathbb{P}_{\theta_0^*}$ only if the model is correctly specified. For any function $f : \mathcal{E} \subset \mathbb{R}^d \to \mathbb{R}$ that is two times continuously differentiable, set

$$\|d^{(2)} f\|_\infty := \sup_{\boldsymbol{x}\in\mathcal{E}} \sup_{k,l=1,\dots,d} \left| \frac{\partial^2 f}{\partial x_k \partial x_l}(\boldsymbol{x}) \right|.$$

We assume in this section that the kernel K_U is symmetrical, i.e. $K_U(\boldsymbol{u},\boldsymbol{v}) = K_U(\boldsymbol{v},\boldsymbol{u})$ for every $\boldsymbol{u}$ and $\boldsymbol{v}$ in $[0,1]^d$ (otherwise, replace K_U by a symmetrized version). We also assume that the kernel is bounded over $[0,1]^2$. Note that the popular Gaussian kernel $K_G(\boldsymbol{u},\boldsymbol{v}) = \exp(-\|\boldsymbol{u}-\boldsymbol{v}\|^2/\gamma^2)$, is characteristic, symmetric and bounded.

2.1 Non-asymptotic guarantees

The first result of this section is a non-asymptotic "universal" upper bound in terms of MMD distance that holds with high probability for any underlying distribution. Our bound is exact, and exhibits clear dimensionality- and kernel-dependent constants. It establishes that the MMD estimator is robust to misspecification, and is consistent at the usual optimal $n^{-1/2}$ rate. Similar results can be found in the literature, both in the i.i.d. (Theorem 1 in [4], Theorem 3.1 in [10]) and in the dependent setting (Theorem 3.2 in [10]), but none of them can be applied to semi-parametric copula models.

Theorem 1. *The kernel K_U is assumed to be two times continuously differentiable on $[0,1]^d$. Then for any $\nu, \delta > 0$ with $\nu + \delta < 1$, with probability larger than $1-\delta-\nu \in (0,1)$,*

$$\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_0) \le \inf_{\theta\in\Theta} \mathbb{D}(\mathbb{P}_\theta, \mathbb{P}_0) + \left(\frac{8}{n} \sup_{\boldsymbol{u}\in[0,1]^d} K_U(\boldsymbol{u},\boldsymbol{u}) \right)^{1/2} \left\{ 1 + \left(-\ln \delta \right)^{1/2} \right\}$$
$$+ \left(\frac{2d^2}{n} \|d^{(2)} K_U\|_\infty \ln\left(\frac{2d}{\nu} \right) \right)^{1/2}.$$

Note that $\inf_{\theta\in\Theta} \mathbb{D}(\mathbb{P}_\theta, \mathbb{P}_0) = \mathbb{D}(\mathbb{P}_{\theta_0^*}, \mathbb{P}_0)$ by definition, and this quantity is zero if the model is correctly specified.

Proof. For every $\theta \in \Theta$, we have

$$\begin{aligned}\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_0) &\leq \mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \hat{\mathbb{P}}_n) + \mathbb{D}(\hat{\mathbb{P}}_n, \mathbb{P}_n) + \mathbb{D}(\mathbb{P}_n, \mathbb{P}_0) \\ &\leq \mathbb{D}(\mathbb{P}_\theta, \hat{\mathbb{P}}_n) + \mathbb{D}(\hat{\mathbb{P}}_n, \mathbb{P}_n) + \mathbb{D}(\mathbb{P}_n, \mathbb{P}_0) \\ &\leq \mathbb{D}(\mathbb{P}_\theta, \mathbb{P}_0) + 2\mathbb{D}(\hat{\mathbb{P}}_n, \mathbb{P}_n) + 2\mathbb{D}(\mathbb{P}_n, \mathbb{P}_0).\end{aligned}$$

With probability greater than $1-\delta$, Lemma 1 in [4] yields

$$\mathbb{D}(\mathbb{P}_n, \mathbb{P}_0) \leq \left(\frac{2}{n} \sup_{\boldsymbol{u}\in[0,1]^d} K_U(\boldsymbol{u}, \boldsymbol{u})\right)^{1/2} \Big\{1 + \big(\ln(1/\delta)\big)^{1/2}\Big\}. \tag{3}$$

Moreover, by some limited expansions of K_U wrt each of its arguments, evaluated at $(\boldsymbol{U}_i, \boldsymbol{U}_j)$ and with matrix notations, we get

$$\begin{aligned}\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n) &= \frac{1}{n^2}\sum_{i,j=1}^n \Big\{K_U(\boldsymbol{U}_i, \boldsymbol{U}_j) - 2K_U(\hat{\boldsymbol{U}}_i, \boldsymbol{U}_j) + K_U(\hat{\boldsymbol{U}}_i, \hat{\boldsymbol{U}}_j)\Big\} \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \Big\{\partial_1 K_U(\boldsymbol{U}_i, \boldsymbol{U}_j)^T(\boldsymbol{U}_i - \hat{\boldsymbol{U}}_i) - \frac{1}{2}(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T\partial_1^2 K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j)(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \\ &- \partial_2 K_U(\hat{\boldsymbol{U}}_i, \boldsymbol{U}_j)^T(\boldsymbol{U}_j - \hat{\boldsymbol{U}}_j) + \frac{1}{2}(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j)^T\partial_2^2 K_U(\hat{\boldsymbol{U}}_i, \tilde{\boldsymbol{U}}_j)(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j)\Big\},\end{aligned}$$

for some random vectors $\boldsymbol{U}_i^*$ (resp. $\tilde{\boldsymbol{U}}_j$) that lie between $\boldsymbol{U}_i$ and $\hat{\boldsymbol{U}}_i$ (resp. between $\boldsymbol{U}_j$ and $\hat{\boldsymbol{U}}_j$). Since the kernel is symmetrical, $\partial_1 K_U(\boldsymbol{u}, \boldsymbol{v}) = \partial_2 K_U(\boldsymbol{v}, \boldsymbol{u})$ for every $(\boldsymbol{u}, \boldsymbol{v})$ in $[0,1]^{2d}$. This yields, with obvious notations,

$$\begin{aligned}\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n) &= \frac{1}{n^2}\sum_{i,j=1}^n \Big\{\frac{(-1)}{2}(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T\partial_1^2 K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j)(\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \\ &- (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T\partial_{12}^2 K_U(\bar{\boldsymbol{U}}_i, \boldsymbol{U}_j)(\boldsymbol{U}_j - \hat{\boldsymbol{U}}_j) + \frac{1}{2}(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j)^T\partial_2^2 K_U(\hat{\boldsymbol{U}}_i, \tilde{\boldsymbol{U}}_j)(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j)\Big\},\end{aligned}$$

and we deduce

$$\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n) \leq 2d^2 \|d^{(2)}K_U\|_\infty \sup_{i=1,\dots,n}\sup_{k=1,\dots,d} |\hat{U}_{ik} - U_{ik}|^2. \tag{4}$$

The DKW inequality (p. 383 in [5]) yields

$$\mathbb{P}\Big(\sup_{i=1,\dots,n}\sup_{k=1,\dots,d} |\hat{U}_{i,k} - U_{i,k}|^2 > \varepsilon\Big) \leq 2d\exp\big(-2n\varepsilon\big),$$

and $\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n)$ is less than $d^2\|d^{(2)}K_U\|_\infty \ln(2d/\nu)/n$ with a probability larger than $1-\nu$. In addition with (3), this proves the result. □

Remark 2. *It is possible to slightly strengthen Theorem 1 at the price of more regularity for K_U. Indeed, assume K_U is three times differentiable and invoke a second-order limited expansion at $(\boldsymbol{U}_i, \boldsymbol{U}_j)$ for all the maps $(\boldsymbol{u}, \boldsymbol{v}) \mapsto K_U(\boldsymbol{u}, \boldsymbol{v}) - 2K_U(\boldsymbol{u}, \boldsymbol{U}_j) + K_U(\boldsymbol{U}_i, \boldsymbol{U}_j)$, $i, j \in \{1, \ldots, n\}$. With the same reasoning as in the proof above, this yields*

$$\begin{aligned}
\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n) = \frac{1}{n^2} \sum_{i,j=1}^{n} \Big\{ & (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T \partial^2_{1,2} K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j^*)(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j) \\
+ \; & (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T \big\{ \partial^2_{1,1} K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j^*) - \partial^2_{1,1} K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j) \big\} (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i) \\
= \; & \frac{1}{n^2} \sum_{i,j=1}^{n} (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^T \partial^2_{1,2} K_U(\boldsymbol{U}_i^*, \boldsymbol{U}_j^*)(\hat{\boldsymbol{U}}_j - \boldsymbol{U}_j) \\
+ \; & \frac{1}{n^2} \sum_{i,j=1}^{n} \partial^3_{1,1,2} K_U(\boldsymbol{U}_i^*, \tilde{\boldsymbol{U}}_j) \cdot (\hat{\boldsymbol{U}}_i - \boldsymbol{U}_i)^{(2)} \cdot (\boldsymbol{U}_j^* - \boldsymbol{U}_j),
\end{aligned}$$

since $\partial^2_{1,1} K_U(\boldsymbol{u}, \boldsymbol{v}) = \partial^2_{2,2} K_U(\boldsymbol{v}, \boldsymbol{u})$, with obvious notations for differentials. Then,

$$\mathbb{D}^2(\hat{\mathbb{P}}_n, \mathbb{P}_n) \leq d^2 \|d^{(2)} K_U\|_\infty \sup_{i=1,\ldots,n} \sup_{k=1,\ldots,d} |\hat{U}_{ik} - U_{ik}|^2 + d^3 \|d^{(3)} K_U\|_\infty \sup_{i=1,\ldots,n} \sup_{k=1,\ldots,d} |\hat{U}_{ik} - U_{ik}|^3.$$

As above, we get with probability larger than $1 - \delta - \nu$,

$$\begin{aligned}
\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_0) \leq & \inf_{\theta \in \Theta} \mathbb{D}(\mathbb{P}_\theta, \mathbb{P}_0) + \left(\frac{8}{n} \sup_{\boldsymbol{u} \in [0,1]^d} K_U(\boldsymbol{u}, \boldsymbol{u}) \right)^{1/2} \Big\{ 1 + \big(- \ln \delta \big)^{1/2} \Big\} \\
+ \; & \left(\frac{d^2}{n} \|d^{(2)} K_U\|_\infty \ln \big(\frac{2d}{\nu} \big) \right)^{1/2} + \left(\frac{d^3}{\sqrt{2} n^{3/2}} \|d^{(3)} K_U\|_\infty \left(\ln \left(\frac{2d}{\nu} \right) \right)^{3/2} \right)^{1/2}.
\end{aligned}$$

Let us emphasize the consequences of Theorem 1 when the data is contaminated by a proportion ε of outliers. Huber proposed a contamination model for which $\mathbb{P}_0 = (1 - \varepsilon)\mathbb{P}_{\theta_0} + \varepsilon \mathbb{Q}$. That is, while the majority of the observations is actually generated from the "true" model, a (small) proportion ε of them is generated by an arbitrary contamination distribution $\mathbb{Q}$. Using this framework, it is possible to upper bound the distance between the MMD estimator and the true parameter directly. To be short, assume here that $\sup_{\boldsymbol{u} \in [0,1]^d} K_U(\boldsymbol{u}, \boldsymbol{u}) \leq 1$, as for the usual Gaussian kernel. Since $\mathbb{D}(\mathbb{P}_0, \mathbb{P}_{\theta_0}) \leq 2\varepsilon$ and $\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_0) \leq 2\varepsilon + \mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_{\theta_0})$ by the triangle inequality, Theorem 1 yields

$$\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_{\theta_0}) \leq 4\varepsilon + \left(\frac{8}{n} \right)^{\frac{1}{2}} \Big\{ 1 + \big(- \ln \delta \big)^{1/2} \Big\} + \left(\frac{2d^2}{n} \|d^{(2)} K_U\|_\infty \ln \big(\frac{2d}{\nu} \big) \right)^{1/2}. \tag{5}$$

In any model where an upper bound on $\|\hat{\theta}_n - \theta_0\|^2$ can be deduced from an upper bound on $\mathbb{D}(\mathbb{P}_{\hat{\theta}_n}, \mathbb{P}_{\theta_0})$, this proves the robustness of $\hat{\theta}_n$.

Example **1.** *As an illustration, let us consider the Gaussian copula model in dimension* $d = 2$*, whose laws* $(\mathbb{P}_\theta)_{\theta\in(-1,1)}$ *are given by their density*

$$c_\theta(u_1, u_2) := \frac{1}{2\pi\sqrt{1-\theta^2}\phi(x_1)\phi(x_2)} \exp\Big(- \frac{1}{2(1-\theta^2)}\big(x_1^2 + x_2^2 - 2\theta x_1 x_2\big)\Big), \tag{6}$$

by setting $x_k = \Phi^{-1}(u_k)$*,* $k = 1, 2$*. We use the Gaussian kernel:*

$$K_U(\boldsymbol{U}, \boldsymbol{V}) = \exp\left(-\|\Phi^{-1}(\boldsymbol{U}) - \Phi^{-1}(\boldsymbol{V})\|^2/\gamma^2\right),$$

where Φ *is the c.d.f of a standard Gaussian random variable, and its inverse* Φ^{-1} *is applied coordinatewise. We prove at the end of Appendix D that, using the latter Gaussian kernel, there is a constant* $c(\gamma) \in (0, +\infty)$ *that depends only on* γ *such that, for any* $(\theta_1, \theta_2) \in (-1, 1)^2$*,* $|\theta_1 - \theta_2| \le c(\gamma)\mathbb{D}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})$*. Together with (5), this gives:*

$$|\hat{\theta}_n - \theta_0| \le c(\gamma)\left[4\varepsilon + \left(\frac{8}{n}\right)^{\frac{1}{2}}\Big\{1 + \big(-\ln\delta\big)^{1/2}\Big\} + \left(\frac{8}{n}\|d^{(2)}K_U\|_\infty \ln\big(\frac{4}{\nu}\big)\right)^{1/2}\right].$$

2.2 Asymptotic guarantees

We denote

$$\ell(\boldsymbol{w}; \theta) := \int K_U(\boldsymbol{u}, \boldsymbol{v})\mathbb{P}_\theta(d\boldsymbol{u})\,\mathbb{P}_\theta(d\boldsymbol{v}) - 2\int K_U(\boldsymbol{u}, \boldsymbol{w})\mathbb{P}_\theta(d\boldsymbol{u}).$$

We assume that the functions $\ell(\cdot; \theta)$ are measurable and $\mathbb{P}_0$-integrable for every $\theta \in \Theta$. The theoretical loss function is

$$L_0(\theta) := \mathbb{E}[\ell(\boldsymbol{U}; \theta)] = \int_{[0,1]^d} \ell(\boldsymbol{w}; \theta)\mathbb{P}_0(d\boldsymbol{w}).$$

Here, it is approximated by the empirical "feasible" loss

$$L_n(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\hat{\boldsymbol{U}}_i; \theta) = \int_{[0,1]^d} \ell(\boldsymbol{w}; \theta)\hat{\mathbb{P}}_n(d\boldsymbol{w}),$$

so that $\hat{\theta}_n = \arg\min_{\theta\in\Theta} L_n(\theta)$ and $\theta_0^* = \arg\min_{\theta\in\Theta} L_0(\theta)$. The asymptotic properties of M-estimators ("Quasi-MLE" particularly) for possibly misspecified models are well established in the literature: see [37, 38] for instance. As usual in the statistical theory of copulas, the main difficulty will come here from unspecified margins.

2.2.1 Consistency

Under classical assumptions, we prove that the MMD estimator is consistent.

Condition 1. *The parameter space Θ is compact. The map $L_0 : \Theta \to \mathbb{R}$ is continuous on Θ and uniquely minimized at θ_0^*.*

Condition 2. *The family $\mathcal{F} := \{\ell(\cdot, \theta);\, \theta \in \Theta\}$ is a collection of measurable functions with an integrable envelope function F. For every $\boldsymbol{w} \in [0,1]^d$, the map $\theta \mapsto \ell(\boldsymbol{w}; \theta)$ is continuous on Θ.*

Theorem 2. *If Conditions 1 and 2 are fulfilled, then $\hat{\theta}_n$ is strongly consistent, i.e.*

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} \theta_0^*.$$

Proof. As Θ is compact, then the δ-bracketing numbers $\mathcal{N}_{[\cdot]}\big(\delta, \mathcal{F}, L^1(\mathbb{P}_0)\big)$ are finite for every $\delta > 0$, invoking Example 19.8 in [36]. Moreover, using Lemma 1(c) in [8], we obtain the strong uniform law of large numbers

$$\sup_{\theta\in\Theta} |L_0(\theta) - L_n(\theta)| \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} 0.$$

Hence, according to Theorem 2.1 in [27] for example, we deduce the strong consistency of the minimizer $\hat{\theta}_n$ of L_n towards the unique minimizer of L_0. □

2.2.2 Asymptotic normality

Although Theorem 2 gives conditions under which we obtain the consistency of the MMD estimator, it does not provide any information on its rate of convergence. Hence, we now state the weak convergence of $\sqrt{n}(\hat{\theta}_n - \theta_0^*)$. First, we need a set of usual regularity conditions to deal with M-estimators. It mainly requires the functions $\ell(\boldsymbol{w}; \cdot)$ to be smooth enough on a small neighborhood of θ_0^* when $\boldsymbol{w} \in [0,1]^d$.

Condition 3. *θ_0^* is an interior point of Θ.*

Condition 4. *There exists an open neighborhood $\mathcal{O} \subset \Theta$ of θ_0^* s.t. the maps $\theta \mapsto \ell(\boldsymbol{w}; \theta)$ are twice continuously differentiable on $\mathcal{O}$, for $\mathbb{P}_0$-almost every $\boldsymbol{w} \in [0,1]^d$. Moreover, all functions $\nabla^2_{\theta,\theta}\ell(\cdot; \theta)$ are measurable on $[0,1]^d$ for any $\theta \in \mathcal{O}$.*

Condition 5. *There exists a compact set $K \subset \mathcal{O}$ whose interior contains θ_0^* such that*

$$\mathbb{E}\left[\sup_{\theta \in K} \left\|\nabla^2_{\theta,\theta} \ell(\boldsymbol{U}; \theta)\right\|\right] < +\infty,$$

for any matrix norm $\|\cdot\|$. Moreover, the map $\theta \mapsto \mathbb{E}[\nabla^2_{\theta,\theta} \ell(\boldsymbol{U}; \theta)]$ is continuous at θ_0^*.

Condition 6. *The matrix $B = \mathbb{E}[\nabla^2_{\theta,\theta} \ell(\boldsymbol{U}; \theta_0^*)]$ is positive definite.*

Condition 7. $\mathbb{E}[\nabla_\theta \ell(\boldsymbol{U}; \theta_0^*)] = 0$.

Second, the asymptotic behavior of our estimator is closely related to the asymptotic distribution of the empirical copula that has been widely studied in the last two decades. The weak convergence in $(\ell^\infty([0,1]^d), \|\cdot\|_\infty)$ of the empirical copula process $\{\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)(\boldsymbol{u}), \boldsymbol{u} \in [0,1]^d\}$ to a Gaussian process was formally stated by [16], by requiring the first-order partial derivatives of the copula $\mathbb{P}_0$ to exist and to be continuous on the entire unit hypercube $[0,1]^d$. Actually, as initially suggested in Theorem 4 of [16], the continuity is not needed on the boundary of the hypercube, but only on the interior of the hypercube. This result was established by [33] under minimal assumptions, rewritten below as Condition 9. With additional smoothness requirements on the loss function ℓ (Condition 8), we will be able to obtain the asymptotic normality of our MMD estimator $\hat{\theta}_n$ from the weak convergence of the empirical copula process.

Condition 8. *The function $\nabla_\theta \ell(\cdot; \theta_0^*)$ is right continuous and of bounded variations.*

Condition 9. *For each $j = 1, ..., d$, the j-th first-order partial derivative $\dot{C}_j$ of the true copula $\mathbb{P}_0$ exists and is continuous on the set $\{\boldsymbol{w} \in [0,1]^d : 0 < \boldsymbol{w}_j < 1\}$.*

Still, it is possible to obtain the weak convergence of the empirical copula process for an even larger class of copulas using semi-metrics on $\ell^\infty([0,1]^d)$ that are weaker than the sup-norm, but the limiting distribution will no longer be Gaussian in general. Indeed, [6] established the hypi-convergence of the empirical copula process $\{\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)(\boldsymbol{u}), \boldsymbol{u} \in [0,1]^d\}$ under the following assumption that is weaker than Condition 9.

Condition 10. *The set $\mathcal{S}$ of points in $[0,1]^d$ where the partial derivatives of the true copula $\mathbb{P}_0$ exist and are continuous has Lebesgue measure 1.*

Condition 11. *For some* $q \in (1, +\infty)$, $\int_{[0,1]^d} \left|\nabla_\theta \ell(d\boldsymbol{w}; \theta_0^*)\right|^q < \infty$.

Now, let us state the weak convergence of $\sqrt{n}(\hat{\theta}_n - \theta_0^*)$.

Theorem 3. *If Conditions 1-9 are fulfilled, then* $\sqrt{n}(\hat{\theta}_n - \theta_0^*)$ *is asymptotically normal. Alternatively, under Conditions 1-8 and 10-11, the weak limit of* $\sqrt{n}\big(\hat{\theta}_n - \theta_0^*\big)$ *still exists.*

The proof has been postponed to Appendix A. In the case of asymptotic normality, the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0^*)$ is $B^{-1}\Sigma B^{-1}$, where

$$\Sigma := \int \mathcal{C}_0(\boldsymbol{w}, \boldsymbol{w}') \nabla_\theta \ell(d\boldsymbol{w}; \theta_0^*) \nabla_\theta \ell(d\boldsymbol{w}'; \theta_0^*)^T,$$

and $\mathcal{C}_0(\cdot, \cdot)$ is the covariance function associated to the limiting law of the empirical copula process, i.e.

$$\mathcal{C}_0(\boldsymbol{w}, \boldsymbol{w}') := \mathbb{E}\Big[\big\{\alpha(\boldsymbol{w}) - \sum_{j=1}^d \dot{C}_j(\boldsymbol{w})\alpha_j(w_j)\big\}\big\{\alpha(\boldsymbol{w}') - \sum_{j=1}^d \dot{C}_j(\boldsymbol{w}')\alpha_j(w_j')\big\}\Big],$$

denoting by α a usual $\mathbb{P}_0$-Brownian bridge on $[0,1]^d$. In particular, note that

$$\mathbb{E}[\alpha(\boldsymbol{w})\alpha(\boldsymbol{w}')] = C_0(\boldsymbol{w} \wedge \boldsymbol{w}') - C_0(\boldsymbol{w})C_0(\boldsymbol{w}'), \ (\boldsymbol{w}, \boldsymbol{w}') \in [0,1]^{2d}.$$

The previous matrices can be empirically estimated: see Remark 2 in [8], or [35]. Note that a more explicit formula of Σ is given in the latter papers, say

$$\Sigma = \mathrm{Var}\Big[\nabla_\theta \ell(\boldsymbol{U}; \theta_0^*) + \sum_{j=1}^d \int \nabla^2_{\theta, u_j} \ell(\boldsymbol{u}; \theta_0^*) \left\{\mathbf{1}(U_j \leq u_j) - u_j\right\} \mathbb{P}_0(d\boldsymbol{u})\Big].$$

Alternatively, the asymptotic variance of $\hat{\theta}_n$ can be estimated by bootstrap resampling (see below).

In canonical maximum likelihood estimation of semi-parametric models, the asymptotic normality of the copula parameter is usually obtained by similar techniques but using slightly different assumptions: see e.g. [17, 8, 35]. In such a situation, the loss function ℓ is the copula log-likelihood and Condition 8 should then hold on the score function rather than on $\nabla_\theta \ell(\,\cdot\,; \theta_0^*)$. Unfortunately, the bounded variation assumption is violated by many popular copula families with unbounded copula score functions such as the Gaussian copula. Hence, it is not possible to establish the asymptotic normality of CML-estimators for

the latter copula family using the same set of assumptions as in Theorem 3. Hopefully, our MMD estimator does not suffer in general from these drawbacks as the derivatives of our loss are bounded most often. Nonetheless, if this is not the case, we can still rely on another set of technical assumptions, as for the CML. Now, we provide the following result adopting this alternative formulation, whose assumptions naturally hold for the Gaussian copula and can be checked by a direct analysis.

Condition 12. *For any* $\boldsymbol{w} \in (0,1)^d$, $\left\|\nabla_\theta \ell(\boldsymbol{w}; \theta_0^*)\right\| \leq C_1 \prod_{k=1}^d \{\boldsymbol{w}_k(1-\boldsymbol{w}_k)\}^{-a_k}$ *for some constants* C_1 *and* $a_k \geq 0$ *such that*

$$\mathbb{E}\Big[\prod_{k=1}^d \{\boldsymbol{U}_k(1-\boldsymbol{U}_k)\}^{-2a_k}\Big] < +\infty.$$

Moreover, for any $\boldsymbol{w} \in (0,1)^d$ *and any* $k = 1, \ldots, d$,

$$\left\|\nabla^2_{\theta,w_k} \ell(\boldsymbol{w}; \theta_0^*)\right\| \leq C_2 \left\{\boldsymbol{w}_k(1-\boldsymbol{w}_k)\right\}^{-b_k} \prod_{j=1, j\neq k}^d \left\{\boldsymbol{w}_j(1-\boldsymbol{w}_j)\right\}^{-a_j},$$

for some constants C_2 *and* $b_k > a_k$ *such that*

$$\mathbb{E}\Big[\left\{\boldsymbol{U}_k(1-\boldsymbol{U}_k)\right\}^{\zeta_k - b_k} \prod_{j=1, j\neq k}^d \left\{\boldsymbol{U}_j(1-\boldsymbol{U}_j)\right\}^{-a_j}\Big] < +\infty,$$

for some $\zeta_k \in (0, 1/2)$.

Under the latter conditions, the partial derivatives of $\ell(\boldsymbol{w}, \theta)$ are allowed to blow up at the boundaries of $[0,1]^d$, but not "too quickly". Therefore, we get the same result as in Theorem 3.

Theorem 4. *If Conditions 1-7 and 12 are fulfilled, then the MMD estimator* $\hat{\theta}_n$ *is asymptotically normal:* $\sqrt{n}(\hat{\theta}_n - \theta_0^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, B^{-1}\Sigma B^{-1})$.

The proof follows the lines of the proof of Theorem 3, adding some arguments from Proposition 2 in [8]. The details of the proof are straightforward and are left to interested readers.

The limiting laws obtained in Theorem 3 and 4 are most often complex, even in the case of Gaussian limit laws. Once pseudo-observations are managed, particularly through

empirical copula processes, it is common practice to rely on bootstrap schemes. In our case, we promote the use of Efron's nonparametric bootstrap and, more generally, the multiplier bootstrap as defined in [6]. The validity of the latter bootstrap scheme for the estimation of θ_0^* in our framework is due to the validity of the corresponding bootstrap scheme, as stated in [16, 6] (see (13) in our proofs). In practical terms, the calculation of a bootstrap estimator requires resampling every observation i in the sample with a convenient weight $W_{i,n}$, independently of the sample. For the nonparametric bootstrap, $(W_{1,1}, \ldots, W_{n,n})$ is drawn following a n multinomial law with success probabilities $(1/n, \ldots, 1/n)$. In the case of the multiplier bootstrap, the weights are i.i.d. with both mean and variance equal to one.

2.3 Examples

Now, let us check that the previous asymptotic results can be applied for two usual bivariate copula families, here the Gaussian and the Marshall-Olkin copulas. In this subsection, we assume that the model is well-specified, i.e. that the law of the observations belongs to the considered parametric family. As a consequence, the pseudo-true parameter θ_0^* is in fact the true underlying parameter and is denoted θ_0.

In both cases, we will use some characteristic Gaussian-type kernel K_U defined by

$$K_h(\boldsymbol{u}, \boldsymbol{v}) := \exp\Big(- \frac{\{h(u_1) - h(v_1)\}^2 + \{h(u_2) - h(v_2)\}^2}{\gamma^2}\Big), \tag{7}$$

for some injective map $h : [0,1] \mapsto \mathbb{R}$ and some tuning parameter $\gamma > 0$ (see [12], Th. 2.2, e.g.). Indeed, the latter function K_h is a kernel: let $\zeta : \mathbb{R}^2 \to \mathcal{F}$ be the feature map that is associated to the usual Gaussian kernel K_G, i.e. $K_G(\boldsymbol{x}, \boldsymbol{y}) = \langle\, \zeta(\boldsymbol{x}), \zeta(\boldsymbol{y}) \,\rangle_{\mathcal{F}}$, where the Gaussian kernel is defined for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ by

$$K_G(\boldsymbol{x}, \boldsymbol{y}) := \exp\Big(- \frac{\{x_1 - y_1\}^2 + \{x_2 - y_2\}^2}{\gamma^2}\Big).$$

Then, the feature map that defines K_h is simply $\psi : [0,1]^2 \to \mathcal{F}$ given by $\psi(\boldsymbol{u}) = \zeta(h(u_1), h(u_2))$ for every $\boldsymbol{u} \in (0,1)^2$, and K_h inherits from K_G its "characteristic" property.

Hereafter, we shall denote by Φ and ϕ the cumulative distribution function and the probability density function of the standard normal distribution, respectively. Then, a

natural choice is to set $h(u) = \Phi^{-1}(u)$. This will be our choice by default in this section, and the latter kernel will simply be denoted by K_U. Even if it is possible to choose the usual Gaussian kernel K_G by setting $h(u) = u$, we have observed that K_U provides better results in some situations. We refer the reader to the simulation study for a detailed comparison. Moreover, it is simpler to check our conditions of regularity with K_U rather than K_G, in the case of Gaussian copulas particularly.

Indeed, it is relatively easy to show that Conditions 8-9 are satisfied for Gaussian copulas using the kernel K_U. As a consequence, its MMD estimator will be asymptotically normal. At the opposite, Marshall-Olkin copulas do not satisfy 9 but rather Condition 10. Hence, it is still possible to define and to analyze the asymptotic behavior of our parameter estimator in the Marshall-Olkin case.

2.3.1 Gaussian copulas

Let us consider two-dimensional Gaussian copulas $C_\theta(\boldsymbol{u}) := \Phi_{2,\theta}\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big)$, indexed by $\theta \in \Theta = [-1, 1]$. Here, $\Phi_{2,\theta}$ denotes the cdf of a bivariate Gaussian centered vector (X_1, X_2), $\mathbb{E}[X_k^2] = 1$, $k = 1, 2$, and $\mathbb{E}[X_1 X_2] = \theta$. Note that $C_1 = C^+$ and $C_{-1} = C^-$, respectively the upper- and lower Frechet bounds. When $\theta \in (-1, 1)$, the associated copula density has been given in Equation (6).

Proposition 1. *For any true parameter $\theta_0 \in [-1, 1]$, the estimator $\hat{\theta}_n$ given by (1) is strongly consistent. When $\theta_0 \in (-1, 1)$, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal.*

The proof is deferred to Appendix B and relies on Theorem 3. In this proof, it is stated that the term B that appears in the asymptotic variance of $\hat{\theta}_0$ has the closed-form expression

$$B = \frac{3\gamma^2\big\{(2 + \gamma^2/2)^2 + 8\theta_0^2\big\}}{2\big\{(2 + \gamma^2/2)^2 - 4\theta_0^2\big\}^{5/2}}.$$

2.3.2 Marshall-Olkin copulas

By definition ([26], Section 3.1.1), the bivariate Marshall-Olkin copula is defined on $[0, 1]^2$ as

$$C_\theta(u, v) = u^{1-\alpha} v \mathbf{1}(u^\alpha \geq v^\beta) + u v^{1-\beta} \mathbf{1}(u^\alpha < v^\beta), \tag{8}$$

for some parameter $\theta := (\alpha, \beta)$, $0 < \alpha, \beta < 1$. This copula has no density w.r.t. the Lebesgue measure on the whole $[0,1]^2$. The absolutely continuous part of C_θ (w.r.t. the Lebesgue measure) is defined on $[0,1]^2 \setminus \mathfrak{C}$, where $\mathfrak{C} := \{(u,v) \in [0,1]^2 \setminus u^\alpha = v^\beta\}$. The singular component is concentrated on the curve $\mathfrak{C}$, and $\mathbb{P}(U^\alpha = V^\beta) = \alpha\beta/(\alpha+\beta-\alpha\beta) =: \kappa$, when $(U,V) \sim C_\theta$. With the same notation as in [26], $C_\theta(u,v) = A_\theta(u,v) + S_\theta(u,v)$, where, for every $(u,v) \in [0,1]^2$, $S_\theta(u,v) = \kappa\big\{\min(u^\alpha, v^\beta)\big\}^{1/\kappa}$ and

$$\begin{aligned} A_\theta(u,v) &= \int_0^u \int_0^v \frac{\partial^2 C_\theta}{\partial u \partial v}(s,t)\, ds\, dt \\ &= \int_0^u \int_0^v \big\{(1-\alpha)s^{-\alpha}\mathbf{1}(s^\alpha > t^\beta) + (1-\beta)t^{-\beta}\mathbf{1}(s^\alpha < t^\beta)\big\}\, ds\, dt. \end{aligned}$$

Let us calculate $\mathbb{E}[\psi\big(U,V\big)]$, $(U,V) \sim C_\theta$, for any measurable map ψ, to be able to calculate $\ell(\boldsymbol{w}, \theta)$ for our bivariate Marshall-Olkin model. Given a small positive real number δ, let us first evaluate the mass along $\mathfrak{C}$, when the abscissa and the ordinate belong to $[u, u+\delta]$ and $[v, v+\delta]$ respectively: if $u^\alpha = v^\beta$ and $\delta \ll 1$,

$$\begin{aligned} & S_\theta(u+\delta, v+\delta) - S_\theta(u+\delta, v) - S_\theta(u, v+\delta) + S_\theta(u,v) \\ &= \kappa \min\big((u+\delta)^\alpha, (v+\delta)^\beta\big)^{1/\kappa} - \kappa u^{\alpha/\kappa} \\ &\simeq \delta\alpha u^{\alpha/\kappa - 1}\mathbf{1}(\alpha v \le \beta u) + \delta\beta v^{\beta/\kappa - 1}\mathbf{1}(\alpha v > \beta u) \\ &\simeq \delta\alpha u^{\alpha/\beta - \alpha}\mathbf{1}(\alpha v \le \beta u) + \delta\beta u^{1-\alpha}\mathbf{1}(\alpha v > \beta u), \end{aligned}$$

providing the density along the curve $\mathfrak{C}$. Therefore, we obtain

$$\mathbb{E}[\psi\big(U,V\big)] = \int \psi(s,t)\frac{\partial^2 C_\theta}{\partial u \partial v}(s,t)\, ds\, dt + \int \psi(u,v)\, S_\theta(du, dv) =: I_1 + I_2, \tag{9}$$

$$I_1 = \int \psi(s,t)\big\{(1-\alpha)s^{-\alpha}\mathbf{1}(s^\alpha > t^\beta) + (1-\beta)t^{-\beta}\mathbf{1}(s^\alpha < t^\beta)\big\}\, ds\, dt. \tag{10}$$

Let $(\bar{u}_{\alpha,\beta}, \bar{v}_{\alpha,\beta})$ be a point of $\mathfrak{C}$ such that $\alpha\bar{v}_{\alpha,\beta} = \beta\bar{u}_{\alpha,\beta}$. It is easy to check that such a point exists in $[0,1]^2$ and is unique, except when $\alpha = \beta$. In the latter case, the couple $(\bar{u}_{\alpha,\beta}, \bar{v}_{\alpha,\beta})$ may be arbitrarily chosen along the main diagonal of $[0,1]^2$. Then, we get

$$I_2 = \int \psi(u,v)\, S_\theta(du, dv) = \int_0^{\bar{u}_{\alpha,\beta}} \psi(u, u^{\alpha/\beta})\, \beta u^{1-\alpha}\, du + \int_{\bar{u}_{\alpha,\beta}}^1 \psi(u, u^{\alpha/\beta})\, \alpha u^{\alpha/\beta - \alpha}\, du, \tag{11}$$

with $\bar{u}_{\alpha,\beta} = \big(\beta/\alpha\big)^{\beta/(\alpha-\beta)}$ when $\alpha \neq \beta$ and $\bar{u}_{\alpha,\alpha} = e^{-1}$. The latter value has been chosen so that the map $(\alpha, \beta) \mapsto \bar{u}_{\alpha,\beta}$ is continuous on the whole set $(0,1)^2$, i.e. even at the main diagonal. For most regular functions ψ, the latter integrals I_1, I_2 and then $\mathbb{E}[\psi\big(U,V\big)]$ are continuous functions of (α, β).

Proposition 2. *For almost any true parameter $\theta_0 = (\alpha_0, \beta_0)$ that belongs to the interior of $\Theta := [\epsilon, 1-\epsilon]^2$ for some $\epsilon \in (0, 1/2)$, the estimator $\hat{\theta}_n$ given by (1) is strongly consistent, using the kernel K_U. Moreover, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is weakly convergent.*

See the proof in Appendix C. In this case, the limiting law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ exists but is not Gaussian in general. It could be numerically evaluated by usual resampling techniques, as the consistent bootstrap scheme in [6, Section 4.2]. Note that the same result applies with the usual Gaussian kernel K_G.

3 Implementation and experimental study

In this section, we compare the MMD estimator to the MLE and the moment estimator on simulated and real data. The MLE and the method of moments by inversion of Kendall's tau are implemented in the R package VineCopula [31]. We implemented the MMD estimator using the stochastic gradient algorithm described in [10]. This procedure requires to sample from the copula model we want to estimate. For this, we used again VineCopula. Note that our implementation of the MMD estimator is itself available as the R package MMDCopula [1].

3.1 Implementation via stochastic gradient and the MMDCopula package

We start by a short description of the algorithm implemented in our R package [1] to compute the MMD estimator. The main idea is differentiating the criterion (2). Under

suitable assumptions on the copula density c_θ w.r.t. the Lebesgue measure on $\mathcal{U}$, we have

$$\begin{aligned}
&\frac{d}{d\theta}\Big[\int K_U(\boldsymbol{u},\boldsymbol{v})c_\theta(\boldsymbol{u})c_\theta(\boldsymbol{v})\,d\boldsymbol{u}\,d\boldsymbol{v} - \frac{2}{n}\sum_{i=1}^n \int K_U(\boldsymbol{u},\hat{\boldsymbol{U}}_i)c_\theta(\boldsymbol{u})\,d\boldsymbol{u}\Big] \\
&\quad = 2\int K_U(\boldsymbol{u},\boldsymbol{v})\frac{d\log c_\theta(\boldsymbol{u})}{d\theta}c_\theta(\boldsymbol{u})c_\theta(\boldsymbol{v})\,d\boldsymbol{u}\,d\boldsymbol{v} - \frac{2}{n}\sum_{i=1}^n \int K_U(\boldsymbol{u},\hat{\boldsymbol{U}}_i)\frac{d\log c_\theta(\boldsymbol{u})}{d\theta}c_\theta(\boldsymbol{u})\,d\boldsymbol{u} \\
&\quad = 2\,\mathbb{E}\Big[\frac{d\log c_\theta(\boldsymbol{U})}{d\theta}\big(K_U(\boldsymbol{U},\boldsymbol{V}) - \frac{1}{n}\sum_{i=1}^n K_U(\boldsymbol{U},\hat{\boldsymbol{U}}_i)\big)\Big],
\end{aligned}$$

where the expectation is taken with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$, that are independently drawn from C_θ (a formal statement can be found in [10]). Even though this expectation is usually not available in closed form, it is possible to estimate it by Monte-Carlo to use a stochastic gradient descent. That is, we fix a starting point, a step size sequence $(\eta_n)_{n\geq 0}$, and iterate:

$$\begin{cases}
\text{draw } \boldsymbol{U}_1,\dots,\boldsymbol{U}_n,\boldsymbol{V}_1,\dots,\boldsymbol{V}_n \sim C_{\theta_n} \text{ i.i.d,} \\
\theta_{n+1} \leftarrow \theta_n - 2\eta_n n^{-2}\sum_{i,j=1}^n \frac{d\log c_\theta(\boldsymbol{U}_j)}{d\theta}_{|\theta=\theta_n}\big(K(\boldsymbol{U}_j,\boldsymbol{V}_i) - K_U(\boldsymbol{U}_j,\hat{\boldsymbol{U}}_i)\big).
\end{cases}$$

The convergence of this algorithm in a general framework is discussed in [10]. Note that the implementation of this algorithm requires 1) to be able to sample from C_θ and 2) to compute c_θ and its partial derivative with respect to θ. A list of densities and differentials can be found in [30] and is implemented in VineCopula [31]. Procedures to sample from C_θ can also be found in VineCopula. The same ideas can be adapted even if the latter copula density does not exist on the whole hypercube, as for the Marshall-Olkin copula. In the latter case with $\alpha = \beta$, we implemented our own sampler and considered the copula density with respect to the measure given by the sum of the Lebesgue measure on $[0,1]^2$ plus the Lebesgue measure on the first diagonal.

Also, note that it is possible to use a quasi-Monte-Carlo rather than a Monte-Carlo sampling scheme. In our package MMDCopula [1], we give the user the possibility to choose the sampling scheme for the $\boldsymbol{U}_j$'s and the $\boldsymbol{V}_i$'s separately. In all our simulations, we observed that the use of Monte-Carlo on the $\boldsymbol{U}_j$ and of quasi-Monte-Carlo on the $\boldsymbol{V}_i$'s led to the best results, so this setting is chosen by default in our package, and it was also used in the following experiments. A important point is that the gradient method is *not*

invariant by reparametrization. In order to deal with gradient descents in compact sets only, we decided to parametrize all the copulas by their Kendall's tau (apart from the Marshall-Olkin copula, implemented in the case $\alpha = \beta$, that is parametrized by α and does not use quasi-Monte Carlo).

Finally, in the MMDCopula package, the estimator $\hat{\theta}_n$ can be computed for five different kernels. In the following simulations, we worked with the Gaussian kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_2^2/\gamma^2)$, the exp-$L_2$ kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_2/\gamma)$ and the exp-L_1 kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_1/\gamma)$, where h is either the identity or Φ^{-1} and is applied coordinatewise. A major question is then: how to calibrate γ, and which kernel to choose? We performed some experiments on synthetic data to answer this question. In Figure 1, we provide the MSE of the estimators based on these three kernels as a function of γ.

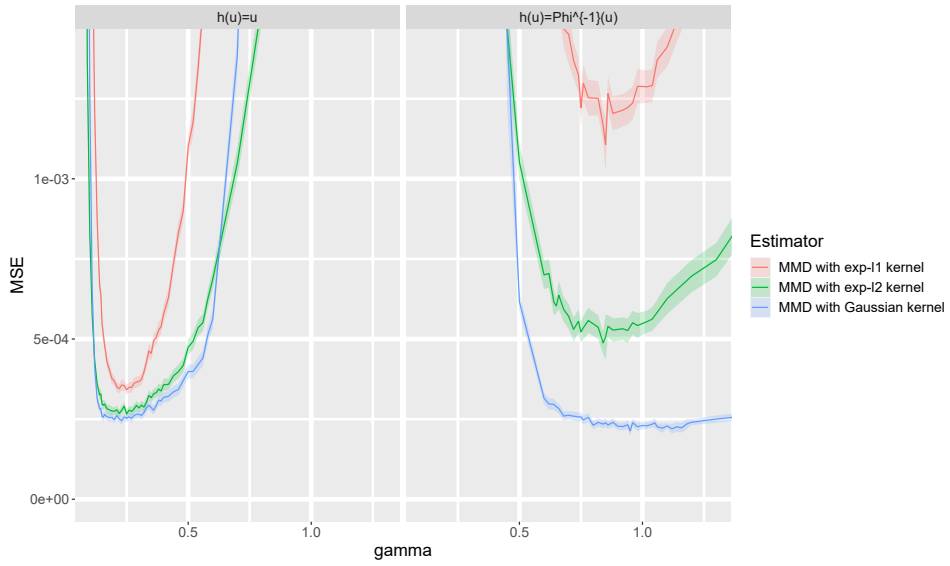

Figure 1: MSE of $\hat{\theta}_n$ based on the Gaussian kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_2^2/\gamma^2)$, the exp-$L_2$ kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_2/\gamma)$ and the exp-L_1 kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-\|h(\boldsymbol{U}) - h(\boldsymbol{V})\|_1/\gamma)$, as functions of γ.

In these experiments, $n = 1000$ observations were sampled from the Gaussian copula, and the objective was to estimate the parameter of this copula. Each experiment was repeated 1000 times.

The take-home message is that, as far as the Gaussian copula is concerned and $n = 1000$, the Gaussian kernel is the best one, whatever the choice of h. When h is the identity map, the optimal γ is $\gamma \simeq 0.23$. (the default choice in our package). For $h(\boldsymbol{u}) = \big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big)$, the optimal value is $\gamma = 0.95$. In the following simulations, we always used the latter values that seemed to perform well in any setting.

3.2 Comparison to the MLE on synthetic data

We now compare the MMD estimators based on the Gaussian kernel (with two choices of h) to the maximum likelihood estimator (MLE) and the estimator based on the inversion of Kendall's tau ("Itau"). We would like to illustrate convergence when the sample size $n \to \infty$ and robustness to the presence of various type of outliers. We designed three types of outliers.

- *uniform*: the outliers are drawn i.i.d from the uniform distribution $\mathcal{U}([0,1]^2)$.
- *top-left*: the outliers belong to the top-left corner of $[0,1]^2$, that is, they are drawn i.i.d from $\mathcal{U}([0,q] \times [1-q,q])$ where $q = 0.001$.
- *bottom-left*: the outliers belong to the bottom-left corner, that is, they are drawn i.i.d from $\mathcal{U}([0,q]^2)$.

In each case, the data are sampled on $[0,1]^2$ from the desired copula. Finally, the contaminated observations are rescaled by their rank in order to keep pseudo-uniform margins.

In a first series of experiments, we use the various estimators to estimate the parameter of the Gaussian copula. We compare their robustness to the presence of a proportion ε of each type of outliers, when ε ranges from 0 to 0.05. In a second time, we go beyond the Gaussian model: we replicate these experiments for the Frank copula, the Clayton copula, the Gumbel copula and the Marshall-Olkin copula. The results being quite similar, we save space by reporting only them for *top-left* outliers. In the last series of experiments, we come back to the Gaussian case, and illustrate the asymptotic theory. In this last experiment, we study the convergence of the estimators when n grows in two situations: no outliers, or a proportion $\varepsilon = 0.1$ of *top-left* outliers.

3.2.1 Robustness to various types of outliers in the Gaussian copula model

For each type of outliers, and for each ε in a grid that ranges from 0 to 0.05, we repeat 1000 times the following experiment: the data are i.i.d from the Gaussian copula, the sample size is $n = 1000$ and the parameter is calibrated so that $\tau = 0.5$. Then, an exact proportion ε of the data is replaced by outliers. We report the mean MSE of each estimator in Figure 2.

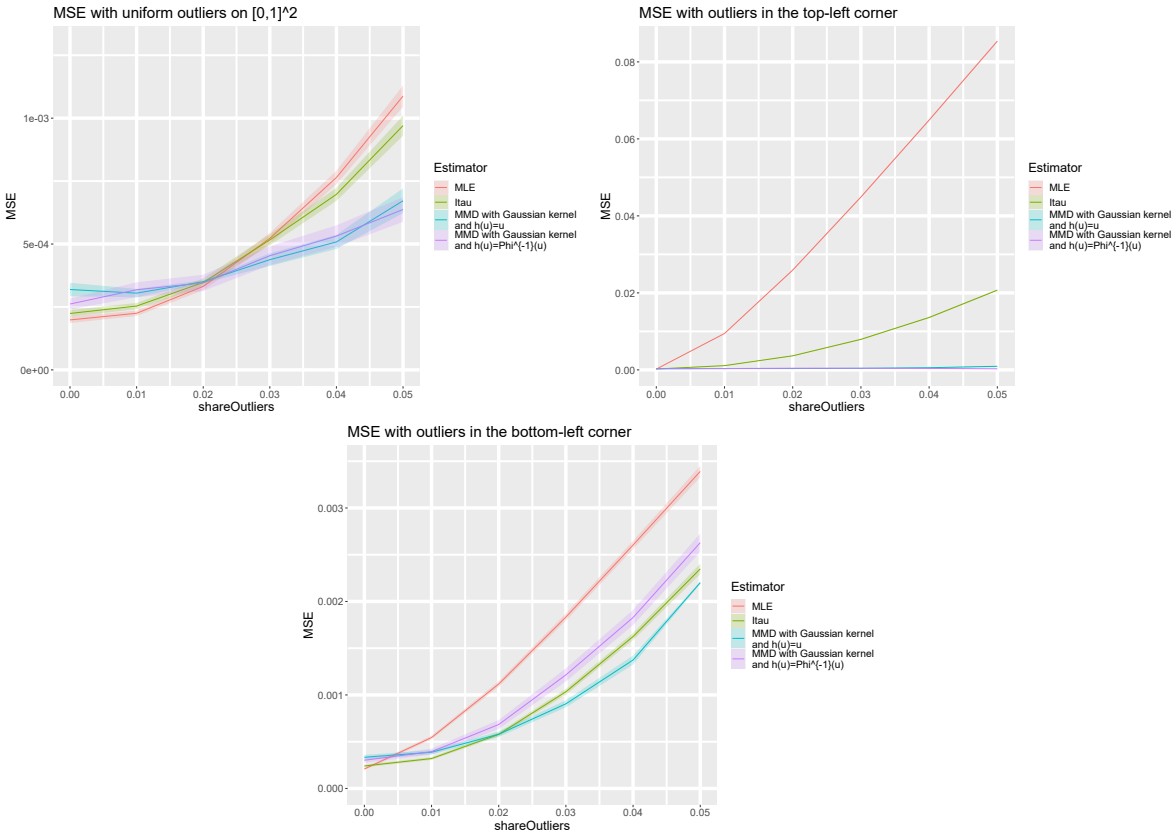

Figure 2: MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the MLE estimator and the method of moment based on Kendall's τ, as a function of the proportion ε of outliers. Sample size: $n = 1000$, model: Gaussian copula. Top-left: *uniform* outliers, top-right: *top-left* outliers, and bottom: *bottom-left* outliers.

When there are no outliers, the MLE is the best estimator. However, as soon as there is more than 2 or 3 percent of outliers, the MMD estimators become much more

reliable. Interestingly, the one based on $h(u) = u$ becomes equivalent to the one based on $h(u) = \Phi^{-1}(u)$ with *uniform* outliers, in terms of MSE.

3.2.2 Robustness in various models

Here, we replicate the previous experiments with other models: Clayton, Gumbel, Frank and Marshall-Olkin. In each case, the parameter was chosen so that $\tau = 0.5$. We report the results in the case of *top-left* outliers in Figure 3.

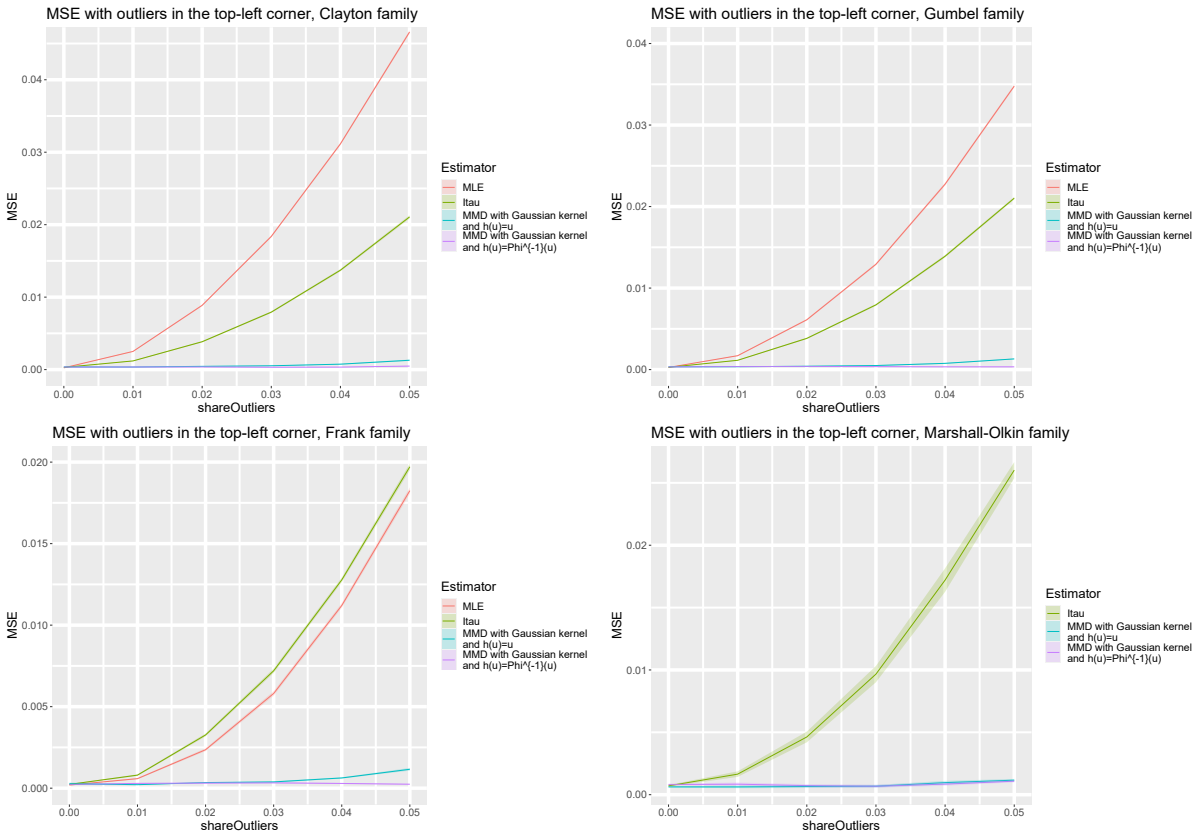

Figure 3: MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the MLE estimator and the method of moment based on Kendall's τ, as a function of the proportion ε of *top-left* outliers. Sample size: $n = 1000$. Top-left: Clayton copula. Top-right: Gumbel copula. Bottom-left: Frank copula. Bottom-right: Marshall-Olkin copula.

The conclusion remains unchanged: in all models, the MMD estimators are far more

robust than the MLE and the method of moments estimators.

3.2.3 Convergence

We finally come back to the Gaussian copula case. This time, we study the influence of the sample size n, ranging from $n = 100$ to $n = 5000$. We report the results of simulations without outliers ($\varepsilon = 0.00$) and with *top-left* outliers ($\varepsilon = 0.10$, independently of the sample size) in Figure 4.

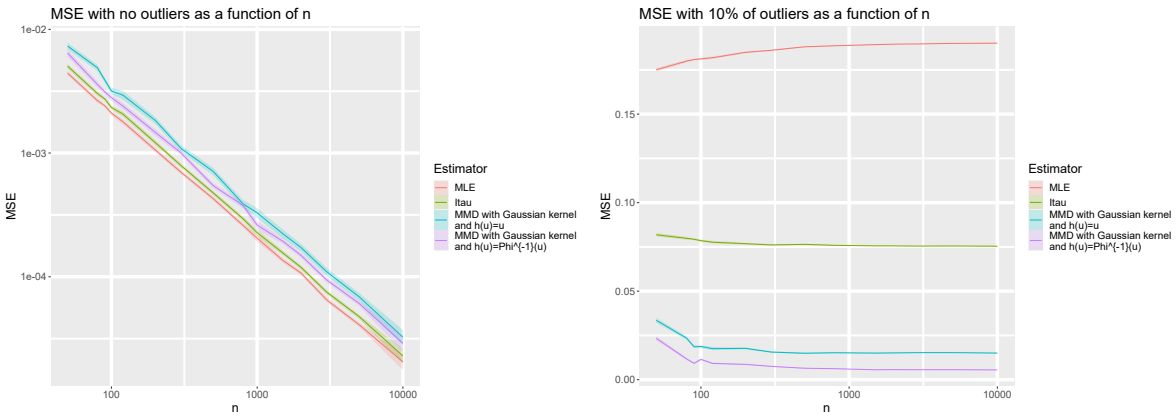

Figure 4: MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the MLE estimator and the method of moment based on Kendall's τ, as a function of the sample size n. Model: Gaussian copula. Left: no outliers. Right: a proportion $\varepsilon = 0.10$ of outliers.

When there are no outliers, we observe the $\sqrt{n}$ consistency of all the estimators, as predicted by the theory. Moreover, as expected, the MLE method yields the best estimator in this case, as expected (it is asymptotically efficient). However, when there are outliers, the situation is dramatically different. All the estimators have an incompressible bias, and only their variances will decrease to 0. However, we already observed that the MMD estimators are a lot more robust to outliers: indeed, here, their bias is (much) smaller than the other competing methods. Note that the hierarchy between the different methods is unaffected by the sample size.

3.3 Real data

In this section, we study the dependence between the daily stock returns of Google and Apple. We consider the "post-Lehmann Brothers" period of time, between 15 September 2008 and 26 August 2012. Using the R package `fGarch` [39], we remove the heteroskedasticity of each time series by ARMA-GARCH filtering, selecting the best lagged model using the BIC criteria. Finally, we obtain a bivariate series of innovations $(\eta_{APPL,i}\,, \eta_{GOOG,i})$ with $n = 995$ observations. A corresponding multivariate Ljung-Box portmanteau test (also called Q-test) of serial independence cannot be rejected at the level 4%. Therefore, we will consider the latter series of bivariate vectors of observations as i.i.d., even if it is probably not the case strictly speaking.

We try to fit several parametric families of bivariate copulas: Gaussian, Clayton and Gumbel. The corresponding implied Kendall's tau and their confidence intervals are displayed in Figure 5. For the MMD estimator, the bootstrap-based confidence intervals are computed as follows:

1. Compute the estimator $\hat{\theta}_{MMD}$ using the original sample.
2. For $j = 1$ to $N = 1000$, independently draw a sample $(\eta^{*,j}_{APPL,i}, \eta^{*,j}_{GOOG,i})_{i=1,\dots,n}$ with replacement from the original sample (usual non-parametric bootstrap).
3. For each of these samples, compute a bootstrapped estimator $\hat{\theta}^{*,j}_{MMD}$.
4. Compute q_{025} as the empirical quantile of $(\hat{\theta}_{MMD} - \hat{\theta}^{*,j}_{MMD})_{j=1,\dots,N}$ at the level 2.5%. Similarly compute q_{975}.
5. Return $[\hat{\theta}_{MMD} + q_{025}\,,\, \hat{\theta}_{MMD} + q_{975}]$.

For the MLE and Itau, we use the asymptotic confidence intervals at level 95% using the corresponding standard error given by the function `BiCopEst` of the package VineCopula.

In the case of the Clayton family, the confidence intervals of the MLE estimator and MMD estimator with $h = \Phi^{-1}$ are disjoint, meaning that their difference is statistically significant. More weakly, we also find that, for other families, the MMD estimator never belongs to the confidence intervals of the MLE and conversely. Such situations in practice

should incite the statistician to use more robust estimators than the MLE, and to try to understand why the MLE acts so differently, for example using some visualization tools to seek outliers in the data.

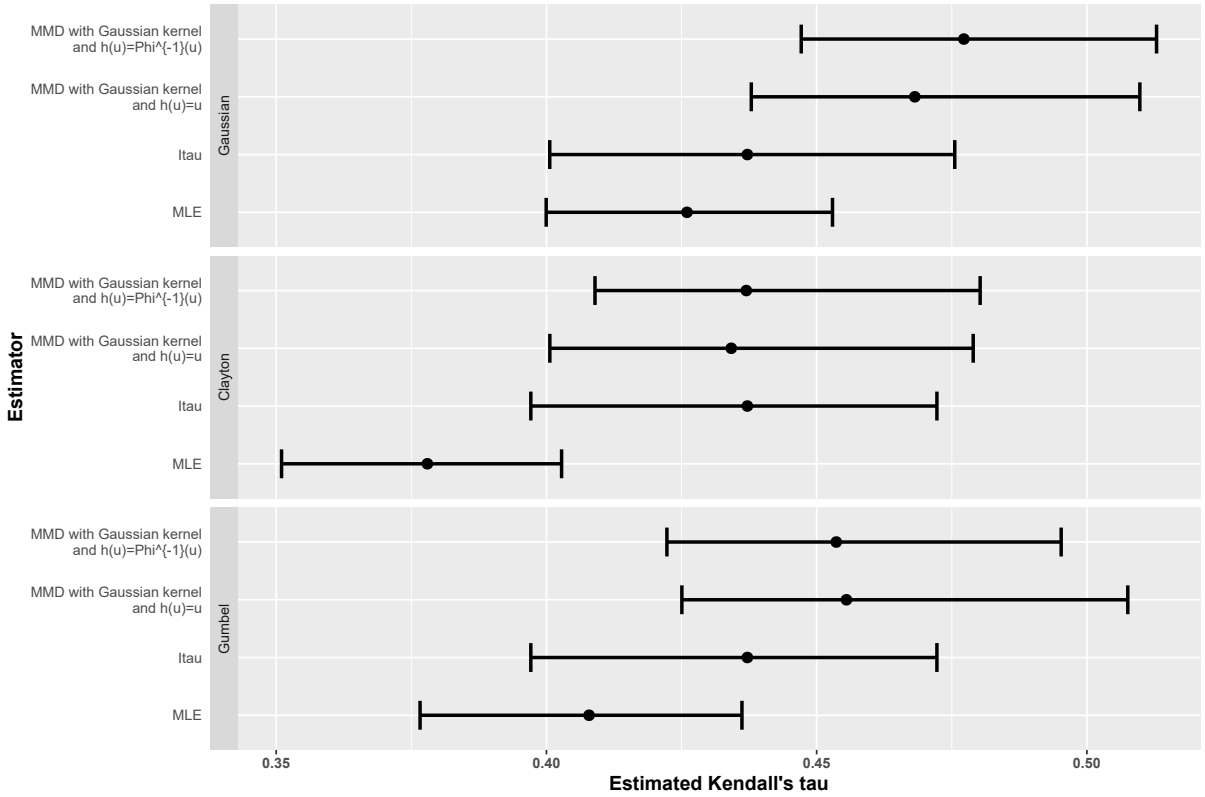

Figure 5: Confidence intervals for the implied Kendall's tau between APPL and GOOG stock returns 2008-2012, estimated by MMD, MLE, and ITau.

4 Conclusion

We have shown that the estimation of semiparametric copula models by MMD methods yields consistent, weakly convergent and robust estimators. In particular, when some outliers contaminate an assumed parametric underlying copula, the comparative advantages of our MMD estimator become patent.

To go further, many open questions would be of interest. For instance, extending our theory to manage time series should be feasible. Indeed, the theory of the weak

convergence of empirical copula processes for dependent data has been established in the literature (see [7], e.g.). Moreover, finding a formal data-driven way of choosing the kernel tuning-parameter γ would be useful. Finally, in the case of highly parameterized models - such as hierarchical Archimedean models (HAC), vines, or reliability models based on Marshall-Olkin copulas also called "fatal shock" models -, it could be interesting to introduce a penalization on θ, for example as

$$\tilde{\theta}_n := \arg\min_{\theta\in\Theta} \int K_U(\boldsymbol{u},\boldsymbol{v})\mathbb{P}_\theta^U(d\boldsymbol{u})\,\mathbb{P}_\theta^U(d\boldsymbol{v}) - \frac{2}{n}\sum_{i=1}^{n}\int K_U(\boldsymbol{u},\hat{\boldsymbol{U}}_i)\mathbb{P}_\theta^U(d\boldsymbol{u}) + \lambda\|\theta\|_1.$$

This idea would be different from the so-called "regularized MMD" in [13] that is reduced to multiplying the first term on the right-hand side of by a scaling factor. To the best of our knowledge, the asymptotic or finite distance theory for the penalized MMD estimator $\tilde{\theta}_n$ still does not exist. An interesting avenue for future research would be to fill this theoretical gap and to adapt this framework to copulas.

References

[1] Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. and Fermanian, J.-D. (2020). R package: MMDCopula. https://github.com/AlexisDerumigny/MMDCopula

[2] Alquier, P. and Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy. ArXiv preprint, arXiv:2006.00840.

[3] Baraud, Y., and Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae*, 2017.

[4] Briol, F.X., Barp, A., Duncan, A.B., and Girolami, M. (2019). Statistical Inference for Generative Models with Maximum Mean Discrepancy. ArXiv preprint, arXiv:1906.05944.

[5] Boucheron, S., Lugosi, G. and Massart, P. (2012). *Concentration inequalities. A nonasymptotic theory of independence.* Oxford University Press.

[6] Bücher, A. and Segers, J. and Volgushev, S. (2012). When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *The Annals of Statistics* 08-4 (42), 1598-1634.

[7] Bücher, A., and Volgushev, S. (2013). Empirical and sequential empirical copula processes under serial dependence. *Journal of Multivariate Analysis*, 119, 61-70.

[8] Chen, X., Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics* 33(2), 389-414.

[9] Chen, X. and Fan, Y. (2006) Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics* 135, 125-54.

[10] Chérief-Abdellatif, B.-E., and Alquier, P. (2019). Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. ArXiv preprint, arXiv:1912.05737.

[11] Chérief-Abdellatif, B.-E., and Alquier, P. (2020). MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy. Proceedings of "The 2nd Symposium on Advances in Approximate Bayesian Inference", PMLR 118:1-21.

[12] Christmann, A., and Steinwart, I. (2010). Universal kernels on non-standard input spaces. In *Advances in neural information processing systems* (pp. 406-414).

[13] Danafar, S., Rancoita, P., Glasmachers, T., Whittingstall, K., and Schmidhuber, J. (2013). Testing hypotheses by regularized maximum mean discrepancy. ArXiv preprint, arXiv:1305.0423.

[14] Denecke, L., and Müller, C.H. (2011). Robust estimators and tests for bivariate copulas based on likelihood depth. *Computational statistics and data analysis* 55(9), 2724-2738.

[15] Dziugaite, G.K., Roy, D.M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence.

[16] Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10 (5), 847–860.

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

[18] Goegebeur, Y., Guillou, A., Le Ho, N.K., and Qin, J. (2020). Robust nonparametric estimation of the conditional tail dependence coefficient. *Journal of Multivariate Analysis* 104607.

[19] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* 13, 723-773.

[20] Guerrier, S., Orso, S., and Victoria-Feser, M.P. (2013). Robust Estimation of Bivariate Copulas. Technical report. University of Geneva.

[21] Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2019). *Elements of copula modeling with R.* Springer.

[22] Kim, B., and Lee, S. (2013). Robust estimation for copula Parameter in SCOMDY models. *Journal of Time Series Analysis* 34(3), 302-314.

[23] Magnus, J.R., Neudecker, H. (1999). *Matrix differential calculus, with applications in statistics and econometrics.* Wiley.

[24] Mendes, B.V., de Melo, E.F., and Nelsen, R.B. (2007). Robust fits for copula models. *Communications in Statistics, Simulation and Computation* 36(5), 997-1017.

[25] Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* 10(1-2), 1-141.

[26] Nelsen, R.B. (2007). *An introduction to copulas.* Springer Science and Business Media.

[27] Newey, K.W., and McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV*, Edited by R.F. Engle and D.L. McFadden, 2112-2245.

[28] Radulović, D., Wegkamp, M., and Zhao, Y. (2017). Weak convergence of empirical copula processes indexed by functions. *Bernoulli* 23(4B), 3346-3384.

[29] Rousseeuw, P.J. and Hubert, M. (1999). Regression depth. *Journal of the American Statistical Association* 94, 388-402.

[30] Schepsmeier, U., and Stöber, J. (2014). Derivatives and Fisher information of bivariate copulas. *Statistical Papers* 55(2), 525-542.

[31] Schepsmeier, U., Stoeber, J., Brechmann, E.C., Graeler, B., Nagler, T., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., and Killiches, M. (2019). Package "VineCopula". R package, version 2.3.0.

[32] Shih, J.H., and Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 1384-1399.

[33] Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli* 18(3), 764-782.

[34] Sklar, A.(1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* 8, 229-231.

[35] Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics* 33(3), 357-375.

[36] Vaart, A.W. van der. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics.

[37] White, H. (1982). Maximum Likelihood estimation of misspecified models. *Econometrica* 50(1), 1-25.

[38] White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge University Press.

[39] Wuertz D., Setz T., Chalabi Y., Boudt C., Chausse P. and Miklovac, M. (2020). fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R package version 3042.83.2.

[40] Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 768-774.

A Proof of Theorem 3

According to Condition 4, L_n is twice differentiable on a neighborhood of θ_0^* and $\partial L_n/\partial\theta_j = n^{-1}\sum_{i=1}^n \partial\ell(\hat{\boldsymbol{U}}_i;\cdot)/\partial\theta_j$. Moreover, due to the the consistency of $\hat{\theta}_n$ (according to Conditions 1 and 2), we can assume that $\hat{\theta}_n$ belongs to such a neighborhood. Using Condition 3, the first-order condition is

$$0 = \nabla_\theta L_n(\hat{\theta}_n) = \nabla_\theta L_n(\theta_0^*) + \nabla_{\theta,\theta^T} L_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0^*), \tag{12}$$

where $\bar{\theta}_{n,j}$ is a vector whose components lie between those of θ_0^* and $\hat{\theta}_n$. Note that $H_n := \nabla_{\theta,\theta^T} L_n(\bar{\theta}_n)$ is an (d,d)-sized Hessian matrix whose (j,k)-th component is $H_{n,jk} = \frac{1}{n}\sum_{i=1}^n \partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_{n,j})/\partial\theta_k\partial\theta_j$, $j,k \in \{1,\dots,d\}$ Let us now study the asymptotic behavior of this Hessian matrix and of $\nabla_\theta L_n(\theta_0^*)$.

For any given coefficient (j,k), the function $\partial^2\ell(\boldsymbol{w};\cdot)/\partial\theta_j\partial\theta_k$ is continuous on the compact set K for $\mathbb{P}_0$ almost every $\boldsymbol{w} \in [0,1]^d$, all second-order functions $\partial^2\ell(\cdot;\theta)/\partial\theta_j\partial\theta_k$ are measurable for any $\theta \in K$ and $\mathbb{E}[\sup_{\theta\in K}|\partial^2\ell(\boldsymbol{U};\theta)/\partial\theta_k\partial\theta_j|] < +\infty$ (Conditions 4 and 5). Therefore, the L^1 bracketing numbers associated to the hessian maps indexed by $\theta \in K$ are finite, invoking Example 19.8 in [36]. Using Lemma 1(c) in [8], we get

$$\sup_{\theta\in K}\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\theta)}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta)}{\partial\theta_k\partial\theta_j}\right]\right| \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} 0.$$

As $\bar{\theta}_{n,j}$ lies between $\hat{\theta}_n$ and θ_0^*, $\bar{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} \theta_0^*$, and then, for n large enough, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right| \\
&\quad\le \left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j}\right]\right| + \left|\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j}\right] - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right| \\
&\quad\le \sup_{\theta\in K}\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\theta)}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta)}{\partial\theta_k\partial\theta_j}\right]\right| + \left|\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j}\right] - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right|.
\end{aligned}$$

The continuity of $\mathbb{E}[\partial^2\ell(\boldsymbol{U};\cdot)/\partial\theta_j\partial\theta_k]$ at θ_0^* (Condition 4) yields

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_{n,j})}{\partial\theta_k\partial\theta_j} \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\theta_j\theta_k}\right].$$

Finally, by definition of H_n and B, we obtain $H_n \xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.} B$.

According to Proposition 3.1 in [33] and under Condition 9, the empirical copula process $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)$ weakly converges to the Gaussian process $\alpha(\boldsymbol{w}) - \sum_{j=1}^{d} \dot{C}_j(\boldsymbol{w})\alpha_j(\boldsymbol{w}_j)$ in $\ell^\infty([0,1]^d)$ where α is a $\mathbb{P}_0$-Brownian bridge. By Condition 8 and an integration by parts argument (see e.g. [16]),

$$\begin{aligned}\sqrt{n}\big\{\nabla_\theta L_n(\theta_0^*) - \mathbb{E}[\nabla_\theta \ell(\boldsymbol{U};\theta_0^*)]\big\} &= \sqrt{n}\int_{[0,1]^d} \nabla_\theta \ell(\boldsymbol{w};\theta_0^*) d(\hat{\mathbb{P}}_n - \mathbb{P}_0)(\boldsymbol{w}) \\ &= (-1)^d \int_{[0,1]^d} \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)(\boldsymbol{w})\, \nabla_\theta \ell(d\boldsymbol{w};\theta_0^*). \qquad (13)\end{aligned}$$

Recalling Condition 7, since a continuous and linear transformation of a Gaussian process is normally distributed, the continuous mapping theorem implies that the weak limit of $\sqrt{n}\nabla_\theta L_n(\theta_0^*)$ exists, is centered and Gaussian:

$$\sqrt{n}\nabla_\theta L_n(\theta_0^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \int \Big\{\alpha(\boldsymbol{w}) - \sum_{j=1}^{d} \dot{C}_j(\boldsymbol{w})\alpha_j(\boldsymbol{w}_j)\Big\}\nabla_\theta \ell(d\boldsymbol{w};\theta_0^*).$$

As the limiting matrix B is invertible, we can infer that the matrix H_n is a.s. invertible for a sufficiently large n. Using Slutsky lemma and Formula (12), we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0^*) = H_n^{-1}\sqrt{n}\nabla_\theta L_n(\theta_0^*) \xrightarrow[n\to+\infty]{\mathcal{L}} B^{-1}\int \Big\{\alpha(\boldsymbol{w}) - \sum_{j=1}^{d} \dot{C}_j(\boldsymbol{w})\alpha_j(\boldsymbol{w}_j)\Big\}\nabla_\theta \ell(d\boldsymbol{w};\theta_0^*).$$

If Condition 9 is replaced by Condition 10, then the empirical process $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)$ weakly converges to the process $\alpha(\boldsymbol{w}) + dC_{(-\alpha_1,\ldots,-\alpha_d)}(\boldsymbol{w})$ in $L_p([0,1]^d)$ for any $1 \le p < \infty$, as detailed in [6] (Theorem 4.5. and the remarks that follow). Due to Condition 11 and Hölder's inequality, the map $h \to \int h(\boldsymbol{w})\,\nabla_\theta \ell(d\boldsymbol{w};\theta_0^*)$ is continuous on $L_p([0,1]^d)$, $1/p + 1/q = 1$. Therefore, by (13) and the continuous mapping theorem, the weak limit of $\sqrt{n}\big\{\nabla_\theta L_n(\theta_0^*) - \mathbb{E}[\nabla_\theta \ell(\boldsymbol{U};\theta_0^*)]\big\}$ exists and is $B^{-1}\int \Big\{\alpha(\boldsymbol{w}) + dC_{(-\alpha_1,\ldots,-\alpha_d)}(\boldsymbol{w})\Big\}\nabla_\theta \ell(d\boldsymbol{w};\theta_0^*)$, proving the result.

B Proof of Proposition 1

For every $\theta \in [-1, 1]$, some integration by parts imply

$$\begin{aligned}\ell(\boldsymbol{w}; \theta) &= \int K_U(\boldsymbol{u}, \boldsymbol{v})\, C_\theta(d\boldsymbol{u})\, C_\theta(d\boldsymbol{v}) - 2 \int K_U(\boldsymbol{u}, \boldsymbol{w})\, C_\theta(d\boldsymbol{u}) \\ &= \int C_\theta(\boldsymbol{u}) C_\theta(\boldsymbol{v})\, K_U(d\boldsymbol{u}, d\boldsymbol{v}) - 2 \int C_\theta(\boldsymbol{u})\, K_U(d\boldsymbol{u}, \boldsymbol{w}). \qquad (14)\end{aligned}$$

Indeed, apply Theorem 15 in [28], by noting that all the terms involving an integration w.r.t. the measure K_U or its derivative (as $K_U(du_1, u_2, \boldsymbol{v})$, $K_U(d\boldsymbol{u}, \boldsymbol{v})$, $K_U(d\boldsymbol{u}, d\boldsymbol{v})$, etc) cancel when one free argument is zero or one. This is a special and nice property of our Gaussian-type kernel K_U. For every θ in a sufficiently small open neighborhood of θ_0, $\theta \in (-1, 1)$, copula densities exist and we have

$$\ell(\boldsymbol{w}; \theta) = \int K_U(\boldsymbol{u}, \boldsymbol{v}) c_\theta(\boldsymbol{u}) c_\theta(\boldsymbol{v})\, d\boldsymbol{u}\, d\boldsymbol{v} - 2 \int K_U(\boldsymbol{u}, \boldsymbol{w}) c_\theta(\boldsymbol{u})\, d\boldsymbol{u}.$$

Let us check that all conditions 1-9 are satisfied in this case, to apply Theorem 3.

- Condition 1: obviously, $\Theta = [-1, 1]^2$ is compact. Use the identity (14) and the dominated convergence theorem to prove that the map $\theta \mapsto L_0(\theta)$ is continuous on Θ. Indeed, $K_U(d\boldsymbol{u}, d\boldsymbol{v}) = K_U(\boldsymbol{u}, \boldsymbol{v}) Q(\boldsymbol{u}, \boldsymbol{v}) d\boldsymbol{u}\, d\boldsymbol{v}$, for some polynomial Q, and then $\int |K_U|(d\boldsymbol{u}, d\boldsymbol{v}) < \infty$. Note that it is even true at the boundaries of Θ. Moreover, $L_0(\cdot)$ is uniquely minimized at θ_0. Indeed, $L_0(\theta)$ is equal to the MMD distance between C_θ and C_{θ_0} (up to a constant), which is minimized at θ_0 and nowhere else due to the identifiability of the Gaussian family and knowing that our kernel is characteristic.

- Condition 2: The envelope function of the family of functions $\boldsymbol{w} \mapsto \ell(\boldsymbol{w}, \theta)$ is integrable: for every $\theta \in \Theta$ and since any copula is less than one,

$$\sup_{\theta \in \Theta} |\ell(\boldsymbol{w}; \theta)| \leq \int |K_U|(d\boldsymbol{u}, d\boldsymbol{v}) + 2 \int |K_U(d\boldsymbol{u}, \boldsymbol{w})|,$$

that is integrable because K_U and its partial derivatives are integrable. As before, use again the identity (14) and the dominated convergence theorem to show that $\theta \mapsto \ell(\boldsymbol{w}, \theta)$ is continuous on Θ.

- Condition 3 is satisfied with our choice $-1 < \theta_0 < 1$.

- Condition 4 is satisfied because $\ell(\boldsymbol{w};\theta) = \mathcal{I}(\theta,\theta) - 2\mathcal{J}(\theta, \Phi^{-1}(w_1), \Phi^{-1}(w_2))$, with the notations of Section D. Such analytic expression are clearly two times continuously differentiable w.r.t. $\theta \in \mathcal{O}$, for some $\mathcal{O} :=]\theta_0 - \eta, \theta_0 + \eta[\subset [-1,1]$, $\eta > 0$ and for every $\boldsymbol{w} \in (0,1)^2$. Moreover, with the notations of Section D, we get

$$\mathbb{E}[\nabla^2_{\theta,\theta}\ell(\boldsymbol{U};\theta)] = \nabla^2_{\theta,\theta}\mathbb{E}[\ell(\boldsymbol{U};\theta)] = \nabla^2_{\theta,\theta}\big\{\mathcal{I}(\theta,\theta) - 2\mathcal{I}(\theta,\theta_0)\big\}_{\theta=\theta_0}, \tag{15}$$

  that can be analytically calculated for every $\theta \in (-1,1)$. The latter function is obviously a continuous map of $\theta \in (-1,1)$, and then particularly at θ_0. Condition 5 is then a consequence of (15) and the formulas of Section D.

- Condition 6 is $0 < B = \mathbb{E}[\nabla^2_{\theta,\theta}\ell(\boldsymbol{U};\theta_0)] < +\infty$. The calculation of B is of interest, because it would yield an analytic form for the asymptotic variance of $\hat{\theta}_n$. As noted before, B can be deduced from the map $\theta \mapsto \mathbb{E}[\ell(\boldsymbol{U};\theta)]$, after calculating the second derivative of the latter function, evaluated at $\theta = \theta_0$. Since

$$\mathbb{E}[\ell(\boldsymbol{U};\theta)] = \mathcal{I}(\theta,\theta) - 2\mathcal{I}(\theta,\theta_0) = I(\theta) - I\big((\theta+\theta_0)/2\big),$$

  with $I(\theta) = \gamma^2\{(2+\gamma^2/2)^2 - 4\theta^2\}^{-1/2}/2$, we deduce $B = 3I''(\theta_0)/4$. Simple calculations yield

$$I'(\theta) = \frac{2\theta\gamma^2}{\{(2+\gamma^2/2)^2 - 4\theta^2\}^{3/2}} \quad \text{and} \quad I''(\theta) = \frac{2\gamma^2\big\{(2+\gamma^2/2)^2 + 8\theta^2\big\}}{\{(2+\gamma^2/2)^2 - 4\theta^2\}^{5/2}},$$

  that is strictly positive.

- Condition 7 is obviously satisfied (first-order conditions).

- Condition 8: to show that the gradient of the loss $\nabla_\theta\ell(\cdot;\theta_0)$ is of bounded variations, it is sufficient to show that the mixed partial derivative $\boldsymbol{w} \mapsto \nabla^3_{\theta,1,2}\ell(\boldsymbol{w};\theta_0)$ is integrable on $[0,1]^2$, and also that the functions $w_1 \mapsto \nabla^2_{\theta,w_1}\ell(w_1,1;\theta_0)$ and

$w_2 \mapsto \nabla^2_{\theta,w_2}\ell(1,w_2;\theta_0)$ are integrable on $[0,1]$. Direct calculations provide

$$\begin{aligned}
\int |\nabla^3_{\theta,1,2}\ell(\boldsymbol{w};\theta_0)|d\boldsymbol{w} &= 2\int \left|\frac{\partial^2 K_U}{\partial w_1 \partial w_2}(\boldsymbol{w},\boldsymbol{u})\nabla_{\theta_0} c_{\theta_0}(u_1,u_2)\right| d\boldsymbol{u}d\boldsymbol{w} \\
&\leq 2\int \frac{|\{\Phi^{-1}(w_1)-\Phi^{-1}(u_1)\}\{\Phi^{-1}(w_2)-\Phi^{-1}(u_2)\}|}{(\gamma^2/2)^2\phi(\Phi^{-1}(w_1))\phi(\Phi^{-1}(w_2))} K_U(\boldsymbol{w},\boldsymbol{u})|\nabla_{\theta_0} c_{\theta_0}(u_1,u_2)|d\boldsymbol{u}d\boldsymbol{w} \\
&\leq \int \frac{|\{x_1-y_1\}\{x_2-y_2\}|}{\gamma^4\pi\sqrt{1-\theta_0^2/4}} \exp\Big(-\frac{(x_1-y_1)^2+(x_2-y_2)^2}{\gamma^2}\Big) \\
&\quad \times \frac{(1+|x_1x_2|+x_1^2+x_2^2)}{(1-\theta_0^2)^2} \exp\Big(-\frac{x_1^2+x_2^2-2\theta_0 x_1 x_2}{2(1-\theta_0^2)}\Big)\, d\boldsymbol{x}\, d\boldsymbol{y},
\end{aligned}$$

that is finite. Indeed, the latter term is less than the expectation of $Q(X_1,X_2,Y_1,Y_2)$ for a particular fourth-order polynomial Q and a four dimension Gaussian random vector (X_1,X_2,Y_1,Y_2). Moreover, both functions $w_1 \mapsto \nabla^2_{\theta,w_1}\ell(w_1,1;\theta_0)$ and $w_2 \mapsto \nabla^2_{\theta,w_2}\ell(1,w_2;\theta_0)$ are zero on $[0,1]$ as the first and second partial derivatives of the kernel K_U are equal to 0 at $(u_1,1,u_3,u_4)$ and $(1,u_2,u_3,u_4)$ respectively (for any $(u_1,u_2,u_3,u_4)\in[0,1]^4$). Therefore,

$$\int |\nabla^2_{\theta,w_1}\ell(w_1,1;\theta_0)|dw_1 = 2\int \left|\frac{\partial K_U}{\partial w_1}(w_1,1,\boldsymbol{u})\nabla_{\theta_0} c_{\theta_0}(\boldsymbol{u})\right| d\boldsymbol{u}d\boldsymbol{w} = 0,$$

and similarly for $\int |\nabla^2_{\theta,w_2}\ell(1,w_2;\theta_0)|dw_2$.

- Condition 9 is satisfied for the Gaussian copula when $|\theta_0|<1$: see Example 5.1 in [33].

C Proof of Proposition 2

To apply Theorem 3, it is sufficient to check that the conditions 1-8 and 10-11 are satisfied. To calculate $\ell(\cdot;\theta)$ and $L_0(\theta)$, we rely on the formulas (9), (10) and (11). Note that we will restrict ourselves to parameters α and β into $[\epsilon,1-\epsilon]$. Therefore,

$$\bar{u}^* := \max_{(\alpha,\beta)\in\Theta} \bar{u}_{\alpha,\beta} < 1, \text{ and } \bar{u}_* := \min_{(\alpha,\beta)\in\Theta} \bar{u}_{\alpha,\beta} > 0.$$

It can be checked that the map $\bar{u}:(\alpha,\beta)\mapsto \bar{u}_{\alpha,\beta} = (\beta/\alpha)^{\beta/(\alpha-\beta)}$ from Θ to $\mathbb{R}$ is two times continuously differentiable. To this goal, it is necessary to extend the map $\bar{u}$

by continuity, setting $\bar{u}(\alpha,\alpha) = e^{-1}$, $\partial_1\bar{u}(\alpha,\alpha) = e^{-1}/(2\alpha)$, $\partial_2\bar{u}(\alpha,\alpha) = -e^{-1}/(2\alpha)$, $\partial^2_{1,1}\bar{u}(\alpha,\alpha) = -5e^{-1}/(12\alpha^2)$, $\partial^2_{2,2}\bar{u}(\alpha,\alpha) = 7e^{-1}/(12\alpha^2)$ and $\partial^2_{1,2}\bar{u}(\alpha,\alpha) = -5e^{-1}/(12\alpha^2)$.

For any continuous and bounded map $\psi : [0,1]^2 \mapsto \mathbb{R}$, we recall that

$$\mathbb{E}_\theta[\psi(U_1,U_2)] = \int \psi(s,t)\big\{(1-\alpha)s^{-\alpha}\mathbf{1}(s^\alpha > t^\beta) + (1-\beta)t^{-\beta}\mathbf{1}(s^\alpha < t^\beta)\big\}\,ds\,dt$$
$$+ \int_0^{\bar{u}_{\alpha,\beta}} \psi(u,u^{\alpha/\beta})\,\beta u^{1-\alpha}\,du + \int_{\bar{u}_{\alpha,\beta}}^1 \psi(u,u^{\alpha/\beta})\,\alpha u^{\alpha/\beta-\alpha}\,du,$$

that can be seen as the integral of a map $(s,t) \mapsto g_\theta(s,t)$ on $[0,1]^2$ w.r.t. the Lebesgue measure (single integrals are particular cases of double integrals!). Such maps are continuous a.e., and

$$\sup_{\theta\in\Theta}|g_\theta|(s,t) \le \|\psi\|_\infty\big\{s^{\epsilon-1} + t^{\epsilon-1} + 2\big\}.$$

The function on the r.h.s. of the latter equation is integrable on $[0,1]^2$ w.r.t. the Lebesgue measure. By dominated convergence, we deduce the map $\theta \mapsto \ell(\boldsymbol{w},\theta) = \mathbb{E}_\theta\big[K_U(\boldsymbol{U},\boldsymbol{V})\big] - 2\mathbb{E}_\theta\big[K_U(\boldsymbol{U},\boldsymbol{w})\big]$ is continuous on Θ for every $\boldsymbol{w}$. The same arguments apply for $\theta \mapsto L_0(\theta)$ when $\theta \in \Theta$. We deduce that Conditions 1 and 2 are satisfied, and $\hat{\theta}_n$ is consistent.

To check Condition 4, we have to prove that the calculations of derivatives of $\ell(\boldsymbol{w},\theta)$ w.r.t. θ are permitted inside our integral signs. Such integrands are indeed two times continuously differentiable w.r.t. $\theta \in \Theta$ for almost all their other arguments into the interior of their domains. Moreover, they are upper bounded by some integrable envelope functions. Then, the dominated convergence theorem applies. Nonetheless, since these integrands are often integrals themselves, it may be necessary to rely on the dominated convergence theorem again to state continuity. The calculations of such derivatives induce many terms, but the same technique applies to all. We will illustrate the arguments on a few of them.

For instance, the θ-derivative of $\ell(\boldsymbol{w},\theta)$ involves the derivative of

$$\theta \mapsto I_1(\theta) := \int u_1^{-\alpha}v_1^{-\alpha}\Big\{\int K_U(\boldsymbol{u},\boldsymbol{v})\mathbf{1}(u_1^{\alpha/\beta} > u_2, v_1^{\alpha/\beta} > v_2)\,du_2\,dv_2\Big\}\,du_1\,dv_1$$
$$=: \int \tilde{I}_1(\theta;x_1,y_1)\,dx_1\,dy_1,$$

after the change of variables $u_k = \Phi(x_k)$, $v_k = \Phi(y_k)$, $k = 1, 2$, denoting

$$\tilde{I}_1(\theta; x_1, y_1) = J_1(\theta; x_1, y_1) \exp\Big(- \frac{(x_1 - y_1)^2}{\gamma^2}\Big) \Phi(x_1)^{-\alpha} \Phi(y_1)^{-\alpha} \phi(x_1)\phi(y_1),$$

$$J_1(\theta; x_1, y_1) := \int \exp\Big(- \frac{(x_2 - y_2)^2}{\gamma^2}\Big) \mathbf{1}\big(x_2 \leq x_{1,\alpha/\beta}; y_2 \leq y_{1,\alpha/\beta}\big) \phi(x_2)\phi(y_2)\, dx_2\, dy_2,$$

and $t_a := \Phi^{-1}\big(\Phi(t)^a\big)$ for any real numbers a and t. The map $\theta \mapsto J_1(\theta; x_1, y_1)$ is $C^2(\Theta)$ for every $(x_1, y_1) \in (0, 1)^2$. Its α-derivative is

$$\begin{aligned}\partial_\alpha J_1(\theta; x_1, y_1) = {} & \phi(x_{1,\alpha/\beta}) \partial_\alpha \{\Phi^{-1}(\Phi(x_1)^{\alpha/\beta})\} \int \exp\Big(- \frac{(x_{1,\alpha/\beta} - y_2)^2}{\gamma^2}\Big) \mathbf{1}\big(y_2 \leq x_{y,\alpha/\beta}\big) \phi(y_2)\, dy_2 \\ + {} & \phi(y_{1,\alpha/\beta}) \partial_\alpha \{\Phi^{-1}(\Phi(y_1)^{\alpha/\beta})\} \int \exp\Big(- \frac{(x_2 - y_{1,\alpha/\beta})^2}{\gamma^2}\Big) \mathbf{1}\big(x_2 \leq x_{1,\alpha/\beta}\big) \phi(x_2)\, dy_2.\end{aligned}$$

Note that $\phi(t_{\alpha/\beta})\ \partial_\alpha\{\Phi^{-1}(\Phi(t)^{\alpha/\beta})\} = \Phi(t)^{\alpha/\beta} \ln \Phi(t)/\beta$ for every t. Clearly, we have $|J_1(\theta; x_1, y_1)| \leq 1$ and

$$|\partial_\alpha J_1(\theta; x_1, y_1)| \leq \big\{\Phi(x_1)^{\alpha/\beta} |\ln \Phi(x_1)| + \Phi(y_1)^{\alpha/\beta} |\ln \Phi(y_1)|\big\}/\varepsilon,$$

for every (x_1, y_1). This yields

$$\begin{aligned}\partial_\alpha \tilde{I}_1(\theta; x_1, y_1) = {} & \exp\Big(- \frac{(x_1 - y_1)^2}{\gamma^2}\Big) \Phi(x_1)^{-\alpha} \Phi(y_1)^{-\alpha} \phi(x_1)\phi(y_1) \\ \times {} & \Big\{\partial_\theta J_1(\theta; x_1, y_1) - J_1(\theta; x_1, y_1) \ln\big(\Phi(x_1)\Phi(y_1)\big)\Big\}, \text{ and}\end{aligned}$$

$$\sup_{\theta \in \Theta} |\partial_\alpha \tilde{I}_1(\theta; x_1, y_1)| \leq 2\big(1 + |\ln\big(\Phi(x_1)\Phi(y_1)\big|\big) \Phi(x_1)^{\epsilon - 1} \Phi(y_1)^{\epsilon - 1} \phi(x_1)\phi(y_1)/\varepsilon,$$

that is integrable on $\mathbb{R}^2$. The same reasoning can be led for the derivative w.r.t β. This means that $\theta \mapsto I_1(\theta)$ is differentiable and its derivative is given by $\int \nabla_\theta \tilde{I}_1(\theta; x_1, y_1)\, dx_1\, dy_1$.

Another term of $\ell(\boldsymbol{w}; \theta)$ is

$$\theta \mapsto I_2(\theta) := \int u^{1-\alpha} v_1^{-\alpha} K_U(u, u^{\alpha/\beta}, \boldsymbol{v}) \mathbf{1}(u < \bar{u}_{\alpha,\beta}, v_2 < v_1^{\alpha/\beta})\, du\, d\boldsymbol{v} =: \int \tilde{I}_2(\theta; y_1)\, dy_1,$$

after the change of variables $u = \Phi(x)$, $v_k = \Phi(y_k)$, $k = 1, 2$ and setting $\tilde{I}_2(\theta; y_1) := J_2(\theta; y_1)\phi(y_1)\Phi(y_1)^{-\alpha}$ with

$$\begin{aligned}J_2(\theta; y_1) := {} & \int \exp\Big(- \frac{(x - y_1)^2}{\gamma^2} - \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta}) - y_2)^2}{\gamma^2}\Big) \\ \times {} & \Phi(x)^{1-\alpha} \mathbf{1}\big(x < \Phi^{-1}(\bar{u}_{\alpha,\beta}), y_2 < y_{1,\alpha/\beta}\big) \phi(x)\phi(y_2)\, dx\, dy_2.\end{aligned}$$

Clearly, $|J_2(\theta; y_1)|$ is less than $\int \Phi(t)^{\varepsilon-1}\phi(t)\,dt < \infty$. Moreover, $\partial_\alpha J_2(\theta; y_1)$ can be obtained by a derivation inside the integral sign, i.e.

$$\begin{aligned}
\partial_\alpha J_2(\theta; y_1) = &\int \exp\Big(-\frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_1)^2}{\gamma^2} - \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_2)^2}{\gamma^2}\Big)\bar{u}_{\alpha,\beta}^{1-\alpha} \\
&\times \mathbf{1}\big(y_2 < y_{1,\alpha/\beta}\big)\phi(\Phi^{-1}(\bar{u}_{\alpha,\beta}))\phi(y_2)\,dy_2\;\partial_\alpha\{\Phi^{-1}(\bar{u}_{\alpha,\beta})\} \\
+ &\int \exp\Big(-\frac{(x-y_1)^2}{\gamma^2} - \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_{1,\alpha/\beta})^2}{\gamma^2}\Big)\Phi(x)^{1-\alpha} \\
&\times \mathbf{1}\big(x < \Phi^{-1}(\bar{u}_{\alpha,\beta})\big)\phi(x)\phi(y_{1,\alpha/\beta})\,dx\;\partial_\alpha\{\Phi^{-1}(\Phi(y_1)^{\alpha/\beta})\} \\
- &\int \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_2)}{\gamma^2/2}\exp\Big(-\frac{(x-y_1)^2}{\gamma^2} - \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_2)^2}{\gamma^2}\Big)\Phi(x)^{1-\alpha} \\
&\times \mathbf{1}\big(x < \Phi^{-1}(\bar{u}_{\alpha,\beta}), y_2 < y_{1,\alpha/\beta}\big)\phi(x)\phi(y_2)\,dx\,dy_2\;\partial_\alpha\{\Phi^{-1}(\bar{u}_{\alpha,\beta})\} \\
- &\int \exp\Big(-\frac{(x-y_1)^2}{\gamma^2} - \frac{(\Phi^{-1}(\bar{u}_{\alpha,\beta})-y_2)^2}{\gamma^2}\Big)\Phi(x)^{1-\alpha}\ln\Phi(x) \\
&\times \mathbf{1}\big(x < \Phi^{-1}(\bar{u}_{\alpha,\beta}), y_2 < y_{1,\alpha/\beta}\big)\phi(x)\phi(y_2)\,dx\,dy_2, \qquad (16)
\end{aligned}$$

for every $\theta \in \Theta$ and every $y_1 \in \mathbb{R}$. Indeed, the map $\theta \mapsto \partial_\alpha\{\Phi^{-1}(\bar{u}_{\alpha,\beta})\}$ is bounded on the compact subset Θ. Moreover, $\phi(y_{1,\alpha/\beta})\;\partial_\alpha\{\Phi^{-1}(\Phi(y_1)^{\alpha/\beta})\} = \Phi(y_1)^{\alpha/\beta}\ln\Phi(y_1)/\beta$. In other words, the dominated convergence theorem can be applied to prove formula (16), yielding the continuity of $\theta \mapsto \partial_\alpha J_2(\theta; y_1)$ and then of $\theta \mapsto \partial_\alpha \tilde{I}_2(\theta; y_1)$. We deduce

$$\sup_{\theta\in\Theta}|\partial_\alpha \tilde{I}_2(\theta; y_1)| < M\big(|\ln\Phi(y_1)|+1\big)\Phi(y_1)^{2(\varepsilon-1)}\phi(y_1),$$

that is integrable (M denotes a constant). Doing the same task with β-derivatives, we obtain that $\theta \mapsto I_2(\theta)$ is differentiable and its derivative is given by $\int \nabla_\theta \tilde{I}_2(\theta; y_1)\,dy_1$. Then, Condition 4 has been checked.

Conditions 5 and 6 can be obtained by the same type of reasonings. Nonetheless, we do not exclude that B could be not invertible for particularly unhappy choices of (α, β, γ). Since the latter set of parameters is the roots of some analytic expression $H(\alpha, \beta, \gamma) = 0$, its Lebesgue measure is zero. Due to the regularity of L_0 and the correct model specification, Condition 7 is fulfilled.

Again, Condition 8 is satisfied still by the same techniques. Indeed, it is sufficient to show that $\boldsymbol{w} \mapsto \nabla^2_{\theta,\boldsymbol{w}}|\ell(\boldsymbol{w}, \theta)|$ is integrable on $[0,1]^2$ w.r.t. the Lebesgue measure

and that $w_1 \mapsto \nabla^2_{\rho,w_1}\ell(w_1,1;\rho_0)$ and $w_2 \mapsto \nabla^2_{\rho,w_2}\ell(1,w_2;\rho_0)$ are integrable on $[0,1]$ w.r.t. the Lebesgue measure. By differentiating $\boldsymbol{w} \mapsto \ell(\boldsymbol{w},\theta)$, all the calculations above are similar, except that our terms $\exp\big(-\{(x_1-a)^2+(x_2-b)^2\}/(\gamma^2)\big)$ are replaced by $Q(x_1,x_2,a,b)\exp\big(-\{(x_1-a)^2+(x_2-b)^2\}/(\gamma^2)\big)$ for some polynomial Q. This does not significantly deteriorate the upper bounds we have exhibited for the study of $\theta \mapsto \nabla_\theta\ell(\boldsymbol{w};\theta)$. The same arguments apply to check Condition 11.

Finally, Condition 10 is obviously satisfied because the curve $\mathfrak{C}$ has Lebesgue measure zero on the plane.

D MMD criterion for a bivariate Gaussian copula model

Here, we explicitly write our MMD criterion in the case of bivariate Gaussian copulas. Recall that the density of a Gaussian copula in dimension two is

$$c_\theta(u_1,u_2) := \frac{1}{2\pi\sqrt{1-\theta^2}\phi(x_1)\phi(x_2)}\exp\Big(-\frac{1}{2(1-\theta^2)}\big(x_1^2+x_2^2-2\theta x_1x_2\big)\Big),$$

by setting $x_k = \Phi^{-1}(u_k)$, $k=1,2$. Define $\boldsymbol{x} := (x_1,x_2)$. Similarly, $y_k = \Phi^{-1}(v_k)$, $k=1,2$ and $\boldsymbol{y} := (y_1,y_2)$. For obtaining closed form formulas, it is necessary to select an adapted kernel. Here, we use the Gaussian-type kernel (7), with $h = \Phi^{-1}$.

Now, let us analytically specify the criterion in (1). First, let us calculate

$$\mathcal{I}(\theta_1,\theta_2) := \int K_U(\boldsymbol{u},\boldsymbol{v})\, C_{\theta_1}(d\boldsymbol{u})\, C_{\theta_2}(d\boldsymbol{v}),\ (\theta_1,\theta_2)\in(-1,1)^2.$$

By a change of variable, note that

$$\mathcal{I}(\theta_1,\theta_2) := \mathbb{E}\Big[\exp\big(-\frac{(X_1-Y_1)^2+(X_2-Y_2)^2}{\gamma^2}\big)\Big],$$

for a Gaussian centered random vector (X_1,X_2,Y_1,Y_2) whose 4×4 covariance matrix is block-diagonal. Its first (resp. second) 2×2 block is a correlation matrix with an extra-diagonal coefficient θ_1 (resp. θ_2). Therefore, the bivariate random vector $(Z_1,Z_2) := (X_1-Y_1, X_2-Y_2)/\sqrt{2}$ is centered Gaussian and its covariance matrix is a correlation matrix with an extra-diagonal coefficient $s := (\theta_1+\theta_2)/2$. Since the conditional law of Z_1 given

$Z_2 = z_2$ is $\mathcal{N}(sz_2, 1-s^2)$, we can easily calculate $\psi(z) := \mathbb{E}\big[\exp(-Z_1^2/(\gamma^2/2))|Z_2 = z\big]$. Indeed, setting $\tau^2 := \{1/\gamma^2 + 1/(1-s^2)\}^{-1}$, we have

$$\begin{aligned}\psi(z) &= \int \exp\big(-\frac{t^2}{\gamma^2/2}\big)\exp\Big(-\frac{(t-sz)^2}{2(1-s^2)}\Big)\frac{dt}{\sqrt{2\pi}\sqrt{1-s^2}} \\ &= \int \exp\big(-\frac{t^2}{2\tau^2} + \frac{stz}{1-s^2}\big)\frac{dt}{\sqrt{2\pi}\sqrt{1-s^2}}\exp\big(-\frac{s^2z^2}{2(1-s^2)}\big) \\ &= \frac{\gamma/\sqrt{2}}{\sqrt{2(1-s^2)+\gamma^2/2}}\exp\Big(-\frac{s^2z^2}{2(1-s^2)+\gamma^2/2}\Big).\end{aligned}$$

We deduce

$$\begin{aligned}\mathcal{I}(\theta_1,\theta_2) &= \mathbb{E}\Big[\exp\big(-\frac{Z_1^2+Z_2^2}{\gamma^2/2}\big)\Big] = \mathbb{E}_{Z_2}\Big[\exp\big(-\frac{Z_2^2}{\gamma^2/2}\big)\mathbb{E}\big[\exp\big(-\frac{Z_1^2}{\gamma^2/2}\big)|Z_2\big]\Big] \\ &= \int \exp\big(-\frac{t^2}{\gamma^2/2}\big)\psi(t)\phi(t)\,dt = \gamma^2/2\big\{(2+\gamma^2/2)^2 - 4s^2\big\}^{-1/2} =: I(s).\end{aligned}$$

Moreover, the other integrals in (1) are as

$$\int K_U(\boldsymbol{u}, \hat{\boldsymbol{U}}_i)c_\theta(\boldsymbol{u})\,d\boldsymbol{u} = \mathbb{E}\Big[\exp\big(-\frac{(X_1-\Phi^{-1}(\hat{U}_{i,1}))^2 + (X_2-\Phi^{-1}(\hat{U}_{i,2}))^2}{\gamma^2}\big)\Big],$$

for some standardized bivariate Gaussian random vector (X_1, X_2), $\mathbb{E}[X_1X_2] = \theta$. For any real numbers (a, b), standard arguments yield

$$\begin{aligned}\mathcal{J}(\theta,a,b) &= \mathbb{E}\Big[\exp\big(-\frac{(X_1-a)^2+(X_2-b)^2}{\gamma^2}\big)\Big] \\ &= \mathbb{E}\Big[\exp\big(-\frac{(X_2-b)^2}{\gamma^2}\big)\mathbb{E}\big[\exp\big(-\frac{(X_1-a)^2}{\gamma^2}\big)|X_2\big]\Big] \\ &= \frac{\gamma/\sqrt{2}}{\sqrt{1+\gamma^2/2-\theta^2}}\mathbb{E}\Big[\exp\big(-\frac{(X_2-b)^2}{\gamma^2}\big)\exp\big(-\frac{(\theta X_2-a)^2}{2(1+\gamma^2/2-\theta^2)}\big)\Big] \\ &= \frac{\gamma/\sqrt{2}}{\sqrt{1+\gamma^2/2-\theta^2}}\int \exp\big(-\frac{x^2}{2g^2} + \frac{\lambda x}{g^2} - \frac{b^2}{\gamma^2} - \frac{a^2}{2(1+\gamma^2/2-\theta^2)}\big)\frac{dx}{\sqrt{2\pi}} \\ &= \frac{g\gamma/\sqrt{2}}{\sqrt{1+\gamma^2/2-\theta^2}}\exp\Big(\frac{\lambda^2}{2g^2} - \frac{b^2}{\gamma^2} - \frac{a^2}{2(1+\gamma^2/2-\theta^2)}\Big),\end{aligned}$$

by setting

$$\frac{1}{g^2} := \frac{1}{\gamma^2/2} + \frac{\theta^2}{1+\gamma^2/2-\theta^2} + 1,\ \frac{\lambda}{g^2} := \frac{b}{\gamma^2/2} + \frac{a\theta}{1+\gamma^2/2-\theta^2}.$$

Therefore, the estimated parameter of the bivariate Gaussian copula is

$$\hat{\theta}_n = \arg\min_{\theta} \mathcal{I}(\theta,\theta) - 2n^{-1}\sum_{i=1}^{n} \mathcal{J}\big(\theta, \Phi^{-1}(\hat{U}_{i,1}), \Phi^{-1}(\hat{U}_{i,2})\big).$$

Note that generalizations of the latter calculations in larger dimensions would be quite cumbersome. Finally, let us notice that

$$\begin{aligned}\mathbb{D}^2(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &= \mathcal{I}(\theta_1,\theta_1) + \mathcal{I}(\theta_2,\theta_2) - 2\mathcal{I}(\theta_1,\theta_2) \\ &= f(\theta_1) + f(\theta_2) - 2f\left(\frac{\theta_1+\theta_2}{2}\right),\end{aligned}$$

where

$$f(x) = \frac{\gamma^2/2}{\sqrt{(2+\gamma^2/2)^2 - 4x^2}}\cdot$$

As, for any $x \in (-1,1)$,

$$\begin{aligned}f''(x) &= \frac{3x^2\gamma^2/2}{\left((2+\gamma^2/2)^2 - 4x^2\right)^{5/2}} + \frac{\gamma^2/2}{\left((2+\gamma^2/2)^2 - 4x^2\right)^{3/2}} \\ &\geq \frac{\gamma^2/2}{\left(2+\gamma^2/2\right)^3} =: \alpha(\gamma),\end{aligned}$$

we obtain that f is $\alpha(\gamma)$-strongly convex. This leads to

$$f\left(\frac{\theta_1+\theta_2}{2}\right) \leq \frac{f(\theta_1)+f(\theta_2)}{2} - \frac{\alpha(\gamma)}{8}(\theta_1-\theta_2)^2,$$

that implies

$$(\theta_1-\theta_2)^2 \leq \frac{4}{\alpha(\gamma)}\mathbb{D}^2(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}).$$

Therefore, we have obtained $|\theta_1 - \theta_2| \leq 2\mathbb{D}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})/\sqrt{\alpha(\gamma)}$, which proves the claim in Example 1, setting

$$c(\gamma) = \frac{2}{\sqrt{\alpha(\gamma)}} = \frac{(4+\gamma^2)^{3/2}}{\gamma}\cdot$$