Model robust inference for copulae via two-stage maximum likelihood estimation

Vinnie Ko*, Nils Lid Hjort
Department of Mathematics, University of Oslo
PB 1053, Blindern, NO-0316 Oslo, Norway

December 2017

Abstract

This article is concerned with inference in parametric copula setups, where both the marginals and the copula have parametric forms. For such models, two-stage maximum likelihood estimation, often referred to as inference function for margins, is used as an attractive alternative to the full or joint maximum likelihood estimation strategy. Previous studies of the two-stage maximum likelihood estimator have largely been based on the assumption that the chosen parametric model is the true model. We study the impact of dropping this true model assumption, both theoretically and numerically. We first show that the two-stage maximum likelihood estimator is consistent for a well-defined least false parameter value, different from the analogous least false parameter associated with the full maximum likelihood procedure. Then we demonstrate limiting normality of the full vector of estimators, with concise matrix notation for the variance matrices involved. Along with consistent estimators for these, we have built a model-robust machinery for inference on parametric copula models. The special case where the parametric model is assumed to hold corresponds to situations studied earlier in the literature, with simpler formulae for variance matrices.

As a numerical illustration, we perform a set of simulations and find that ignoring model uncertainty often leads to over-confident results. In addition, we observe that the two-stage maximum likelihood estimator is still highly efficient when the true model assumption is dropped and thus the model robust asymptotic variance formulae are used. Additionally, we discover that using highly misspecified models can lead to situations where the asymptotic variance of the two-stage maximum likelihood estimator is lower than that of full maximum likelihood estimator. We also analyse five-dimensional Norwegian precipitation data, with results concordant with that of the simulation study.

Keywords: copula, inference functions for margins, two-stage maximum likelihood. large-sample inference, model misspecification, model robust

1 Introduction and copula models

The popularity of copula modelling and methods has increased rapidly in the last decade and now they are regularly used in fields like biostatistics, hydrology, finance and actuarial science (Embrechts, 2009). One

*Corresponding author.
E-mail addresses: vinniebk@math.uio.no (V. Ko), nils@math.uio.no (N.L. Hjort)

of the more popular estimation techniques for parametric copulae with parametric margins is the two-stage maximum likelihood estimation (two-stage ML estimation), first introduced by Shih & Louis (1995). This two-stage method is also often referred to as inference function for margins (IFM). The asymptotic behaviour and efficiency of this method have been studied both theoretically and numerically by Shih & Louis (1995), Xu (1996), Joe (1997, 2005), Andersen (2005) and Kim *et al.* (2007).

Although the insights and results from these studies are fruitful, they are generally based on the assumption that the chosen copula and marginal distributions are the true model that generated data. We refer to this as the 'true model assumption', and the limitations and impact of this assumption for two-stage ML estimation in copula contexts has not been studied earlier, to our knowledge. Andersen (2004) has worked with model robust versions of asymptotic variances. This was in the context of composite likelihoods, however, and her work does not give attention to the inference and impact of model robustness. In this paper we develop model robust inference methods based on two-stage ML estimators, and study their theoretical and numerical properties under possible model misspecification.

Our technical setting is as follows. Let $(Y_1, \cdots, Y_d)^{\mathrm{T}}$ be a d-variate continuous stochastic variable originating from a joint density $g(y_1, \cdots, y_d)$, and let further $y_i = (y_{i,1}, \cdots, y_{i,d})^{\mathrm{T}}$ for $i = 1, \ldots, n$ be independent observations of this variable. Typically, this true joint distribution g is unknown. Let $f(y_1, \cdots, y_d, \eta)$ be our choice of parametric approximation of g, with η the parameter vector, belonging to some connected subset of the appropriate Euclidean space. Further, G and $F(\cdot, \eta)$ indicate cumulative distribution functions corresponding to g and $f(\cdot, \eta)$, respectively. In addition, $G_j(y_j)$ and $F_j(y_j, \alpha_j)$ indicate j-th marginal distribution functions corresponding to G and $F(\cdot, \eta)$ respectively, with α_j the parameter vector pertaining to modelling margin component j.

Starting from the joint distribution the parametric $F(\cdot, \eta)$, Sklar's theorem (Sklar, 1959) implies that there is a copula $C(u_1, \ldots, u_d, \theta)$ that satisfies

$$F(y_1, \cdots, y_d, \eta) = C(F_1(y_1, \alpha_1), \cdots, F_d(y_d, \alpha_d), \theta),$$

with θ the vector of parameters for the copula. For the full parameter vector, now conveniently blocked into parameters for margins and the part for the copula, we use

$$\eta = (\alpha^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}} = (\alpha_1^{\mathrm{T}}, \cdots, \alpha_d{}^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}.$$

When $F(y_1, \cdots, y_d, \eta)$ is continuous, $C(\cdot, \theta)$ is unique. If we assume that $F_1(y_1, \alpha_1), \cdots, F_d(y_d, \alpha_d)$ are absolutely continuous and strictly increasing, $C(\cdot, \theta)$ can be differentiated,

$$f(y_1, \cdots, y_d, \eta) = c\left(F_1(y_1, \alpha_1), \cdots, F_d(y_d, \alpha_d), \theta\right) \prod_{j=1}^{d} f_j(y_j, \alpha_j), \tag{1}$$

where $c(u_1, \ldots, u_d) = \partial^d C(u_1, \ldots, u_d, \theta)/\partial u_1 \cdots \partial u_d$ and $f_j(y_j, \alpha_j) = \partial F_j(y_j, \alpha_j)/\partial y_j$ (see Nelsen (2006)). Analogously, we can decompose the true density g into marginal densities and copula density and obtain

$$g(y_1, \cdots, y_d) = c_0\left(G_1(y_1), \cdots, G_d(y_d)\right) \prod_{j=1}^{d} g_j(y_j),$$

with $c_0(\cdot)$ the true copula.

One of the most important families of copulae is the Archimedean copula family. An Archimedean family copula has the form

$$C(u_1, \cdots, u_d) = \phi^{-1}(\phi(u_1) + \cdots + \phi(u_d))$$

where ϕ is its generator. One of the most common Archimedean copulae is the Gumbel copula, which is defined as

$$C(u_1, \cdots, u_d) = \exp\left[-\left\{(-\log u_1)^\theta + \cdots + (-\log u_d)^\theta\right\}^{\frac{1}{\theta}}\right].$$

Another member of the family is the Frank copula, defined as

$$C(u_1, \cdots, u_d) = -\frac{1}{\theta}\log\left[1 + \frac{\prod_{i=1}^{d}\{\exp(-\theta u_i) - 1\}}{\{\exp(-\theta - 1)\}^{d-1}}\right].$$

For further parametric copula constructions, see e.g. Joe (1997); Nelsen (2006); Genest & Rivest (1993); Genest & Favre (2007), and with several multiparameter copula models exhibited and discussed in Coles *et al.* (1999); Nikoloulopoulos *et al.* (2012).

The further structure of this paper is as follows. In Section 2, we briefly describe how the two-stage ML estimation method works. In Section 3, we first derive the limit distribution for two-stage ML estimators outside model conditions, including proving consistency towards the relevant least false parameter. After this, we examine consequences of the true model assumption for the asymptotic distribution. The resulting model non-robust asymptotic variance formula is essentially similar to that given in Joe (2005), but one difference is that we choose to use quantities that are more in line with the classical ML estimation theory. Our results give rise to clear recipes for confidence intervals, confidence curves, and hypothesis tests, based on two-stage ML estimators, discussed in Section 4.

In Sections 5 and 6 we study the numerical behaviour of our model robust inference methods by using a set of simulations and a data set of Norwegian precipitation. Generally speaking, the two-stage ML estimator $\widetilde{\eta}$ has a variance matrix, say Σ, for which our apparatus provides two estimators, say $\widetilde{\Sigma}_A$ computed under model conditions and $\widetilde{\Sigma}_B$ using the model-robust machinery. We find first that when data really stem from the model used, then both variance matrix estimators tend to be in agreement, but with more variability for the second method. Next, when the data generating mechanism lies outside the model, the second method aims at the correct matrix, whereas the first is not consistent, and often will be too small. Furthermore, we observe that a high degree of model misspecification may lead to situations where the limiting variance matrix of the two-stage ML estimator is smaller than the corresponding variance matrix of the full ML estimator. This can not happen when the data generating mechanism is inside the parametric model. By classical theorems on the optimality of ML estimators, such as the Hájek–Le Cam convolution theorem (Hájek, 1970), each competing sequence of estimators will have a limiting variance matrix at least as large as the one for ML estimation (the matrix difference is nonnegative definite), under model conditions and a few further regularity assumptions on competitors. But these optimality theorems for ML estimation do not

apply when the data generating mechanism is outside the parametric model.

In our final Section 7 we offer a list of concluding remarks, some pointing to further research work. In particular, we mention model selection criteria, and explain briefly that our two-stage ML machinery may be extended to classes of conditional copula regression models.

2 Two-stage maximum likelihood

2.1 Maximum likelihood and Kullback–Leibler divergence

With observations from a model parametrised via a parameter vector η, the ML estimator $\widehat{\eta}$ is the maximiser of $\ell(\eta)$, the log-likelihood function of the model for the given observations. Properties of ML estimators are extensively covered by the classic literature in statistics, including Le Cam (1990) and Casella & Berger (2002). In this paper, we use '$\widehat{\ }$' to indicate that a quantity is estimated by full or joint ML, and '$\widetilde{\ }$' to indicate that the quantity in question is estimated by two-stage ML, covered in Section 2.2.

The Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) measures the extent to which one probability distribution diverges from another. The KL divergence from g to f is defined as

$$\mathrm{KL}(g, f(\cdot, \eta)) = \int g(y) \log \frac{g(y)}{f(y, \eta)} \, \mathrm{d}y.$$

In our technical setting, the true density g is the same for all models. Thus, minimising the KL divergence is equivalent to maximising $\int g(y) \log f(y, \eta) \, \mathrm{d}y$.

Under the usual regularity assumptions and assuming that the integral is finite, the law of large numbers gives

$$\frac{1}{n}\ell = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i, \eta) \ \xrightarrow{p} \ \int g(y) \log f(y, \eta) \, \mathrm{d}y = \mathrm{E}_G \, \log f(y, \eta).$$

Here and later it is understood that the convergence relates to the sample size n growing beyond bounds. The above result implies under mild and standard regularity conditions that $\widehat{\eta}$, the ML estimator of η, will tend asymptotically to η_0, the least false parameter value, which is the minimiser of the KL divergence and the maximiser of $\int g(y) \log f(y, \eta) \, \mathrm{d}y$. Thus,

$$\widehat{\eta} \ \xrightarrow{p} \ \eta_0 = \arg\min_{\eta} \, \mathrm{KL}(g, f(\cdot, \eta)) = \arg\max_{\eta} \, \int g(y) \log f(y, \eta) \, \mathrm{d}y.$$

When the parametric model is correctly specified (i.e. $g(y) = f(y, \eta_0)$, for a suitable η_0), the minimum of the KL divergence is zero and η_0 is called true parameter value.

By applying the decomposition (1) on our setting and $f(y, \eta)$, we have that $\widehat{\eta}$ is a consistent estimator of the least false parameter η_0, say $\eta_{0,\mathrm{ml}} = (\alpha_{0,\mathrm{ml}}^{\mathrm{T}}, \theta_{0,\mathrm{ml}}^{\mathrm{T}})^{\mathrm{T}}$.

2.2 Two-stage maximum likelihood estimator

When the dimension of the copula (d) gets higher, the ML estimator becomes computationally more demanding and not even feasible when the number of parameters is high (Joe, 1997). To avoid this problem,

the two-stage ML is a natural alternative estimation strategy, first proposed, for the two-dimensional case, by Shih & Louis (1995).

When parametric families for the copula and the margins are chosen, the two-stage ML estimator works as follows. Stage 1: For each j in $1 \leq j \leq d$, obtain $\widetilde{\alpha}_j$, the marginal ML estimate of α_j, by maximising $\ell_{f_j} = \sum_{i=1}^n \log f_j(y_{i,j}, \alpha_j)$ with respect to α_j. Stage 2: Plug in $\widetilde{\alpha}_1, \cdots, \widetilde{\alpha}_d$ from stage 1, to get

$$\ell\left(\widetilde{\alpha}, \theta\right) = \sum_{i=1}^{n} \left\{\log f_1(y_{i,1}, \widetilde{\alpha}_1) + \cdots + \log f_d(y_{i,d}, \widetilde{\alpha}_d) + \log c\left(F_1(y_{i,1}, \widetilde{\alpha}_1), \cdots, F_d(y_{i,d}, \widetilde{\alpha}_d), \theta\right)\right\}.$$

Then, θ is estimated by maximising $\ell(\widetilde{\alpha}, \theta)$ with respect to θ, yielding

$$\widetilde{\theta} = \arg\max_{\theta} \ell\left(\widetilde{\alpha}, \theta\right) = \arg\max_{\theta} \ell_c(\widetilde{\alpha}, \theta)$$

where $\ell_c(\alpha, \theta) = \sum_{i=1}^n \log c\left(F_1(y_{i,1}, \alpha_1), \cdots, F_d(y_{i,d}, \alpha_d), \theta\right)$. The two-stage ML estimate $\widetilde{\eta} = (\widetilde{\alpha}^{\mathrm{T}}, \widetilde{\theta}^{\mathrm{T}})^{\mathrm{T}} = (\widetilde{\alpha}_1^{\mathrm{T}}, \cdots, \widetilde{\alpha}_d^{\mathrm{T}}, \widetilde{\theta}^{\mathrm{T}})^{\mathrm{T}}$ obtained in this way satisfies

$$\left(\frac{\partial \ell_{f_1}}{\partial \widetilde{\alpha}_d}, \cdots, \frac{\partial \ell_{f_d}}{\partial \widetilde{\alpha}_d}, \frac{\partial \ell_c}{\partial \widetilde{\theta}}\right) = 0.$$

There are $p_1 + \cdots + p_d + q$ parameters and equations here, where p_j is the dimension of α_j and q the dimension of θ. In the next section we demonstrate limiting normality for the full $p_1 + \cdots + p_d + q$-dimensional $\widetilde{\eta}$.

3 Large-sample behaviour of two-stage maximum likelihood estimators

To work out clear limit theorems for the two-stage ML estimators we are helped by the following regularity assumptions:

A1 $\eta \in \Theta$, where Θ is compact.

A2 $\log f_j(y, \alpha_j)$ is twice differentiable with respect to α_j.

A3 $\log c(F_1(y_1, \alpha_1), \ldots, F_d(y_d, \alpha_d), \theta)$ is thrice differentiable with respect to α and θ.

A4 There exists a function $R(y_j)$ such that $\mathrm{E}_G[R(Y_{i,j})] < \infty$ and $|\log f_j(y_j, \alpha_j)| \leq R(y)$ for all y_j, α_j.

A5 There exists a function $R(y, \alpha)$ such that $\mathrm{E}_G[R(Y_i, \alpha)] < \infty$ and $|\log c(y, \alpha, \theta)| \leq R(y, \alpha)$ for all y, α, θ.

In stage 1 of two-stage ML estimation, the parameters of each margin are estimated by using separate ML estimation. We may consequently use large-sample results from ML estimation theory directly, regarding $\widetilde{\alpha}_j$ behaviour. In particular, the estimator $\widetilde{\alpha}_j$ is now aiming for the least false value instead of the usual true parameter value under the true model assumption.

3.1 Large-sample results for stage 1 of two-stage ML estimation

Lemma 1. *Let $\widetilde{\alpha}$ be the ML estimator of the margin parameter vector $\alpha = (\alpha_1, \cdots, \alpha_d)^{\mathrm{T}}$ from stage 1 and let $\alpha_0 = (\alpha_{0,1}, \cdots, \alpha_{0,d})^{\mathrm{T}}$ be the least false value. Then we have*

$$\begin{aligned}\sqrt{n}\,(\widetilde{\alpha} - \alpha_0) &= \sqrt{n}\,\mathcal{I}_\alpha^{-1}\, U_{n,\alpha}(\alpha_0) + o_p(1) \\ &\xrightarrow{d}\ \mathcal{I}_\alpha^{-1}\,\Lambda_\alpha \sim \mathrm{N}\left(0, \mathcal{I}_\alpha^{-1} K_\alpha\, \mathcal{I}_\alpha^{-1}\right),\end{aligned}$$

where

$$U_{n,\alpha}(\alpha) = \begin{pmatrix} U_{n,\alpha_1}(\alpha_1) \\ \vdots \\ U_{n,\alpha_d}(\alpha_d) \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n U_{\alpha_1}(y_{i,1}, \alpha_1) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n U_{\alpha_d}(y_{i,d}, \alpha_d) \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \partial \log f_1(y_{i,1}, \alpha_1)/\partial\alpha_1 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \partial \log f_d(y_{i,d}, \alpha_d)/\partial\alpha_d \end{pmatrix},$$

$$K_\alpha = \begin{pmatrix} K_{\alpha_1} & K_{\alpha_1,\alpha_2} & \cdots & K_{\alpha_1,\alpha_d} \\ K_{\alpha_2,\alpha_1} & K_{\alpha_2} & \cdots & K_{\alpha_2,\alpha_d} \\ \vdots & \vdots & \ddots & \vdots \\ K_{\alpha_d,\alpha_1} & K_{\alpha_d,\alpha_2} & \cdots & K_{\alpha_d} \end{pmatrix},$$

$$K_{\alpha_j,\alpha_k} = \mathrm{Cov}_G\left(U_{\alpha_j}(y_j, \alpha_{0,j}), U_{\alpha_k}(y_k, \alpha_{0,k})\right) = \mathrm{E}_G\left[U_{\alpha_j}(y_j, \alpha_{0,j})\, U_{\alpha_k}(y_k, \alpha_{0,k})^{\mathrm{T}}\right],$$

$$\mathcal{I}_\alpha = \begin{pmatrix} \mathcal{I}_{\alpha_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{I}_{\alpha_d} \end{pmatrix},$$

$$\mathcal{I}_{\alpha_j} = -\int g_j\, \frac{\partial^2 \log f_j(y_j, \alpha_{0,j})}{\partial\alpha_{0,j}\partial\alpha_{0,j}^{\mathrm{T}}}\, \mathrm{d}y_j = -\mathrm{E}_{G_j}\left[H_{\alpha_j}(y_j, \alpha_{0,j})\right].$$

The proof is given in Appendix A.1. We have used $H_{\alpha_j}(y_j, \alpha_j)$ for the Hessian matrix operation on $\log f_j(y_j, \alpha_j)$.

Under the assumption that the margins are correctly specified, it holds that $K_{\alpha_j} = \mathcal{I}_{\alpha_j}$. When $j \neq k$, however, in general it holds that $K_{\alpha_j,\alpha_k} \neq 0$. The asymptotic variance matrix will hence keep the sandwich form even under the assumption that margins are correctly specified.

3.2 Large-sample results for stage 2 of two-stage ML estimation

Under the assumption that all margins and the copula are correctly specified, it is well known that $\widetilde{\theta}$ is a consistent estimator of the true parameter value θ_0 (i.e. $\mathrm{KL}(c_0, c(\cdot, \theta_0)) = 0$). However, without this assumption, the consistency of $\widetilde{\theta}$ needs more care, also since it needs to be clarified precisely what it is aiming for.

Lemma 2. *Consider the function*

$$M(\alpha_0, \theta) = \int g \log c\left(F_1(y_1, \alpha_{0,1}), \cdots, F_d(y_d, \alpha_{0,d}), \theta\right)\, \mathrm{d}y$$

and assume that there is θ_0, a unique and well-separated point of maximum of $M(\alpha_0, \theta)$, which satisfies

$\sup\{M(\alpha_0,\theta)\colon d(\theta,\theta_0) < \varepsilon\} < M(\alpha_0,\theta_0)$ *for every* $\varepsilon > 0$*; here* $d(\theta,\theta_0)$ *refers to Euclidean distance. Let* $M_n(\widetilde{\alpha},\theta)$ *be the stage 2 sample version of* $M(\alpha_0,\theta)$*,*

$$M_n(\widetilde{\alpha},\theta) = \frac{1}{n}\sum_{i=1}^{n} \log c\left(F_1(y_{i,1},\widetilde{\alpha}_1),\cdots,F_d(y_{i,d},\widetilde{\alpha}_d),\theta\right),$$

where the $\widetilde{\alpha}_j$ *are the estimates from the stage 1. Then* $\widetilde{\theta}$*, the maximiser of* $M_n(\widetilde{\alpha},\theta)$*, is a consistent estimator of* θ_0*, the least false parameter value.*

The proof is given in Appendix A.2.

So, under margin and copula misspecification, $\widetilde{\theta}$ is still consistent, but for the appropriate least false parameter value, rather than for any 'true' parameter value. The divergence in question is of the KL type form

$$\int g \log \frac{c_0(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}))}{c(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}),\theta)}\,\mathrm{d}y.$$

This leads to a precise notion of the least false parameter vector $\eta_0 = (\alpha_0^{\mathrm{T}},\theta_0^{\mathrm{T}})^{\mathrm{T}}$ associated with two-stage ML estimation.

Based on the consistency lemma above, we now derive the model robust asymptotic distribution of the two-stage ML estimator.

Proposition 1. *With* $\widetilde{\eta}$ *the two-stage ML estimator of* η*, we have*

$$\begin{aligned}\sqrt{n}\,(\widetilde{\eta}-\eta_0) &= \sqrt{n}\,\mathcal{I}_\eta^{-1}\begin{pmatrix} U_{n,\alpha}(\alpha_0) \\ U_{n,\theta}(\alpha_0,\theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1}U_{n,\alpha}(\alpha_0)\end{pmatrix} + \begin{pmatrix} o_p(1) \\ o_p(1)\end{pmatrix} \\ &\xrightarrow{d}\ \mathcal{I}_\eta^{-1}L\,\Lambda_\eta \sim \mathrm{N}\left(0,\,V_\eta\right)\end{aligned}$$

where

$$
\begin{aligned}
&V_\eta = \mathcal{I}_\eta^{-1} L\, K_\eta L^{\mathrm{T}} \mathcal{I}_\eta^{-1},\\
&U_{n,\theta}(\alpha,\theta) = \frac{1}{n}\sum_{i=1}^{n} U_\theta(y_i,\alpha,\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log c\left(F_1(y_{i,1},\alpha_1),\cdots,F_d(y_{i,d},\alpha_d),\theta\right)}{\partial\theta},\\
&K_\eta = \begin{pmatrix} K_\alpha & K_{\alpha,\theta} \\ K_{\alpha,\theta}^{\mathrm{T}} & K_\theta \end{pmatrix},\\
&K_\theta = \mathrm{Var}_G\, U_\theta(y,\alpha_0,\theta_0) = \mathrm{E}\left[U_\theta(y,\alpha_0,\theta_0)\, U_\theta(y,\alpha_0,\theta_0)^{\mathrm{T}}\right],\\
&K_{\alpha,\theta} = \mathrm{Cov}_G\left(U_\alpha(y,\alpha_0), U_\theta(y,\alpha_0,\theta_0)\right) = \mathrm{E}\left[U_\alpha(y,\alpha_0)\, U_\theta(y,\alpha_0,\theta_0)^{\mathrm{T}}\right],\\
&\mathcal{I}_\eta = \begin{pmatrix} \mathcal{I}_\alpha & 0 \\ 0 & \mathcal{I}_\theta \end{pmatrix},\\
&\mathcal{I}_\theta = -\int g\, \frac{\partial^2 \log c\left(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}),\theta\right)}{\partial\theta_0 \partial\theta_0^{\mathrm{T}}}\,\mathrm{d}y = -\mathrm{E}_G\left[H_\theta(y,\alpha_0,\theta_0)\right],\\
&\mathcal{I}_{\alpha,\theta} = -\int g\, \frac{\partial^2 \log c\left(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}),\theta\right)}{\partial\alpha_0 \partial\theta_0^{\mathrm{T}}}\,\mathrm{d}y = \mathrm{E}_G\left[-H_{\alpha,\theta}(y,\alpha_0,\theta_0)\right],\\
&L = \begin{pmatrix} I & 0 \\ -\mathcal{I}_{\alpha,\theta}^{\mathrm{T}} \mathcal{I}_\alpha^{-1} & I \end{pmatrix}.
\end{aligned}
$$

The proof is given in Appendix A.3.

Compared to the results from Xu (1996) and Joe (2005) where margins and the copula are assumed to be correctly specified, the asymptotic variance expression in Proposition 1 has a more general form. Writing it in block matrix form gives

$$
V_\eta = \begin{pmatrix} V_\alpha & V_{\alpha,\theta} \\ V_{\alpha,\theta}^{\mathrm{T}} & V_\theta \end{pmatrix} = \begin{pmatrix} \mathcal{I}_\alpha^{-1} & 0 \\ 0 & \mathcal{I}_\theta^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\mathcal{I}_{\alpha,\theta}^{\mathrm{T}} \mathcal{I}_\alpha^{-1} & I \end{pmatrix} \begin{pmatrix} K_\alpha & K_{\alpha,\theta} \\ K_{\alpha,\theta}^{\mathrm{T}} & K_\theta \end{pmatrix} \begin{pmatrix} I & -\mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha,\theta} \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathcal{I}_\alpha^{-1} & 0 \\ 0 & \mathcal{I}_\theta^{-1} \end{pmatrix}, \quad (2)
$$

where

$$
\begin{aligned}
&V_\alpha = \mathcal{I}_\alpha^{-1} K_\alpha\, \mathcal{I}_\alpha^{-1},\\
&V_\theta = \mathcal{I}_\theta^{-1} K_\theta\, \mathcal{I}_\theta^{-1} + \mathcal{I}_\theta^{-1}\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\, \mathcal{I}_\alpha^{-1} K_\alpha\, \mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha,\theta}\, \mathcal{I}_\theta^{-1} - \mathcal{I}_\theta^{-1} K_{\alpha,\theta}^{\mathrm{T}}\, \mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha,\theta}\, \mathcal{I}_\theta^{-1} - \mathcal{I}_\theta^{-1}\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\, \mathcal{I}_\alpha^{-1} K_{\alpha,\theta}\, \mathcal{I}_\theta^{-1},\\
&V_{\alpha,\theta} = \mathcal{I}_\alpha^{-1} K_{\alpha,\theta}\, \mathcal{I}_\theta^{-1} - \mathcal{I}_\alpha^{-1} K_\alpha\, \mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha,\theta}\, \mathcal{I}_\theta^{-1}.
\end{aligned}
$$

Andersen (2004) has a similar result, but in the context of a composite likelihood construction. Assuming the true model (i.e. $f = g$) results in a set of simplifications of Proposition 1. As already mentioned in Section 3.1, the true model assumption in stage 1 gives equality $K_{\alpha_j} = \mathcal{I}_{\alpha_j}$ for all $j = 1,\ldots,d$, where d indicates the dimension of the data.

Joe (2005) implements this result by replacing each $\mathcal{I}_{\alpha_j}$ with K_{α_j} in his asymptotic variance formula. Although this method is theoretically correct, it is not the most economical way of stating results, since V_η contains more $\mathcal{I}_{\alpha_j}$'s than K_{α_j}'s (note that every K_{α_j} is 'sandwiched' by $\mathcal{I}_{\alpha_j}$ in (2)). In addition, this is in opposition to the usual practice of maximum likelihood theory, where the asymptotic variance is defined and used as the inverse of Fisher information, as opposed to the variance of the score function vector.

Thus, under the true model assumption, we choose to simplify V_η by replacing K_α with

$$K_\alpha^{\mathrm{TMA}} = \begin{pmatrix} \mathcal{I}_{\alpha_1} & K_{\alpha_1,\alpha_2} & \cdots & K_{\alpha_1,\alpha_d} \\ K_{\alpha_2,\alpha_1} & \mathcal{I}_{\alpha_2} & \cdots & K_{\alpha_2,\alpha_d} \\ \vdots & \vdots & \ddots & \vdots \\ K_{\alpha_d,\alpha_1} & K_{\alpha_d,\alpha_2} & \cdots & \mathcal{I}_{\alpha_d} \end{pmatrix}.$$

Although Joe (2005)'s and our method are theoretically the same, it will make a difference in practice since it is almost never the case that the two matrices K_{α_j} and $\mathcal{I}_{\alpha_j}$, though identical qua population quantities, have empirical estimates that are very close.

Lemma 3. *Under the assumption that the margins and copula are correctly specified, it holds that* $K_{\alpha,\theta} = 0$.

The proof is given in Appendix A.4.

In two-stage ML estimation, the log-likelihood functions are different in stages 1 and 2. This implies that the true model assumption does not give $\mathcal{I}_{\alpha,\theta} = K_{\alpha,\theta}$. Instead, the true model assumption for two-stage ML estimation yields a different relationship, as follows.

Lemma 4. *Let*

$$U_\alpha^*(y,\alpha) = \begin{pmatrix} U_{\alpha_1}^*(y,\alpha_1) \\ \vdots \\ U_{\alpha_d}^*(y,\alpha_d) \end{pmatrix} = \begin{pmatrix} \partial \log c\left(F_1(y_1,\alpha_1),\cdots,F_d(y_d,\alpha_d),\theta\right)/\partial\alpha_1 \\ \vdots \\ \partial \log c\left(F_1(y_1,\alpha_1),\cdots,F_d(y_d,\alpha_d),\theta\right)/\partial\alpha_d \end{pmatrix}$$

and

$$K_{\alpha,\theta}^* = \mathrm{Cov}_G\left(U_\alpha^*(y,\alpha_0), U_\theta(y,\alpha_0,\theta_0)\right) = \mathrm{E}_G\left[U_\alpha^*(y,\alpha_0)\, U_\theta(y,\alpha_0,\theta_0)^{\mathrm{T}}\right].$$

Under the assumption that the margins and copula are correctly specified, it holds that $\mathcal{I}_{\alpha,\theta} = K_{\alpha,\theta}^*$.

The proof is given in Appendix A.5.

Lemma 5. *Under the assumption that the margins and copula are correctly specified, it holds that* $\mathcal{I}_\theta = K_\theta$.

The proof is given in Appendix A.6.

3.3 Impact of model misspecification on the copula parameter

The most apparent consequence of margin misspecification is that the two-stage ML copula parameter estimate $\widetilde{\theta}$ is no longer consistent for the true parameter value. Another consequence is the change in asymptotic variance V_η. Applying the lemmas resulting from true model assumption to (2) gives

$$\begin{aligned} V_\eta^{\mathrm{TMA}} &= \begin{pmatrix} V_\alpha^{\mathrm{TMA}} & V_{\alpha,\theta}^{\mathrm{TMA}} \\ \left(V_{\alpha,\theta}^{\mathrm{TMA}}\right)^{\mathrm{T}} & V_\theta^{\mathrm{TMA}} \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{I}_\alpha^{-1} & 0 \\ 0 & \mathcal{I}_\theta^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\mathcal{I}_\alpha^{-1} & I \end{pmatrix} \begin{pmatrix} K_\alpha^{\mathrm{TMA}} & 0 \\ 0 & \mathcal{I}_\theta \end{pmatrix} \begin{pmatrix} I & -\mathcal{I}_\alpha^{-1}\mathcal{I}_{\alpha,\theta} \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathcal{I}_\alpha^{-1} & 0 \\ 0 & \mathcal{I}_\theta^{-1} \end{pmatrix}, \end{aligned} \tag{3}$$

where

$$V_\alpha^{\rm TMA} = \mathcal{I}_\alpha^{-1} K_\alpha^{\rm TMA} \mathcal{I}_\alpha^{-1}$$
$$V_\theta^{\rm TMA} = \mathcal{I}_\theta^{-1} + \mathcal{I}_\theta^{-1} \mathcal{I}_{\alpha,\theta}^{\rm T} \mathcal{I}_\alpha^{-1} K_\alpha^{\rm TMA} \mathcal{I}_\alpha^{-1} \mathcal{I}_{\alpha,\theta} \mathcal{I}_\theta^{-1}$$
$$V_{\alpha,\theta}^{\rm TMA} = -\mathcal{I}_\alpha^{-1} K_\alpha^{\rm TMA} \mathcal{I}_\alpha^{-1} \mathcal{I}_{\alpha,\theta} \mathcal{I}_\theta^{-1}.$$

In Section 5 we illustrate the difference between (2) and (3) in practice by using simulated data. In Section 6, we illustrate the same difference by using real-life data.

4 Inference

Having established limiting normality of the two-stage ML estimator $\widetilde{\eta}$ in Proposition 1, we can derive limiting normality also for smooth parameter functions $\widetilde{\mu} = \mu(\widetilde{\eta})$, via the delta method. To use such results in practice we need consistent estimators of all variances.

The general formula for the variance matrix of the limiting distribution of $\widetilde{\eta}$ has the form

$$V_\eta = \mathcal{I}_\eta^{-1} L K_\eta L^{\rm T} \mathcal{I}_\eta^{-1}.$$

These components are population quantities, defined as means and variance matrices of random variables. Consistent estimators for these components emerge generally speaking by using plug-in sample averages $\bar{h}_n = (1/n)\sum_{i=1}^n h(y_i, \widetilde{\eta})$ for estimating a required $h_0 = {\rm E}_G\, h(y, \eta_0)$. For regularity conditions that secure convergence in probability of such $\bar{h}_n$ to h_0, see Jullum & Hjort (2017). This general recipe leads to consistent estimators of $\mathcal{I}_\eta^{-1}$, L, K_η above, and hence of $\widetilde{V}_\eta = \widetilde{\mathcal{I}_\eta}^{-1} \widetilde{L} \widetilde{K}_\eta \widetilde{L}^{\rm T} \widetilde{\mathcal{I}_\eta}^{-1}$.

Consider now such a focus parameter $\mu = \mu(\eta) = \mu(\alpha, \theta)$, a parameter of primary interest, a smooth function of the model parameters. In addition to the ML estimator $\widehat{\mu} = \mu(\widehat{\alpha}, \widehat{\theta})$, we may define the associated two-stage ML estimator $\widetilde{\mu} = \mu(\widetilde{\alpha}, \widetilde{\theta})$, for which the delta method yields

$$\sqrt{n}(\widetilde{\mu} - \mu_0) \ \xrightarrow{d}\ {\rm N}(0, \tau^2), \tag{4}$$

with $\tau^2 = c^{\rm T} V_\eta c$, and $c = \partial\mu(\eta_0)/\partial\eta$. Then $\widetilde{\tau}^2 = \widetilde{c}^{\rm T} \widetilde{V}_\eta \widetilde{c}$ is consistent for τ^2, with $\widetilde{c} = \partial\mu(\widetilde{\eta})/\widetilde{\eta}$. This leads to confidence intervals, for any parameter of interest, as with

$$\widetilde{\mu} \pm 1.96\, \widetilde{\tau}/\sqrt{n} \tag{5}$$

for the approximate 95% interval. There are different versions of (5), corresponding to different ways of applying the variance matrix formula V_η above. One may use components for $\widetilde{V}_\eta$ in a model-robust fashion, or the version where the parametric model assumption is trusted, leading to say $\widetilde{V}_\eta^{\rm TMA}$ and to consequent $\widetilde{\tau}^{\rm TMA}$. Also, results similar to and in fact more familiar than those of (4) and (5) are easy to write down and use for the full ML estimation method, leading to $\widehat{\mu} \pm 1.96\, \widehat{\tau}/\sqrt{n}$ etc.

We note that full confidence distributions can be computed and displayed, to supplement the two-stage ML based point estimate $\widetilde{\mu}$ and estimated standard deviation $\widetilde{\tau}/\sqrt{n}$. These are random curves ${\rm cc}(\mu)$, one such for each focus parameter, constructed post data, with the property that $P\{{\rm cc}(\mu) \le \alpha\}$ is equal to or

approximately equal to α, for all confidence levels α. In addition to the easy to use first-order large-sample confidence curve

$$\mathrm{cc}(\mu) = \Phi(\sqrt{n}(\mu - \widetilde{\mu})/\widetilde{\tau}),$$

somewhat more elaborate and better approximations may be constructed via methods of Schweder & Hjort (2016, Chapters 3, 4), involving generalisations of the Wilks type theorems. In Section 6 such confidence curves are computed and displayed for relevant parameters pertaining to Norwegian precipitation data.

Our methodology lends itself nicely also to hypothesis testing. If $\mu = \mu(\alpha, \theta)$ is a parameter where a certain null value μ_0 is of interest, then we may test $H_0 \colon \mu = \mu_0$ via inspection of the associated confidence interval, or via some fully or nearly equivalent route. This in particular applies for testing independence in the model structure, for the full d-dimensional vector or for a subset, if this corresponds to a null value for the θ parameter.

We may also use the developed machinery to test whether aspects of two or more sets of data are identical or different. If one has data from two groups thought to be not very different, say A and B, one may fit the same copula model to both, yielding parameter estimates with precision for $\eta_A = (\alpha_A, \theta_A)$ and $\eta_B = (\alpha_B, \theta_B)$. With variance estimation etc. as developed above, one may then test the hypothesis $\theta_A = \theta_B$. Specifically, with sample sizes n_A and n_B, we would have

$$\widetilde{\theta}_A - \widetilde{\theta}_B \approx_d \mathrm{N}(\theta_A - \theta_B, W),$$

say, with $W = \Sigma_A/n_A + \Sigma_B/n_B$. Here Σ_A and Σ_B are the variance matrices appearing in the limit distributions for $n_A^{1/2}(\widetilde{\theta}_A - \theta_A)$ and $n_B^{1/2}(\widetilde{\theta}_B - \theta_B)$. The test statistic

$$Z = (\widetilde{\theta}_A - \widetilde{\theta}_B)^{\mathrm{T}} \widetilde{W}^{-1} (\widetilde{\theta}_A - \widetilde{\theta}_B),$$

with $\widetilde{W}$ involving consistent estimators for Σ_A and Σ_B, would then follow an approximate χ^2_q null distribution, with q the dimension of θ.

5 Simulation study

To study the impact of the true model assumption for two-stage ML estimation, we have performed a set of simulations. Each simulation is based on 100 randomly generated datasets of $n = 1000$ observations. Table 1 and Table 2 contain descriptions of the models that were used to generate the data and the models that were used to fit the generated data. For both simulations, model 1 represents the situation where the copula is correctly specified, but margins are misspecified. Model 2 represents the situation where only the copula is misspecified. Model 3 has both misspecified copula and misspecified margins. Model 4 has correctly specified copula and margins. The misspecified margins and copula are chosen in a way such that they are close to the true margins and copula. This is to mimic a realistic model misspecification situation. However, in simulation 2, we chose Frank copula to illustrate the situation where the degree of misspecification is high.

Tables 3 and 4 give the results of these simulations. For each model, ML and two-stage ML estimation were performed both assuming and not assuming that the model is the true model that generated data. The column 'True model assumption' indicates the presence of this assumption. Dropping the true model

assumption leads to the model robust asymptotic variance formulae. For the ML estimator, this is the so-called 'sandwich estimator'. For a copula model, this estimator can be obtained straightforwardly by applying classical theory covered in Hardin (2003) and Claeskens & Hjort (2008). When we make the true model assumption, the sandwich estimator simplifies to the inverse of the Fisher information. For two-stage ML estimators, dropping the true model assumption yields (2) as asymptotic variance, and assuming the true model, this simplifies into (3).

Below, $\widehat{\theta}$ and $\widetilde{\theta}$ indicate the ML and the two-stage ML estimate of the copula parameter, respectively. Also, $\sqrt{n}\,\mathrm{SE}(\widehat{\theta})$ is the square root of estimated asymptotic variance of $\widehat{\theta}$. It is estimated by choosing the relevant formula, depending on the presence of true model assumption, and, in general terms, replacing $\mathrm{E}_G\, h(y, \eta_0)$ by $(1/n)\sum_{i=1}^{n} h(y_i, \widehat{\eta})$ or $(1/n)\sum_{i=1}^{n} h(y_i, \widetilde{\eta})$; see the discussion in Section 4.

Next, let $\mathrm{b}_{0.8}$ indicate the vector containing 0.8-quantile values of each fitted marginal distribution. With $\mathrm{P}(\mathrm{b}_{0.8} < y)$ we mean the joint probability that each marginal variable has larger value than the corresponding 0.8-quantile value. The $\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$ and $\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$ indicate ML and two-stage ML estimates of this joint probability, respectively. Also, $\sqrt{n}\,\mathrm{SE}(\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y))$ is the square root of the estimated asymptotic variance of $\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$, and $\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$ is the two-stage ML estimator analogue. These asymptotic variances are obtained by writing $\mathrm{P}(\mathrm{b}_{0.8} < y)$ as a function of η and applying the delta method.

When estimating the above mentioned quantities, the main computational bottleneck is estimating the K matrices. For instance, the matrix

$$\mathcal{I}_\theta = -\int g\, \frac{\partial^2 \log c\left(F_1(y_1, \alpha_{0,1}), \cdots, F_d(y_d, \alpha_{0,d}), \theta\right)}{\partial\theta_0 \partial\theta_0^{\mathrm{T}}}\, \mathrm{d}y$$

is estimated by

$$\widetilde{\mathcal{I}}_\theta = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log c(F_1(y_{i,1}, \widetilde{\alpha}_1), \cdots, F_d(y_{i,d}, \widetilde{\alpha}_d), \widetilde{\theta})}{\partial\theta \partial\theta^{\mathrm{T}}}.$$

When calculating this matrix, a for-loop can be avoided by swapping the order of summation and differentiation. When computing

$$\widetilde{K}_\theta = \frac{1}{n}\sum_{i=1}^{n} U_\theta(y_i, \widetilde{\alpha}, \widetilde{\theta}) U_\theta(y_i, \widetilde{\alpha}, \widetilde{\theta})^{\mathrm{T}},$$

however, the same trick can not be used and we are forced to use a time-consuming for-loop.

The results from model 4 in both simulations show that when both the copula and margins are correctly specified, the presence of the true model assumption makes virtually no difference. This confirms the finding from Section 3 that when the model is correctly specified, the model robust variance formula (2) and the non-robust version (3) become identical. Further, for the ML estimation part, it affirms the classical finding that the 'sandwich estimator' becomes the inverse of the Fisher information when the model is correctly specified.

From the results for model 1 to model 3, we see that assuming the true model (i.e. using (3) instead of (2)) often results in the decrease of the asymptotic variance. This seems logical because assuming the true model entails ignoring the model uncertainty and thus lower variance. In other words, ignoring the model uncertainty leads to over-confident confidence intervals. The degree of over-confidence will be bigger when sample size is small, because the asymptotic variance is divided by the sample size when a confidence interval is computed. This also implies that when a hypothesis testing is carried out on the copula parameter or on

a function that takes the copula parameter as input, the choice between model robust and model non-robust variance formula can lead to different outcomes of the test.

Although the true model assumption often results in smaller variance of the copula parameter, there are cases that this does not hold. e.g. the model with precipitation data in Section 6. This is in line with the fact that we could not find any analytical evidence for the inequality $V_\theta^{\mathrm{TMA}} < V_\theta$. In addition, the true model assumption changes the interpretation of $\widetilde{\theta}$ from 'estimate of the least false parameter value' to 'estimate of the true value that generated the data'.

Joe (2005) carried out an extensive study on the asymptotic efficiency for two-stage ML estimators and concluded that two-stage estimation is highly efficient in most cases. By comparing the results from ML and two-stage ML estimation, we can largely confirm that this is also the case when the model robust variance formula is used (particularly when sample size n is large, as the two variance matrix estimators involved are estimating the same quantity).

Another notable result regarding efficiency is that the Hájek–Le Cam convolution theorem, and related theorems on the optimality of ML estimation, do not apply when the model is wrongly specified. The Hájek–Le Cam convolution theorem (Hájek, 1970), combined with the Cramér–Rao lower bound (Cramér, 2016) theory, states that no asymptotically unbiased and normal competing estimator can have a smaller limiting variance matrix than that of the full ML estimator. Models 2 and 3 in both the two- and five-dimensional cases have a relatively large degree of model misspecification (model 2 has misspecified margins and model 3 has misspecified margins and copula). For these models, it frequently happens that the asymptotic variance of the two-stage ML estimator is smaller than that of the full ML estimator. This occurs both when the model-robust and model non-robust variance formulae are used. The largest difference occurs for model 3 from simulation 2 where the degree of model misspecification is high (the copula and all five margins are misspecified).

When it comes to the asymptotic variance of the upper tail probability $\mathrm{P}(\mathrm{b}_{0.8} < y)$, we observe often that the true model assumption decreases the asymptotic variance. Yet, in some cases, we observe that the true model assumption increases the asymptotic variance. We suspect that this is due to the fact that the whole V_η including its non-diagonal elements are used through the delta method formula to compute $V_{\mathrm{P}(\mathrm{b}_{0.8}<y)}$. To check whether this also happens in non-copula models, we simulated a large number of univariate data and fitted well-known distributions and computed upper tail probability from them. We could observe that the true model assumption can increase the limiting variance of the upper tail probability while the limiting variance of parameters decreases.

Further, we notice that the difference between $\widehat{\theta}$ and $\widetilde{\theta}$ gets larger as dimension increases and as the degree of misspecification increases. This is in line with the earlier result from Kim *et al.* (2007) where they discovered that two-stage ML estimation is highly non-robust against misspecification of the margins.

Table 1: Description of the models used in simulation 1.

	Copula	Margin 1	Margin 2
Data generating model	Gaussian $\theta = 0.3$	Weibull $\alpha_1 = (1.5, 4)^{\mathrm{T}}$ (shape, scale)	gamma $\alpha_2 = (2, 1)^{\mathrm{T}}$ (shape, rate)
Model 1	Gaussian	log-normal	log-normal
Model 2	Frank	Weibull	gamma
Model 3	Frank	log-normal	log-normal
Model 4	Gaussian	Weibull	gamma

Table 2: Description of the models used in simulation 2.

	Copula	Margin 1	Margin 2	Margin 3	Margin 4	Margin 5
Data generating model	Gumbel $\theta = 3$	Weibull $\alpha_1 = (1.5, 4)^{\mathrm{T}}$ (shape, scale)	Weibull $\alpha_2 = (2, 3)^{\mathrm{T}}$ (shape, scale)	gamma $\alpha_3 = (2, 1)^{\mathrm{T}}$ (shape, rate)	gamma $\alpha_4 = (3, 1)^{\mathrm{T}}$ (shape, rate)	gamma $\alpha_5 = (4, 2)^{\mathrm{T}}$ (shape, rate)
Model 1	Gumbel	log-normal	log-normal	log-normal	log-normal	log-normal
Model 2	Frank	Weibull	Weibull	gamma	gamma	gamma
Model 3	Frank	log-normal	log-normal	log-normal	log-normal	log-normal
Model 4	Gumbel	Weibull	Weibull	gamma	gamma	gamma

Table 3: Result from simulation 1. $\sqrt{n}\,\mathrm{SE}(\widehat{\theta})$ is the square root of estimated asymptotic variance of $\widehat{\theta}$. $\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$ is the two-stage ML estimator analogue. $\mathrm{P}(\mathrm{b}_{0.8} < y)$ indicates the joint probability that each marginal variable has larger value than its 0.8-quantile value, defined by each marginal model. Next, $\sqrt{n}\,\mathrm{SE}(\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y))$ is the square root of the estimated asymptotic variance of $\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$. $\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$ is the two-stage ML estimation analogue. The 'true model assumption' indicates whether it is assumed that the model is the true model that generated data. If 'True model assumption = No', the model robust variance formulae were used. When 'True model assumption = Yes', the model non-robust variance formulae were used.

Simulation 1

		MLE					
	True model assumption	$\widehat{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\theta})$	$\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
Model 1	No	0.2884	(0.2296, 0.3472)	0.9497	0.0651	(0.0558, 0.0744)	0.1503
	Yes		(0.2317, 0.3452)	0.9159		(0.0544, 0.0757)	0.1718
Model 2	No	1.7975	(1.4055, 2.1894)	6.3265	0.0644	(0.0532, 0.0755)	0.1803
	Yes		(1.4108, 2.1841)	6.2391		(0.0537, 0.0750)	0.1718
Model 3	No	1.9375	(1.5077, 2.3672)	6.9414	0.0663	(0.0564, 0.0762)	0.1594
	Yes		(1.5266, 2.3484)	6.6307		(0.0552, 0.0774)	0.1797
Model 4	No	0.2992	(0.2429, 0.3554)	0.9079	0.0661	(0.0554, 0.0768)	0.1729
	Yes		(0.2429, 0.3555)	0.9083		(0.0554, 0.0768)	0.1723
		Two-stage MLE					
	True model assumption	$\widetilde{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\theta})$	$\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
Model 1	No	0.2884	(0.2296, 0.3472)	0.9497	0.0651	(0.0558, 0.0744)	0.1503
	Yes		(0.2303, 0.3465)	0.9376		(0.0553, 0.0748)	0.1568
Model 2	No	1.7922	(1.4023, 2.1822)	6.2938	0.0643	(0.0533, 0.0753))	0.1783
	Yes		(1.4061, 2.1783)	6.2301		(0.0535, 0.0751)	0.1744
Model 3	No	1.9240	(1.4993, 2.3486)	6.8569	0.0661	(0.0560, 0.0762)	0.1634
	Yes		(1.5092, 2.3388)	6.6945		(0.0560, 0.0762)	0.1634
Model 4	No	0.2992	(0.2429, 0.3554)	0.9081	0.0661	(0.0554, 0.0768)	0.1729
	Yes		(0.2429, 0.3555)	0.9086		(0.0554, 0.0768)	0.1726

Table 4: Result of simulation 2. For the description of column labels, see the caption of Table 3.

Simulation 2

		MLE					
	True model assumption	$\widehat{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\theta})$	$\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
Model 1	No	2.7555	(2.6324, 2.8786)	1.9929	0.1205	(0.1084, 0.1326)	0.1952
	Yes		(2.6447, 2.8664)	1.7892		(0.1069, 0.1341)	0.2193
Model 2	No	11.4996	(10.5764, 12.4228)	15.0082	0.1010	(0.0667, 0.1353)	0.5580
	Yes		(10.9296, 12.0696)	9.2117		(0.0800, 0.1221)	0.3400
Model 3	No	14.7851	(13.6473, 15.9229)	18.3947	0.1187	(0.0877, 0.1498)	0.5031
	Yes		(14.0435, 15.5267)	11.9683		(0.0966, 0.1409)	0.3569
Model 4	No	2.9905	(2.8601, 3.1210)	2.1059	0.1265	(0.1121, 0.141)	0.2330
	Yes		(2.8598, 3.1213)	2.1103		(0.1121, 0.141)	0.2330
		Two-stage MLE					
	True model assumption	$\widetilde{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\theta})$	$\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
Model 1	No	2.9420	(2.8054, 3.0787)	2.2073	0.1254	(0.1126, 0.1381)	0.2057
	Yes		(2.8113, 3.0727)	2.1104		(0.1130, 0.1377)	0.1995
Model 2	No	9.8192	(9.2465, 10.3918)	9.2457	0.0895	(0.0705, 0.1084)	0.3059
	Yes		(9.3525, 10.2859)	7.5353		(0.0714, 0.1076)	0.2922
Model 3	No	9.5410	(8.9495, 10.1325)	9.5514	0.0873	(0.0716, 0.1031)	0.2542
	Yes		(9.0907, 9.9913)	7.2710		(0.0728, 0.1019)	0.2354
Model 4	No	2.9952	(2.8444, 3.1461)	2.4361	0.1266	(0.1109, 0.1424)	0.2546
	Yes		(2.8445, 3.1460)	2.4348		(0.1109, 0.1424)	0.2544

6 Precipitation data

The precipitation data consist of daily measurements of precipitation in mm at five different meteorological stations in Norway (Vestby, Ski, Lørenskog, Nannestad and Hurdal) from Jan 1 1990 to Dec 31 2006. These data were provided by the Norwegian Meteorological Institute and used previously in Aas & Berg (2009) (with one station less) and Hobæk Haff (2013). Following the example of two previous papers that used these data, we modeled only positive precipitation by removing all observations for which at least one of the stations has recorded zero precipitation, resulting in 5536 observations. The main advantage of this is that we remove time dependence.

To choose an adequate model, we first fitted a set of well-known distributions for each margins by using ML estimation and evaluated them by AIC. After the marginal distributions were chosen, we performed probability integral transformation for each margin and fitted a set of well-known copulae by using ML estimation and evaluated these by AIC. The best model obtained in this fashion is described in Table 5. The choice of the Gumbel copula corresponds well with the fact that there are indications of strong upper, but not lower, tail dependence. This is visible in Figure 3, which contains simulated scatter plots between the first 2 variables of 5-dimensional Gumbel copula. Since the Gumbel copula is a member of the Archimedean

copula family, the scatter plots between other possible combinations among five variables will be virtually the same as the one between variable 1 and variable 2. They are therefore not displayed.

Figure 1 shows that the marginal distributions of the fitted model are highly non-normal and can differ between meteorological stations.

Table 6 shows the result from the model described in Table 5. With ML estimation, ignoring model uncertainty gives smaller values of the asymptotic variances. However, with two-stage ML estimators, this is not the case. This phenomenon was discussed earlier in Section 5.

Furthermore, we observe that the asymptotic variance of the two-stage ML estimator is smaller than that of the ML estimator. As already discussed in Section 5, this happens when the degree of misspecification is high and consequently the Hájek–Le Cam convolution theorem is not applicable. In our case, we used the five-dimensional Gumbel copula, which is a member of the Archimedean copula family, to model the precipitation in five locations. The advantage of the Archimedean family copulae is that we can model high-dimensional dependency by using only one parameter. The disadvantage is that the dependency structure and strength is the same between all different pairs of variables. The dependency relationship between u_1 and u_2, for example, is the same as the dependency relationship between u_3 and u_5, in such models. Here $u_j = F_j(y_j)$, for $j = 1, \ldots, 5$, and these are uniform. The pairwise pseudo-observations plot (Figure 3) in Supplement B of Hobæk Haff (2013), however, shows that the pairwise dependencies among the stations are not the same.

This rigidness of Archimedean copulae is causing a misspecification, and therefore we see that the asymptotic variance of two-stage ML estimators is smaller than that of the full ML. Aas & Berg (2009) suggest pair-copula constructions as a method to overcome this limitation.

Figure 2 visualises Table 6 by using confidence curves, see Section 4. For details about confidence curves, see Schweder & Hjort (2016). In Figure 2, the confidence intervals from the model non-robust formula (dashed curves) and the confidence intervals from the model robust formula (non-dashed lines) are quite close to each other. This is partly due to the fact that the sample size ($n = 5536$) is relatively big and thus scales down the asymptotic variance when a confidence interval is computed. For datasets with smaller number of observations, the choice of model (non-)robust formula will make a bigger difference.

The joint upper tail probability $\mathrm{P}(\mathrm{b}_{0.8} < y)$ can be interpreted as the probability that, given that there is precipitation in all 5 locations, there is precipitation higher than the 0.8-quantile values in all 5 locations within the same day.

Table 5: Description of the model for precipitation data.

	Copula	Margin 1 (Vestby)	Margin 2 (Ski)	Margin 3 (Lørenskog)	Margin 4 (Nannestad)	Margin 5 (Hurdal)
Model	Gumbel	gamma	gamma	log-normal	gamma	log-normal

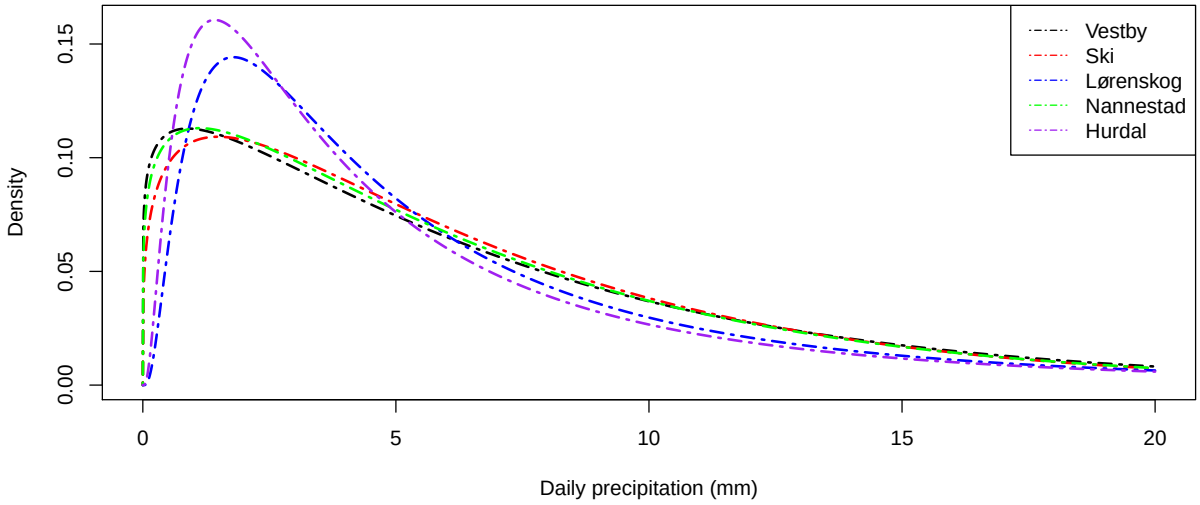

Figure 1: Density plot of fitted marginal distributions of precipitation data.

Table 6: Result from the model described in Table 5 fitted with the precipitation data. For the description of column labels, see the caption of Table 4.

	MLE					
True model assumption	$\widehat{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\theta})$	$\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widehat{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
No	2.3511	(2.2985, 2.4037)	1.9958	0.1075	(0.1010, 0.1140)	0.2460
Yes		(2.3038, 2.3984)	1.7960		(0.1016, 0.1135)	0.2256
	Two-stage MLE					
True model assumption	$\widetilde{\theta}$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\theta})$	$\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$	(95% CI)	$\sqrt{n}\,\mathrm{SE}(\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y))$
No	2.1959	(2.1544, 2.2375)	1.5779	0.1013	(0.0959, 0.1068)	0.2071
Yes		(2.1513, 2.2406)	1.6958		(0.0954, 0.1073)	0.2252

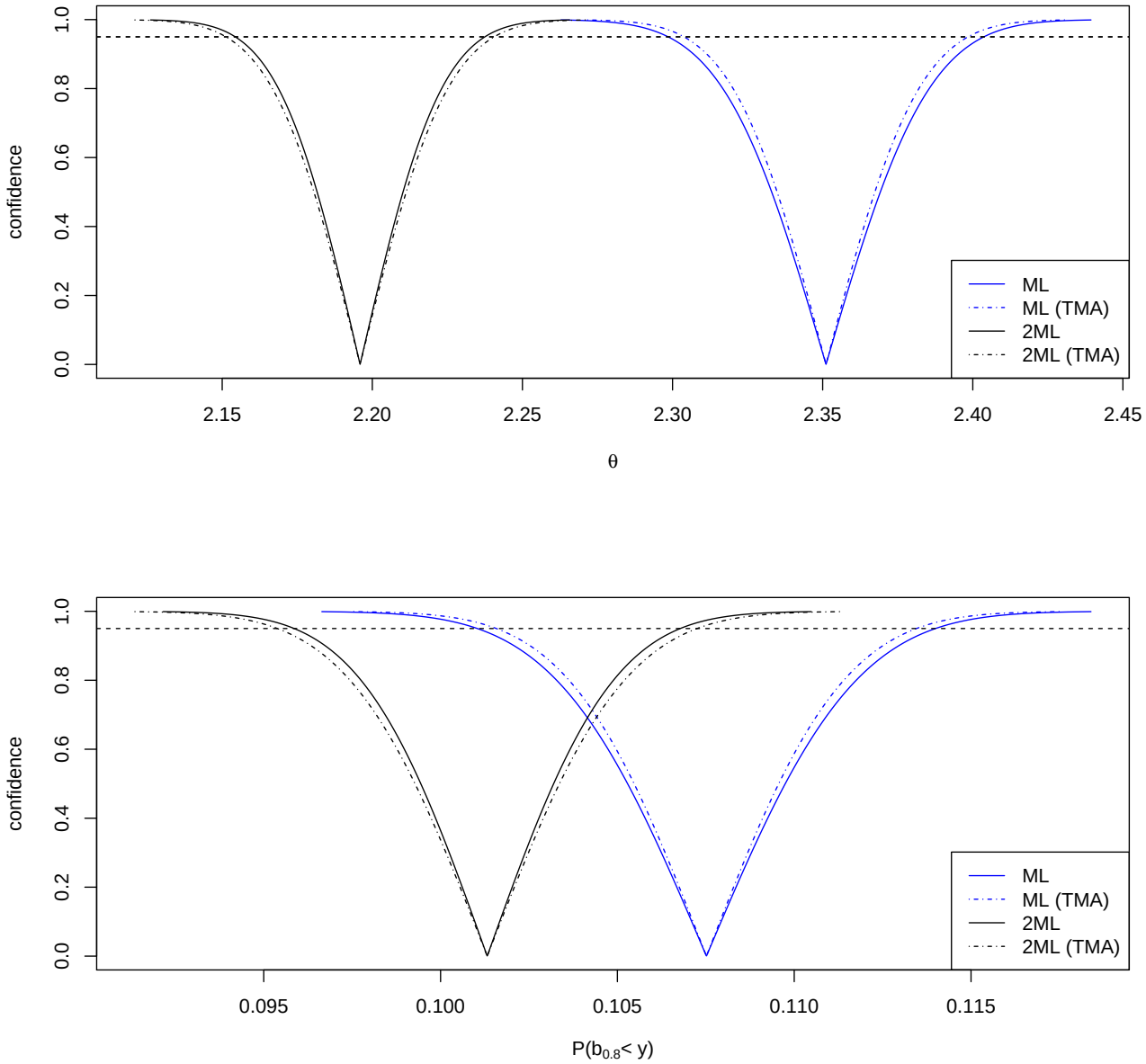

Figure 2: Confidence curves of θ and $\mathrm{P}(\mathrm{b}_{0.8} < y)$. Blue curves are from the model fitted with ML and black curves are from the model fitted with two-stage ML. Dashed curves correspond to the true model assumption. The dashed horizontal lines indicate 95% confidence.

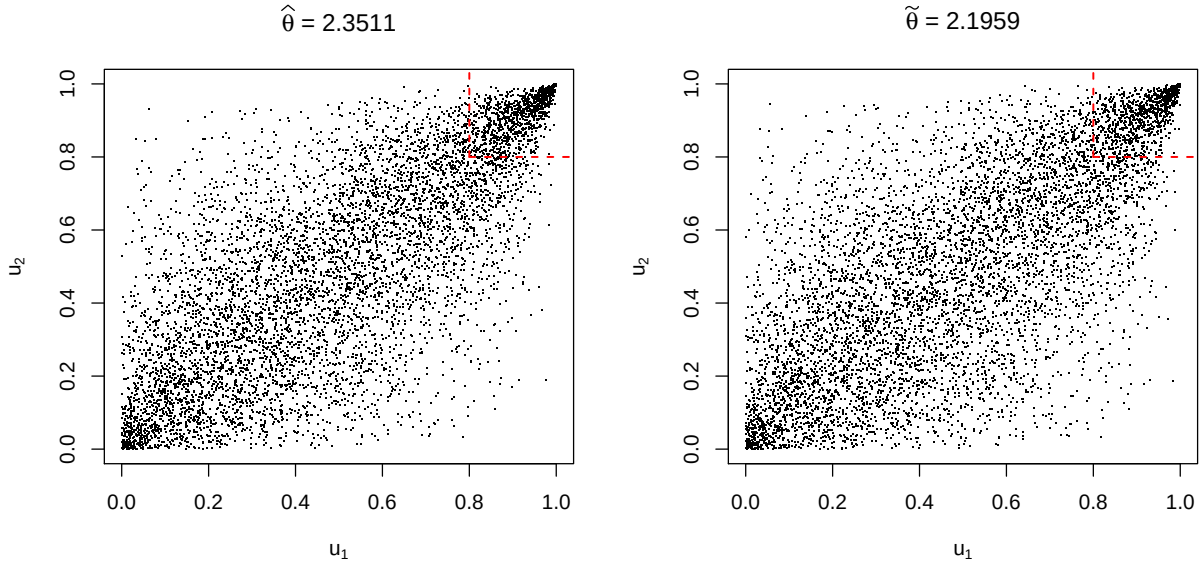

Figure 3: Scatter plots between the first 2 variables of 5-dimensional Gumbel copula. The plots are generated by simulating from the fitted models with precipitation data. The left side is fitted with ML and the right side fitted with two-stage ML. The areas in the upper corners indicated by the red dashed lines are the regions where both variables have higher values than their 0.8-quantiles.

7 Conclusions and further research

In our paper we have reached precise results for the behaviour of two-stage ML estimators for parametric copula models, both under and outside model conditions. This has then led to a full and flexible machinery for inference, including confidence intervals and curves for focus parameters, testing hypotheses, etc. Here we offer some concluding remarks, some pointing to further research.

Assuming that the parametric model holds leads to a set of simplifications in our apparatus. For instance, the model robust asymptotic variance (2) simplifies to the model non-robust asymptotic variance formula (3), which is essentially the same as the asymptotic variance formula from Joe (2005). The difference is that we choose to use the inverse of the Fisher information matrix instead of the covariance of score vectors, with consequences for how this matrix is estimated from data. We believe that this is more concordant with the current practice of ML theory. Although these two definitions of asymptotic variances are theoretically the same under the true model assumption, the practical difference between them can be non-negligible when a real-life dataset is used. A further study on the impact of making one vs. the other choice for estimating the population Fisher information matrix, either in the usual fashion based on the Hessian matrix at the ML position, or using the variance matrix for the score vectors, would be useful not only for the copula community, but more generally for statisticians applying ML theory in other contexts.

Our simulation study shows that assuming true model (i.e. using the model non-robust variance formula) in general leads to a smaller asymptotic variance for the copula parameter estimate. So, ignoring the model

uncertainty often leads to an overconfident confidence interval. The degree of overconfidence will get smaller as sample size increases because the asymptotic standard deviation is divided by $\sqrt{n}$ when computing a confidence interval. When carrying out hypothesis testing for the copula parameter, or on a function that takes the copula parameter as an input, the choice of model-robust variance formula or not can therefore lead to different decisions.

Joe (2005) compares asymptotic relative efficiency (ARE) of the two-stage ML estimator with that of full ML and concludes that two-stage ML method typically has good ARE. We could observe that the two-stage ML estimator is still highly efficient when the true model assumption is dropped and the model robust asymptotic variance formulae are used.

When models are highly misspecified, however, we see that the asymptotic variance of two-stage ML estimators can be smaller than that of the ML estimator. This happens both when the model-robust and model non-robust variance formulae are used. This effect increases as the degree of model misspecification and dimension increases. This relates to the fact that the Hájek–Le Cam convolution theorem, along with theorems of a similar nature for the optimality of ML estimation for parametric models, do not apply when the models are misspecified.

When a fitted copula model is used to compute joint upper tail probability, we observe sometimes that the true model assumption does not decrease the asymptotic variance of this joint probability. A possible reason is that this asymptotic variance is computed with the delta method by using the whole variance matrix of η, including the covariances between copula and margin parameters. A further theoretical and numerical study on the property of variance transformation through the delta method would be fruitful, especially considering the fact that this also happens in various other types of models.

The authors of this study are currently developing model selection criteria for two-stage ML estimators that utilise the model robust large-sample distribution from Section 3 of this article. These model selection criteria will complement earlier model selection methodology efforts, such as the Copula Information Criterion (CIC) developed by Grønneberg & Hjort (2014). One particular goal of these extended efforts will be to aid model building and selection for pair-copula constructions (Aas & Berg, 2009; Aas *et al.*, 2009).

The methodology for two-stage ML estimation for copula models, inside and outside model conditions, has been developed in this paper primarily aiming for the case of i.i.d. sequences of vector observations. Importantly, in the presence of relevant covariate information, the large-sample results and ensuing inference methods can be extended to various classes of conditional copula regression models, with the required extra efforts. This will then lead to further estimation and inference tools for models worked in e.g. Acar *et al.* (2013); Veraverbeke *et al.* (2011). The covariates may influence the margins, the copula mechanism, or both. As one such example, we might fit the Norwegian precipitation data of Section 6 using the Gumbel copula, but now with

$$\theta_i = \theta \exp\{\gamma_1(x_i - 1990) + \gamma_2(x_i - 1990)^2\} \quad \text{for } i = 1, \ldots, n,$$

where x_i is the calendar year for observation i. This leads to estimates and confidence curves for γ_1, γ_2 and related parameters, via the two-stage construction

$$\ell_c(\widetilde{\alpha}, \theta, \gamma) = \sum_{i=1}^{n} \log c(F_1(y_{i,1}, \widetilde{\alpha}_1), \cdots, F_d(y_{i,d}, \widetilde{\alpha}_d), \theta_i),$$

and allows one the opportunity to assess any changes of the copula mechanism over time.

Acknowledgements

The authors would like to thank Ingrid Hobæk Haff for her valuable comments and fruitful discussions. They also acknowledge partial funding from the Norwegian Research Council supported research group FocuStat: Focus Driven Statistical Inference With Complex Data, and from the Department of Mathematics at the University of Oslo.

References

Aas, K. & Berg, D. (2009). Models for construction of multivariate dependence – a comparison study. *The European Journal of Finance* **15**, 639–659.

Aas, K., Czado, C., Frigessi, A. & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics* **44**, 182–195.

Acar, E. F., Craiu, R. V., Yao, F. *et al.* (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics* **7**, 2822–2850.

Andersen, E. W. (2004). Composite likelihood and two-stage estimation in family studies. *Biostatistics* **5**, 15–30.

Andersen, E. W. (2005). Two-stage estimation in copula models used in family studies. *Lifetime Data Analysis* **11**, 333–350.

Casella, G. & Berger, R. L. (2002). *Statistical Inference*. Duxbury Pacific Grove, CA.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Coles, S., Heffernan, J. & Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes* **2**, 339–365.

Cramér, H. (2016). *Mathematical Methods of Statistics [re-issue of the 1946 classic]*. Princeton University Press.

Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance* **76**, 639–650.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall London.

Genest, C. & Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering* **12**, 347–368.

Genest, C. & Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association* **88**, 1034–1043.

Grønneberg, S. & Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics* **41**, 436–459.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Probability Theory and Related Fields* **14**, 323–330.

Hardin, J. W. (2003). The sandwich estimate of variance. In *Maximum likelihood estimation of misspecified models: Twenty years later*. Emerald Group Publishing Limited, pp. 45–73.

Hobæk Haff, I. (2013). Parameter estimation for pair-copula constructions. *Bernoulli* **19**, 462–491.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* **94**, 401–419.

Jullum, M. & Hjort, N. (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* **27**, 951–981.

Kim, G., Silvapulle, M. J. & Silvapulle, P. (2007). Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis* **51**, 2836–2850.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Le Cam, L. (1990). Maximum likelihood: an introduction. *International Statistical Review* **58**, 153–171.

Lehmann, E. L. (2004). *Elements of Large-Sample Theory*. Springer Science & Business Media.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Science & Business Media.

Nikoloulopoulos, A. K., Joe, H. & Li, H. (2012). Vine copulas with asymmetric tail dependence and applications to financial return data. *Computational Statistics & Data Analysis* **56**, 3659–3673.

Schweder, T. & Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Inference With Confidence Distributions*. Cambridge University Press.

Shih, J. H. & Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–1399.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8**, 229–231.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Veraverbeke, N., Omelka, M. & Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics* **38**, 766–780.

Xu, J. J. (1996). *Statistical modelling and inference for multivariate and longitudinal discrete response data*. Ph.D. thesis, University of British Columbia.

A Appendix

A.1 Proof of Lemma 1

Proof. From Van der Vaart (2000, Theorems 5.41 and 5.42), we have

$$\widetilde{\alpha}_j - \alpha_{0,j} = \mathcal{I}_{\alpha_j}^{-1} \frac{1}{n} \sum_{i=1}^{n} U_{\alpha_j}(y_{i,j}, \alpha_{0,j}) + o_p(n^{-1/2})$$

for all $j = 1, \ldots, d$. Applying the multivariate central limit theorem along with the Slutsky theorem gives

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} \widetilde{\alpha}_1 - \alpha_{0,1} \\ \vdots \\ \widetilde{\alpha}_d - \alpha_{0,d} \end{pmatrix} &= \sqrt{n}\begin{pmatrix} \mathcal{I}_{\alpha_1}^{-1} \frac{1}{n} \sum_{i=1}^{n} U_{\alpha_1}(y_{i,1}, \alpha_{0,1}) + o_p(n^{-1/2}) \\ \vdots \\ \mathcal{I}_{\alpha_d}^{-1} \frac{1}{n} \sum_{i=1}^{n} U_{\alpha_d}(y_{i,d}, \alpha_{0,d}) + o_p(n^{-1/2}) \end{pmatrix} \\
&= \sqrt{n}\begin{pmatrix} \mathcal{I}_{\alpha_1}^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{I}_{\alpha_d}^{-1} \end{pmatrix} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} U_{\alpha_1}(y_{i,1}, \alpha_{0,1}) \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} U_{\alpha_d}(y_{i,d}, \alpha_{0,d}) \end{pmatrix} + \begin{pmatrix} o_p(1) \\ \vdots \\ o_p(1) \end{pmatrix} \\
&\xrightarrow{d} \begin{pmatrix} \mathcal{I}_{\alpha_1}^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{I}_{\alpha_d}^{-1} \end{pmatrix} \begin{pmatrix} \Lambda_{\alpha_1} \\ \vdots \\ \Lambda_{\alpha_d} \end{pmatrix} \\
&\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \mathcal{I}_{\alpha_1}^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{I}_{\alpha_d}^{-1} \end{pmatrix} \begin{pmatrix} K_{\alpha_1} & K_{\alpha_1,\alpha_2} & \cdots & K_{\alpha_1,\alpha_d} \\ K_{\alpha_2,\alpha_1} & K_{\alpha_2} & \cdots & K_{\alpha_2,\alpha_d} \\ \vdots & \vdots & \ddots & \vdots \\ K_{\alpha_d,\alpha_1} & K_{\alpha_d,\alpha_2} & \cdots & K_{\alpha_d} \end{pmatrix} \begin{pmatrix} \mathcal{I}_{\alpha_1}^{-1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathcal{I}_{\alpha_d}^{-1} \end{pmatrix} \right).
\end{aligned}$$

Or, by using more compact notation,

$$\begin{aligned}
\sqrt{n}\,(\widetilde{\alpha} - \alpha_0) &= \sqrt{n}\,\mathcal{I}_{\alpha}^{-1}\, U_{n,\alpha}(\alpha_0) + o_p(1) \\
&\xrightarrow{d} \mathcal{I}_{\alpha}^{-1}\,\Lambda_{\alpha} \sim \mathrm{N}\left(0, \mathcal{I}_{\alpha}^{-1} K_{\alpha}\, \mathcal{I}_{\alpha}^{-1}\right)
\end{aligned}$$

. □

A.2 Proof of Lemma 2

Proof. For the estimator $\widetilde{\alpha}$ from stage 1, we have $\tilde{\alpha} \xrightarrow{p} \alpha_0$. The Taylor series expansion of $M_n(\widetilde{\alpha}, \theta)$ around α_0, with θ fixed as an arbitrary constant such that $\theta \in \Theta$, gives

$$M_n(\widetilde{\alpha}, \theta) = M_n(\alpha_0, \theta) + \frac{\partial M_n(\alpha_0, \theta)}{\partial \alpha} (\widetilde{\alpha} - \alpha_0) + o_p(n^{-1/2}).$$

Since $\widetilde{\alpha} \xrightarrow{p} \alpha_0$, we have, for every $\varepsilon > 0$ and all $\theta \in \Theta$, that

$$|M_n(\widetilde{\alpha}, \theta) - M_n(\alpha_0, \theta)| < \varepsilon. \tag{6}$$

Le Cam's uniform convergence theorem (Ferguson (1996, Theorem 16(a))) gives directly:

$$\Pr\left\{\lim_{n\to\infty} \sup_{\theta\in\Theta} |M_n(\alpha_0, \theta) - M(\alpha_0, \theta)| = 0\right\} = 1 \tag{7}$$

By combining (6) and (7), we have

$$\sup_{\theta\in\Theta} |M_n(\widetilde{\alpha}, \theta) - M(\alpha_0, \theta)| \xrightarrow{p} 0. \tag{8}$$

In other words, $M_n(\widetilde{\alpha}, \theta)$ converges uniformly to $M(\alpha_0, \theta)$.

Now, we mimic the proof of Theorem 5.7 in Van der Vaart (2000). Consider the estimator $\widetilde{\theta}_n$ that nearly maximises $M_n(\widetilde{\alpha}, \theta)$ and satisfies

$$M_n(\widetilde{\alpha}, \widetilde{\theta}_n) \geq \sup_{\theta} M_n(\widetilde{\alpha}, \theta) - o_p(1).$$

Then we have certainly $M_n(\widetilde{\alpha}, \widetilde{\theta}_n) \geq M_n(\widetilde{\alpha}, \theta_0) - o_p(1)$. By adding $-M(\alpha_0, \widetilde{\theta}_n) + o_p(1)$ on both sides, we get

$$\begin{aligned} M_n(\widetilde{\alpha}, \theta_0) - M(\alpha_0, \widetilde{\theta}_n) &\leq M_n(\widetilde{\alpha}, \widetilde{\theta}_n) - M(\alpha_0, \widetilde{\theta}_n) + o_p(1) \\ &\leq \sup_{\theta} |M_n(\widetilde{\alpha}, \theta) - M(\alpha_0, \theta)| + o_p(1) \xrightarrow{p} 0 \end{aligned}$$

by (8).
Because of the assumption that is made in the beginning of the lemma, for every $\varepsilon > 0$, there exists $\eta > 0$ such that $M(\alpha_0, \theta) < M(\alpha_0, \theta_0) - \eta$ for every θ satisfying $d(\theta, \theta_0) \geq \varepsilon$. This implies that the event $\{d(\theta, \theta_0) \geq \varepsilon\}$ is contained in the event $\{M(\alpha_0, \theta) < M(\alpha_0, \theta_0) - \eta\}$. The probabilty of the latter event converges to 0. Thus, $d(\widetilde{\theta}_n, \theta_0) \xrightarrow{p} 0$. □

A.3 Proof of Proposition 1

Proof.

Lemma 6. *Let $\widetilde{\theta}$ be the two-stage ML estimator of θ and θ_0 be the least false value of this parameter. Then we have*

$$\widetilde{\theta} - \theta_0 = \mathcal{I}_\theta^{-1} \frac{1}{n} \sum_{i=1}^{n} U_\theta(y_i, \widetilde{\alpha}, \theta_0) + o_p(n^{-1/2}).$$

Proof. The Taylor series expansion of $U_{n,\theta}(\widetilde{\alpha}, \widetilde{\theta})$ around θ_0 yields

$$0 = U_{n,\theta}(\widetilde{\alpha}, \widetilde{\theta}) = U_{n,\theta}(\widetilde{\alpha}, \theta_0) + H_{n,\theta}(\widetilde{\alpha}, \theta_0)(\tilde{\theta} - \theta_0) + o_p(n^{-1/2}),$$

which gives

$$\widetilde{\theta} - \theta_0 = -H_{n,\theta}(\widetilde{\alpha}, \theta_0)^{-1} U_{n,\theta}(\widetilde{\alpha}, \theta_0) + o_p(n^{-1/2}). \tag{9}$$

From stage 1, $\widetilde{\alpha}$ is a consistent estimator of α_0, i.e. $\tilde{\alpha} \xrightarrow{p} \alpha_0$. Since $H_\theta(y, \alpha, \theta_0)$ is differentiable at α_0 and there exists a Jacobian $\dot{H}_\theta(y, \alpha_0, \theta_0)$, we can apply the multivariate delta method (Lehmann (2004, Theorem 3.7)) and obtain $H_{n,\theta}(\widetilde{\alpha}, \theta_0) \xrightarrow{p} H_{n,\theta}(\alpha_0, \theta_0)$. Further, by the law of large numbers, we have

$$-H_{n,\theta}(\alpha_0, \theta_0) \xrightarrow{p} -\mathrm{E}_G\left[H_\theta(y, \alpha_0, \theta_0)\right] = \mathcal{I}_\theta.$$

Thus, we have $-H_{n,\theta}(\widetilde{\alpha}, \theta_0) \xrightarrow{p} \mathcal{I}_\theta$ and (9) becomes

$$\tilde{\theta} - \theta_0 = \mathcal{I}_\theta^{-1} U_{n,\theta}(\widetilde{\alpha}, \theta_0) + o_p(n^{-1/2}).$$

□

Lemma 7. *We have*

$$\begin{aligned}\widetilde{\theta} - \theta_0 &= \mathcal{I}_\theta^{-1}\left\{U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,(\widetilde{\alpha} - \alpha_0)\right\} + o_p(n^{-1/2}) \\ &= \mathcal{I}_\theta^{-1}\left\{U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1} U_{n,\alpha}(\alpha_0)\right\} + o_p(n^{-1/2}).\end{aligned}$$

Proof. The Taylor series expansion of $U_{n,\theta}(\widetilde{\alpha}, \theta_0)$ around α_0 yields

$$\begin{aligned}U_{n,\theta}(\widetilde{\alpha}, \theta_0) &= U_{n,\theta}(\alpha_0, \theta_0) + \left(\frac{\partial U_{n,\theta}(\alpha_0, \theta_0)}{\partial \alpha_0^{\mathrm{T}}}\right)^{\mathrm{T}} (\widetilde{\alpha} - \alpha_0) + o_p(n^{-1/2}) \\ &= U_{n,\theta}(\alpha_0, \theta_0) + H_{n,\alpha,\theta}^{\mathrm{T}}(\alpha_0, \theta_0)\,(\widetilde{\alpha} - \alpha_0) + o_p(n^{-1/2}) \\ &= U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,(\widetilde{\alpha} - \alpha_0) + o_p(n^{-1/2}) \\ &= U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1} U_{n,\alpha}(\alpha_0) + o_p(n^{-1/2}).\end{aligned}$$

Plugging this result into Lemma 6 gives

$$\begin{aligned}\widetilde{\theta} - \theta_0 &= \mathcal{I}_\theta^{-1}\left\{U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,(\widetilde{\alpha} - \alpha_0)\right\} + o_p(n^{-1/2}) \\ &= \mathcal{I}_\theta^{-1}\left\{U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1} U_{n,\alpha}(\alpha_0)\right\} + o_p(n^{-1/2}).\end{aligned}$$

□

Finally, we obtain Proposition 1 by combining the result from Lemma 7 with the result from Lemma 1 by

using Slutsky's theorem:

$$
\begin{aligned}
\sqrt{n}\,(\widetilde{\eta}-\eta_0) &= \sqrt{n}\begin{pmatrix}\widetilde{\alpha}-\alpha_0\\ \widetilde{\theta}-\theta_0\end{pmatrix}\\
&= \sqrt{n}\begin{pmatrix}\mathcal{I}_\alpha^{-1}\,U_{n,\alpha}(\alpha_0)+o_p(n^{-1/2})\\ \mathcal{I}_\theta^{-1}\left(U_{n,\theta}(\alpha_0,\theta_0)-\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1}U_{n,\alpha}(\alpha_0)\right)+o_p(n^{-1/2})\end{pmatrix}\\
&= \sqrt{n}\begin{pmatrix}\mathcal{I}_\alpha^{-1} & 0\\ 0 & \mathcal{I}_\theta^{-1}\end{pmatrix}\begin{pmatrix}U_{n,\alpha}(\alpha_0)\\ U_{n,\theta}(\alpha_0,\theta_0)-\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1}U_{n,\alpha}(\alpha_0)\end{pmatrix}+\begin{pmatrix}o_p(1)\\ o_p(1)\end{pmatrix}\\
&\xrightarrow{d} \begin{pmatrix}\mathcal{I}_\alpha^{-1} & 0\\ 0 & \mathcal{I}_\theta^{-1}\end{pmatrix}\begin{pmatrix}\Lambda_\alpha\\ \Lambda_\theta-\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1}\Lambda_\alpha\end{pmatrix}\\
&= \begin{pmatrix}\mathcal{I}_\alpha & 0\\ 0 & \mathcal{I}_\theta\end{pmatrix}^{-1}\begin{pmatrix}1 & 0\\ -\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1} & 1\end{pmatrix}\begin{pmatrix}\Lambda_\alpha\\ \Lambda_\theta\end{pmatrix}\\
&= \mathcal{I}_\eta^{-1}L\,\Lambda_\eta \sim \mathrm{N}\left(0,\,\mathcal{I}_\eta^{-1}L\,K_\eta\,L^{\mathrm{T}}\mathcal{I}_\eta^{-1}\right).
\end{aligned}
$$

□

A.4 Proof of Lemma 3

Proof. We assume that margins and copula are correctly specified, i.e. $f(\cdot,\eta_0)=g$, or simply $g=f$ in concise notation. Then

$$
\begin{aligned}
K_{\alpha_1,\theta} &= \mathrm{Cov}_G\left(U_{\alpha_1}(y_1,\alpha_{0,1}),U_\theta(y,\alpha_0,\theta_0)\right)\\
&= \mathrm{E}_G\left[U_{\alpha_1}(y_1,\alpha_{0,1})\,U_\theta(y,\alpha_0,\theta_0)^{\mathrm{T}}\right]\\
&= \int_{y_1}\cdots\int_{y_d} g\,U_{\alpha_1}(y_1,\alpha_{0,1})\,U_\theta(y,\alpha_{0,1},\theta_0)^{\mathrm{T}}\,\mathrm{d}y_1\cdots\mathrm{d}y_d\\
&= \int_{y_1} U_{\alpha_1}(y_1,\alpha_{0,1})\left(\int_{y_2}\cdots\int_{y_d} g\,U_\theta(y,\alpha_{0,1},\theta_0)^{\mathrm{T}}\,\mathrm{d}y_2\cdots\mathrm{d}y_d\right)\mathrm{d}y_1\\
&= \int_{y_1} U_{\alpha_1}(y_1,\alpha_{0,1})\left(\int_{y_2}\cdots\int_{y_d} g\left(\frac{\partial\log c\left(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}),\theta_0\right)}{\partial\theta_0}\right)^{\mathrm{T}}\mathrm{d}y_2\cdots\mathrm{d}y_d\right)\mathrm{d}y_1\\
&= \int_{y_1} U_{\alpha_1}(y_1,\alpha_{0,1})\left(\int_{y_2}\cdots\int_{y_d} g\left(\frac{\partial\log f(y,\alpha_{0,1},\theta_0)}{\partial\theta_0}\right)^{\mathrm{T}}\mathrm{d}y_2\cdots\mathrm{d}y_d\right)\mathrm{d}y_1\\
&= \int_{y_1} U_{\alpha_1}(y_1,\alpha_{0,1})\left(\int_{y_2}\cdots\int_{y_d} f\left(\frac{\partial\log f(y,\alpha_{0,1},\theta_0)}{\partial\theta_0}\right)^{\mathrm{T}}\mathrm{d}y_2\cdots\mathrm{d}y_d\right)\mathrm{d}y_1.
\end{aligned}
$$

By using that $\partial \log f / \partial \theta_0 = (1/f) \partial f / \partial \theta_0$, this becomes

$$
\begin{aligned}
&= \int_{y_1} U_{\alpha_1}(y_1, \alpha_{0,1}) \left(\int_{y_2} \cdots \int_{y_d} \left(\frac{\partial f(y, \alpha_{0,1}, \theta_0)}{\partial \theta_0} \right)^{\mathrm{T}} \mathrm{d}y_2 \cdots \mathrm{d}y_d \right) \mathrm{d}y_1 \\
&= \int_{y_1} U_{\alpha_1}(y_1, \alpha_{0,1}) \left(\frac{\partial}{\partial \theta_0} \int_{y_2} \cdots \int_{y_d} f \, \mathrm{d}y_2 \cdots \mathrm{d}y_d \right)^{\mathrm{T}} \mathrm{d}y_1 \\
&= \int_{y_1} U_{\alpha_1}(y_1, \alpha_{0,1}) \left(\frac{\partial}{\partial \theta_0} f_1 \right)^{\mathrm{T}} \mathrm{d}y_1 \\
&= \int_{y_1} U_{\alpha_1}(y_1, \alpha_{0,1}) \cdot 0 \, \mathrm{d}y_1 \\
&= 0.
\end{aligned}
$$

Similarly, $K_{\alpha_j, \theta} = 0$ for all $j = 1, \ldots, d$. Thus, $K_{\alpha, \theta} = 0$.

□

A.5 Proof of Lemma 4

Proof. We assume again that margins and copula are correctly specified, i.e. $g = f(\cdot, \eta_0)$, or $g = g$, and have

$$
\begin{aligned}
\mathrm{E}_G \left[U_\theta(y, \alpha_0, \theta_0) \right] &= \int g \, \frac{\partial \log c}{\partial \theta_0} \, \mathrm{d}y \\
&= \int f \, \frac{\partial \log c}{\partial \theta_0} \, \mathrm{d}y \\
&= \int f \, \frac{1}{c} \frac{\partial c}{\partial \theta_0} \, \mathrm{d}y \\
&= \int \left(\prod_{j=1}^{d} f_j \right) \frac{\partial c}{\partial \theta_0} \, \mathrm{d}y \\
&= \int \frac{\partial c \prod_{j=1}^{d} f_j}{\partial \theta_0} \, \mathrm{d}y \\
&= \int \frac{\partial f}{\partial \theta_0} \, \mathrm{d}y \\
&= \frac{\partial}{\partial \theta_0} \int f \, \mathrm{d}y,
\end{aligned}
$$

which is zero. Also,

$$
\begin{aligned}
0 &= \frac{\partial \, \mathrm{E}_G \left[U_\theta(y, \alpha_0, \theta_0) \right]^{\mathrm{T}}}{\partial \alpha_0} \\
&= \frac{\partial}{\partial \alpha_0} \int f \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \, \mathrm{d}y \\
&= \int f \, \frac{\partial^2 \log c}{\partial \alpha_0 \partial \theta_0^{\mathrm{T}}} + \frac{\partial f}{\partial \alpha_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \, \mathrm{d}y \\
&= \int f \, \frac{\partial^2 \log c}{\partial \alpha_0 \partial \theta_0^{\mathrm{T}}} + f \, \frac{\partial \log f}{\partial \alpha_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \, \mathrm{d}y \\
&= \int f \, \frac{\partial^2 \log c}{\partial \alpha_0 \partial \theta_0^{\mathrm{T}}} + f \, \left(\frac{\partial \sum_{j=1}^{d} \log f_j}{\partial \alpha_0} + \frac{\partial \log c}{\partial \alpha_0} \right) \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \, \mathrm{d}y \\
&= -\mathrm{E}_G \left[-\frac{\partial^2 \log c}{\partial \alpha_0 \partial \theta_0^{\mathrm{T}}} \right] + \mathrm{E}_G \left[\frac{\partial \sum_{j=1}^{d} \log f_j}{\partial \alpha_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \right] + \mathrm{E}_G \left[\frac{\partial \log c}{\partial \alpha_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \right] \\
&= -\mathrm{E}_G \left[-H_{\alpha,\theta}(y, \alpha_0, \theta_0) \right] + \mathrm{E}_G \left[U_\alpha(y, \alpha_0) \, U_\theta(y, \alpha_0, \theta_0)^{\mathrm{T}} \right] + \mathrm{E}_G \left[U_\alpha^*(y, \alpha_0) \, U_\theta(y, \alpha_0, \theta_0)^{\mathrm{T}} \right] \\
&= -\mathcal{I}_{\alpha,\theta} + K_{\alpha,\theta} + K_{\alpha,\theta}^* \\
&= -\mathcal{I}_{\alpha,\theta} + K_{\alpha,\theta}^*.
\end{aligned}
$$

So, $\mathcal{I}_{\alpha,\theta} = K_{\alpha,\theta}^*$. □

A.6 Proof of Lemma 5

Proof. We assume that margins and copula are correctly specified, i.e. $g = f(\cdot, \eta_0)$. Then

$$
\begin{aligned}
0 &= \frac{\partial \, \mathrm{E}_G \left[U_\theta(y, \alpha_0, \theta_0) \right]^{\mathrm{T}}}{\partial \theta_0} \\
&= \frac{\partial}{\partial \theta_0} \int f \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \, \mathrm{d}y \\
&= \int f \Big(\frac{\partial^2 \log c}{\partial \theta_0 \partial \theta_0^{\mathrm{T}}} + \frac{\partial f}{\partial \theta_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \Big) \, \mathrm{d}y \\
&= \int \Big(f \, \frac{\partial^2 \log c}{\partial \theta_0 \partial \theta_0^{\mathrm{T}}} + f \, \frac{\partial \log f}{\partial \theta_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \Big) \, \mathrm{d}y \\
&= \int \Big(f \, \frac{\partial^2 \log c}{\partial \theta_0 \partial \theta_0^{\mathrm{T}}} + f \, \frac{\partial \log c}{\partial \theta_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \Big) \, \mathrm{d}y \\
&= -\mathrm{E}_G \left[-\frac{\partial^2 \log c}{\partial \theta_0 \partial \theta_0^{\mathrm{T}}} \right] + \mathrm{E}_G \left[\frac{\partial \log c}{\partial \theta_0} \, \frac{\partial \log c}{\partial \theta_0^{\mathrm{T}}} \right] \\
&= -\mathrm{E}_G \left[-H_\theta(y, \alpha_0, \theta_0) \right] + \mathrm{E}_G \, U_\theta(y, \alpha_0, \theta_0) \, U_\theta(y, \alpha_0, \theta_0)^{\mathrm{T}} \\
&= -\mathcal{I}_\theta + K_\theta.
\end{aligned}
$$

So, $\mathcal{I}_\theta = K_\theta$. □