Copula information criterion for model selection with two-stage maximum likelihood estimation

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

February 2018

Abstract

In parametric copula setups, where both the margins and copula have parametric forms, two-stage maximum likelihood estimation, often referred to as inference functions for margins, is used as an attractive alternative to the full maximum likelihood estimation strategy. Exploiting basic results derived earlier by the present authors, we develop a copula information criterion (CIC) for model selection. The CIC is defined as $\text{CIC} = 2\ell_n(\widetilde{\eta}) - 2\widetilde{p}^*_\eta$, where $\ell_n(\widetilde{\eta})$ is the maximized log-likelihood under the two-stage maximum likelihood estimation scheme, with η the full parameter vector for the candidate model in question, and $\widetilde{p}^*_\eta$ is a certain penalization factor. In a nutshell, CIC aims for the model that minimizes the Kullback–Leibler divergence from the real data generating mechanism. CIC does not assume that the chosen parametric model captures this true model, unlike what is assumed for AIC. In this sense CIC is analogous to the Takeuchi Information Criterion (TIC), which is defined for the full maximum likelihood. If we make an additional assumption that a candidate model is correctly specified, then CIC for that model simplifies to AIC. Further, since both CIC and TIC are estimating the same part of the Kullback–Leibler divergence, they are compatible, in the sense that they can be used to compare the performance of full maximum-likelihood and two-stage maximum likelihood for a given model. Additionally, we show that CIC can easily be extended to the conditional copula setup where covariates are parametrically linked to the copula model.

As a numerical illustration, we perform a simulation and find that CIC outperforms AIC in terms of prediction performance from the selected models. However, as sample size grows, the difference between CIC and AIC becomes minimal because the log-likelihood part outgrows the bias correction part. Further, we learn from the simulation that $\widetilde{p}^*_\eta$, the bias correction term of CIC, has a strong positive relationship with the prediction performance of the model. So, a model with bad prediction performance is being penalized more by CIC.

Keywords: copula, Akaike information criterion, copula information criterion, model robust, two-stage maximum likelihood, inference functions for margins,

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

1 Introduction and copula models

One of the main practical issues in copula modeling is model selection. In the full parametric setup, where both the copula and margins are assumed to have a parametric form, one often has multiple candidates for both the copula and margins. As the dimension of the model increases, a list of possible combinations of margins and copula grows rapidly. Hence, there is a need for a model selection criterion that can evaluate each model systematically according to certain philosophy or criteria and assign a score to each model. In the end, one would choose the model with the best score.

Throughout this paper, we consider the full parametric setup. In this setup, one can simultaneously estimate all parameters of the model (i.e. both copula parameters and margin parameters) by using maximum likelihood (ML) estimation. In this ML estimation framework, one can for instance use AIC_{ML} (Akaike, 1974) or TIC (also known as model-robust AIC_{ML}) (Takeuchi, 1976) as model selection criterion and select the model with the best score. (Note that we denote the AIC under ML estimation as AIC_{ML} to distinguish it from the two-stage ML based AIC_{2ML}, which we will derive in Section 2.3.) However, when the dimension of the copula model gets high, the number of parameters increases quickly and the ML estimation is not always feasible in terms of speed and numerical stability. Two-stage maximum likelihood (two-stage ML) estimation, also often referred to as inference functions for margins (IFM), is a popular alternative estimation strategy that is designed to overcome these drawbacks of the ML estimation. In stage 1 of the two-stage ML estimation, the parameter vectors of each marginal distribution are estimated separately by ML. In stage 2, the estimates from stage 1 are plugged into the log-likelihood of the model. Then, the parameters of the copula, which are now the only unknown parameters, are estimated by using ML estimation again. One of the advantages of this multi-stage approach is that it is computation-wise much faster than estimating all parameters simultaneously, because it does not have to search for the global maximum in high-dimensional space. A drawback of the two-stage ML estimation method, however, is that we cannot use the classical results based on ML estimation, which include model selection criteria such as TIC and BIC.

In practice, different sorts of goodness-of-fit testing are often used as substitutes, to choose the best model (Genest & Favre, 2007). Another often used model selection strategy for the two-stage ML is that one first evaluates candidates of each marginal distribution with AIC_{ML} and consequently chooses the best distribution for each margin. Once the margins are chosen, one fits different copulae and evaluates the copula part with AIC_{ML}. However, this piecewise model evaluation cannot evaluate the model as a whole.

In this paper, we develop the copula information criterion (CIC) for two-stage ML estimation, which has the form

$$\text{CIC} = 2\ell_n(\widetilde{\eta}) - 2\widetilde{p}^*_\eta.$$

Here $\ell_n(\widetilde{\eta})$ is the maximized log-likelihood with the two-stage ML estimation method, in terms of the full parameter vector η of the model in question, and $\widetilde{p}^*_\eta$ is a suitable penalization factor, worked out in Section 2.2. The main advantage of CIC is that it can evaluate a parametric copula with parametric margins as a whole. CIC is also a model-robust model selection criterion which means that it does not assume that the candidate model contains the true model. As the overlap of the name already suggests, our CIC is analogous to CIC from Grønneberg & Hjort (2014), which is designed for copulae estimated with pseudo maximum likelihood (PML). In PML framework, margins are estimated empirically, while two-stage ML assumes parametric forms of margins.

Our technical setting, identical to Ko & Hjort (2018), 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 $y_i = (y_{i,1}, \cdots, y_{i,d})^{\mathrm{T}}$, for $i = 1, \ldots, n$, be independent observations of this variable. The true joint distribution g is typically unknown. Let $f(y_1, \cdots, y_d, \eta)$ be our parametric approximation of g, with the parameter vector η, belonging to some connected subset of the appropriate Euclidean space. In addition, G and $F(\cdot, \eta)$ indicate cumulative distribution functions corresponding to g and $f(\cdot, \eta)$, respectively. Here $G_j(y_j)$ and $F_j(y_j, \alpha_j)$ indicate j-th marginal distribution functions corresponding to G and $F(\cdot, \eta)$ respectively, with α_j as the parameter vector belonging to margin component j.

According to Sklar's theorem (Sklar, 1959), there always exists a copula $C(u_1, \ldots, u_d, \theta)$ that satisfies

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

where the full parameter vector η is now blocked as

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

By assuming the regulatory conditions from Ko & Hjort (2018), $C(\cdot, \theta)$ can be differentiated,

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

where $c(u_1, \ldots, u_d) = \partial^d C(u_1, \ldots, u_d, \theta)/\partial u_1 \cdots \partial u_d$ and $f_j(y_j, \alpha_j) = \partial F_j(y_j, \alpha_j)/\partial y_j$. For further details of copula modeling, see Joe (1997) and Nelsen (2006). Analogously, the true density g can also be decomposed into marginal densities and the copula density

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

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

The further structure of this paper is as follows. In Section 2.1, we briefly explain Kullback–Leibler divergence and its relationship to TIC and $\mathrm{AIC_{ML}}$. In Section 2.2, we derive and define our copula information criterion. In Section 2.3, we prove that the $\mathrm{AIC_{2ML}}$ formula holds under the two-stage ML estimation. In Section 2.4, we summarize the relationship between TIC, CIC, $\mathrm{AIC_{ML}}$ and $\mathrm{AIC_{2ML}}$. In Section 2.5, we illustrate what CIC looks like in the two-dimensional setting and show how CIC easily can be extended to the conditional copula setting. In Section 3, we study the numerical behavior of those model selection criteria. In our final Section 4, we offer a few concluding remarks and suggestions for future research.

2 The copula information criterion for two-stage maximum likelihood estimation

2.1 Kullback–Leibler divergence

The Kullback–Leibler (KL) divergence from g to f measures how the density f diverges from g (Kullback & Leibler, 1951) and is defined as

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

It is well known that the ML estimator aims for parameter values that minimize the Kullback–Leibler divergence (Akaike, 1998).

Consider a case where one has competing models for certain data and the parameters are estimated by ML. An arbitrary candidate model with ML parameter estimates can be denoted as $f(y, \widehat{\eta})$. (Throughout this paper, we use '$\widehat{\ }$' to indicate that a quantity is estimated with ML and '$\widetilde{\ }$' to indicate that a quantity is estimated with two-stage ML.) Since $g(y)$ is the same across all candidate models, minimizing KL divergence is equal to maximizing $Q(\widehat{\eta}) = \int g(y) \log f(y, \widehat{\eta}) \mathrm{d}y$. This quantity is, however, not directly observable since $g(y)$ is unknown. As an alternative, one may use the empirical equivalent of $Q(\widehat{\eta})$:

$$\widehat{Q}(\widehat{\eta}) = \frac{1}{n} \ell(\widehat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \Big[\log f_1(y_{i,1}, \widehat{\alpha}_1) + \cdots + \log f_d(y_{i,d}, \widehat{\alpha}_d) + \log c\big(F_1(y_{i,1}, \widehat{\alpha}_1), \cdots, F_d(y_{i,d}, \widehat{\alpha}_d), \widehat{\theta}\big) \Big].$$

Yet, this estimator of $Q(\widehat{\eta})$ is a biased estimator. By identifying and subtracting the bias, we obtain the unbiased estimator of $Q(\widehat{\eta})$:

$$\widehat{Q}^*(\widehat{\eta}) = \frac{1}{n} \ell(\widehat{\eta}) - \frac{1}{n} \mathrm{tr}\big(\widehat{\mathcal{I}}_\eta^{-1} \widehat{K}_\eta\big).$$

where $\widehat{\mathcal{I}}_\eta$ is the observed information and $\widehat{K}_\eta$ is the estimated covariance matrix of η.

The TIC (Takeuchi, 1976) aims for the model that maximizes $\widehat{Q}^*(\widehat{\eta})$ and is defined as

$$\mathrm{TIC} = 2\ell(\widehat{\eta}) - 2\widehat{p}^*_{\eta,\mathrm{TIC}}, \tag{1}$$

where $\widehat{p}^*_{\eta,\mathrm{TIC}} = \mathrm{tr}\big(\widehat{\mathcal{I}}_\eta^{-1} \widehat{K}_\eta\big)$. This shows that TIC is basically a scaled version of $\widehat{Q}^*(\widehat{\eta})$.

When one boldly makes the assumption that the candidate model is correct, i.e. contains the true data generating mechanism, TIC simplifies to $\mathrm{AIC}_{\mathrm{ML}}$, possibly the most well known model selection criterion in statistics, which is defined as

$$\mathrm{AIC}_{\mathrm{ML}} = 2\ell(\widehat{\eta}) - 2p_\eta, \tag{2}$$

where p_η is the length of the parameter vector η. Note that the formula of TIC and $\mathrm{AIC}_{\mathrm{ML}}$ are only valid if the parameters are estimated by the ML estimator. For more details about TIC and $\mathrm{AIC}_{\mathrm{ML}}$, see chapter 2 of Claeskens & Hjort (2008).

2.2 Derivation of the copula information criterion

When the copula model is estimated with the two-stage ML, the bias correction term of TIC, i.e. $\widehat{p}^*_{\eta,\mathrm{TIC}}$, is not valid. We derive the copula information criterion (CIC) which is analogous to TIC and is made for copula models estimated with the two-stage ML.

When the copula model is estimated with the two-stage ML, the non-constant part of the KL divergence is

$$Q(\widetilde{\eta}) = \int g(y) \left\{ \log f_1(y_1, \widetilde{\alpha}_1) + \cdots + \log f_d(y_d, \widetilde{\alpha}_d) + \log c\big(F_1(y_1, \widetilde{\alpha}_1), \cdots, F_d(y_d, \widetilde{\alpha}_d), \widetilde{\theta}\big) \right\} \mathrm{d}y.$$

The empirical equivalent is

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

Now we check the bias of $\widetilde{Q}(\widetilde{\eta})$:

$$\begin{aligned}
\mathrm{E}_G\big[\widetilde{Q}(\widetilde{\eta})\big] - Q(\widetilde{\eta}) = {} & \mathrm{E}_{G_1}\left[\frac{1}{n}\sum_{i=1}^{n} \log f_1(y_{i,1}, \widetilde{\alpha}_1)\right] - \int g(y_1) \log f_1(y_1, \widetilde{\alpha}_1) \mathrm{d}y_1 \\
& + \cdots \\
& + \mathrm{E}_{G_d}\left[\frac{1}{n}\sum_{i=1}^{n} \log f_d(y_{i,d}, \widetilde{\alpha}_d)\right] - \int g(y_d) \log f_d(y_d, \widetilde{\alpha}_d) \mathrm{d}y_d \\
& + \mathrm{E}_{G}\left[\frac{1}{n}\sum_{i=1}^{n} \log c\left(F_1(y_{i,1}, \widetilde{\alpha}_1), \cdots, F_d(y_{i,d}, \widetilde{\alpha}_d), \widetilde{\theta}\right)\right] \\
& - \int g(y) \log c\left(F_1(y_1, \widetilde{\alpha}_1), \cdots, F_d(y_d, \widetilde{\alpha}_d), \widetilde{\theta}\right) \mathrm{d}y.
\end{aligned}$$

Since the parameters of marginals from stage 1 ($\widetilde{\alpha}_j$s) are obtained with ML estimation, we can directly use the results from the derivation of TIC and obtain

$$\mathrm{E}_{G_j}\left[\frac{1}{n}\sum_{i=1}^{n} \log f_j(y_{i,j}, \widetilde{\alpha}_j)\right] - \int g(y_j) \log f(y_j, \widetilde{\alpha}_j) \mathrm{d}y_j = \frac{1}{n}\mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_j} K_{\alpha_j}\right) + o(n^{-1}) = \frac{1}{n}\widehat{p}^*_{\alpha_j} + o(n^{-1}),$$

with $\mathcal{I}^{-1}_{\alpha_j}$ and K_{α_j} as defined in Lemma 1 of Ko & Hjort (2018). In a nutshell, $\mathcal{I}_{\alpha_j}$ is the Fisher information of j-th margin and K_{α_j} is the covariance matrix of the score vector that belongs to j-th margin.

Further, let

$$Q_c(\widetilde{\eta}) = \int g(y) \log c\big(F_1(y_1, \widetilde{\alpha}_1), \cdots, F_d(y_d, \widetilde{\alpha}_d), \widetilde{\theta}\big) \mathrm{d}y$$

and

$$\widetilde{Q}_c(\widetilde{\eta}) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log c\big(F_1(y_{i,1}, \widetilde{\alpha}_1), \cdots, F_d(y_{i,d}, \widetilde{\alpha}_d), \widetilde{\theta}\big)\right].$$

So, we can write

$$\mathrm{E}_G\left[\widetilde{Q}(\widetilde{\eta})\right] - Q(\widetilde{\eta}) = \frac{1}{n}\sum_{j=1}^{d} \mathrm{tr}\left(\mathcal{I}_{\alpha_j}^{-1} K_{\alpha_j}\right) + \mathrm{E}_G\left[\widetilde{Q}_c(\widetilde{\eta})\right] - Q_c(\widetilde{\eta}) + o(n^{-1}).$$

Now, $\mathrm{E}_G\left[\widetilde{Q}_c(\widetilde{\eta})\right] - Q_c(\widetilde{\eta})$ is the only element that should be evaluated. Let

$$\begin{aligned}
Q_c(\eta_0) &= \int g(y) \log c\left(F_1(y_1, \alpha_{0,1}), \cdots, F_d(y_d, \alpha_{0,d}), \theta_0\right) \mathrm{d}y, \\
Z_i &= \log c\left(F_1(y_{i,1}, \alpha_{0,1}), \cdots, F_d(y_{i,d}, \alpha_{0,d}), \theta_0\right) - Q_c(\eta_0), \\
A_\eta &= \sqrt{n}\left(\widetilde{\eta} - \eta_0\right) = \sqrt{n}\,\mathcal{I}_\eta^{-1} \begin{pmatrix} U_{n,\alpha}(\alpha_0) \\ U_{n,\theta}(\alpha_0, \theta_0) - \mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\,\mathcal{I}_\alpha^{-1} U_{n,\alpha}(\alpha_0) \end{pmatrix} + \begin{pmatrix} o_p(1) \\ o_p(1) \end{pmatrix},
\end{aligned}$$

which stems from Proposition 1 in Ko & Hjort (2018), and furthermore

$$\begin{aligned}
U_{n,\eta}(\eta) &= \frac{1}{n}\sum_{i=1}^{n} U_\eta(y_i, \eta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log c\left(F_1(y_{i,1}, \alpha_1), \cdots, F_d(y_{i,d}, \alpha_d), \theta\right)}{\partial \eta}, \\
H_{n,\eta}(\eta) &= \frac{1}{n}\sum_{i=1}^{n} H_\eta(y_i, \eta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log c\left(F_1(y_{i,1}, \alpha_1), \cdots, F_d(y_{i,d}, \alpha_d), \theta\right)}{\partial \eta \partial \eta^{\mathrm{T}}}, \\
\mathcal{I}_\eta^* &= -\mathrm{E}_G\left[H_\eta(y, \eta_0)\right] = -\int g(y) H_\eta(y, \eta_0)\, \mathrm{d}y,
\end{aligned}$$

which, incidentally, should not be confused with $\mathcal{I}_\eta$ in Proposition 1 of Ko & Hjort (2018). Then we have $\mathrm{E}[\overline{Z}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Z_i] = 0$, along with

$$\begin{aligned}
\widetilde{Q}_c(\widetilde{\eta}) &= \frac{1}{n}\sum_{i=1}^{n} \log c(y_i, \widetilde{\eta}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\log c(y_i, \eta_0) - Q_c(\eta_0) + Q_c(\eta_0) + (\widetilde{\eta} - \eta_0)^{\mathrm{T}} U_\eta(y_i, \eta_0) + \frac{1}{2}(\widetilde{\eta} - \eta_0)^{\mathrm{T}} H_\eta(y_i, \eta_0)(\widetilde{\eta} - \eta_0)\right] + o_p(n^{-1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} [Z_i] + Q_c(\eta_0) + (\widetilde{\eta} - \eta_0)^{\mathrm{T}} U_{n,\eta}(\eta_0) + \frac{1}{2}(\widetilde{\eta} - \eta_0)^{\mathrm{T}} H_{n,\eta}(\eta_0)(\widetilde{\eta} - \eta_0) + o_p(n^{-1}) \\
&= Q_c(\eta_0) + \overline{Z}_n + \frac{1}{\sqrt{n}} A_\eta^{\mathrm{T}} U_{n,\eta}(\eta_0) + \frac{1}{2n} A_\eta^{\mathrm{T}} H_{n,\eta}(\eta_0) A_\eta + o_p(n^{-1}),
\end{aligned}$$

$$
\begin{aligned}
Q_c(\widetilde{\eta}) &= \int g(y) \log c\left(F_1(y_1, \widetilde{\alpha}_1), \cdots, F_d(y_d, \widetilde{\alpha}_d), \widetilde{\theta}\right) \mathrm{d}y \\
&= \int g(y) \left[\log c(y, \eta_0) + (\widetilde{\eta} - \eta_0)^{\mathrm{T}} U_\eta(y, \eta_0) + \frac{1}{2}(\widetilde{\eta} - \eta_0)^{\mathrm{T}} H_\eta(y, \eta_0)(\widetilde{\eta} - \eta_0)\right] \mathrm{d}y + o_p(n^{-1}) \\
&= \int g(y) \log c(y, \eta_0)\, \mathrm{d}y + (\widetilde{\eta} - \eta_0)^{\mathrm{T}} \int g(y) U_\eta(y, \eta_0)\, \mathrm{d}y + \frac{1}{2}(\widetilde{\eta} - \eta_0)^{\mathrm{T}} \int g(y) H_\eta(y, \eta_0)\, \mathrm{d}y \cdot (\widetilde{\eta} - \eta_0) + o_p(n^{-1}) \\
&= Q_c(\eta_0) + (\widetilde{\eta} - \eta_0)^{\mathrm{T}} \cdot 0 + \frac{1}{2n}\sqrt{n}(\widetilde{\eta} - \eta_0)^{\mathrm{T}} \int g(y) H_\eta(y, \eta_0)\, \mathrm{d}y \cdot \sqrt{n}(\widetilde{\eta} - \eta_0) + o_p(n^{-1}) \\
&= Q_c(\eta_0) - \frac{1}{2n} A_\eta^{\mathrm{T}} \mathcal{I}_\eta^* A_\eta + o_p(n^{-1})
\end{aligned}
$$

and

$$
n\{\widetilde{Q}_c(\widetilde{\eta}) - Q_c(\widetilde{\eta})\} = n\overline{Z}_n + \sqrt{n} A_\eta^{\mathrm{T}} U_{n,\eta}(\eta_0) + \frac{1}{2} A_\eta^{\mathrm{T}} H_{n,\eta}(\eta_0) A_\eta + \frac{1}{2} A_\eta^{\mathrm{T}} \mathcal{I}_\eta^* A_\eta + o_p(1).
$$

Further, note that

$$
U_{n,\eta}(\eta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log c\left(F_1(y_{i,1}, \alpha_{0,1}), \cdots, F_d(y_{i,d}, \alpha_{0,d}), \theta_0\right)}{\partial \eta_0}
$$

has the following convergence, by the central limit theorem:

$$
\sqrt{n} U_{n,\eta}(\eta_0) = \sqrt{n}\left\{U_{n,\eta}(\eta_0) - \mathrm{E}\left[U_\eta(Y, \eta_0)\right]\right\} \xrightarrow{d} \Lambda_\eta^* \sim N\left(0, K_\eta^*\right)
$$

where $K_\eta^* = \mathrm{Var}\left(U_\eta(y, \eta_0)\right) = \mathrm{E}\left[U_\eta(y, \eta_0)\, U_\eta(y, \eta_0)^{\mathrm{T}}\right]$. (Here Λ_η^* and K_η^* should not be confused with Λ_η and K_η in Proposition 1 of Ko & Hjort (2018).)

Now we evaluate $\mathrm{E}_G[n\{\widetilde{Q}_c(\widetilde{\eta}) - Q_c(\widetilde{\eta})\}]$:

$$
\begin{aligned}
\mathrm{E}_G\left[n\left(\widetilde{Q}_c(\widetilde{\eta}) - Q_c(\widetilde{\eta})\right)\right] &= \mathrm{E}_G\left[n\overline{Z}_n + \sqrt{n} A_\eta^{\mathrm{T}} U_{n,\eta}(\eta_0) + \frac{1}{2} A_\eta^{\mathrm{T}} H_{n,\eta}(\eta_0) A_\eta + \frac{1}{2} A_\eta^{\mathrm{T}} \mathcal{I}_\eta^* A_\eta\right] + o(1) \\
&\xrightarrow{p} \mathrm{E}_G\left[\left(\mathcal{I}_\eta^{-1} L\, \Lambda_\eta\right)^{\mathrm{T}} \Lambda_\eta^*\right] \\
&= \mathrm{E}_G\left[\mathrm{tr}\left(\mathcal{I}_\eta^{-1} L\, \Lambda_\eta \left(\Lambda_\eta^*\right)^{\mathrm{T}}\right)\right] \\
&= \mathrm{tr}\left(\mathcal{I}_\eta^{-1} L\, \mathrm{E}_G\left[\Lambda_\eta (\Lambda_\eta^*)^{\mathrm{T}}\right]\right) = \mathrm{tr}\left(\mathcal{I}_\eta^{-1} L\, K_\eta^\circ\right) = p_\theta^*,
\end{aligned}
$$

where

$$
K_\eta^\circ = \mathrm{E}_G\left[\Lambda_\eta (\Lambda_\eta^*)^{\mathrm{T}}\right] = \mathrm{E}_G\left[\begin{pmatrix} \Lambda_\alpha \left(\Lambda_\alpha^*\right)^{\mathrm{T}} & \Lambda_\alpha \Lambda_\theta^{\mathrm{T}} \\ \Lambda_\theta \left(\Lambda_\alpha^*\right)^{\mathrm{T}} & \Lambda_\theta \Lambda_\theta^{\mathrm{T}} \end{pmatrix}\right] = \begin{pmatrix} K_\alpha^\circ & K_{\alpha,\theta} \\ \left(K_{\alpha,\theta}^*\right)^{\mathrm{T}} & K_\theta \end{pmatrix}.
$$

It is practical to note that

$$\begin{aligned}
\operatorname{tr}\left(\mathcal{I}_\eta^{-1} L\, K_\eta^\circ\right) &= \operatorname{tr}\left(\begin{pmatrix} \mathcal{I}_\alpha^{-1} & 0 \\ 0 & \mathcal{I}_\theta^{-1} \end{pmatrix}\begin{pmatrix} I & 0 \\ -\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\mathcal{I}_\alpha^{-1} & I \end{pmatrix}\begin{pmatrix} K_\alpha^\circ & K_{\alpha,\theta} \\ \left(K_{\alpha,\theta}^*\right)^{\mathrm{T}} & K_\theta \end{pmatrix}\right) \\
&= \operatorname{tr}\left(\begin{pmatrix} \mathcal{I}_\alpha^{-1}K_\alpha^\circ & \mathcal{I}_\alpha^{-1}K_{\alpha,\theta} \\ -\mathcal{I}_\theta^{-1}\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\mathcal{I}_\alpha^{-1}K_\alpha^\circ + \mathcal{I}_\theta^{-1}\left(K_{\alpha,\theta}^*\right)^{\mathrm{T}} & -\mathcal{I}_\theta^{-1}\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\mathcal{I}_\alpha^{-1}K_{\alpha,\theta} + \mathcal{I}_\theta^{-1}K_\theta \end{pmatrix}\right) \\
&= \operatorname{tr}\left(\begin{pmatrix} \mathcal{I}_\alpha^{-1}K_\alpha^\circ & 0 \\ 0 & -\mathcal{I}_\theta^{-1}\mathcal{I}_{\alpha,\theta}^{\mathrm{T}}\mathcal{I}_\alpha^{-1}K_{\alpha,\theta} + \mathcal{I}_\theta^{-1}K_\theta \end{pmatrix}\right).
\end{aligned}$$

Consistent estimators for $\mathcal{I}_{\alpha_j}^{-1}$, K_{α_j}, $\mathcal{I}_\eta^{-1}$, L and K_η° can be obtained by using plug-in sample averages. The consequent estimators are denoted as $\widetilde{\mathcal{I}}_{\alpha_j}^{-1}$, $\widetilde{K}_{\alpha_j}$, $\widetilde{\mathcal{I}}_\eta^{-1}$, $\widetilde{L}$, $\widetilde{K}_\eta^\circ$. For regularity conditions that ensure convergence in probability, see Jullum & Hjort (2017).

Thus, the unbiased estimator of $Q(\widetilde{\eta})$ is

$$\widetilde{Q}^*(\widetilde{\eta}) = \frac{1}{n}\ell_n(\widetilde{\eta}) - \frac{1}{n}\left\{\sum_{j=1}^{d}\operatorname{tr}\left(\widetilde{\mathcal{I}}_{\alpha_j}^{-1}\widetilde{K}_{\alpha_j}\right) + \operatorname{tr}\left(\widetilde{\mathcal{I}}_\eta^{-1}\widetilde{L}\,\widetilde{K}_\eta^\circ\right)\right\}.$$

By defining $\widetilde{p}_{\alpha_j}^* = \operatorname{tr}(\widetilde{\mathcal{I}}_{\alpha_j}^{-1}\widetilde{K}_{\alpha_j})$, $\widetilde{p}_\theta^* = \operatorname{tr}(\widetilde{\mathcal{I}}_\eta^{-1}\widetilde{L}\,\widetilde{K}_\eta^\circ)$, $\widetilde{p}_\eta^* = \sum_{j=1}^d \widetilde{p}_{\alpha_j}^* + \widetilde{p}_\theta^*$ and scaling $\widetilde{Q}^*(\widetilde{\eta})$, we can finally define the copula information criterion as

$$\mathrm{CIC} = 2\ell_n(\widetilde{\eta}) - 2\widetilde{p}_\eta^*. \tag{3}$$

2.3 AIC for two-stage maximum likelihood estimator

The CIC, derived in Section 2.2, is a model robust model selection criterion. This means that the CIC does not assume that the parametric model includes the true model that generated data. In this section we show that, if we do make such a true model assumption, the CIC simplifies to the $\mathrm{AIC}_{\mathrm{2ML}}$. To our knowledge, this is the first time that the validity of the $\mathrm{AIC}_{\mathrm{2ML}}$ formula is proven for the two-stage ML estimator.

Lemma 1. *Under the assumption that the margins and copula are correctly specified, it holds that* $K_\alpha^\circ = \mathcal{I}_\alpha - K_\alpha$.

Proof. Assume the candidate model worked with contains the true data generating mechanism, i.e. that

$f = g$ at the relevant parameter point. Then

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

□

If we make the true model assumption (i.e. $f = g$), we have from the classical results of maximum likelihood theory that $\mathcal{I}_{\alpha_j} = K_{\alpha_j}$. This implies that we have for the bias correction term in the marginal parameters $p_{\alpha_j}^* = \mathrm{tr}(\mathcal{I}_{\alpha_j}^{-1} K_{\alpha_j}) = \dim(\alpha_j)$, for each j.

For the bias correction term in the copula parameters part, the true model assumption results in Lemma 1. Combining Lemma 1 with Lemma 3 and Lemma 5 from Ko & Hjort (2018) gives

$$
\begin{aligned}
\mathrm{tr} \left(\mathcal{I}_\eta^{-1} L \, K_\eta^\circ \right) &= \mathrm{tr} \left(\begin{pmatrix} \mathcal{I}_\alpha^{-1} K_\alpha^\circ & 0 \\ 0 & -\mathcal{I}_\theta^{-1} \mathcal{I}_{\alpha,\theta}^{\mathrm{T}} \mathcal{I}_\alpha^{-1} K_{\alpha,\theta} + \mathcal{I}_\theta^{-1} K_\theta \end{pmatrix} \right) \\
&= \mathrm{tr} \left(\begin{pmatrix} I - \mathcal{I}_\alpha^{-1} K_\alpha & 0 \\ 0 & I \end{pmatrix} \right) \\
&= \mathrm{tr} \left(\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \right) = \dim(\theta).
\end{aligned}
$$

Thus, the unbiased estimator $\widetilde{Q}^*(\widetilde{\eta})$ can be simplified as

$$
\widetilde{Q}^*(\widetilde{\eta}) = \frac{1}{n} \ell_n(\widetilde{\eta}) - \frac{1}{n} \left(\sum_{j=1}^{d} \dim(\alpha_j) + \dim(\theta) \right).
$$

By defining $p_\eta = \sum_{j=1}^{d} \dim(\alpha_j) + \dim(\theta) = \dim(\eta)$ and scaling $\widetilde{Q}^*(\widetilde{\eta})$, we obtain $\mathrm{AIC}_{\mathrm{2ML}}$ for the two-stage ML estimated copula models

$$
\mathrm{AIC}_{\mathrm{2ML}} = 2\ell_n(\widetilde{\eta}) - 2p_\eta. \tag{4}
$$

Compared to AIC_{ML} from Section 2.1, the only difference is that the log-likelihood is now estimated under the two-stage ML instead of ML. i.e. we use $\widetilde{\eta}$ instead of $\widehat{\eta}$.

2.4 Relationship between the model selection criteria

So far, we have discussed four model selection criteria for copula models. Table 1 shows an overview of the relationship between them. When the model (i.e. both copula and margins) is correctly specified, CIC and AIC_{2ML} become equal, and the same happens between TIC and AIC_{ML}. Thus, one can compare CIC and AIC_{2ML} (or TIC and AIC_{ML} in case of ML estimation) to check whether the model is correctly specified.

Further, since both CIC and TIC are estimating the same part of the KL divergence under the same model robust environment, they are compatible. This implies that one can compare CIC to TIC to measure how much one looses in terms of KL divergence by using two-stage ML estimation instead of ML estimation. The same can be done by comparing AIC_{2ML} to AIC_{ML} when one believes in the model. However, one should not compare CIC with AIC_{ML} (or TIC with AIC_{2ML}) since they are based on two different model beliefs (i.e. presence of the true model assumption). Figuratively speaking, one is in this situation comparing apples with pears.

		Model robust	
		Yes	No
Estimation	ML	TIC	AIC_{ML}
	2ML	CIC	AIC_{2ML}

Table 1: An overview of the relationship between model selection criteria discussed in this paper.

2.5 Illustration for two-dimensional case and extension to the conditional copula regression.

To make things more concrete, we now give an example of the two-dimensional case with (Y_1, Y_2) from the unknown $g(y)$. As candidate margins we choose a two-parameter distribution (e.g. normal, gamma, Weibull, etc.) for both F_1 and F_2. For the copula part, we choose a one-parameter copula (e.g. Gumbel, Frank, Clayton, etc.). The candidate model then has the form

$$f(y_1, y_2, \eta) = c\left(F_1(y_1, \alpha_1), F_2(y_2, \alpha_2), \theta\right) \cdot f_1(y_1, \alpha_1) \cdot f_2(y_2, \alpha_2)$$

where $\eta = (\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2}, \theta)^{\text{T}}$. The ingredients for the CIC can then be written down by using the block matrix form, in line with Ko & Hjort (2018):

$$K_\eta = \begin{pmatrix} K_{\alpha_1} & K_{\alpha_1,\alpha_2} & K_{\alpha_1,\theta} \\ K_{\alpha_2,\alpha_1} & K_{\alpha_2} & K_{\alpha_2,\theta} \\ K_{\alpha_1,\theta}^{\text{T}} & K_{\alpha_2,\theta}^{\text{T}} & K_\theta \end{pmatrix}, \; \mathcal{I}_\eta = \begin{pmatrix} \mathcal{I}_{\alpha_1} & 0 & 0 \\ 0 & \mathcal{I}_{\alpha_2} & 0 \\ 0 & 0 & \mathcal{I}_\theta \end{pmatrix},$$

$$L = \begin{pmatrix} I_{2\times 2} & 0 & 0 \\ 0 & I_{2\times 2} & 0 \\ -\mathcal{I}_{\alpha_1,\theta}^{\text{T}}\mathcal{I}_{\alpha_1}^{-1} & -\mathcal{I}_{\alpha_2,\theta}^{\text{T}}\mathcal{I}_{\alpha_2}^{-1} & I_{1\times 1} \end{pmatrix}, \; K_\eta^\circ = \begin{pmatrix} K_{\alpha_1}^\circ & K_{\alpha_1,\alpha_2}^\circ & K_{\alpha_1,\theta} \\ K_{\alpha_2,\alpha_1}^\circ & K_{\alpha_2}^\circ & K_{\alpha_2,\theta} \\ \left(K_{\alpha_1,\theta}^*\right)^{\text{T}} & \left(K_{\alpha_2,\theta}^*\right)^{\text{T}} & K_\theta \end{pmatrix}.$$

We consequently have

$$\begin{aligned} p^*_\eta &= p^*_{\alpha_1} + p^*_{\alpha_2} + p^*_\theta \\ &= \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_1} K_{\alpha_1}\right) + \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_2} K_{\alpha_2}\right) + \mathrm{tr}\left(\mathcal{I}^{-1}_{\eta} L\, K^\circ_\eta\right) \\ &= \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_1} K_{\alpha_1}\right) + \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_2} K_{\alpha_2}\right) + \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_1} K^\circ_{\alpha_1}\right) + \mathrm{tr}\left(\mathcal{I}^{-1}_{\alpha_2} K^\circ_{\alpha_2}\right) \\ &\quad + \mathrm{tr}\left(-\mathcal{I}^{-1}_{\theta} \mathcal{I}^{\mathrm{T}}_{\alpha_1,\theta} \mathcal{I}^{-1}_{\alpha_1} K_{\alpha_1,\theta} - \mathcal{I}^{-1}_{\theta} \mathcal{I}^{\mathrm{T}}_{\alpha_2,\theta} \mathcal{I}^{-1}_{\alpha_2} K_{\alpha_2,\theta} + \mathcal{I}^{-1}_{\theta} K_\theta\right). \end{aligned}$$

More generally, we can consider the conditional copula regression where all marginal distributions and copula are conditioned on the k-variate covariate X parametrically. Patton (2002) extends the existing theories of copula to the conditional copula setting, including the conditional version of Sklar's theorem, which gives conditional copula density

$$f(y_1, \cdots, y_d, \eta|x) = c\left(F_1(y_1, \alpha_1|x), F_2(y_2, \alpha_2|x), \theta|x\right) \cdot f_1(y_1, \alpha_1|x) \cdot f_2(y_2, \alpha_2|x).$$

For simplicity, we consider the case where the copula parameter θ is modeled by the linear calibration function with $\theta = X\beta$ where $\beta = (\beta_0, \cdots, \beta_k)^{\mathrm{T}}$ is a $k+1$ dimensional parameters. We can then consider β as the copula parameter instead of θ, which results in $\eta = (\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2}, \beta_0, \cdots, \beta_k)^{\mathrm{T}}$. One can easily define matrices K_η, $\mathcal{I}_\eta$, L and K°_η accordingly. For details about conditional copula regression, see Patton (2006), Acar *et al.* (2013) and Palaro & Hotta (2006).

3 Simulation study

To study the behavior of CIC, we have performed a simulation study.

3.1 Simulation 1

In simulation 1, we generated datasets of 3 sizes ($n = 100, 1000, 10000$) with the data generating model described in Table 2. We then, for each dataset, fitted 18 different copula models, based on the possible combinations of the candidate copulas and margins described in Table 3. We repeated this process 1000 times and the averaged results from the fitted models can be found in Table 7, Table 8 and Table 9 in the Appendix.

Table 2: Description of the data generating model used in simulation 1.

	Copula	Margin 1	Margin 2
Data generating model	Gumbel $\theta = 3$	Weibull $\alpha_1 = (1.5, 4)^{\mathrm{T}}$ (shape, scale)	Gamma $\alpha_2 = (2, 1)^{\mathrm{T}}$ (shape, rate)

Table 3: List of the candidate copulae and margins used in simulation 1

	Candidates
Copula	Gumbel, Gaussian
Margin 1	Weibull, Gamma, Log-normal
Margin 2	Weibull, Gamma, Log-normal

From all three tables, we can see that CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ result in similar model ranks. This is expected since both are aiming for the model that minimizes KL divergence. When the model is correctly specified (model 1), the estimated value of CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ are essentially the same. This confirms that CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ are equal under the true model assumption (analytically proven in Section 2.3). Further, we can observe that 'better models' according to CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ have smaller value of $\mathrm{MSE}(\widetilde{\mathrm{P}})$ in general. To clarify, $\mathrm{MSE}(\widetilde{\mathrm{P}})$ indicates mean squared error of two-stage ML estimated $\mathrm{P}(\mathrm{b}_{0.8} < y)$. Here $\mathrm{P}(\mathrm{b}_{0.8} < y)$ is the joint probability that each marginal variable has larger value than its 0.8-quantile value, defined by each marginal model. Similarly, $\mathrm{MSE}(\widehat{\mathrm{P}})$ is mean squared error of ML estimated $\mathrm{P}(\mathrm{b}_{0.8} < y)$.

When the model is correctly specified (model 1), $\widetilde{p}^*_\eta$ is virtually equal to p_η, the length of the parameter vector η. When the model has misspecification, the CIC value is penalized more as $\widetilde{p}^*_\eta$ increases. However, this is not the case for $\mathrm{AIC}_{\mathrm{2ML}}$ since $p_\eta = 5$ across all models. Thus, CIC has higher chance of choosing a less wrong model. As n increases, the influence of $\widetilde{p}^*_\eta$ on CIC decreases since the absolute value of log-likelihood grows much faster than the penalty term. This is observable in Table 4. When $n = 100$, the best models chosen by the model robust model selection criteria (TIC and CIC) result in smaller MSE values. However, when $n = 10000$, the penalty term of these model selection criteria is very small compared to the log-likelihood value. Consequently, CIC and TIC choose the same models as their model non-robust variants ($\mathrm{AIC}_{\mathrm{2ML}}$ and $\mathrm{AIC}_{\mathrm{ML}}$) do. This results in the same MSE performance.

Table 4: Result from simulation 1. For each dataset, the best model was chosen among 18 candidate models by using CIC, TIC, $\mathrm{AIC}_{\mathrm{2ML}}$ or $\mathrm{AIC}_{\mathrm{ML}}$. Then, $\mathrm{P}(\mathrm{b}_{0.8} < y)$ was computed from the best models. The table contains mean squared error of the estimated $\mathrm{P}(\mathrm{b}_{0.8} < y)$ multiplied by 10^5.

n	TIC	$\mathrm{AIC}_{\mathrm{ML}}$	CIC	$\mathrm{AIC}_{\mathrm{2ML}}$
100	3.7871	3.8723	3.9079	4.0122
1000	0.2859	0.2859	0.2858	0.2858
10000	0.0251	0.0251	0.0251	0.0251

3.2 Simulation 2

In simulation 2, we generated datasets of size $n = 1000$ with the data generating model described in Table 5. We then fitted 486 different copula models, which are based on the possible combinations of the candidate copulae and margins described in Table 6. We repeated this process 100 times and averaged the results. Like in simulation 1, $\mathrm{P}(\mathrm{b}_{0.8} < y)$ was computed from every fitted models. Figure 1 displays the relationship between mean squared error of estimated $\mathrm{P}(\mathrm{b}_{0.8} < y)$ and CIC and $\mathrm{AIC}_{\mathrm{2ML}}$. We can see that both model selection criteria evaluate the models that have lower mean squared error as better models. The difference

between CIC and AIC_{2ML} in this perspective is minimal. This is because the log-likelihood, the element that is shared by both model selection criteria, has much bigger absolute value than the bias correction term.

Table 5: Description of the data generating model used in simulation 2.

	Copula	Margin 1	Margin 2	Margin 3	Margin 4	Margin 5
Data generating model	Gumbel $\theta = 3$	Weibull $\alpha_1 = (1.5, 4)^{\text{T}}$ (shape, scale)	Weibull $\alpha_2 = (2, 3)^{\text{T}}$ (shape, scale)	Gamma $\alpha_3 = (2, 1)^{\text{T}}$ (shape, rate)	Gamma $\alpha_4 = (3, 1)^{\text{T}}$ (shape, rate)	Gamma $\alpha_5 = (4, 2)^{\text{T}}$ (shape, rate)

Table 6: List of the candidate copulae and margins used in simulation 2

	Candidates
Copula	Gumbel, Gaussian
Margin 1	Weibull, Gamma, Log-normal
Margin 2	Weibull, Gamma, Normal
Margin 3	Weibull, Gamma, Log-normal
Margin 4	Weibull, Gamma, Log-normal
Margin 5	Weibull, Gamma, Log-normal

Figure 2 plots the same as Figure 1, but the x-axis is now the bias correction term ($\widetilde{p}^*_\eta$ for CIC and p_η for AIC_{2ML}). The difference between CIC and AIC_{2ML} is now more clear. While p_η (dimension of the parameter vector η) for AIC_{2ML} is fixed at 11 across all models, $\widetilde{p}^*_\eta$ tend to penalize misspecified models more and forms a strong relationship with the mean squared error of estimated $\text{P}(\text{b}_{0.8} < y)$.

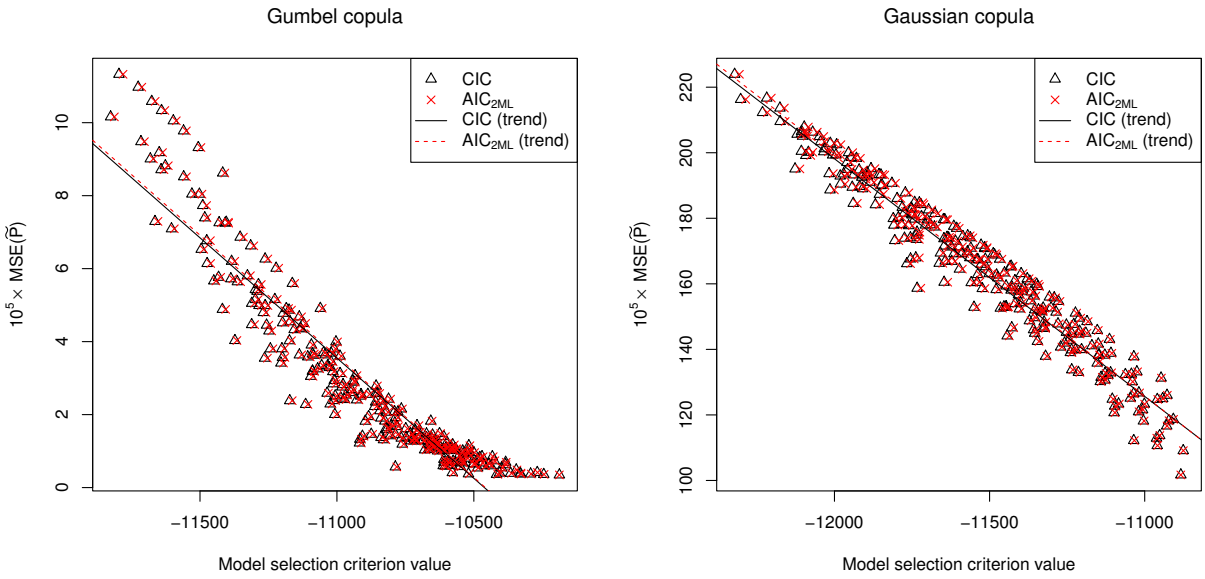

Figure 1: Result from simulation 2. On the x-axis is the value of CIC or $\mathrm{AIC}_{\mathrm{2ML}}$. The 486 different copula models, defined by using the candidate copulae and margins described in Table 6, are fitted to the dataset generated from the data generating model described in Table 5. The y-axis is the mean squared error of $\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$, which indicates the two-stage ML estimated joint probability that each marginal variable has larger value than its 0.8-quantile value, defined by each marginal model. Since different choices of copulae leads to a big difference in model selection score, the result is separately displayed for each copula. The left plot contains 283 models with the Gumbel copula and the right plot contains 283 models with the Gaussian copula. The data generating model had Gumbel copula.

As mentioned in Section 2.4, one can compare CIC to TIC, or $\mathrm{AIC}_{\mathrm{2ML}}$ to $\mathrm{AIC}_{\mathrm{ML}}$, to measure how much one loses in terms of KL divergence by performing two-stage ML estimation instead of ML estimation. Figure 3 shows that $\mathrm{TIC} - \mathrm{CIC}$ or $\mathrm{AIC}_{\mathrm{2ML}} - \mathrm{AIC}_{\mathrm{ML}}$ has a relationship with the loss of MSE of $\mathrm{P}(\mathrm{b}_{0.8} < y)$ caused by two-stage ML estimation. However, this relationship seems weaker when the copula is misspecified (right panel). We tried to identify any sub-pattern in the plot that can cause this, for example by plotting only a subset of models that have specific margins. Yet, we weren't able to detect any sub-pattern.

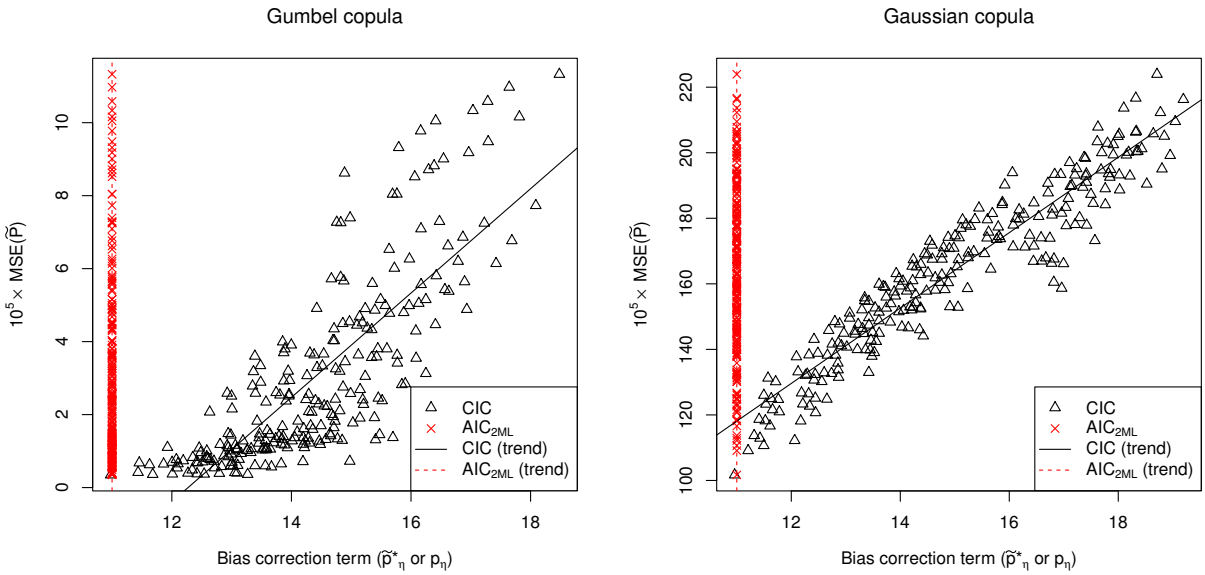

Figure 2: Result from simulation 2. The 486 different copula models, defined by using the candidate copulae and margins described in Table 6, are fitted to the dataset generated from data generating model described in Table 5. The y-axis is the mean squared error of $\widetilde{\mathrm{P}}(\mathrm{b}_{0.8} < y)$. The x-axis is the value of the bias correction terms in model selection criteria ($\widetilde{p}^*_\eta$ for CIC and p_η for $\mathrm{AIC}_{2\mathrm{ML}}$). Like for Figure 1, the result is separately displayed for each copula. The data generating model had Gumbel copula.

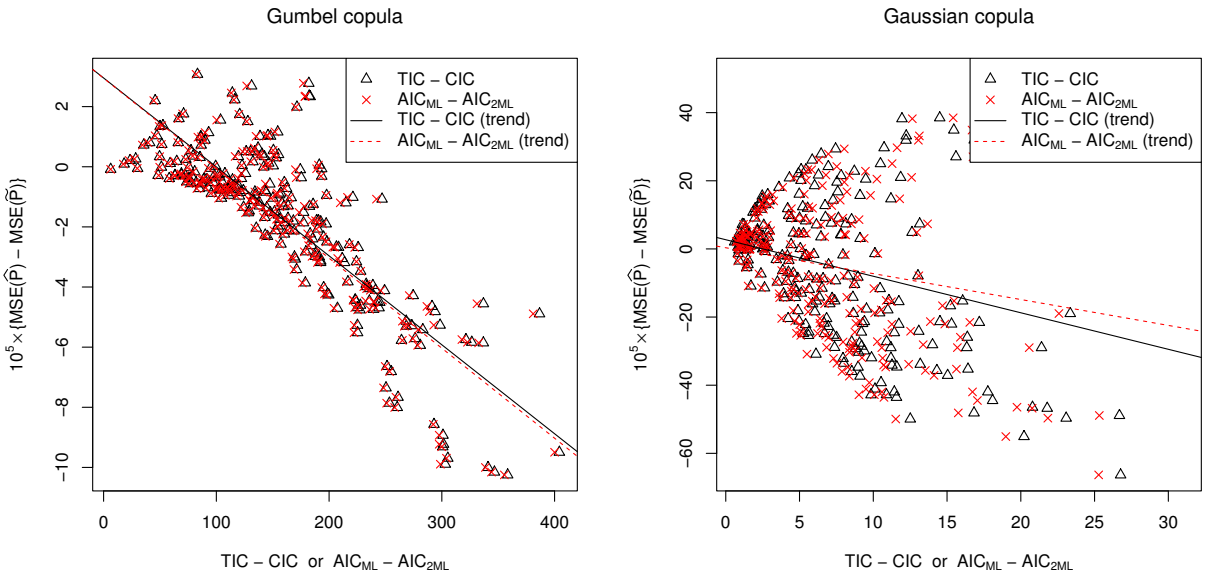

Figure 3: Result from simulation 2. The 486 different copula models, defined by using the candidate copulae and margins described in Table 6, are fitted to the dataset generated from data generating model described in Table 5. The y-axis is the difference between $\mathrm{MSE}(\widehat{\mathrm{P}})$ (MSE of $\mathrm{P}(\mathrm{b}_{0.8} < y)$ estimated by ML) and $\mathrm{MSE}(\widetilde{\mathrm{P}})$ (MSE of $\mathrm{P}(\mathrm{b}_{0.8} < y)$ estimated by two-stage ML). Like in Figure 1 and Figure 2, the result is displayed separately for each copula. The data generating model had Gumbel copula.

4 Conclusions and further research

In this paper, we have developed the copula information criterion (CIC), which is a TIC-like model robust model selection criterion for two-stage ML estimated copulae. When we make an assumption that the parametric candidate model contains the true model, CIC becomes equal to $\mathrm{AIC}_{2\mathrm{ML}}$. This validates the use of $\mathrm{AIC}_{2\mathrm{ML}}$ for the two-stage ML estimated copula models. To our knowledge, this is the first time that $\mathrm{AIC}_{2\mathrm{ML}}$ formula is analytically justified. Further, since both TIC and CIC are estimating the same part of the KL divergence, without the presence of the true model assumption, they are compatible to each other and can be used to check possible disadvantages caused by the two-stage ML estimation. The same can be done by comparing $\mathrm{AIC}_{\mathrm{ML}}$ and $\mathrm{AIC}_{2\mathrm{ML}}$, when one believes in the model.

Regarding the assumption that a candidate model is correct, one can compare $\widetilde{p}^*_\eta$ (bias correction term of CIC) and p_η (bias correction term of $\mathrm{AIC}_{2\mathrm{ML}}$) to check whether the model severely diverges from the data generating model, i.e. as a separate goodness-of-fit test. It may be noted that the job of the CIC is to rank models according to a sensible criterion, and to identify the best ones, but doing well in this ranking is not the same as claiming that the model passes goodness-of-fit tests. In yet other words, the winning model, using the CIC, may still not be a perfect model, perhaps since the list of candidate models has not been the best.

We performed a simulation study. For relatively small sample sizes, CIC outperforms $\mathrm{AIC}_{2\mathrm{ML}}$ in terms

of prediction performance from the selected models. When the sample size is large, the log-likelihood term grows much faster than the bias correction term and the difference between CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ is minimal. Naturally, for large n, the best models are those with high values of the maximized (two-stage) log-likelihoods, which means richly parametrised models.

From our simulation, cf. Figure 2, we can see that $\widetilde{p}^*_\eta$ alone has a strong correlation with the prediction performance (measured in MSE). So, one can consider to use $\widetilde{p}^*_\eta$ (without the log-likelihood part) to judge the model. In addition, $\mathrm{TIC} - \mathrm{CIC}$ and $\mathrm{AIC}_{\mathrm{ML}} - \mathrm{AIC}_{\mathrm{2ML}}$ turn out to have high correlation with the loss of prediction performance (measured as difference in MSE) caused by the switch from ML estimation to two-stage ML estimation.

The results from the simulation study hold mostly both when the copula is correctly specified and misspecified. Yet, in Figure 3, although the overall tendency is similar, we observe that the result from the misspecified Gaussian copula seems to consist of two different sub-patterns. However, we were not able to find a possible cause of this.

Because of the large number of possible models in high dimensional setting, the number of situations that we could examine, was limited. (In case of a 5-dimensional copula model with 2 candidate copulae and 3 candidate for each margin, there are 486 models that we have to test, and each model has to be fitted by 2 different estimation schemes.) Another problem was that we could not try all copulae and margins on the simulated data since fitting a heavily misspecified copula and margins would cause numerical problems. A further large-scale simulation study that examines the behavior of different types of copulae in variety of situations would be fruitful.

Furthermore, CIC is computationally expensive mainly because $\widetilde{K}^\circ_\alpha$, $\widetilde{K}_{\alpha,\theta}$ and $\widetilde{K}_\theta$ require score functions for every data point separately. For example, for $\widetilde{\mathcal{I}}^{-1}_\alpha$ and $\widetilde{\mathcal{I}}_\theta$, one can avoid this by swapping the order of summation and differentiation, but for $\widetilde{K}^\circ_\alpha$, $\widetilde{K}_{\alpha,\theta}$ and $\widetilde{K}_\theta$, this is not possible. A numerical technique that can make CIC less computationally extensive would be appreciated.

Although CIC performs decently well in selecting a good model that fits the data best in terms of KL divergence, there are situations where one is interested in a model that is suitable for specific tasks. The task of interest could be for example estimating tail probabilities, the mean, or the median. A model selection criterion for copula models under the two-stage ML scheme that can take this into account would be useful.

Acknowledgments

The authors would like to thank Ingrid Hobæk Haff and Steffen Grønneberg for their valuable comments and fruitful discussions. The authors 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

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, 716–723.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*. Springer, pp. 199–213.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*, vol. 330. Cambridge University Press Cambridge.

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.

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

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

Jullum, M. & Hjort, N. L. (2017). Parametric or nonparametric: the fic approach. *Statistica Sinica* .

Ko, V. & Hjort, N. L. (2018). Model robust inference for copulae via two-stage maximum likelihood estimation. Submitted to Journal of Multivariate Analysis.

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

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

Palaro, H. & Hotta, L. (2006). Using conditional copula to estimate value at risk. *Journal of Data Science* **4**, 93–115.

Patton, A. J. (2002). *Applications of copula theory in financial econometrics*. Ph.D. thesis, University of California, San Diego.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International economic review* **47**, 527–556.

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

Takeuchi, K. (1976). The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science* **153**, 12–18.

Appendix

	Model no.	TIC	AIC_{ML}	CIC	AIC_{2ML}	$\widehat{p}^*_{\eta,\text{TIC}}$	$\widetilde{p}^*_{\eta}$	p_η	$10^5 \cdot \text{MSE}(\widehat{\text{P}})$	$10^5 \cdot \text{MSE}(\widetilde{\text{P}})$	Copula	Margin 1	Margin 2
n = 100	1	-609.92	-610.10	-610.77	-610.93	4.91	4.92	5	2.6695	2.8444	Gumbel	Weibull	Gamma
	2	-613.04	-612.93	-614.75	-614.26	5.05	5.24	5	2.6092	2.9347	Gumbel	Weibull	Weibull
	4	-613.64	-613.54	-615.04	-614.93	5.05	5.05	5	2.7188	2.8276	Gumbel	Gamma	Gamma
	11	-621.56	-620.75	-621.61	-621.12	5.41	5.25	5	27.3463	22.8859	Gaussian	Weibull	Weibull
	10	-622.25	-621.41	-622.33	-621.69	5.42	5.32	5	29.5786	26.3657	Gaussian	Weibull	Gamma
	5	-619.30	-618.63	-623.51	-622.01	5.34	5.75	5	2.7269	3.3693	Gumbel	Gamma	Weibull
	6	-624.39	-623.56	-627.26	-626.22	5.41	5.52	5	2.5957	2.7652	Gumbel	Gamma	Log-normal
	13	-628.24	-627.03	-628.26	-627.05	5.60	5.60	5	31.3167	30.7712	Gaussian	Gamma	Gamma
	14	-629.62	-628.39	-629.99	-628.63	5.62	5.68	5	26.9486	28.9915	Gaussian	Gamma	Weibull
	3	-627.54	-626.46	-634.34	-632.53	5.54	5.90	5	2.6429	3.8486	Gumbel	Weibull	Log-normal
	9	-640.53	-638.38	-643.09	-640.44	6.07	6.32	5	2.7582	2.7121	Gumbel	Log-normal	Log-normal
	12	-646.70	-643.93	-646.92	-644.09	6.39	6.41	5	41.5681	43.6130	Gaussian	Weibull	Log-normal
	15	-647.74	-644.76	-647.75	-644.77	6.49	6.49	5	43.4360	43.2919	Gaussian	Gamma	Log-normal
	7	-647.85	-645.24	-654.37	-651.88	6.31	6.24	5	3.0737	6.1935	Gumbel	Log-normal	Gamma
	8	-654.09	-651.41	-665.48	-661.68	6.34	6.90	5	2.5910	9.1325	Gumbel	Log-normal	Weibull
	16	-666.52	-662.25	-666.53	-662.25	7.14	7.14	5	57.9316	58.2482	Gaussian	Log-normal	Gamma
	17	-671.15	-666.76	-672.01	-667.17	7.19	7.42	5	51.5845	61.5621	Gaussian	Log-normal	Weibull
	18	-674.16	-668.48	-674.16	-668.48	7.84	7.84	5	56.7535	56.7548	Gaussian	Log-normal	Log-normal

Table 7: Result of simulation 1 with $n = 100$. The simulation was repeated 1000 times and the results were averaged.

	Model no.	TIC	AIC_{ML}	CIC	AIC_{2ML}	$\widehat{p}^*_{\eta,\text{TIC}}$	$\widetilde{p}^*_{\eta}$	p_η	$10^5 \cdot \text{MSE}(\widehat{\text{P}})$	$10^5 \cdot \text{MSE}(\widetilde{\text{P}})$	Copula	Margin 1	Margin 2
n = 1000	1	-6060.83	-6060.84	-6061.78	-6061.79	4.99	4.99	5	0.2861	0.2996	Gumbel	Weibull	Gamma
	2	-6093.59	-6092.60	-6102.19	-6101.23	5.50	5.48	5	0.2981	0.3218	Gumbel	Weibull	Weibull
	4	-6095.45	-6095.16	-6105.59	-6105.31	5.15	5.14	5	0.3009	0.3331	Gumbel	Gamma	Gamma
	11	-6171.40	-6170.13	-6174.20	-6173.32	5.63	5.44	5	23.3046	18.7508	Gaussian	Weibull	Weibull
	10	-6178.38	-6177.18	-6179.74	-6178.74	5.60	5.50	5	25.5380	22.1326	Gaussian	Weibull	Gamma
	5	-6151.67	-6150.39	-6183.69	-6181.42	5.64	6.13	5	0.3913	0.7139	Gumbel	Gamma	Weibull
	6	-6204.66	-6203.44	-6230.19	-6228.67	5.61	5.76	5	0.2753	0.3321	Gumbel	Gamma	Log-normal
	13	-6235.93	-6234.28	-6236.12	-6234.49	5.82	5.82	5	27.2102	26.6318	Gaussian	Gamma	Gamma
	14	-6250.52	-6248.66	-6252.09	-6250.10	5.93	6.00	5	22.6996	24.9377	Gaussian	Gamma	Weibull
	3	-6237.75	-6236.14	-6300.84	-6298.28	5.80	6.28	5	0.3561	1.3135	Gumbel	Weibull	Log-normal
	9	-6355.65	-6352.79	-6369.91	-6366.28	6.43	6.81	5	0.5153	0.4199	Gumbel	Log-normal	Log-normal
	12	-6425.80	-6421.62	-6426.35	-6422.09	7.09	7.13	5	38.1316	40.5651	Gaussian	Weibull	Log-normal
	15	-6432.74	-6428.42	-6432.78	-6428.47	7.16	7.16	5	39.9678	39.8287	Gaussian	Gamma	Log-normal
	7	-6427.95	-6424.60	-6495.95	-6492.60	6.68	6.67	5	0.6189	3.5215	Gumbel	Log-normal	Gamma
	8	-6490.98	-6487.46	-6597.27	-6592.10	6.76	7.58	5	0.2722	6.3816	Gumbel	Log-normal	Weibull
	16	-6612.77	-6606.21	-6612.80	-6606.24	8.28	8.28	5	55.3695	55.7180	Gaussian	Log-normal	Gamma
	17	-6660.51	-6653.56	-6664.39	-6656.88	8.47	8.76	5	48.5595	59.4736	Gaussian	Log-normal	Weibull
	18	-6686.70	-6678.29	-6686.70	-6678.29	9.20	9.20	5	52.9131	52.9139	Gaussian	Log-normal	Log-normal

Table 8: Result of simulation 1 with $n = 1000$. The simulation was repeated 1000 times and the results were averaged.

	Model no.	TIC	AIC_{ML}	CIC	AIC_{2ML}	$\widehat{p}^*_{\eta,\text{TIC}}$	$\widetilde{p}^*_{\eta}$	p_η	$10^5 \cdot \text{MSE}(\widehat{\text{P}})$	$10^5 \cdot \text{MSE}(\widetilde{\text{P}})$	Copula	Margin 1	Margin 2
n = 10000	1	-60528.25	-60528.25	-60529.19	-60529.18	5.00	5.00	5	0.0251	0.0268	Gumbel	Weibull	Gamma
	2	-60848.99	-60847.94	-60929.30	-60928.31	5.52	5.49	5	0.0381	0.0433	Gumbel	Weibull	Weibull
	4	-60869.96	-60869.66	-60966.13	-60965.85	5.15	5.14	5	0.0419	0.0699	Gumbel	Gamma	Gamma
	11	-61624.69	-61623.39	-61655.65	-61654.74	5.65	5.45	5	22.8801	18.3281	Gaussian	Weibull	Weibull
	10	-61696.56	-61695.34	-61710.64	-61709.62	5.61	5.51	5	25.1114	21.7084	Gaussian	Weibull	Gamma
	5	-61427.65	-61426.32	-61734.40	-61732.07	5.67	6.17	5	0.1282	0.4364	Gumbel	Gamma	Weibull
	6	-61969.26	-61967.99	-62223.80	-62222.22	5.63	5.79	5	0.0292	0.0867	Gumbel	Gamma	Log-normal
	13	-62265.03	-62263.36	-62266.99	-62265.33	5.83	5.83	5	26.7925	26.2132	Gaussian	Gamma	Gamma
	14	-62409.19	-62407.28	-62422.48	-62420.44	5.95	6.02	5	22.2643	24.5218	Gaussian	Gamma	Weibull
	3	-62301.64	-62299.98	-62929.03	-62926.39	5.83	6.32	5	0.1043	1.0768	Gumbel	Weibull	Log-normal
	9	-63458.15	-63455.22	-63582.75	-63579.03	6.47	6.86	5	0.2704	0.1838	Gumbel	Log-normal	Log-normal
	12	-64168.16	-64163.79	-64171.83	-64167.38	7.18	7.23	5	37.7631	40.2916	Gaussian	Weibull	Log-normal
	15	-64230.67	-64226.18	-64231.09	-64226.61	7.24	7.24	5	39.6239	39.4905	Gaussian	Gamma	Log-normal
	7	-64177.36	-64173.93	-64854.87	-64851.44	6.71	6.71	5	0.3752	3.2669	Gumbel	Log-normal	Gamma
	8	-64807.87	-64804.26	-65853.68	-65848.39	6.80	7.64	5	0.0245	6.1275	Gumbel	Log-normal	Weibull
	16	-65997.87	-65991.08	-65998.13	-65991.33	8.40	8.40	5	54.9718	55.3255	Gaussian	Log-normal	Gamma
	17	-66475.00	-66467.78	-66507.88	-66500.08	8.61	8.90	5	48.1140	59.1384	Gaussian	Log-normal	Weibull
	18	-66730.29	-66721.58	-66730.29	-66721.58	9.36	9.36	5	52.3035	52.3041	Gaussian	Log-normal	Log-normal

Table 9: Result of simulation 1 with $n = 10000$. The simulation was repeated 1000 times and the results were averaged.