STATISTICAL INFERENCE FOR COPULAS IN HIGH DIMENSIONS: A SIMULATION STUDY

BY

PAUL EMBRECHTS AND MARIUS HOFERT

ABSTRACT

Statistical inference for copulas has been addressed in various research papers. Due to the complicated theoretical results, studies have been carried out mainly in the bivariate case, be it properties of estimators or goodness-of-fit tests. However, from a practical point of view, higher dimensions are of interest. This work presents the results of large-scale simulation studies with particular focus on the question to what extent dimensionality influences point and interval estimators.

KEYWORDS

Maximum likelihood estimator, inference functions for margins, maximum pseudo-likelihood estimator, Kendall's tau estimator, confidence intervals, high dimensions.

1. INTRODUCTION

It is well known that distributions of test statistics of goodness-of-fit tests for copulas may be influenced when the test is based on pseudo-observations; see (Dobrić and Schmid, 2007; Genest *et al.*, 2009; Kojadinovic and Yan, 2010a, and references therein). A parametric bootstrap is typically used to compute approximate p-values.

The influence of estimating the marginal distributions on the precision of point and interval estimators is treated much less in the literature; available references are Genest *et al.* (1995), Tsukahara (2005), Kim *et al.* (2007) and Kojadinovic and Yan (2010b). Moreover, none of the existing work on investigating this influence (on either estimators or goodness-of-fit tests) has been carried out in higher dimensions d (say, $d = 100$). However, from a practical point of view, applications are often high-dimensional; specific examples are Arbenz *et al.* (2012) in an insurance context, Hofert and Scherer (2011) in the context of finance, or, in a more general context of risk management, McNeil *et al.* (2005, Chapter 2).

Astin Bulletin 43(2), 81-95. doi: 10.1017/asb.2013.6

Typically, statements about the precision of estimators in higher dimensions are made based on investigations in small dimensions, say, $d \in \{2, 3, 4, 5\}$, and conclusions are often drawn by extrapolating beyond the investigated dimensions. The problem is that this might not give a full picture of the behavior of estimators in truly high dimensions. Additionally, without paying attention to numerical issues (such as cancellation, rounding errors, underflows, overflows, flat objective functions, and non-convergence of optimizers), it might very well happen that computational results obtained are not reliable. These issues can (sometimes invisibly) appear already in small dimensions, such as $d = 5$, and may lead to doubtful statements about the precision of estimators in higher dimensions; see Weiß (2010) for such a statement and Hofert *et al.* (2012) for a correction. Once computations are carried out in higher dimensions, the numerics involved become increasingly important in implementations, and careful checks have to be conducted in order to obtain solid results. The common approach "let's compute it and see if we obtain anything reasonable" is by far not good enough.

In the present work, our goal is to investigate the impact of using estimated marginal distributions on the precision of common point and interval estimators of copula parameters. We focus particularly on high dimensions and assess numerically the magnitude of this impact. The asymptotic theory of the estimators we consider is quite demanding, difficult to check at best (see for instance Kojadinovic and Yan, 2010b), and sometimes not developed yet. We therefore conduct large-scale simulation studies in order to compare the performance of the considered estimators in small ($d = 2$) to high dimensions ($d = 100$) for popular copulas in finance and insurance.

Before we continue, let us remark that the term *high-dimensional* is not new in the statistics literature. In contrast to what we consider as high-dimensional (d being large; see (1) below), there is also a well-known body of literature on *high-dimensional statistics* for the problem of a high-dimensional parameter space (in the notation below, this would mean that p was large). Although this is different from our setup, we would like to point out, for example, Bühlmann and van de Geer (2011), Hastie *et al.* (2009, Chapter 9), or Portnoy (1988) for the reader interested in this area of statistics.

The paper is organized as follows. In Section 2, the notation and the estimators considered are introduced. Section 3 presents the simulation studies conducted, and Section 4 concludes.

2. The parametric and semi-parametric estimators considered

Let $\boldsymbol{X}_i$, $i \in \{1, \dots, n\}$, be independent and identically distributed (i.i.d.) copies of a d-dimensional random vector $\boldsymbol{X}$ following the distribution function H with continuous margins F_j, $j \in \{1, \dots, d\}$, and corresponding copula C, that is (by

Sklar's Theorem),

$$H(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \ \boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d. \tag{1}$$

In the following, we assume the jth margin F_j to belong to a parametric family of continuous distribution functions

$$\mathcal{F}_j = \{F_j(\cdot\,; \boldsymbol{\theta}_j) \,|\, \boldsymbol{\theta}_j \in \Theta_j\}$$

with true but unknown parameter vector $\boldsymbol{\theta}_{0,j}$, where Θ_j is a non-empty, open subset of $\mathbb{R}^{p_j}$, $j \in \{1, \ldots, d\}$. Furthermore, we assume that C belongs to a parametric family of copulas

$$\mathcal{C} = \{C(\cdot\,; \boldsymbol{\theta}_C) \,|\, \boldsymbol{\theta}_C \in \Theta_C\}$$

with true but unknown parameter vector $\boldsymbol{\theta}_{0,C}$, where Θ_C is a non-empty, open subset of $\mathbb{R}^{p_C}$. Based on realizations of $\boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$, the main statistical problem is to estimate the parameter vector

$$\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,C}, \boldsymbol{\theta}_{0,1}, \ldots, \boldsymbol{\theta}_{0,d})^\top$$

of H and to construct confidence regions for it; here and in the following, we drop the transpose inside vectors for the reader's convenience.

In this work, we focus particularly on $\boldsymbol{\theta}_{0,C}$. We assume that the margins F_j, $j \in \{1, \ldots, d\}$, and the copula C are absolutely continuous with corresponding densities f_j, $j \in \{1, \ldots, d\}$, and c, respectively.

2.1. Popular point estimators

In this section, we briefly present the estimators considered in this work. Although rather unrealistic from a practical point of view, the known margins maximum likelihood estimator, see Section *2.1.4* below, is included as a benchmark.

2.1.1. *The maximum likelihood estimator.* If H admits a density h which can be expressed via (1) by

$$h(\boldsymbol{x}; \boldsymbol{\theta}) = c(F_1(x_1; \boldsymbol{\theta}_1), \ldots, F_d(x_d; \boldsymbol{\theta}_d); \boldsymbol{\theta}_C) \prod_{j=1}^{d} f_j(x_j; \boldsymbol{\theta}_j), \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_C, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)^\top \in \Theta \subseteq \mathbb{R}^p$, $\Theta = \Theta_C \times \Theta_1 \times \cdots \times \Theta_d$, $p = p_C + \sum_{j=1}^{d} p_j$. Let $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log h(\boldsymbol{x}; \boldsymbol{\theta})$. The log-likelihood corresponding to

(2) based on $\boldsymbol{X}_i$, $i \in \{1, \dots, n\}$, is thus

$$\begin{aligned}\ell(\boldsymbol{\theta}; \boldsymbol{X}_1, \dots, \boldsymbol{X}_n) &= \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{X}_i) \\ &= \sum_{i=1}^{n} \ell_C(\boldsymbol{\theta}_C; F_1(X_{i1}; \boldsymbol{\theta}_1), \dots, F_d(X_{id}; \boldsymbol{\theta}_d)) \\ &\quad + \sum_{i=1}^{n} \sum_{j=1}^{d} \ell_j(\boldsymbol{\theta}_j; X_{ij}), \end{aligned} \tag{3}$$

where $\ell_C(\boldsymbol{\theta}_C; u_1, \dots, u_d) = \log c(u_1, \dots, u_d; \boldsymbol{\theta}_C)$ and $\ell_j(\boldsymbol{\theta}_j; x) = \log f_j(x; \boldsymbol{\theta}_j)$, $j \in \{1, \dots, d\}$. The *maximum likelihood estimator (MLE)* $\hat{\boldsymbol{\theta}}_n^{\mathrm{MLE}}$ of $\boldsymbol{\theta}_0$ is then defined by

$$\hat{\boldsymbol{\theta}}_n^{\mathrm{MLE}} = \operatorname*{argsup}_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \boldsymbol{X}_1, \dots, \boldsymbol{X}_n). \tag{4}$$

The optimization in (4) is typically done by numerical means. Note that this can be quite demanding, especially in high dimensions.

2.1.2. *The inference functions for margins estimator.* Due to the form of (3), Joe and Xu (1996) suggested to first estimate $\boldsymbol{\theta}_{0,j}$ by its MLE $\hat{\boldsymbol{\theta}}_{n,j}^{\mathrm{MLE}}$ for each $j \in \{1, \dots, d\}$ individually (step 1), and then (step 2) estimate $\boldsymbol{\theta}_{0,C}$ via

$$\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{IFME}} = \operatorname*{argsup}_{\boldsymbol{\theta}_C \in \Theta_C} \ell\left(\boldsymbol{\theta}_C, \hat{\boldsymbol{\theta}}_{n,1}^{\mathrm{MLE}}, \dots, \hat{\boldsymbol{\theta}}_{n,d}^{\mathrm{MLE}}; \boldsymbol{X}_1, \dots, \boldsymbol{X}_n\right).$$

The corresponding *inference functions for margins estimator (IFME)* $\hat{\boldsymbol{\theta}}_n^{\mathrm{IFME}}$ of $\boldsymbol{\theta}_0$ is thus

$$\hat{\boldsymbol{\theta}}_n^{\mathrm{IFME}} = \left(\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{IFME}}, \hat{\boldsymbol{\theta}}_{n,1}^{\mathrm{MLE}}, \dots, \hat{\boldsymbol{\theta}}_{n,d}^{\mathrm{MLE}}\right)^{\top},$$

which is typically much easier to compute than $\hat{\boldsymbol{\theta}}_n^{\mathrm{MLE}}$ from a numerical point of view.

2.1.3. *The maximum pseudo-likelihood estimator.* The so-called *maximum pseudo-likelihood estimator (MPLE)* is based on an idea similar to $\hat{\boldsymbol{\theta}}_n^{\mathrm{IFME}}$, but for the former the jth margin is estimated non-parametrically by its empirical distribution function $\hat{F}_{n,j}$, $j \in \{1, \dots, d\}$. This means that the margins are estimated based on ranks, while the copula parameter $\boldsymbol{\theta}_C$ is estimated based on the

pseudo-observations $\hat{\boldsymbol{U}}_i$, $i \in \{1, \dots, n\}$, given by

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1},$$

where R_{ij} denotes the *rank* of X_{ij} among all X_{ij}, $i \in \{1, \dots, n\}$. To be more precise, after computing the pseudo-observations (step 1), the copula parameter $\boldsymbol{\theta}_{0,C}$ is estimated (step 2) via

$$\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MPLE}} = \operatorname*{argsup}_{\boldsymbol{\theta}_C \in \Theta_C} \sum_{i=1}^{n} \ell_C(\boldsymbol{\theta}_C; \hat{U}_{i1}, \dots, \hat{U}_{id}) = \operatorname*{argsup}_{\boldsymbol{\theta}_C \in \Theta_C} \sum_{i=1}^{n} \log c(\hat{\boldsymbol{U}}_i; \boldsymbol{\theta}_C).$$

This estimator is introduced by Genest *et al.* (1995), who show consistency and asymptotic normality under suitable regularity conditions (see also Tsukahara, 2005). Genest and Werker (2002) show that $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MPLE}}$ is not asymptotically efficient in general. Kim *et al.* (2007) compare $\hat{\boldsymbol{\theta}}_n^{\mathrm{MLE}}$, $\hat{\boldsymbol{\theta}}_n^{\mathrm{IFME}}$, and $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MPLE}}$ in a simulation study and argue in favor of the latter overall in terms of the (simulated) mean squared error, especially with respect to robustness against misspecification of the margins. Note that this study, as many others in the literature, is only carried out in the bivariate case. Although from the notation Kim *et al.* (2007) treat the general d-dimensional case, their conclusions are drawn only from the examined bivariate case.

2.1.4. *Estimation under known margins.* Although $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MPLE}}$ resembles an MLE based on data from the copula C itself, in practical applications one never observes data from C directly. Indeed, one should keep in mind that the pseudo-observations do not form a perfect random sample from C (even if the $\boldsymbol{X}_i$'s form a random sample) as they are neither independent nor perfectly following a univariate standard uniform distribution; see, for example, Genest *et al.* (2009). If, however, the $\boldsymbol{X}_i$'s form a random sample and we know all marginal distribution functions F_j, $j \in \{1, \dots, d\}$, we can build

$$U_{ij} = F_j(X_{ij}), \ i \in \{1, \dots, n\}, \ j \in \{1, \dots, d\},$$

and compute an estimator for $\boldsymbol{\theta}_{0,C}$ based on the i.i.d. sample $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, from C. Let us stress again that the assumption of known margins is unrealistic in practical applications. However, we include this *known margins maximum likelihood estimator (KMMLE)* $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{KMMLE}}$ in our simulation study to numerically address the question to what amount "replacing the $\boldsymbol{U}_i$'s by the $\hat{\boldsymbol{U}}_i$'s" affects the precision of the estimators considered, that is, how large is the estimation error (and how is the estimation error affected by the number of dimensions) when going from known margins to rank-based estimated margins. Note that a more interesting question would be how much information is lost by going from

$\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MLE}}$ to $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MPLE}}$, but computing $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MLE}}$ in the high-dimensional cases we consider is challenging at best.

2.1.5. *The pairwise Kendall's tau estimator.* Similar to method-of-moment estimators, Genest and Rivest (1993) suggested to estimate the parameter $\theta_{0,C}$ of one-parameter bivariate Archimedean copulas by matching Kendall's tau $\tau = \tau(\theta_C)$ as a function of θ_C with its sample version $\hat{\tau}_n$, that is, by solving $\tau(\theta_C) = \hat{\tau}_n$ with respect to θ_C. Kendall's tau is available for many copulas in closed or semi-closed form and, thus, computing this estimator is often straightforward. It has been extended to the multivariate case ($d > 2$) in two ways. First, for elliptical copulas, it has been applied in the multi-parameter case to estimate the correlation matrix through pairwise inverting Kendall's tau; see, for example, McNeil *et al.* (2005, p. 231). Second, it has been applied to estimate parameters in one-parameter exchangeable copula models where the bivariate margins depend on $\theta_{0,C}$; see Berg (2009), Kojadinovic and Yan (2010b) and Savu and Trede (2010). In the latter case, if $\hat{\tau}_{n,j_1j_2}$ denotes the sample version of Kendall's tau based on the random sample $(X_{ij_1}, X_{ij_2})^\top$, $i \in \{1, \ldots, n\}$, then $\theta_{0,C}$ is estimated by the *pairwise Kendall's tau estimator (PKTE)*

$$\hat{\theta}_{n,C}^{\mathrm{PKTE}} = \tau^{-1}\left(\binom{d}{2}^{-1} \sum_{1 \le j_1 < j_2 \le d} \hat{\tau}_{n,j_1j_2}\right), \tag{5}$$

that is, by matching Kendall's tau with the mean over all different pairwise sample versions of Kendall's tau (which seems intuitive in exchangeable models). In the bivariate case, properties of this estimator follow from Kendall's tau being a U-statistic (see the book by Lee, 1990 for more details). The multivariate case is more difficult; see Kojadinovic and Yan (2010b).

Note that we can base the computation of the pairwise sample versions of Kendall's tau on either the random sample $\boldsymbol{X}_i$, $i \in \{1, \ldots, n\}$, or the pseudo-observations $\hat{\boldsymbol{U}}_i$, $i \in \{1, \ldots, n\}$, because, as a measure of concordance, Kendall's tau is invariant under strictly increasing transformations on the ranges of the underlying random variables.

Finally, let us remark that there exist multivariate versions of measures of association such as Kendall's tau (being applicable to $d > 2$); see, for example, Jaworski *et al.* (2010, pp. 209). Since they are less widely used and numerically more challenging to evaluate, they are not considered here.

2.2. Likelihood-based confidence-interval estimators

One way to obtain confidence intervals for $\theta_{0,C}$ (in the one-parameter case) or confidence regions for $\boldsymbol{\theta}_{0,C}$ (in the multi-parameter case) is by using asymptotic normality. In the one-dimensional case, this always gives symmetric confidence intervals about the estimator of $\theta_{0,C}$. Another disadvantage comes from a numerical point of view. In many copula models, computing derivatives of the

likelihood function is significantly more demanding than computing the likelihood itself; see, for example, the (one-parameter) Gumbel family in Hofert *et al.* (2012).

The approach we consider here is based on likelihoods only and does not require to compute derivatives of the likelihood. Confidence regions for $\boldsymbol{\theta}_{0,C}$, based on a random sample $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, from C, can be obtained via the *likelihood ratio statistic*, defined by

$$W(\boldsymbol{\theta}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n) = 2\left(\ell_C\left(\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MLE}}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n\right) - \ell_C(\boldsymbol{\theta}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n)\right).$$

As Davison (2003, p. 126) notes, the likelihood ratio statistic asymptotically follows a chi-square distribution:

$$W(\boldsymbol{\theta}_{0,C}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n) \xrightarrow{d} \chi^2_{p_C} \ (n \to \infty).$$

Based on this result, an asymptotic $1 - \alpha$ *confidence region* for $\boldsymbol{\theta}_{0,C}$ is given by

$$\left\{\boldsymbol{\theta}_C \in \Theta_C \,\middle|\, \ell_C(\boldsymbol{\theta}_C; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n) \geq \ell_C\left(\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MLE}}; \boldsymbol{U}_1, \dots, \boldsymbol{U}_n\right) - q_{\chi^2_{p_C}}(1-\alpha)/2\right\}. \tag{6}$$

In Section 3 we investigate how the coverage probability of $1 - \alpha$ is affected if we replace $\hat{\boldsymbol{\theta}}_{n,C}^{\mathrm{MLE}}$ and, correspondingly, the non-observable random sample $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, from C in (6) by the estimators presented in Section 2.1.

3. Simulation studies

The theoretical conditions involved in the asymptotics of the estimators presented in the last section are difficult to compute theoretically (see Kojadinovic and Yan, 2010b). We therefore conduct two simulation studies to assess how point and interval estimators are affected by different substitutes for a perfect random sample $\boldsymbol{U}_i$, $i \in \{1, \dots, d\}$, from the underlying copula. To limit the computational burden, we only consider exchangeable copulas, but go up to rather high dimensions to get a realistic picture of how the dimension enters the equations.

3.1. A word concerning the implementation

The results presented in this section are based on the following setup. All our computations are done in R 2.15.0 with the package `copula` (version $\geq$ 0.99.0) as can be obtained from http://nacopula.r-forge.r-project.org/. The required R scripts to reproduce our studies may be obtained from the authors upon request. The computations are carried out on the computer cluster *Brutus* of ETH Zurich which runs CentOS 5.4. The batch jobs are run on nodes with four quad-core AMD Opteron 8380 CPUs and 32 GB of RAM.

The procedures are numerically and computationally challenging in many ways and much effort has gone into accurate and efficient implementation in R. We refer the interested reader to the package source code, documentation, and examples.

3.2. Study 1: Bias, root mean-squared error

In our first simulation study, we consider the sample sizes $n \in \{20, 50, 100, 250\}$, the dimensions $d \in \{2, 5, 10, 20, 50, 100\}$, dependencies corresponding to the Kendall's tau $\tau \in \{0.25, 0.5, 0.75\}$, the homogeneous (elliptical) Gaussian and t_4 copulas, as well as the one-parameter (Archimedean) copula families of Ali–Mikhail–Haq (AMH), Clayton, Frank, Gumbel and Joe (parameters are chosen such that the above Kendall's tau are matched). For each of these combinations (the exception being AMH for which the range of admissible Kendall's tau is bounded from above by 1/3), we generate a random sample of the corresponding size and compute $\hat{\theta}_{n,C}^{\mathrm{KMMLE}}$, $\hat{\theta}_{n,C}^{\mathrm{IFME}}$, $\hat{\theta}_{n,C}^{\mathrm{MPLE}}$, and $\hat{\theta}_{n,C}^{\mathrm{PKTE}}$. We repeat this procedure $N = 500$ times and compute the bias and the root mean-squared error for each of the four estimators.

Consider one of the N runs. The computations for the KMMLE $\hat{\theta}_{n,C}^{\mathrm{KMMLE}}$ are based on the random sample $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, from the corresponding copula. This case, although unrealistic in real-life examples (especially in high dimensions), is included as a benchmark.

For IFME $\hat{\theta}_{n,C}^{\mathrm{IFME}}$, the same copula samples $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, are taken. However, they are first transformed to standard normal margins (that is, transformed to a random sample $\boldsymbol{X}_i$, $i \in \{1, \dots, n\}$, where $X_{ij} = \Phi^{-1}(U_{ij})$ for each i, j; Φ denotes the distribution function of the standard normal distribution) and the sample mean and sample standard deviation (asymptotically equivalent to the MLE for the standard deviation) are used as estimators of the parameters of the normal margins. We hereby assume that we already know the distributional families for all margins, but not their true underlying parameters. After the margins have been estimated, the sample $\boldsymbol{X}_i$, $i \in \{1, \dots, n\}$, is transformed componentwise via the (estimated) marginal probability integral transforms, and the estimator of the copula parameter is computed based on this transformed sample. One could interpret this sample as "parametric(ally estimated) pseudo-observations", hence we refer to it as *para-pseudo-observations*.

For MPLE $\hat{\theta}_{n,C}^{\mathrm{MPLE}}$, we also consider the same copula samples $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, as for $\hat{\theta}_{n,C}^{\mathrm{KMMLE}}$, but we first transform the samples to pseudo-observations $\hat{\boldsymbol{U}}_i$, $i \in \{1, \dots, n\}$. Clearly, the same realizations would have been obtained by starting with any marginal distributions due to the fact that ranks are built.

The PKTE $\hat{\theta}_{n,C}^{\mathrm{PKTE}}$ is also computed based on the same simulated copula samples $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, as the other estimators. Since Kendall's tau is a measure of concordance, we could have transformed the samples to any other margins before and would have received exactly the same value of the estimator.

3.3. Study 1: Results

Overall, the bias was quite small so that the sample standard deviation was close to the root mean squared error. We therefore only report on the latter.

Figure 1 shows the root mean squared error of all four estimators as a function in the dimension for the considered copula families and Kendall's tau (for AMH, note that $\tau \in [0, 1/3)$). The sample size was fixed to $n = 100$. For $\hat{\theta}_{n,C}^{\mathrm{KMMLE}}$, the root mean squared error seems linearly decreasing in d in log–log scale (with the only exception being the t_4 copula). For the exchangeable Archimedean families, this behavior has already been found numerically by Hofert *et al.* (2012) and translates to the approximate behavior MSE $\propto 1/(nd)$ for the mean squared error (MSE). The theoretical verification of this result is still an open research problem.

To gain more insight in the realized values, Figure 2 shows, for fixed sample size $n = 100$, boxplots of the root absolute deviations $\sqrt{|\hat{\theta}_{n,C}^{\cdot} - \theta_{0,C}|}$ of the different estimators $\hat{\theta}_{n,C}^{\cdot}$ (being $\hat{\theta}_{n,C}^{\mathrm{KMMLE}}$, $\hat{\theta}_{n,C}^{\mathrm{IFME}}$, $\hat{\theta}_{n,C}^{\mathrm{MPLE}}$ and $\hat{\theta}_{n,C}^{\mathrm{PKTE}}$) for the different dimensions considered.

Overall, the simulations show several important results. First note that the root mean squared error decreases for increasing dimensions d (uniformly for all considered Kendall's tau and copula families). However, as visible already in the bivariate case (especially for larger Kendall's tau), the impact of not knowing the marginal distribution functions (in comparison to knowing them) on the precision of point estimators for the parameter in the exchangeable copula models considered is increasingly severe in higher dimensions. This is important in that it stresses that we cannot expect results about statistical inference in small dimensions to carry over to higher dimensions. Furthermore, in high dimensions, the performance of the different estimators under unknown margins does not differ very much. Rather, computational aspects will play a role which estimator to choose. For example, note that the complexity of computing $\hat{\theta}_{n,C}^{\mathrm{PKTE}}$ is at least of the order of $d^2 n \log n$, hence quadratic in d; see (5). In large dimensions, this becomes computationally demanding. The likelihood-based estimators might be computed significantly faster in high dimensions, although one of the reasons why $\hat{\theta}_{n,C}^{\mathrm{PKTE}}$ has been considered has been computational simplicity.

3.4. Study 2: Coverage probabilities of likelihood-based confidence intervals

Besides estimation, in a realistic setup, we typically would like to compute confidence intervals for the true underlying copula parameter. As mentioned earlier, it is often convenient to compute confidence intervals based on likelihoods, thus not requiring to compute possibly complicated derivatives of the likelihood or score functions. To compute such confidence intervals, one, theoretically, would like to use a sample $\boldsymbol{U}_i$, $i \in \{1, \ldots, n\}$, from the underlying copula C. Since such a sample is available only if the data are a random sample from H with copula C and (a) known margins, one would resort to either (b)

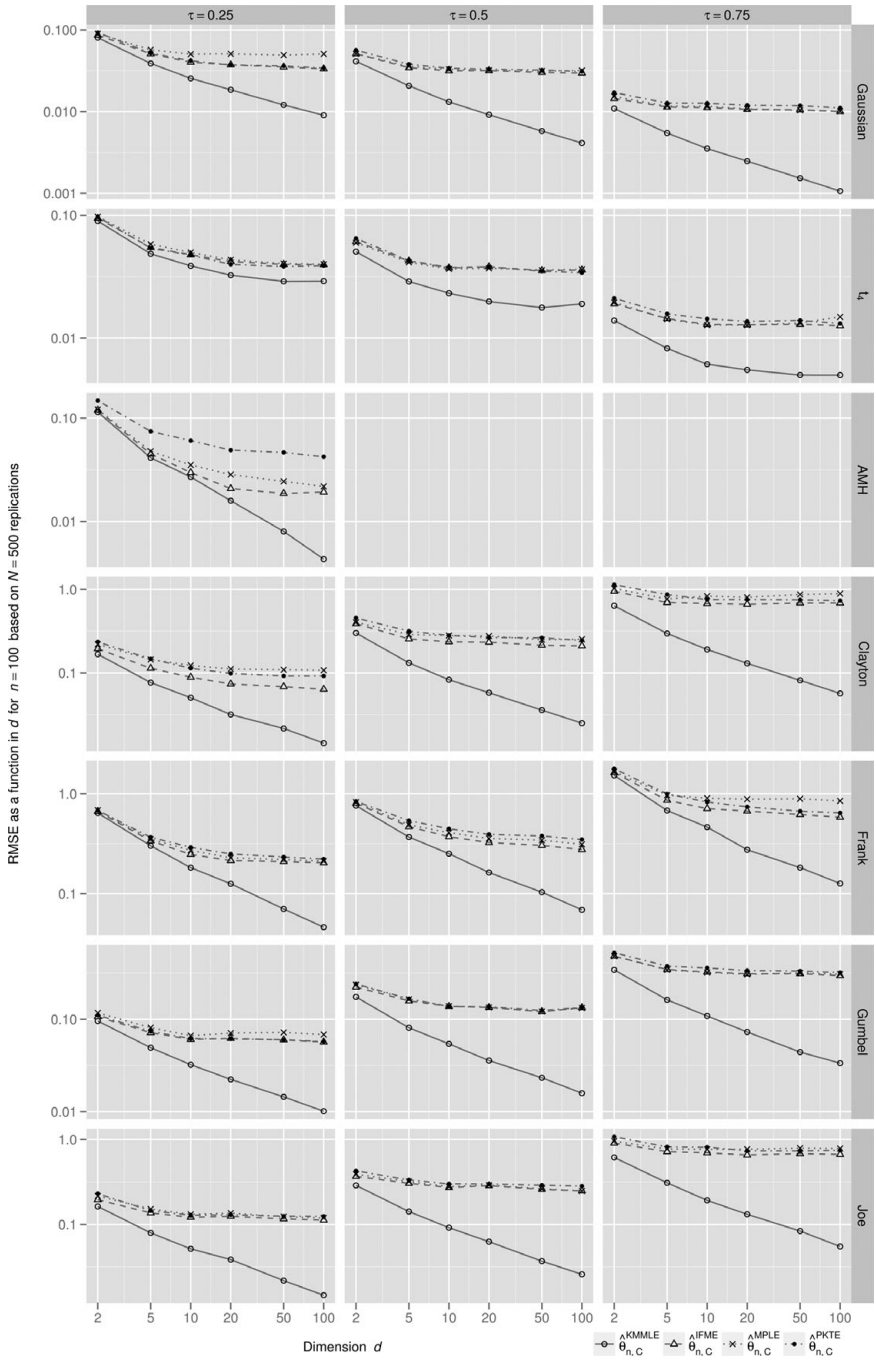

FIGURE 1: Study 1: root mean squared error ($N = 500$ replications) as a function in d in log–log scale for $n = 100$. Note that the family of AMH is limited to $\tau \in [0, 1/3)$.

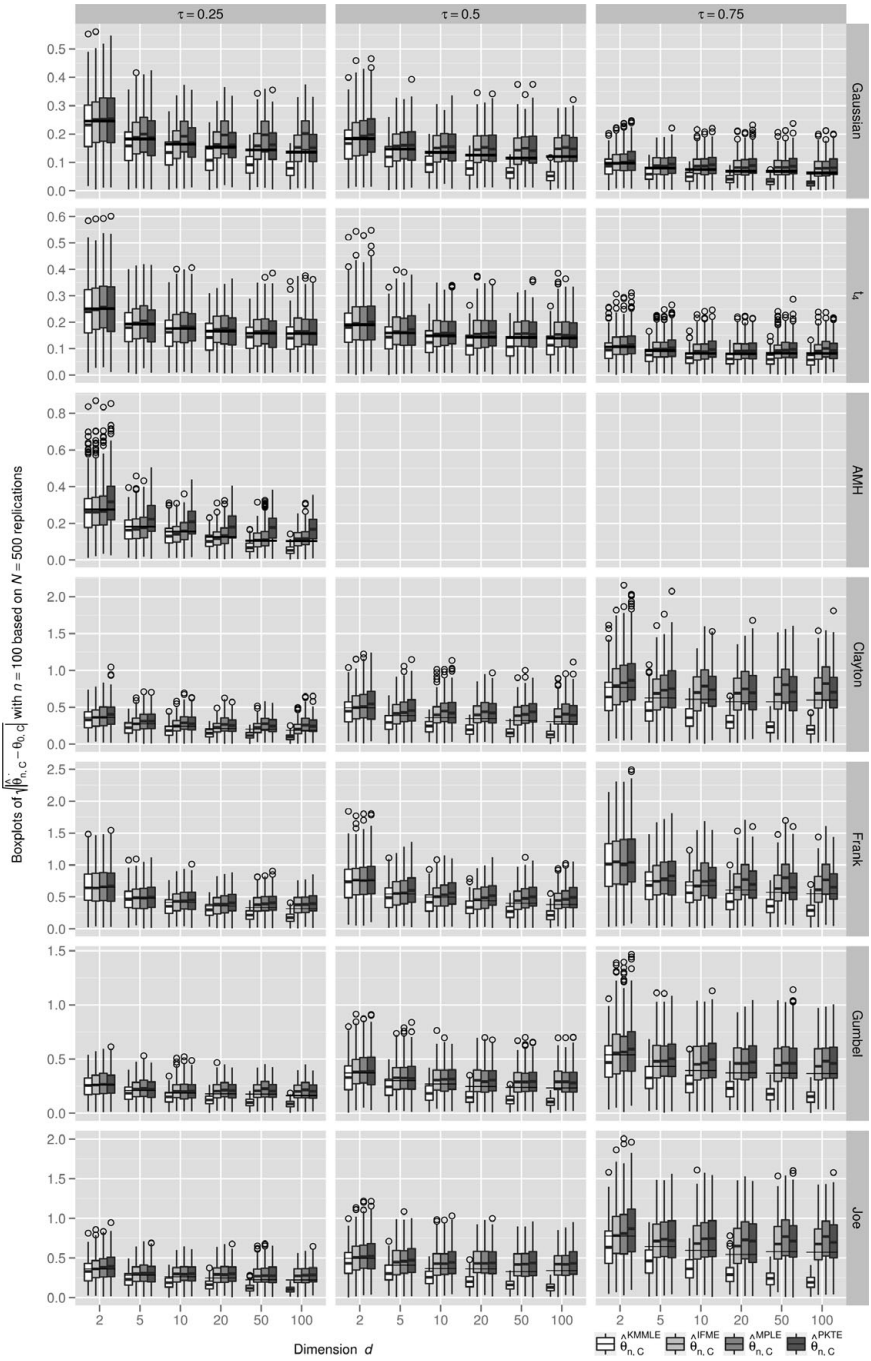

FIGURE 2: Study 1: boxplots of the $N = 500$ root absolute deviations $\sqrt{|\hat{\theta}_{n,C}^{\cdot} - \theta_{0,C}|}$.

para-pseudo-observations or (c) pseudo-observations. To what amount is the coverage probability of likelihood-based confidence intervals affected by not having perfect copula observations available? How does this error behave as a function in the number of dimensions? These are questions Study 2 addresses.

We base our computations on $N = 500$ repetitions, the same dimensions, copula families and dependencies measured in terms of Kendall's tau as before. For all combinations of these variables to our study and each of the N runs, we draw a sample of size $n = 100$, transform it by either (a) the identity, (b) para-pseudo-observations, or (c) pseudo-observations, and determine if the true underlying parameter is covered by the likelihood-based confidence interval as given in (6). Based on all runs, we thus obtain estimated coverage probabilities of likelihood-based confidence intervals. We consider $1-\alpha$ confidence intervals for $\alpha \in \{0.01, 0.05\}$.

3.5. Study 2: Results

Figure 3 shows the estimated coverage probabilities in percent of likelihood-based confidence intervals obtained under (a)–(c). To be more precise, (6) is used to numerically determine the confidence interval for each of the N runs. The coverage probabilities are then determined as the percentage of runs for which the confidence intervals contain the true underlying parameter $\theta_{0,C}$.

Similar to the results from Study 1, the impact of not knowing the margins is increasingly severe in higher dimensions. Also, if the margins are not known exactly, the differences in the approaches (b) (parametrically estimated margins) and (c) (non-parametrically estimated margins) do not seem to make a big difference. Finally, let us remark that from a practical point of view, these problems are severe and there is no easy solution known yet.

4. Conclusion

We conducted two simulation studies with particular focus on high dimensions. In the first study, we investigated the precision of the following point estimators: the maximum likelihood estimator under known margins (included as a benchmark), the inference functions for margins estimator the maximum pseudo-likelihood estimator and the pairwise Kendall's tau estimator. In the second study, we considered how coverage probabilities of likelihood-based confidence intervals are affected by not knowing the marginal distribution functions (which is typically, if not always, the case in practical applications, especially in high dimensions).

Overall, the results of our studies indicate that the root mean squared error decreases with increasing dimensions. However, not knowing the margins has an increasingly severe effect on both point and interval estimators the higher the considered dimension. Although there is no simple solution to this problem yet, we believe it is important to be aware of it when working in higher dimensions,

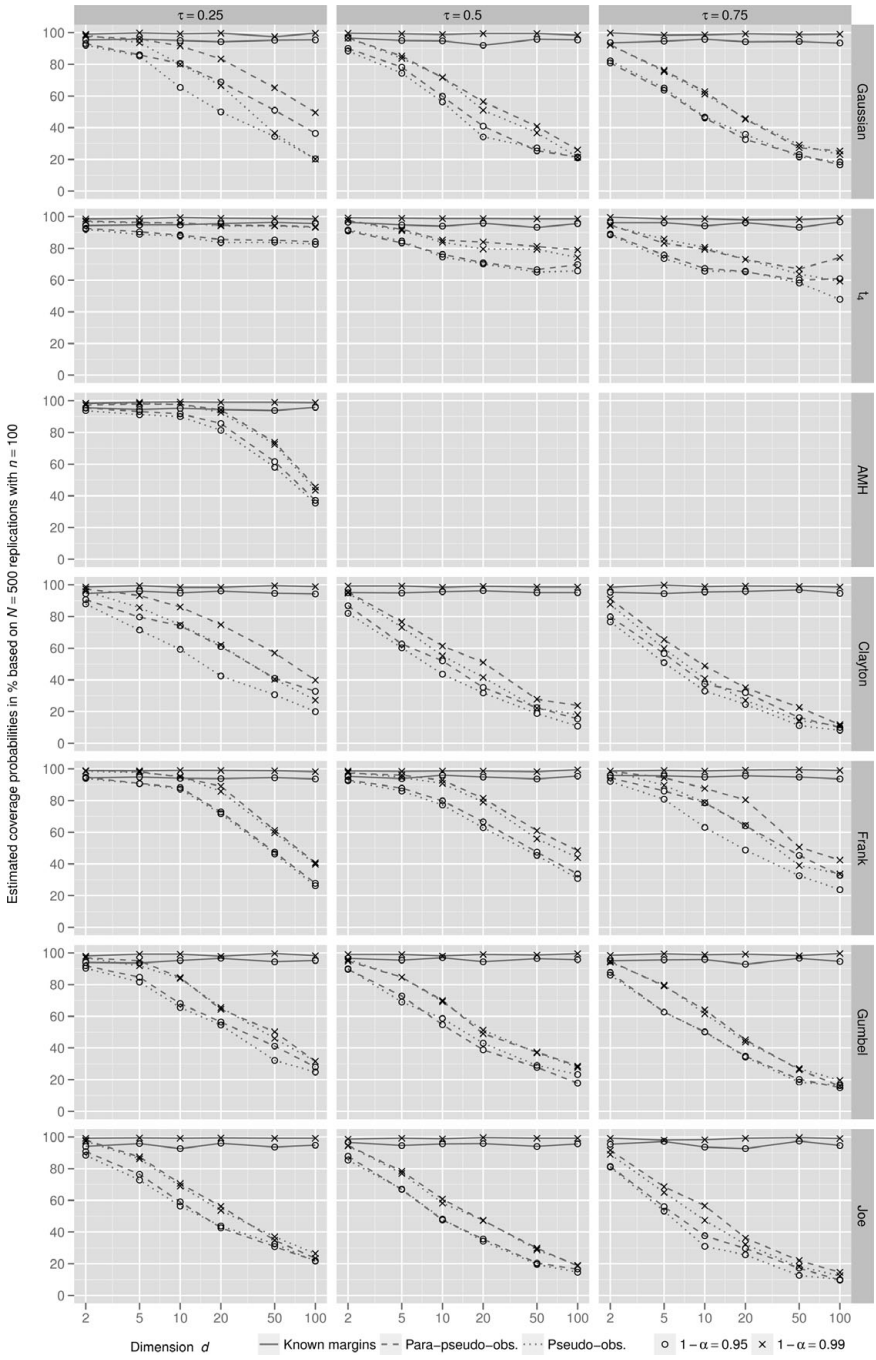

FIGURE 3: Study 2: estimated coverage probabilities in percent ($N = 500$ replications) of likelihood-based confidence intervals under known, parametrically, and non-parametrically estimated margins.

especially since problems of this kind have been partly ignored in the existing literature to this day.

Let us finish with some remarks on possible future research in this direction. The performance of the inference functions for margins estimator might depend on the choice of the marginal distributions. It is unclear how the performance in terms of the root mean squared error, for example, differs in the case of other marginals (such as t, log-normal or log-t distributions). Concerning the rate of decrease of the root mean squared error for increasing dimensions under known margins, this effect contrasts the expectations according to the curse of dimensionality (see Bellman, 1957). It can be motivated through the exchangeability of the considered models and similar results can be seen in partially exchangeable models such as nested Archimedean copulas, where each sector or child copula is Archimedean and thus the above results directly apply to the sectors. An interesting question would be in what way one can move away from exchangeability such that the root mean squared error still decreases in the dimension and thus the curse of dimensionality does not strike yet. As can already be seen from our first study here, even in the case of perfect symmetry, the story looks quite a bit different when the marginal distributions are not known. From a practical point of view, this therefore does not leave too much hope that one can move away from symmetry much, although, theoretically, it would be interesting to investigate.

Acknowledgements

We would like to thank Niels Hagenbuch (Karolinska Institutet, Stockholm) for giving us feedback on this work. Furthermore, Marius Hofert (Willis Research Fellow) thanks Willis Re for financial support while this work was being completed. The authors also take pleasure in thanking the referees for their constructive comments on an earlier version of the paper.

References

Arbenz, P., Hummel, C. and Mainik, G. (2012) Copula based hierarchical risk aggregation through sample reordering. *Insurance: Mathematics and Economics*, **51**, 122–133.

Bellman, R.E. (1957) *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Berg, D. (2009) Copula goodness-of-fit testing: An overview and power comparison. *The European Journal of Finance*, **15**, 675–701.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-Dimensional Data*. Springer.

Davison, A.C. (2003) *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Dobrić, J. and Schmid, F. (2007) A goodness of fit test for copulas based on Rosenblatt's transformation. *Computational Statistics & Data Analysis*, **51**, 4633–4642.

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

Genest, C., Rémillard, B. and Beaudoin, D. (2009) Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, **44**, 199–213.

GENEST, C. and RIVEST, L.-P. (1993) Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, **88**(423), 1034–1043.
GENEST, C. and WERKER, B.J.M. (2002) Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In *Distributions with Given Marginals and Statistical Modelling* (ed. C.M. Cuadras, J. Fortiana and J.A. Rodríguez-Lallena), pp. 103–112. Dordrecht: Kluwer.
HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer.
HOFERT, M., MÄCHLER, M. and MCNEIL, A.J. (2012) Likelihood inference for Archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis*, **110**, 133–150.
HOFERT, M. and SCHERER, M. (2011) CDO pricing with nested Archimedean copulas. *Quantitative Finance*, **11**(5), 775–787.
JAWORSKI, P., DURANTE, F., HÄRDLE, W.K. and RYCHLIK, T. (eds.) (2010) *Copula Theory and Its Applications*. Lecture Notes in Statistics – Proceedings, vol. 198. Springer.
JOE, H. and XU, J.J. (1996) *The Estimation Method of Inference Functions for Margins for Multivariate Models*. Technical Report 166, Department of Statistics, University of British Columbia.
KIM, G., SILVAPULLE, M.J. and SILVAPULLE, P. (2007) Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis*, **51**, 2836–2850.
KOJADINOVIC, I. and YAN, J. (2010a) Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics*, **47**, 52–63.
KOJADINOVIC, I. and YAN, J. (2010b) Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, **34**(9), 1–20.
LEE, A.J. (1990). *U-Statistics: Theory and Practice*. Dekker.
MCNEIL, A.J., FREY, R. and EMBRECHTS, P. (2005) *Quantitative Risk Management: Concepts, Techniques, Tools*. Princeton, NJ: Princeton University Press.
PORTNOY, S. (1988) Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, **16**(1), 356–366.
SAVU, C. and TREDE, M. (2010) Hierarchies of Archimedean copulas. *Quantitative Finance*, **10**(3), 295–304.
TSUKAHARA, H. (2005) Semiparametric estimation in copula models. *The Canadian Journal of Statistics*, **33**(3), 357–375.
WEIß, G.N.F. (2010) Copula parameter estimation: Numerical considerations and implications for risk management. *The Journal of Risk*, **13**(1), 17–53.

PAUL EMBRECHTS (CORRESPONDING AUTHOR)
RiskLab, Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland,
E-Mail: embrechts@math.ethz.ch

MARIUS HOFERT
RiskLab, Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland,
E-Mail: marius.hofert@math.ethz.ch