# Estimation of Copulas via Maximum Mean Discrepancy

Alquier, Pierre; Chérief-Abdellatif, Badr Eddine; Derumigny, Alexis; Fermanian, Jean David

**DOI**
10.1080/01621459.2021.2024836

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Estimation of Copulas via Maximum Mean Discrepancy

Pierre Alquier, Badr-Eddine Chérief-Abdellatif, Alexis Derumigny & Jean-David Fermanian

**To cite this article:** Pierre Alquier, Badr-Eddine Chérief-Abdellatif, Alexis Derumigny & Jean-David Fermanian (2022): Estimation of Copulas via Maximum Mean Discrepancy, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.2024836

**To link to this article:** https://doi.org/10.1080/01621459.2021.2024836

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2021.2024836



# Estimation of Copulas via Maximum Mean Discrepancy

Pierre Alquier[a], Badr-Eddine Chérief-Abdellatif[b], Alexis Derumigny[c], and Jean-David Fermanian[d]

[a]RIKEN AIP, Tokyo, Japan; [b]Department of Statistics, University of Oxford, Oxford, UK; [c]Department of Applied Mathematics, Delft University of Technology, Delft, The Netherlands; [d]CREST, ENSAE, Institut Polytechnique de Paris, Palaiseau, France

**ABSTRACT**

This article 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 nonasymptotic 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. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received October 2020
Accepted December 2021

**KEYWORDS**
Algorithms semiparametric inference; Copula; Kernel methods and RKHS; Robust procedures

## 1. Introduction

### 1.1. Context

Since the seminal work of Sklar (1959), 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 user 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 Nelsen (2007) and Hofert et al. (2019) for a general introduction and background on copula models.

Often, 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 $\theta$. 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 (CML), shorter (Genest, Ghoudi, and Rivest 1995; Shih and Louis 1995). This is a semiparametric 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. The latter estimators have been implemented in the R package VineCopula (Schepsmeier et al. 2019) and attain the usual $\sqrt{n}$ rate of convergence as if the margins were known: see Tsukahara (2005) for the asymptotic theory.

Nevertheless, all the aforementioned estimation approaches suffer from drawbacks. In particular, they are not robust statistically speaking. More specifically, 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 $\theta_0$ when $\varepsilon \neq 0$, since this problem still occurs in the case of most usual M-estimators generally speaking.

In the literature, there are very few attempts to build robust estimation methods for semiparametric 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, Mendes, de Melo, and Nelsen (2007) 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 (Chen and Fan 2006), Kim and Lee (2013) built a minimum density power divergence estimator which shows some resistance to some types of outliers. Denecke and Müller (2011) proposed a parametric robust estimation method based on likelihood depth (Rousseeuw and Hubert 1999). Recently, Goegebeur et al. (2020) have considered robust and nonparametric estimation of the coefficient of tail dependence in presence of random covariates,

**CONTACT** Pierre Alquier pierre.alquier.stat@gmail.com RIKEN AIP, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027 Japan.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

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 article is an attempt to fill this gap.

To this end, 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 a 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, that is, $\mathbb{D}(\mathbb{P},\mathbb{Q}) = ||\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}||_{\mathcal{H}}$; see Muandet et al. (2017, sec. 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, for example, Danafar et al. (2013) and Gretton et al. (2012) for the two-sample test problem.

As a tool for parametric estimation, MMD has been studied as a general method for inference only recently (Briol et al. 2019; Alquier and Gerber 2020; Chérief-Abdellatif and Alquier 2020, 2022), even though it was implicitly used in specific examples in machine learning (Dziugaite, Roy, and Ghahramani 2015). In the latter articles, it appeared that MMD criteria lead to consistent estimators that are robust to model misspecification, for most models and without any assumption on the actual distribution of the data. Moreover, the flexibility offered by the choice of the tuning parameter of the kernel, which can be used to build a tradeoff 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 can sometimes be important.

In this article, we will study a general semiparametric inference procedure for copulas that is robust with respect to 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 (Yatracos 1985) or the Hellinger distance (Baraud, Birgé, and Sart 2017). However, the estimation procedures proposed in these articles are not computable. Also, we refer the reader to Baraud, Birgé, and Sart (2017) for a thorough discussion on why the MLE, based on the Kullback-Leibler divergence, cannot enjoy the same robustness properties.

The rest of the article is organized as follows: the remaining of the introduction yields notations and the definition of our estimators. Section 2 contains our theoretical results: nonasymptotic oracle inequalities, consistency and asymptotic distributions of our estimators. Section 3 provides experimental results. A simulation study confirms the robustness of MMD. We also provide an R Package, called MMDCopula (Alquier et al. 2020), 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 (Briol et al. 2019; Chérief-Abdellatif and Alquier 2022), 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 (Schepsmeier et al. 2019), 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 iid 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. We assume these margins are continuous. This standard assumption will allow to invoke powerful results from the theory of empirical copula processes (Bücher, Segers, and Volgushev 2012 in particular). Let us define the unobservable 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 $\boldsymbol{X} = (X_1,\dots,X_d)$ whose underlying copula is $C_0$ and underlying margins are $F_1,\dots,F_d$. Obviously, the cdf of $\boldsymbol{U}$ is $C_0$, whose law is denoted by $\mathbb{P}_0$. The empirical measure associated to $(\boldsymbol{X}_i)_{i=1,\dots,n}$ 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 $\theta_0 \in \Theta$ that is, $C_0 = C_{\theta_0}$. 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 $C_\theta$ 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 $C_\theta$ on the hypercube $\mathcal{U} = [0,1]^d$, a pseudo-true value is formally defined as

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

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

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

for every $i \in \{1,\dots,n\}$, $k \in \{1,\dots,d\}$ and every real number $t$, denoting by $\mathbf{1}(\cdot)$ the usual indicator function. Our goal will be to evaluate the pseudo-true parameter $\theta_0^*$ with MMD techniques, from the initial sample $(\boldsymbol{X}_i)_{i=1,\dots,n}$ or from the pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\dots,n}$. The empirical distribution of the latter pseudo-sample is called the empirical copula $C_n$ (Fermanian, Radulovic, and Wegkamp 2004).

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 $n$ iid realizations of $\boldsymbol{U}$, called $\boldsymbol{U}_1,\dots,\boldsymbol{U}_n$, and let $\mathbb{P}_n^U$ be the associated empirical measure on $\mathcal{U}$. To obtain an estimator of $\theta$, 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 Briol et al. (2019), we have

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

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, an estimator of $\theta_0^*$ is defined as

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

If $C_\theta$ has a density $c_\theta$ with respect to the Lebesgue measure on $[0,1]^d$, this criterion may be rewritten

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

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 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 cdf of a standard Gaussian random variable, applied coordinatewise). Interestingly, it empirically seems that the value of $\gamma$ that leads to the smallest MSE mainly depends on the kernel, and not really on the sample size nor the true value of the parameter. This is shown in Figure 1, and in additional plots in the supplementary material. Actually, this fact was rigorously proven in Chérief-Abdellatif and Alquier (2022) for the Gaussian mean model, and we conjecture that it holds more generally. This allows to calibrate $\gamma$ once and for all through a preliminary set of simulations. Note that Dziugaite, Roy, and Ghahramani (2015) proposed a median heuristic to calibrate $\gamma$ that yields good results in practice. Alternatively, Briol et al. (2019) 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 in Briol et al. (2019), p. 14.

*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 semiparametric, because its margins are nonparametrically estimated. To obtain an estimator of $\theta$, 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 $F_1, \dots, F_d$ and $C_\theta$. Applying Sklar's theorem, note that, for every $\boldsymbol{x} = (x_1, \dots, x_d)$, $\mathbb{P}_\theta^X(\boldsymbol{X} \le \boldsymbol{x}) = C_\theta\big(F_1(x_1), \dots, F_d(x_d)\big)$. 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 $\boldsymbol{x} = (x_1, \dots, x_d)$, we define $\hat{\mathbb{P}}_\theta^X(\boldsymbol{X} \le \boldsymbol{x}) = C_\theta\big(F_{n,1}(x_1), \dots, F_{n,d}(x_d)\big)$. Then, this provides another estimator

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

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

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

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

## 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 omit 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 twice 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, that is,$K_U(\boldsymbol{u},\boldsymbol{v}) = K_U(\boldsymbol{v},\boldsymbol{u})$ 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. We recall that, when $K$ is a characteristic kernel, the divergence

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

induces a true distance between probability measures on $[0,1]^d$.

### 2.1. Nonasymptotic Guarantees

The first result of this section is a nonasymptotic "universal" upper bound in terms of MMD distance that holds with high probability for any underlying distribution. Our bound 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 iid (Briol et al. 2019, Theorem 1; Chérief-Abdellatif and Alquier 2022, Theorem 3.1) and in the dependent setting (Chérief-Abdellatif and Alquier 2022, Theorem 3.2), 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 $1-\delta-\nu \in (0,1)$,

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

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

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

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

With probability greater than $1-\delta$, Lemma 1 in Briol et al. (2019) yields

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

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

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

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,

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

and we deduce

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

The Dvoretzky-Kiefer-Wolfowitz inequality (Boucheron, Lugosi, and Massart 2012, p. 383) yields

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

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.* Note that if an exact minimizer $\hat{\theta}_n$ of (1) does not exist, we can simply define $\hat{\theta}_n$ as any value that reaches the infimum up to $1/n$. The extension of Theorem 1 to this case is direct.

It is possible to slightly strengthen Theorem 1 at the price of more regularity for $K_U$, details are provided in Appendix A, supplementary material.

Let us emphasize the consequences of Theorem 1 when the data are contaminated by a proportion $\varepsilon$ 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 $\varepsilon$ 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 $\sup_{\boldsymbol{u}\in[0,1]^d} K_U(\boldsymbol{u},\boldsymbol{u}) \le 1$, as for the usual Gaussian kernel. Since $\mathbb{D}(\mathbb{P}_0,\mathbb{P}_{\theta_0}) \le 2\varepsilon$ and $\mathbb{D}(\mathbb{P}_{\hat\theta_n},\mathbb{P}_{\theta_0}) \le 2\varepsilon + \mathbb{D}(\mathbb{P}_{\hat\theta_n},\mathbb{P}_0)$ by the triangle inequality, Theorem 1 yields

$$\mathbb{D}(\mathbb{P}_{\hat\theta_n},\mathbb{P}_{\theta_0}) \le 4\varepsilon + \left(\frac{8}{n}\right)^{\frac12}\left\{1+\left(-\ln\delta\right)^{1/2}\right\} + \left\{\frac{2d^2}{n}||d^{(2)}K_U||_\infty \ln\left(\frac{2d}{\nu}\right)\right\}^{1/2}. \quad (4)$$

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

$$c_\theta(u_1,u_2) = \frac{1}{2\pi\sqrt{1-\theta^2}\phi(x_1)\phi(x_2)} \exp\left(-\frac{1}{2(1-\theta^2)}\left(x_1^2+x_2^2-2\theta x_1x_2\right)\right), \quad (5)$$

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 $\Phi$ is the cdf of a standard Gaussian random variable, and its inverse $\Phi^{-1}$ is applied coordinatewise. We prove at the end of Appendix F, supplementary material that, using the latter Gaussian kernel, there is a constant $c(\gamma) \in (0,+\infty)$ that depends only on $\gamma$ 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 (4), this gives:

$$|\hat\theta_n - \theta_0| \le c(\gamma)\left[4\varepsilon + \left(\frac{8}{n}\right)^{\frac12}\left\{1+\left(-\ln\delta\right)^{1/2}\right\} + \left(\frac{8}{n}||d^{(2)}K_U||_\infty \ln\left(\frac{4}{\nu}\right)\right)^{1/2}\right].$$

In the general case, we can use the following proposition.

*Proposition 1.* Assume that the map $\theta \mapsto \mathbb{D}^2(\mathbb{P}_\theta,\mathbb{P}_{\theta_0})$ is twice continuously differentiable in a neighborhood of $\theta_0$. Denoting by $\lambda_{\min}(\theta)$ the smallest eigenvalue of $\nabla^2_{\theta,\theta}\mathbb{D}^2(\mathbb{P}_\theta,\mathbb{P}_{\theta_0})$, assume that $\lambda_{\min}(\theta) \ge \lambda_{\min}(\theta_0)/2 > 0$ when $||\theta - \theta_0|| < r$, for some $r > 0$. Set $\alpha = \inf_{\{\theta;||\theta-\theta_0||\ge r\}} \mathbb{D}(\mathbb{P}_\theta,\mathbb{P}_{\theta_0})$ and assume $\alpha > 0$.

Then, for any contamination distribution $\mathbb{Q}$, when the data are drawn from $(1-\varepsilon)\mathbb{P}_{\theta_0} + \varepsilon\mathbb{Q}$ for some $\varepsilon \in [0,\alpha/8]$, for any $\nu > 0$ and $\delta > 0$ with $\nu + \delta < 1$, as soon as

$$\sqrt{n}\alpha \ge \left\{32 \sup_{\boldsymbol{u}\in[0,1]^d} K_U(\boldsymbol{u},\boldsymbol{u})\right\}^{1/2}\left\{1+\left(-\ln\delta\right)^{1/2}\right\} + \left\{8\,d^2||d^{(2)}K_U||_\infty \ln\left(\frac{2d}{\nu}\right)\right\}^{1/2},$$

we have, with probability at least $1-\nu-\delta$,

$$||\hat\theta_n - \theta_0|| \le \frac{2}{\sqrt{\lambda_{\min}(\theta_0)}}\left[4\varepsilon + \left\{\frac{8}{n}\sup_{\boldsymbol{u}\in[0,1]^d} K_U(\boldsymbol{u},\boldsymbol{u})\right\}^{\frac12} \times\left\{1+\left(-\ln\delta\right)^{\frac12}\right\} + \left\{\frac{2d^2}{n}||d^{(2)}K_U||_\infty \ln\left(\frac{2d}{\nu}\right)\right\}^{\frac12}\right].$$

The proof is provided in Appendix B, supplementary material.

### 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

$$L_n(\theta) = \frac1n\sum_{i=1}^n \ell(\hat{\boldsymbol{U}}_i;\theta) = \int_{[0,1]^d} \ell(\boldsymbol{w};\theta)\hat{\mathbb{P}}_n(d\boldsymbol{w}),$$

so that $\hat\theta_n \in \arg\min_{\theta\in\Theta} L_n(\theta)$ and $\theta_0^* \in \arg\min_{\theta\in\Theta} L_0(\theta)$. The asymptotic properties of M-estimators ("Quasi-MLE" particularly) for possibly misspecified models are well established in the literature: see White (1982, 1994) 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 $\Theta$ is compact. The map $L_0 : \Theta \to \mathbb{R}$ is continuous on $\Theta$ and uniquely minimized at $\theta_0^*$.

*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 $\Theta$.

*Theorem 2.* If Conditions 1 and 2 are fulfilled, then $\hat\theta_n$ is strongly consistent, that is,

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

*Proof.* As $\Theta$ is compact, then the $\delta$-bracketing numbers $\mathcal{N}_{[\cdot]}\left(\delta,\mathcal{F},L^1(\mathbb{P}_0)\right)$ are finite for every $\delta > 0$, invoking Example 19.8 in Vaart (1998). Moreover, using Lemma 1(c) in Chen and Fan (2005), 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 Newey and McFadden (1994), for example, we deduce the strong consistency of the minimizer $\hat\theta_n$ of $L_n$ toward 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 $\theta_0^*$ when $\boldsymbol{w} \in [0,1]^d$.

*Condition 3.* $\theta_0^*$ is an interior point of $\Theta$.

*Condition 4.* There exists an open neighborhood $\mathcal{O} \subset \Theta$ of $\theta_0^*$ such that 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 $K_0 \subset \mathcal{O}$ whose interior contains $\theta_0^*$ such that

$$\mathbb{E}\left[\sup_{\theta\in K_0} \left\|\nabla^2_{\theta,\theta}\ell(\boldsymbol{U};\theta)\right\|\right] < +\infty,$$

for any matrix norm $||\cdot||$. Moreover, the map $\theta \mapsto \mathbb{E}[\nabla^2_{\theta,\theta}\ell(\boldsymbol{U};\theta)]$ is continuous at $\theta_0^*$.

*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 Fermanian, Radulovic, and Wegkamp (2004), 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 Fermanian, Radulovic, and Wegkamp (2004), the continuity is not needed on the boundary of the hypercube, but only on the interior of the hypercube. This result was established by Segers (2012) under minimal assumptions, rewritten below as Condition 9. With additional smoothness requirements on the loss function $\ell$ (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, that is, it is coordinatewise right-continuous in each coordinate, and is of bounded variation in the sense of Hardy-Krause (see, Radulović, Wegkamp, and Zhao 2017, sec. 2).

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

Still, it is possible to obtain the weak convergence of the empirical copula process for an even larger class of copulas using semimetrics 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, Bücher, Segers, and Volgushev (2012) established the hyper-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.

Note a related regularity assumption in Genest, Nešlehová, and Rémillard (2017), Condition 1. Hereafter, $(\boldsymbol{w}_I, \mathbf{1}_{-I})$ denotes a vector in $[0,1]^d$ whose $j$th component is $w_j$ when $j \in I$ and is one otherwise.

*Condition 11.* For any $I \subset \{1,\dots,d\}$, $I \neq \emptyset$, there exists some $q_I \in (1,+\infty)$ such that $\int_{[0,1]^{|I|}} \left|\nabla_\theta \ell(d\boldsymbol{w}_I, \mathbf{1}_{-I};\theta_0^*)\right|^{q_I} < \infty$.

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.

*Proof.* According to Condition 4, $L_n$ is twice differentiable on a neighborhood of $\theta_0^*$ 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 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

$$0 = \nabla_\theta L_n(\hat{\theta}_n) = \nabla_\theta L_n(\theta_0^*) + \nabla_{\theta,\theta^\top} L_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0^*), \qquad (6)$$

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

For any pair $(j,k)$, 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 Vaart (1998). Using Lemma 1(c) in Chen and Fan (2005), we get

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

As $\bar{\theta}_n$ lies between $\hat{\theta}_n$ and $\theta_0^*$ componentwise, $\bar{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}_0-\text{a.s.}} \theta_0^*$. Moreover, taking expectations with respect to $(\boldsymbol{U},\bar{\theta}_n)$ or $(\hat{\boldsymbol{U}},\bar{\theta}_n)$, respectively, we have for $n$ large enough

$$\left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_n)}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right|$$
$$\leq \left|\frac{1}{n}\sum_{i=1}^n \frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\bar{\theta}_n)}{\partial\theta_k\partial\theta_j} - \mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_n)}{\partial\theta_k\partial\theta_j}\right]\right|$$

$$+\left|\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_n)}{\partial\theta_k\partial\theta_j}\right]-\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right|$$

$$\leq \sup_{\theta\in K}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\ell(\hat{\boldsymbol{U}}_i;\theta)}{\partial\theta_k\partial\theta_j}-\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta)}{\partial\theta_k\partial\theta_j}\right]\right|$$

$$+\left|\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\bar{\theta}_n)}{\partial\theta_k\partial\theta_j}\right]-\mathbb{E}\left[\frac{\partial^2\ell(\boldsymbol{U};\theta_0^*)}{\partial\theta_k\partial\theta_j}\right]\right|.$$

The continuity of $\mathbb{E}[\partial^2\ell(\boldsymbol{U};\cdot)/\partial\theta_j\partial\theta_k]$ at $\theta_0^*$ (Condition 4) yields

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

Finally, by definition of $H_n$ and $B$ (see Condition 6), we obtain $H_n\xrightarrow[n\to+\infty]{\mathbb{P}_0-a.s.}B$.

According to Proposition 3.1 in Segers (2012) 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 $\alpha$ is a $\mathbb{P}_0$-Brownian bridge. By Condition 8 and an integration by parts argument (Proposition 3 in Radulović, Wegkamp, and Zhao 2017), we have with obvious notations

$$\begin{aligned}&\sqrt{n}\left\{\nabla_\theta L_n(\theta_0^*)-\mathbb{E}[\nabla_\theta\ell(\boldsymbol{U};\theta_0^*)]\right\}\\&=\sqrt{n}\int_{(0,1]^d}\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)d(\hat{\mathbb{P}}_n-\mathbb{P}_0)(\boldsymbol{w})\\&=\sum_{I\subset\{1,\ldots,d\};I\neq\emptyset}(-1)^{|I|}\int_{(\mathbf{0}_I,\mathbf{1}_I]}\sqrt{n}(\hat{\mathbb{P}}_n-\mathbb{P}_0)(\boldsymbol{w}_I,\mathbf{1}_{-I})\,\nabla_\theta\ell(d\boldsymbol{w}_I,\mathbf{1}_{-I};\theta_0^*).\end{aligned}\tag{7}$$

Since all the maps $\boldsymbol{w}_I\mapsto\nabla_\theta\ell(\boldsymbol{w}_I,\mathbf{1}_{-I};\theta_0^*)$ are of bounded variation, the maps

$$g\mapsto\int_{(\mathbf{0}_I,\mathbf{1}_I]}g(\boldsymbol{w}_I)\,\nabla_\theta\ell(d\boldsymbol{w}_I,\mathbf{1}_{-I};\theta_0^*)$$

are continuous on $\ell^\infty\big([0,1]^{|I|},||\cdot||_\infty\big)$ for any $I\neq\emptyset$. Recalling Condition 7, the continuous mapping theorem implies that the weak limit of $\sqrt{n}\nabla_\theta L_n(\theta_0^*)$ exists, is centered and Gaussian:

$$\begin{aligned}&\sqrt{n}\nabla_\theta L_n(\theta_0^*)\xrightarrow[n\to+\infty]{\mathcal{L}}\\&\sum_{I\subset\{1,\ldots,d\};I\neq\emptyset}(-1)^{|I|}\int_{(\mathbf{0}_I,\mathbf{1}_I]}\left\{\alpha(\boldsymbol{w}_I,\mathbf{1}_{-I})-\sum_{j\in I}\dot{C}_j(\boldsymbol{w}_I,\mathbf{1}_{-I})\alpha_j(\boldsymbol{w}_j)\right\}\\&\nabla_\theta\ell(d\boldsymbol{w}_I,\mathbf{1}_{-I};\theta_0^*).\end{aligned}$$

Invoking the integration by parts again, this yields

$$\begin{aligned}&\sqrt{n}\nabla_\theta L_n(\theta_0^*)\xrightarrow[n\to+\infty]{\mathcal{L}}\int\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)d\\&\left\{\alpha(\boldsymbol{w})-\sum_{j\in I}\dot{C}_j(\boldsymbol{w})\alpha_j(\boldsymbol{w}_j)\right\}.\end{aligned}$$

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's lemma and Formula (6), we get

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

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 $1\leq p<\infty$, as detailed in Bücher, Segers, and Volgushev (2012) (Theorem 4.5. and the remarks that follow). Due to Condition 11 and Hölder's inequality, the maps $h\to\int h(\boldsymbol{w}_I)\,\nabla_\theta\ell(d\boldsymbol{w}_I,\mathbf{1}_{-I};\theta_0^*)$ are continuous on $L_{p_I}([0,1]^{|I|})$, $1/p_I+1/q_I=1$. Therefore, by (7) and the continuous mapping theorem, the weak limit of $\sqrt{n}\left\{\nabla_\theta L_n(\theta_0^*)-\mathbb{E}[\nabla_\theta\ell(\boldsymbol{U};\theta_0^*)]\right\}$ exists and is $B^{-1}\int\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)\,d\left\{\alpha(\boldsymbol{w})+dC_{(-\alpha_1,\ldots,-\alpha_d)}(\boldsymbol{w})\right\}$, proving the result. □

In the case of asymptotic normality, the asymptotic variance of $\sqrt{n}(\hat{\theta}_n-\theta_0^*)$ is $B^{-1}\Sigma B^{-1}$, where

$$\Sigma=\int\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)\nabla_\theta\ell(\boldsymbol{w}';\theta_0^*)^\top\,\mathcal{C}_0(d\boldsymbol{w},d\boldsymbol{w}'),$$

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

$$\begin{aligned}\mathcal{C}_0(\boldsymbol{w},\boldsymbol{w}')=\mathbb{E}\Bigg[&\left\{\alpha(\boldsymbol{w})-\sum_{j=1}^{d}\dot{C}_j(\boldsymbol{w})\alpha_j(w_j)\right\}\\&\left\{\alpha(\boldsymbol{w}')-\sum_{j=1}^{d}\dot{C}_j(\boldsymbol{w}')\alpha_j(w_j')\right\}\Bigg],\end{aligned}$$

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

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

denoting $\boldsymbol{w}\wedge\boldsymbol{w}'=\big(\min(w_1,w_1'),\ldots,\min(w_d,w_d')\big)$. The previous matrices can be empirically estimated: see Remark 2 in Chen and Fan (2005), or Tsukahara (2005). Note that a more explicit formula of $\Sigma$ is given in the latter articles, say

$$\begin{aligned}\Sigma=\mathrm{var}\Bigg[&\nabla_\theta\ell(\boldsymbol{U};\theta_0^*)+\sum_{j=1}^{d}\int\nabla^2_{\theta,u_j}\ell(\boldsymbol{u};\theta_0^*)\\&\mathbf{1}(U_j\leq u_j)\mathbb{P}_0(d\boldsymbol{u})\Bigg].\end{aligned}\tag{8}$$

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

*Remark 3.* If the map $\boldsymbol{w}\mapsto\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)$ is "sufficiently regular", the limiting law of $\sqrt{n}\big(\hat{\theta}_n-\theta_0^*\big)$ may still be Gaussian under Condition 10 even if 9 is not fulfilled. Indeed, this law is deduced from the weak convergence of integrals as $\int\sqrt{n}(\hat{\mathbb{P}}_n-\mathbb{P}_0)(\boldsymbol{w})\,\nabla_\theta\ell(d\boldsymbol{w};\theta_0^*)$ (Equation (7) in the proofs). It is well-known that integration is a way of regularizing potentially discontinuous processes. In particular, $\sqrt{n}\big(\hat{\theta}_n-\theta_0^*\big)$ is asymptotically normal if $\boldsymbol{w}\mapsto\nabla_\theta\ell(\boldsymbol{w};\theta_0^*)$ has an integrable density with respect to the Lebesgue measure on $[0,1]^d$: apply Theorem 4.5 in Bücher, Segers, and Volgushev (2012) and the remark that follows, using the fact that the limiting copula process has bounded trajectories in every case.

*Remark 4.* Theorem 3 relies on the weak convergence of the usual empirical copula process and an integration by parts trick.

If the map $\boldsymbol{w} \mapsto \nabla_\theta \ell(\boldsymbol{w}; \theta_0^*)$ is not of bounded variation, as required in Condition 8, an alternative method would be to invoke the weak convergence of the weighted empirical process (Berghaus, Bücher, and Volgushev 2017, Theorem 2.2) in Equation (7). This is relevant when $g(\boldsymbol{w})|\nabla_\theta \ell(d\boldsymbol{w}; \theta_0^*)|$ defines a finite measure, by setting

$$g(\boldsymbol{w}) = \min\Big\{ \bigwedge_{k=1}^{d} u_k, \bigwedge_{k=1}^{d} \big(1 - \min_{j \neq k} u_j\big)\Big\}^{\omega}, \ \omega \in (0, 1/2).$$

The price to be paid for this strategy would be to require the existence and a certain amount of regularity for the second order derivatives of $C_0$, say

$$\partial^2 C_0(\boldsymbol{u})/\partial u_j \partial u_k \leq K\Big[\max\big\{u_j(1-u_j), u_k(1-u_k)\big\}\Big]^{-1}, \ K > 0,$$

for every $\boldsymbol{u} \in V_j \cap V_k$ and every indices $j$ and $k$ in $\{1, \ldots, d\}$. Both ways of reasoning seem to be complementary, but without any clear hierarchy between them.

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, for example, Genest, Ghoudi, and Rivest (1995), Chen and Fan (2005), and Tsukahara (2005). In such a situation, the loss function $\ell$ 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. Our MMD estimator is most often less demanding. Indeed, its loss function is typically obtained by integrating copula densities, inducing a "regularization procedure". In other words, conditions of regularity as Condition 8 should be satisfied more easily in the MMD case compared to the usual CML method (even if this statement is not a universal rule).

Nonetheless, in every case, we can still rely on another set of technical assumptions, as for the CML method. 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 $\boldsymbol{w} \in (0,1)^d$, $\big\|\nabla_\theta \ell(\boldsymbol{w}; \theta_0^*)\big\| \leq C_1 \prod_{k=1}^{d} \{w_k(1-w_k)\}^{-a_k}$ for some constants $C_1$ and $a_k \geq 0$ such that

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

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

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

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

$$\mathbb{E}\Big[\{U_k(1-U_k)\}^{\zeta_k - b_k} \prod_{j=1, j\neq k}^{d} \{U_j(1-U_j)\}^{-a_j}\Big] < +\infty,$$

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". Such conditions are well-known in the copula literature: see Assumption A.3 in Chen and Fan (2005) or Assumption A.1 in Tsukahara (2005). 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 beginning of the proof involves a first-order decomposition as in the proof of Theorem 3. Nonetheless, instead of invoking integration by parts, it relies on some results about multivariate rank statistics that have been obtained by Ruymgaart and his coauthors in the 70's: see Proposition 2 in Chen and Fan (2005).

*Proof.* As in the proof of Theorem 3, we have under Conditions 1–6:

$$0 = \nabla_\theta L_n(\theta_0^*) + H_n(\hat{\theta}_n - \theta_0^*), \quad \text{and} \quad H_n \xrightarrow[n\to+\infty]{\mathbb{P}_0 - a.s.} B.$$

Moreover, according to Lemma 2 in Chen and Fan (2005) applied to $J = \nabla_\theta \ell(\cdot; \theta_0^*)$ and $w_j(v) = (v(1-v))^{\zeta_j}$, Condition 12 directly leads to:

$$\sqrt{n}\Big(\nabla_\theta L_n(\theta_0^*) - \mathbb{E}\big[\nabla_\theta \ell(\boldsymbol{W}; \theta_0^*)\big]\Big) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma), \text{ with}$$

where $\Sigma$ is given in (8). Condition 7 yields $\sqrt{n}\nabla_\theta L_n(\theta_0^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)$. Finally, as previously, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = H_n^{-1}\sqrt{n}\nabla_\theta L_n(\theta_0^*) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, B^{-1}\Sigma B^{-1}).$$

□

The limiting laws obtained in Theorems 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.

Any bootstrap scheme can be invoked as long as it is valid to evaluate the limiting law of the empirical copula process: see (7) in our proofs. Under Condition 9 (resp. Conditions 10–11), its weak convergence in $\ell^\infty([0,1]^d)$ (resp. $L^q([0,1]^d)$ for some $q > 1$) is sufficient. In the former case, we can rely on Efron's nonparametric bootstrap (Fermanian, Radulovic, and Wegkamp 2004), the multiplier bootstrap in Rémillard and Scaillet (2009), among others. In the latter case, apply another version of the multiplier bootstrap as defined in Bücher, Segers, and Volgushev (2012) (see the remark at the top of p. 1611). And, in the case of a correctly specified copula model, the parametric bootstrap (Genest and Rémillard 2008) could surely be invoked too.

To be specific, the calculation of our nonparametric bootstrap estimator requires resampling every observation in the initial sample with replacement, yielding a bootstrap

sample $\mathcal{S}_n^* = \big(\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_n^*\big)$. The associated empirical measure is

$$\mathbb{P}_n^* = n^{-1} \sum_{i=1}^{n} \delta_{\boldsymbol{X}_i^*} = n^{-1} \sum_{i=1}^{n} W_{i,n} \delta_{\boldsymbol{X}_i},$$

where the vector of weights $(W_{1,1}, \ldots, W_{n,n})$ is drawn following a $n$ multinomial law with success probabilities $(1/n, \ldots, 1/n)$. We deduce the bootstrapped empirical process as $\sqrt{n}\big(\hat{\mathbb{P}}_n^* - \hat{\mathbb{P}}_n\big)$, where $\hat{\mathbb{P}}_n^*$ denotes the empirical measure of the pseudo-sample obtained from $\mathcal{S}_n^*$. Exactly as for $\hat{\theta}_n$, one gets a bootstrapped estimator $\hat{\theta}_n^*$, but working on $\mathcal{S}_n^*$ instead of the initial sample. The asymptotic laws of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ will be the same because the limiting laws of $\sqrt{n}(\hat{\mathbb{P}}_n^* - \hat{\mathbb{P}}_n)$ and $\sqrt{n}\big(\hat{\mathbb{P}}_n - \mathbb{P}_0\big)$ are similar in (7).

For the multiplier bootstrap (Bücher, Segers, and Volgushev 2012, sec. 4.2), consider iid weights $(\xi_i)_{1 \le i \le n}$, with both mean and variance equal to one. These weights satisfy $\int \sqrt{\mathbb{P}(\xi_i > t)}\, dt < \infty$ and are independent of the sample. Introduce the cdf $G_n^*(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} \xi_i \mathbf{1}(\boldsymbol{X}_i \le \boldsymbol{x})$ on $\mathbb{R}^d$ and its margins $G_{n,k}^*$, $k \in \{1, \ldots, d\}$. Build the pseudo-sample $(\boldsymbol{V}_i^*)_{i=1,\ldots,n}$ where $V_{i,k}^* = G_{n,k}^*(X_{i,k})$ for any $k$, the associated empirical copula $\tilde{C}_n^*$ and the empirical copula process $\sqrt{n}\big(\tilde{C}_n^* - C_n\big)$. The bootstrapped estimator of $\theta_0$ is then obtained by MMD minimization, but replacing the initial pseudo-sample $(\hat{\boldsymbol{U}}_i)_{i=1,\ldots,n}$ by $(\boldsymbol{V}_i^*)_{i=1,\ldots,n}$, and the same arguments as above apply.

Recently, subsampling has been proposed as an interesting alternative to bootstrap estimates of functionals of many empirical copula processes, possibly smoothed or weighted (Kojadinovic and Stemikovskaya 2019). This technique is valid when our Condition 9 is satisfied and when the usual empirical process of $(\boldsymbol{U}_i)_{i=1,\ldots,n}$ is weakly convergent in $\ell^\infty([0,1]^d)$ to a tight centered Gaussian process. In particular, the latter result applies when our $\boldsymbol{X}$-sample is a stretch from a strongly mixing stationary sequence.

## 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 section, when we assume that the model is well-specified, that is, that the law of the observations belongs to the considered parametric family, the pseudo-true parameter $\theta_0^*$ is simply the true underlying parameter and is denoted by $\theta_0$.

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

$$K_h(\boldsymbol{u}, \boldsymbol{v}) = \exp\left\{ - \frac{(h(u_1) - h(v_1))^2 + (h(u_2) - h(v_2))^2}{\gamma^2} \right\}, \tag{9}$$

for some injective map $h : [0,1] \mapsto \mathbb{R}$ and some tuning parameter $\gamma > 0$ (see, e.g., Christmann and Steinwart 2010, Th. 2.2). Indeed, the latter function $K_h$ is a kernel: let $\zeta : \mathbb{R}^2 \to \mathcal{F}$ be the feature map that is associated with the usual Gaussian kernel $K_G$, that is, $K_G(\boldsymbol{x}, \boldsymbol{y}) = \langle \zeta(\boldsymbol{x}), \zeta(\boldsymbol{y}) \rangle_{\mathcal{F}}$, where the Gaussian kernel is defined for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ by

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

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

Hereafter, we shall denote by $\Phi$ and $\phi$ the cumulative distribution function and the probability density function of the standard normal distribution, respectively. Then, a natural choice is to set $h(u) = \Phi^{-1}(u)$. The latter kernel will simply be denoted by $K_U$. Even if it is possible to always choose the usual Gaussian kernel $K_G$ by setting $h(u) = u$, we have observed that $K_U$ provides better numerical results in some situations. We refer the reader to the simulation study for a detailed comparison. Moreover, it is sometimes simpler to use $K_U$ rather than $K_G$. For example, in the case of Gaussian copulas, the criterion $L_0$ can be analytically calculated when $K = K_U$ (see Appendix F, supplementary material), contrary to $K = K_G$. Note that it is not so surprising that $K_U$ provides better empirical results than $K_G$. Indeed, it is a common procedure in copula modeling to push back the sample observations on $\mathbb{R}^d$ using Gaussian quantile functions componentwise. This trick spreads the data cloud and often improves inference. At the opposite, our conditions of regularity for Marshall-Olkin copulas can be checked only when the kernel is $K_G$.

### 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 $\theta \in (-1,1)$. 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 $\mathbb{E}[X_1 X_2] = \theta$. The associated copula density has been given in Equation (5).

*Proposition 2.* Assume that the true underlying copula is $C_{\theta_0}$ for some parameter $\theta_0 \in (-1,1)$. Then, when $K \in \{K_U, K_G\}$ and $\gamma^2 < 2$, the estimator $\hat{\theta}_n$ given by (1) is strongly consistent and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal.

The proof is deferred to Appendix C, supplementary material. For the sake of illustration, we will verify the conditions of validity of Theorem 3, even if those of Theorem 4 can be checked too. In this proof, it is stated that the term $B$ that appears in the asymptotic variance of $\hat{\theta}_0$ when $K = K_U$ has the closed-form expression

$$B_G(\theta_0) = \frac{3\gamma^2\{(2 + \gamma^2/2)^2 + 8\theta_0^2\}}{2\{(2 + \gamma^2/2)^2 - 4\theta_0^2\}^{5/2}} > 0.$$

Now, let us deal with the general case of misspecification.

*Corollary 1.* Assume that the true underlying copula is $C_0$ and $K \in \{K_U, K_G\}$ with $\gamma^2 < 2$. If the estimator $\hat{\theta}_n$ given by (1) is strongly consistent to $\theta_0^* \in (-1,1)$ that satisfies the first-order Condition 7 and if $B > 0$, then $\sqrt{n}(\hat{\theta}_n - \theta_0^*)$ is asymptotically normal.

The proof is given in Appendix D, supplementary material. When a Gaussian copula is contaminated by a fixed bivariate

copula $\bar{C}$, then $C_0 = (1-\varepsilon)C_{\theta_0} + \varepsilon\bar{C}$, and the real number $B$ is now

$$B = \int \nabla^2_{\theta,\theta}\ell(\boldsymbol{u};\theta_0^*)C_0(d\boldsymbol{u}) = (1-\varepsilon)B_G(\theta_0^*) + \varepsilon\int \nabla^2_{\theta,\theta}\ell(\boldsymbol{u};\theta_0^*)\bar{C}(d\boldsymbol{u}).$$

Here, we have assumed the consistency of $\hat{\theta}_n$ because we cannot exclude the existence of several minimizers of $L_0$ in general, even if it is a very unlikely situation.

### 2.3.2. Marshall-Olkin Copulas

By definition (Nelsen 2007, sec. 3.1.1), the bivariate Marshall-Olkin copula is defined on $[0,1]^2$ as

$$C_\theta(u,v) = u^{1-\alpha}v\mathbf{1}(u^\alpha \geq v^\beta) + uv^{1-\beta}\mathbf{1}(u^\alpha < v^\beta), \qquad (10)$$

for some parameter $\theta = (\alpha,\beta)$, $0 < \alpha,\beta < 1$. This copula has no density with respect to the Lebesgue measure on the whole $[0,1]^2$. The absolutely continuous part of $C_\theta$ (with respect to 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 Nelsen (2007), $C_\theta(u,v) = A_\theta(u,v) + S_\theta(u,v)$, where, for every $(u,v) \in [0,1]^2$, $S_\theta(u,v) = \kappa\left\{\min(u^\alpha, v^\beta)\right\}^{1/\kappa}$ and

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

Let us calculate $\mathbb{E}[\psi(U,V)]$, $(U,V) \sim C_\theta$, for any measurable map $\psi$, to be able to calculate $\ell(\boldsymbol{w},\theta)$ for our bivariate Marshall-Olkin model. Given a small positive real number $\delta$, 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 $\delta \ll 1$,

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

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

$$\begin{aligned} \mathbb{E}[\psi(U,V)] &= \int \psi(s,t)\frac{\partial^2 C_\theta}{\partial u\partial v}(s,t)\,ds\,dt \\ &\quad + \int \psi(u,v)\,S_\theta(du,dv) =: I_1 + I_2, \qquad (11) \\ I_1 &= \int \psi(s,t)\left\{(1-\alpha)s^{-\alpha}\mathbf{1}(s^\alpha > t^\beta)\right. \\ &\quad \left. + (1-\beta)t^{-\beta}\mathbf{1}(s^\alpha < t^\beta)\right\} ds\,dt. \qquad (12) \end{aligned}$$

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

$$\begin{aligned} I_2 &= \int \psi(u,v)\,S_\theta(du,dv) = \int_0^{\bar{u}_{\alpha,\beta}} \psi(u,u^{\alpha/\beta})\,\beta u^{1-\alpha}\,du \\ &\quad + \int_{\bar{u}_{\alpha,\beta}}^1 \psi(u,u^{\alpha/\beta})\,\alpha u^{\alpha/\beta-\alpha}\,du, \qquad (13) \end{aligned}$$

with $\bar{u}_{\alpha,\beta} = (\beta/\alpha)^{\beta/(\alpha-\beta)}$ 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 $(0,1)^2$, that is, even at the main diagonal. For most regular functions $\psi$, the latter integrals $I_1$, $I_2$ and then $\mathbb{E}[\psi(U,V)]$ are continuous functions of $(\alpha,\beta)$.

*Proposition 3.* 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 $K_U$ or $K_G$. Moreover, when $K = K_G$ and $B$ is positive definite, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is weakly convergent.

See the proof in Appendix E, supplementary material. When the latter limiting law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ exists, it is not Gaussian in general. It could be numerically evaluated by usual resampling techniques, as the consistent bootstrap scheme in (Bücher, Segers, and Volgushev 2012, sec. 4.2).

*Remark 5.* The difficulty to state the limiting law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ with $K = K_U$ arises from the second-order derivatives of $\ell(\boldsymbol{w},\theta)$ with respect to $\theta$. To be short, at some stage, one has to deal with integrals as

$$\int \exp\left\{-\frac{(x-y)^2}{\gamma^2} - \frac{(x_{\alpha/\beta} - y_{\alpha/\beta})^2}{\gamma^2}\right\} \times \frac{x^a y^{\bar{a}}\Phi(x)^b\Phi(y)^{\bar{b}}}{\phi^2(x_{\alpha/\beta})}\ln^c\Phi(x)\ln^{\bar{c}}\Phi(y)\phi(x)\phi(y)\,dx\,dy$$

by setting $t_\nu = \Phi^{-1}(\Phi(t)^\nu)$ for any $\nu \geq 0$ and any real number $t$. Here, $(a,b,c,\bar{a},\bar{b},\bar{c})$ denotes a vector of nonnegative real numbers. It can be proved that the latter integral is not convergent, even in the simplest case $\alpha = \beta$ and all the other constants are zero.

In the case of general misspecification, a similar result is still valid.

*Corollary 2.* Assume that the true underlying copula $C_0$ is arbitrary. If the estimator $\hat{\theta}_n$ given by (1) is strongly consistent to $\theta_0^* \in (-1,1)$ that satisfies the first-order Condition 7, then the same results as in Proposition 3 apply, replacing $\theta_0$ by $\theta_0^*$.

The arguments of the proof are exactly the same as in Appendix E, supplementary material.

## 3. Implementation and Experimental Study

In this section, we compare the MMD estimator to the CML and the moment estimator on simulated data. The CML and the method of moments by inversion of Kendall's tau are implemented in the R package VineCopula (Schepsmeier

et al. 2019). We implemented the MMD estimator using the stochastic gradient algorithm described in Chérief-Abdellatif and Alquier (2022). This procedure requires sampling 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 (Alquier et al. 2020).

### 3.1. Implementation via Stochastic Gradient and the MMDCopula Package

We start by a short description of the algorithm implemented in our R package (Alquier et al. 2020) to compute the MMD estimator in the bivariate case. It is of course possible to use the vine-copula procedure to decompose higher-dimensional copulas into bivariate ones. The main idea is differentiating the criterion (2). Under suitable regularity assumptions on the copula density $c_\theta$ with respect to the Lebesgue measure on $\mathcal{U}$, we have

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

where the expectation is taken with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$, that are independently drawn from $C_\theta$ (a formal statement can be found in Chérief-Abdellatif and Alquier 2022). 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 $(\eta_t)_{t\geq 0}$, and iterate:

$$\left\{\begin{array}{l}\text{draw } \boldsymbol{U}_1^\star,\ldots,\boldsymbol{U}_n^\star,\boldsymbol{V}_1^\star,\ldots,\boldsymbol{V}_n^\star\sim C_{\theta_t}\text{ iid},\\ \theta_{t+1}\leftarrow\theta_t-2\eta_t n^{-2}\sum_{i,j=1}^{n}\frac{d\ln c_\theta(\boldsymbol{U}_j^\star)}{d\theta}_{|\theta=\theta_t}\big\{K(\boldsymbol{U}_j^\star,\boldsymbol{V}_i^\star)-K_U(\boldsymbol{U}_j^\star,\hat{\boldsymbol{U}}_i)\big\}.\end{array}\right.$$

In practice, we take $\eta_t = 1/\sqrt{t}$ as recommended in Chérief-Abdellatif and Alquier (2022). We perform 200 iterations, and return the average of $\theta_t$ over the last 100 iterations.

The implementation of this algorithm requires (i) to be able to sample from $C_\theta$ and (ii) to compute $c_\theta$ and its partial derivative with respect to $\theta$. A list of copula densities and their differentials can be found in Schepsmeier and Stöber (2014) and is implemented in VineCopula (Schepsmeier et al. 2019). Some procedures to sample from $C_\theta$ 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.

In theory, the criterion in (1) has no reason to be convex in $\theta$. Therefore, it is possible that the algorithm gets stuck in a local minimum. In order to avoid this situation, we propose two possible strategies: (a) starting from a random initialization and (b) starting from the empirical Kendall's tau and the associated $\theta$ values. We compared these two strategies in a set of experiments in the supplementary material. In the noncontaminated case, the Kendall's tau initialization is slightly better (especially for small $\gamma$'s) but both strategies are comparable. However, in a contaminated case, the random initialization becomes better. We suspect Kendall's tau might be close to a local minimizer of the MMD in the latter case. In our package and in our simulations, the random initialization is the default mode. The convergence of stochastic gradient algorithms for MMD mininimization in a general framework is discussed in Chérief-Abdellatif and Alquier (2022).

Also, note that it is possible to use a quasi Monte Carlo rather than a Monte Carlo sampling scheme. In our package MMDCopula (Alquier et al. 2020), 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. An important point is that the gradient method is *not* invariant by reparameterization. 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 $\alpha$ 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 $\Phi^{-1}$ and is applied coordinatewise. A major question is then: how to calibrate $\gamma$, 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 $\gamma$. A more complete study of the dependence of the MSE with respect to $\gamma$ in various models is provided in Appendix I, supplementary material.

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 200 times. Except in some experiments in the supplement used to calibrate $\gamma$, the true Kendall's tau was fixed as $\tau = 0.5$.

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 $\gamma$ is $\gamma \simeq 0.25$. For $h(\boldsymbol{u}) = \big(\Phi^{-1}(u_1),\Phi^{-1}(u_2)\big)$, the optimal value is $\gamma = 0.80$. We performed similar experiments for other copula families. The results can be found in Appendix I, supplementary material. The optimal values for each family are set as default values in our package, and used in the following experiments.

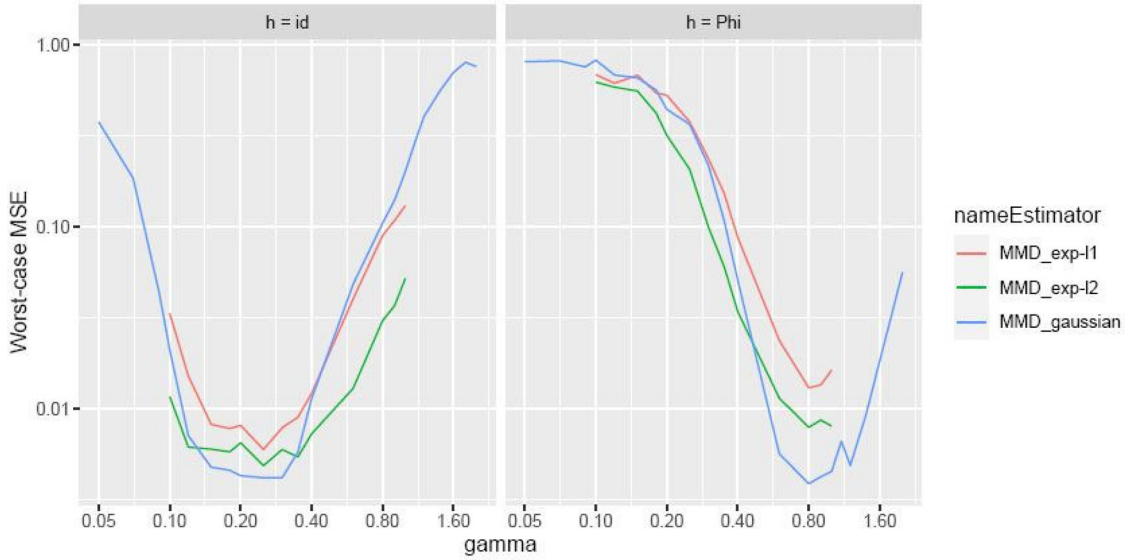


**Figure 1.** MSE of $\hat{\theta}_n$ based on the Gaussian kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-||h(\boldsymbol{U}) - h(\boldsymbol{V})||_2^2/\gamma^2)$, the exp-$L_2$ kernel $k_U(\boldsymbol{U}, \boldsymbol{V}) = \exp(-||h(\boldsymbol{U}) - h(\boldsymbol{V})||_2/\gamma)$ and the exp-$L_1$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

Finally, note that we also discuss the computational cost in Appendix G, supplementary material (for n = 1000, a MMD estimation takes around 4–7 sec for most copula families).

### 3.2. Comparison to CML on Synthetic Data

We now compare the MMD estimators based on the Gaussian kernel (with two choices of $h$) to the canonical maximum likelihood (CML) estimator 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 nine types of outliers.

- *Uniform*: the outliers are drawn iid 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 iid 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 iid from $\mathcal{U}([0,q]^2)$.
- *Diagonal*: the outliers are uniform on the first diagonal.
- *Gauss 0.2*: the outliers are drawn from the Gaussian copula with a Kendall's tau equal to 0.2.
- *Gauss* $-0.8$: the outliers are drawn from the Gaussian copula with a Kendall's tau equal to $-0.8$.
- *Frank* $-0.8$: the outliers are drawn from the Frank copula with a Kendall's tau equal to $-0.8$.
- *Clayton 0.5*: the outliers are drawn from the Clayton copula with a Kendall's tau equal to 0.5.
- *Student 0.5 3df*: the outliers are drawn from the Student copula with a Kendall's tau equal to 0.5 and 3 degrees of freedom.

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 $\varepsilon$ of each type of outliers, when $\varepsilon$ 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 $\varepsilon \in \{0.05, 0.1\}$ 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 $\varepsilon$ in a grid that ranges from 0 to 0.05, we repeat 1000 times the following experiment: the data are iid from the Gaussian copula, the sample size is $n = 1000$ and the parameter is calibrated so that $\tau = 0.5$. Then, an exact proportion $\varepsilon$ of the data is replaced by outliers. We report the mean MSE of each estimator in Figure 2.

When there are no outliers, CML yields the best estimator. However, as soon as there is more than 2% or 3% of outliers, the MMD estimators become much more reliable when contamination arises from a distribution that significantly differs from the reference Gaussian copula. 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.

The conclusion remains unchanged: in all models, the MMD estimators are far more robust than the CML and the method of moments estimators.

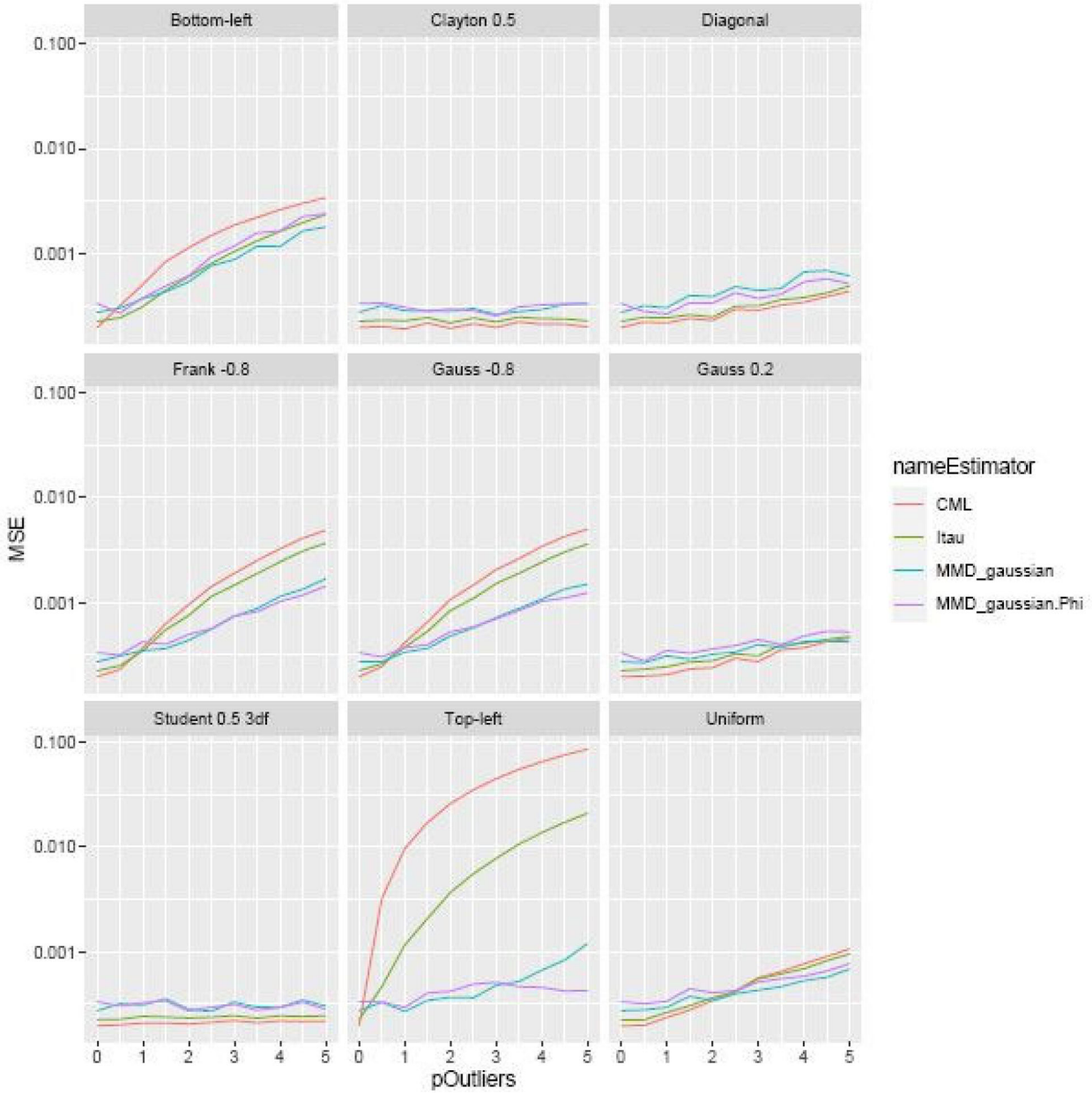


**Figure 2.** MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the CML estimator and the method of moment based on Kendall's $\tau$, as a function of the proportion $\varepsilon$ of outliers. Sample size: $n = 1000$, model: Gaussian copula. The title of each box gives the distribution of the contamination.

### 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 ($\varepsilon = 0.05$ and $\varepsilon = 0.1$, independently of the sample size) in Figure 4.

When there are no outliers, we observe the $\sqrt{n}$ consistency of all the estimators, as predicted by the theory. The CML method yields the best estimator in this case. 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.

## 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

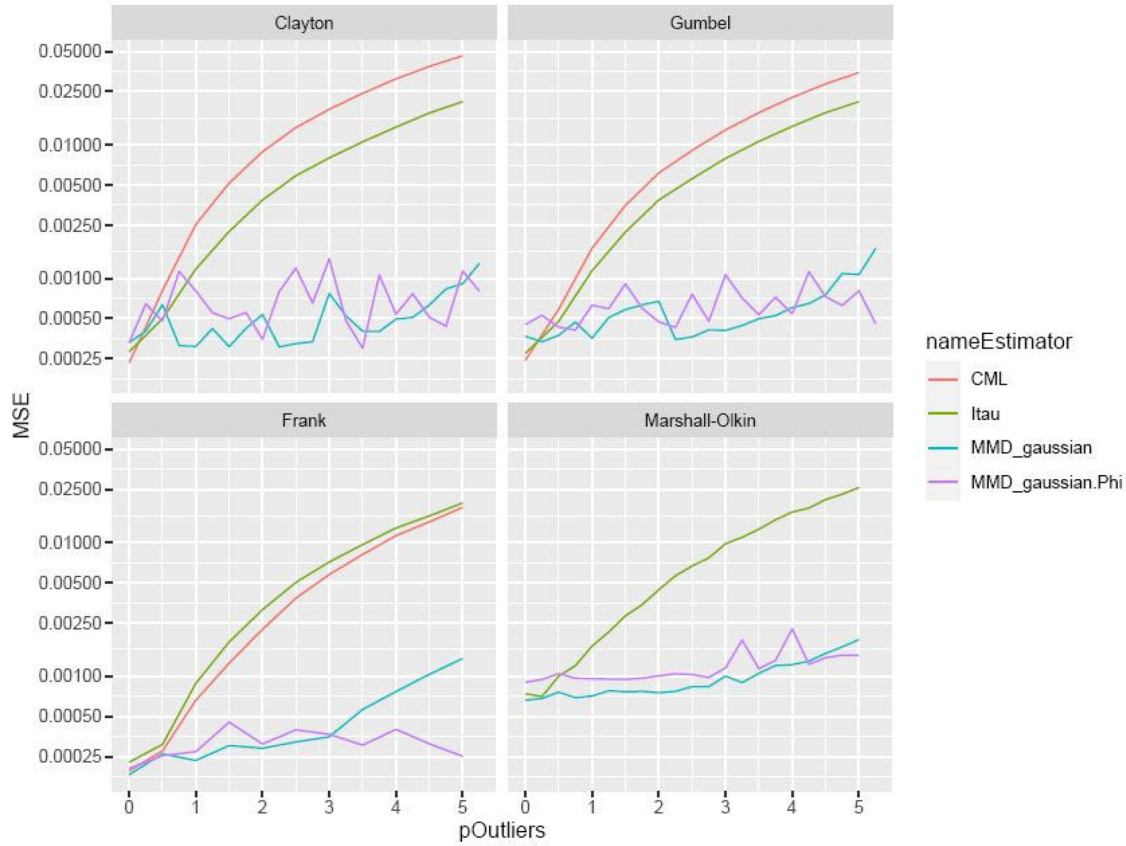


**Figure 3.** MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the CML estimator and the method of moment based on Kendall's $\tau$, as a function of the proportion $\varepsilon$ of *top-left* outliers. Sample size: $n = 1000$.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

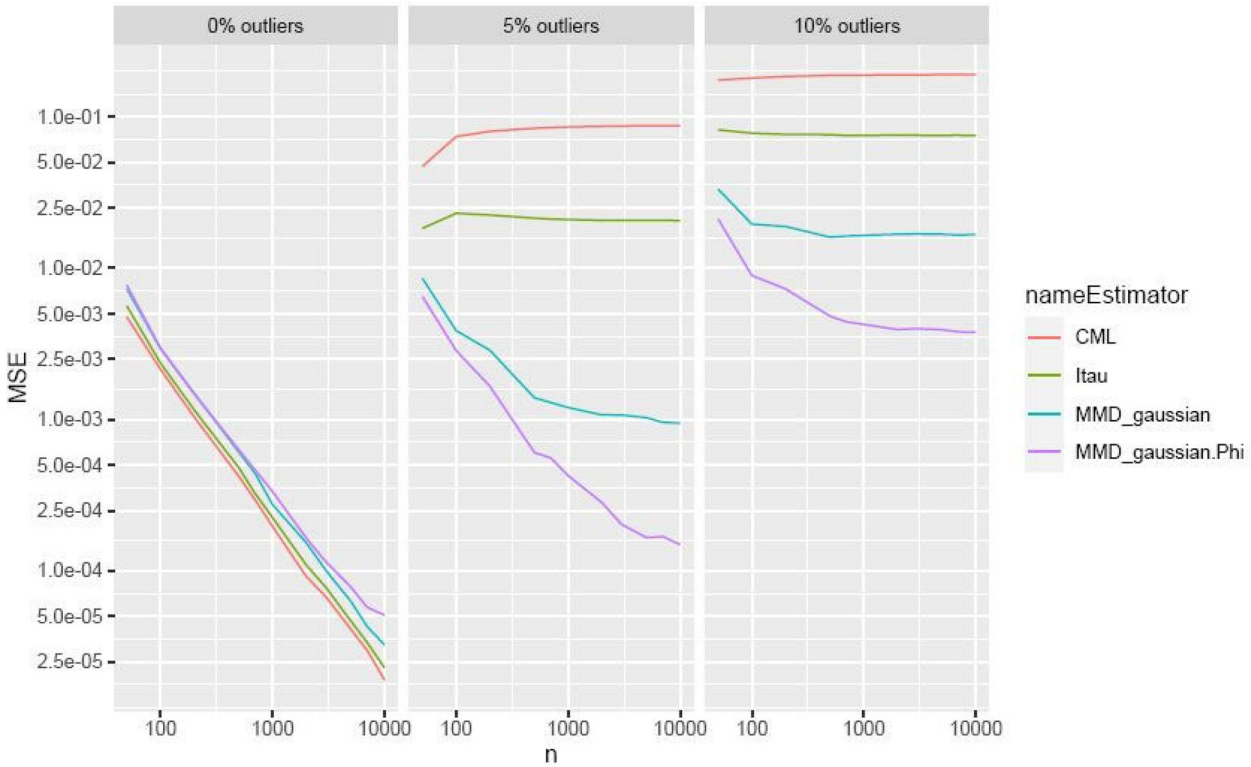


**Figure 4.** MSE of the MMD estimator with Gaussian kernel and $h(u) = u$, the MMD estimator with Gaussian kernel and $h(u) = \Phi^{-1}(u)$, the CML estimator and the method of moment based on Kendall's $\tau$, as a function of the sample size $n$. Model: Gaussian copula. Left: no outliers. Middle: a proportion $\varepsilon = 0.05$ of outliers. Right: a proportion $\varepsilon = 0.1$ of outliers.

the literature; see, for example, Bücher and Volgushev (2013). Moreover, finding a formal data-driven way of choosing the kernel tuning-parameter $\gamma$ 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 $\theta$, for example as

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

This idea would be different from the so-called "regularized MMD" in Danafar et al. (2013) that is reduced to multiplying the first term on the right-hand side of the latter equation 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.

## Supplementary Materials

Plot 3D Marshall-Olkin.html : contains the interactive plot of the MSE for the Marshall-Olkin family of copulas.
Plot 3D parametric families.html : contains the interactive plot of the MSE for parametric families of copulas.
Reproducibility : this folder contains the code and instructions to reproduce the figures of the article.

## Acknowledgments

The authors thank both anonymous Referees for their insightful comments that led to many improvements of the article.

## Funding

Badr-Eddine Chérief-Abdellatif acknowledges support of the UK Defence Science and Technology Laboratory (DSTL) and EPSRC under grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. Jean-David Fermanian has been supported by the labex Ecodec (reference project ANR-11-LABEX-0047).

## ORCID

Pierre Alquier http://orcid.org/0000-0003-4249-7337
Alexis Derumigny http://orcid.org/0000-0002-6163-8097
Jean-David Fermanian http://orcid.org/0000-0001-5960-5555

## References

Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A., and Fermanian, J.-D. (2020), "R package: MMDCopula." Available at *https://github.com/AlexisDerumigny/MMDCopula* [2,11]

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

Baraud, Y., and Birgé, L., and Sart, M. (2017), "A New Method for Estimation and Model Selection: $\rho$-estimation," *Inventiones Mathematicae*, 207, 425–517. [2]

Berghaus, B., Bücher, A., and Volgushev, S. (2017), "Weak Convergence of the Empirical Copula Process with Respect to Weighted Metrics," *Bernoulli*, 23, 743–772. [8]

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. [2,3,4]

Boucheron, S., Lugosi, G., and Massart, P. (2012), *Concentration Inequalities. A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press. [4]

Bücher, A., 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* 42, 1598–1634. [2,6,7,8,9,10]

Bücher, A., and Volgushev, S. (2013), "Empirical and Sequential Empirical Copula Processes Under Serial Dependence," *Journal of Multivariate Analysis*, 119, 61–70. [15]

Chen, X., and Fan, Y. (2005), "Pseudo-Likelihood Ratio Tests for Semiparametric Multivariate Copula Model Selection," *The Canadian Journal of Statistics*, 33, 389–414. [5,6,7,8]

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. [1]

Chérief-Abdellatif, B.-E., and Alquier, P. (2022), "Finite Sample Properties of Parametric MMD Estimation: Robustness to Misspecification and Dependence," *Bernoulli*, 28, 181–213. [2,3,4,11]

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

Christmann, A., and Steinwart, I. (2010), "Universal Kernels on Non-standard Input Spaces," in *Advances in Neural Information Processing Systems*, pp. 406–414. [9]

Danafar, S., Rancoita, P., Glasmachers, T., Whittingstall, K., and Schmidhuber, J. (2013), "Testing Hypotheses by Regularized Maximum Mean Discrepancy," ArXiv preprint, arXiv:1305.0423. [2,15]

Denecke, L., and Müller, C. H. (2011), "Robust Estimators and Tests for Bivariate Copulas based on Likelihood Depth," *Computational Statistics and Data Analysis*, 55, 2724–2738. [1]

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015), "Training Generative Neural Networks via Maximum Mean Discrepancy Optimization," in *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. [2,3]

Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004), "Weak Convergence of Empirical Copula Processes," *Bernoulli*, 10, 847–860. [2,6,8]

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995), "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions," *Biometrika*, 82, 543–552. [1,8]

Genest, C., and Rémillard, B. (2008), "Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models," *Annales de l'IHP Probabilités et statistiques*, 44, 1096–1127. [8]

Genest, C., Nešlehová, J. G., and Rémillard, B. (2017), "Asymptotic Behavior of the Empirical Multilinear Copula Process Under Broad Conditions," *Journal of Multivariate Analysis*, 159, 82–110. [6]

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* 178, 104607. [1]

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. [2]

Hofert, M., Kojadinovic, I., Mächler, M., and Yan, J. (2019), *Elements of Copula Modeling with R*, Cham: Springer. [1]

Kim, B., and Lee, S. (2013), "Robust Estimation for Copula Parameter in SCOMDY Models," *Journal of Time Series Analysis*, 34, 302–314. [1]

Kojadinovic, I., and Stemikovskaya, K. (2019), "Subsampling (Weighted Smooth) Empirical Copula Processes," *Journal of Multivariate Analysis*, 173, 704–723. [9]

Mendes, B. V. M., de Melo, E. F. L., and Nelsen, R. B. (2007), "Robust Fits for Copula Models," *Communications in Statistics, Simulation and Computation*, 36, 997–1017. [1]

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, 1–141. [2]

Nelsen, R. B. (2007), *An Introduction to Copulas*, New-York: Springer. [1,10]

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, IV*, eds. R. F. Engle and D. L. McFadden, pp. 2112–2245, Amsterdam: North Holland. [5]

Radulović, D., Wegkamp, M., and Zhao, Y. (2017), "Weak Convergence of Empirical Copula Processes Indexed by Functions," *Bernoulli*, 23, 3346–3384. [6,7]

Rémillard, B., and Scaillet, O. (2009), "Testing for Equality between Two Copulas," *Journal of Multivariate Analysis*, 100, 377–386. [8]

Rousseeuw, P. J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association*, 94, 388–402. [1]

Schepsmeier, U., and Stöber, J. (2014), "Derivatives and Fisher Information of Bivariate Copulas," *Statistical Papers*, 55, 525-542. [11]

Schepsmeier, U., Stöber, J., Brechmann, E.C., Gräler, 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. [1,2,11]

Shih, J. H., and Louis, T. A. (1995), "Inferences on the Association Parameter in Copula Models for Bivariate Survival Data," *Biometrics*, 51, 1384–1399. [1]

Segers, J. (2012), "Asymptotics of Empirical Copula Processes Under Non-restrictive Smoothness Assumptions," *Bernoulli*, 18, 764–782. [6,7]

Sklar, A.(1959), "Fonctions de Répartition à *n* Dimensions et leurs Marges," *Publications de l'Institut de statistique de l'Université de Paris*, 8, 229–231. [1]

Tsukahara, H. (2005), "Semiparametric Estimation in Copula Models," *Canadian Journal of Statistics*, 33, 357–375. [1,7,8]

Vaart, A. W. van der. (1998), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press. [5,6]

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [5]

White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge, UK: Cambridge University Press. [5]

Yatracos, Y. G. (1985), "Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy," *The Annals of Statistics*, 13, 768–774. [2]