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

Vinnie Ko[a,*], Nils Lid Hjort[a]

[a] *Department of Mathematics, University of Oslo*
PB 1053, Blindern, NO-0316 Oslo, Norway

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 the existing model robust inference of two-stage maximum likelihood estimation, copula information criterion (CIC) for model selection is developed. 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 the additional assumption that a candidate model is correctly specified is made, then CIC for that model simplifies to AIC. Additionally, CIC can easily be extended to the conditional copula setup where covariates are parametrically linked to the copula model. As a numerical illustration, simulation studies were performed to show that the better model according to CIC also has better prediction performance in general. The result also shows that the bias correction term of CIC penalizes the misspecified model more heavily. This bias correction term 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. Although this behavior of the bias correction part is an important conceptual advance of CIC, this is not sufficient to make CIC outperform AIC in practice. This is because each misspecified model has the bias correction term and they grow at different speeds, depending on the model. The difference between CIC and AIC becomes minimal as sample size grows because the log-likelihood part outgrows the bias correction part.

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

*Corresponding author. Email address: `vinniebk@math.uio.no`

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 a 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 $\mathrm{AIC}_{\mathrm{ML}}$ (Akaike, 1974) or TIC (also known as model-robust $\mathrm{AIC}_{\mathrm{ML}}$) (Takeuchi, 1976) as model selection criterion and select the model with the best score. (Note that we denote the AIC under ML estimation as $\mathrm{AIC}_{\mathrm{ML}}$ to distinguish it from the two-stage ML based $\mathrm{AIC}_{2\mathrm{ML}}$, 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 and Favre, 2007). Another often used model selection strategy for the two-stage ML is that one first evaluates candidates of each marginal distribution with $\mathrm{AIC}_{\mathrm{ML}}$ and consequently chooses the best distribution for each margin. Once the margins are chosen, one fits different copulas and evaluates the copula part with $\mathrm{AIC}_{\mathrm{ML}}$. 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

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

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 the CIC from Grønneberg and Hjort (2014), which is designed for copulas estimated with pseudo maximum likelihood (PML). In the PML framework, margins are estimated empirically, i.e. nonparametrically, while two-stage ML assumes parametric forms of margins.

Our technical setting, identical to Ko and Hjort (2019), 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 regularity conditions from Ko and Hjort (2019), $C(\cdot, \theta)$ 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.

Note that the theories of CIC and $\mathrm{AIC}_{\mathrm{2ML}}$, which we will develop in Section 2, should hold for both continuous and discrete variables since the two-stage ML estimation keeps its properties in case of discrete

variables. In this paper, however, we only consider the case of continuous variables for simpler notation and exposition.

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 $\mathrm{AIC}_{\mathrm{ML}}$. In Section 2.2, we derive and define our copula information criterion. In Section 2.3, we prove that the $\mathrm{AIC}_{\mathrm{2ML}}$ formula holds under the two-stage ML estimation. In Section 2.4, we summarize the relationship between TIC, CIC, $\mathrm{AIC}_{\mathrm{ML}}$ and $\mathrm{AIC}_{\mathrm{2ML}}$. 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 and 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 $Q(\widehat{\eta})$:

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

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

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 and 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 $\widetilde{Q}(\widetilde{\eta})$:

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

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

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

where

$$
\begin{aligned}
K_{\alpha_j} &= \mathrm{E}_{G_j}\left[U_{\alpha_j}(y_j,\alpha_{0,j})\,U_{\alpha_j}(y_j,\alpha_{0,j})^{\mathrm{T}}\right], \\
U_{\alpha_j}(y_j,\alpha_j) &= \frac{\partial \log f_j(y_j,\alpha_j)}{\partial \alpha_j}
\end{aligned}
$$

and

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

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

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

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

$$\begin{aligned} Q_c(\eta_0) &= \int g(y)\log c\left(F_1(y_1,\alpha_{0,1}),\cdots,F_d(y_d,\alpha_{0,d}),\theta_0\right)\mathrm{d}y, \\ Z_i &= \log c\left(F_1(y_{i,1},\alpha_{0,1}),\cdots,F_d(y_{i,d},\alpha_{0,d}),\theta_0\right) - Q_c(\eta_0), \end{aligned}$$

and

$$A_\eta = \sqrt{n}\left(\widetilde{\eta}-\eta_0\right) = \sqrt{n}\begin{pmatrix}\mathcal{I}_\alpha & 0\\ \mathcal{I}_{\alpha,\theta}^{\mathrm{T}} & \mathcal{I}_\theta\end{pmatrix}^{-1}\begin{pmatrix}U_{n,\alpha}(\alpha_0)\\ U_{n,\theta}(\alpha_0,\theta_0)\end{pmatrix} + \begin{pmatrix}o_p(1)\\ o_p(1)\end{pmatrix},$$

which stems from Proposition 1 in Ko and Hjort (2019), and furthermore

$$\begin{aligned} U_{n,\eta}(\eta) &= \frac{1}{n}\sum_{i=1}^{n}U_\eta(y_i,\eta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log c\left(F_1(y_{i,1},\alpha_1),\cdots,F_d(y_{i,d},\alpha_d),\theta\right)}{\partial\eta}, \\ H_{n,\eta}(\eta) &= \frac{1}{n}\sum_{i=1}^{n}H_\eta(y_i,\eta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\log c\left(F_1(y_{i,1},\alpha_1),\cdots,F_d(y_{i,d},\alpha_d),\theta\right)}{\partial\eta\partial\eta^{\mathrm{T}}}, \end{aligned}$$

and

$$\mathcal{I}_\eta^* = -\mathrm{E}_G\left[H_\eta(y,\eta_0)\right] = -\int g(y)H_\eta(y,\eta_0)\,\mathrm{d}y,$$

which, incidentally, should not be confused with $\mathcal{I}_\eta$ in Proposition 1 of Ko and Hjort (2019). 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

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

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 and Hjort (2019).)

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

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

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

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

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

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

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

By defining $\widetilde{p}_{\alpha_j}^* = \operatorname{tr}(\widetilde{\mathcal{I}}_{\alpha_j}^{-1}\widetilde{K}_{\alpha_j})$, $\widetilde{p}_\theta^* = \operatorname{tr}(\widetilde{\mathcal{I}}_\eta^{-1}\widetilde{K}_\eta^\circ)$, $\widetilde{p}_\eta^* = \sum_{j=1}^d \widetilde{p}_{\alpha_j}^* + \widetilde{p}_\theta^*$ 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 Lemmas 3 and 5 from Ko and Hjort (2019) gives

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

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 AIC_{2ML} for the two-stage ML estimated copula models

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

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 offers 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 loses 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 would in that situation be comparing 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 $\eta = (\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2}, \theta)^{\mathrm{T}}$. The ingredients for the CIC can then be written down by using the block matrix form, in line with Ko and Hjort (2019):

$$K_\eta = \begin{pmatrix} K_{\alpha_1} & K_{\alpha_1,\alpha_2} & K_{\alpha_1,\theta} \\ K_{\alpha_2,\alpha_1} & K_{\alpha_2} & K_{\alpha_2,\theta} \\ K^{\mathrm{T}}_{\alpha_1,\theta} & K^{\mathrm{T}}_{\alpha_2,\theta} & K_\theta \end{pmatrix}, \ \mathcal{I}_\eta = \begin{pmatrix} \mathcal{I}_{\alpha_1} & 0 & 0 \\ 0 & \mathcal{I}_{\alpha_2} & 0 \\ \mathcal{I}^{\mathrm{T}}_{\alpha_1,\theta} & \mathcal{I}^{\mathrm{T}}_{\alpha_2,\theta} & \mathcal{I}_\theta \end{pmatrix}, \ K^\circ_\eta = \begin{pmatrix} K^\circ_{\alpha_1} & K^\circ_{\alpha_1,\alpha_2} & K_{\alpha_1,\theta} \\ K^\circ_{\alpha_2,\alpha_1} & K^\circ_{\alpha_2} & K_{\alpha_2,\theta} \\ \left(K^*_{\alpha_1,\theta}\right)^{\mathrm{T}} & \left(K^*_{\alpha_2,\theta}\right)^{\mathrm{T}} & K_\theta \end{pmatrix}.$$

We consequently have

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

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, $\mathcal{I}_\eta$ and K°_η accordingly. For details about conditional copula regression, see Patton (2006), Acar et al. (2013) and Palaro and Hotta (2006).

3. Simulation studies

To study the behavior of CIC, we have performed a set of simulation studies. Simulation 1 studies the similarity between CIC and $\mathrm{AIC}_{\mathrm{2ML}}$ and whether the models with higher CIC or $\mathrm{AIC}_{\mathrm{2ML}}$ scores are indeed 'better' models in practice. Simulation 2 focuses on the difference in the behavior of CIC and $\mathrm{AIC}_{\mathrm{2ML}}$. In particular, we analyze whether CIC outperforms $\mathrm{AIC}_{\mathrm{2ML}}$ in practice.

3.1. Simulation 1

In simulation 1, we generated a dataset of size $n = 1000$ with the model described in Table 2. We then, with this dataset, fitted 324 different copula models, which are based on the possible combinations of the candidate copulas and margins described in Table 3. Further, $\mathrm{P}(q_{0.7} < Y)$ was computed from every fitted model. Here, $q_{0.7}$ is a vector that contains the 0.7-quantile value of each margin according to the true model. i.e. $q_{0.7} = \left(G_1^{-1}(0.7), \cdots, G_4^{-1}(0.7)\right)$. We repeated this process 1000 times and averaged the results.

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

	Copula	Margin 1	Margin 2	Margin 3	Margin 4
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)

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

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

Figure 1 displays the relationship between mean squared error of estimated $\mathrm{P}(q_{0.7} < Y)$ and CIC and $\mathrm{AIC}_{2\mathrm{ML}}$. 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 $\mathrm{AIC}_{2\mathrm{ML}}$ 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 and dominate the criteria. The difference between CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ (the misalignment between the black triangles and the red crosses in the figure) is caused by the difference in their bias correction part alone. Further, the model ranks determined by CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ are highly similar and both model selection criteria successfully picked the true data generating model as the best model in all 1000 repetitions. In short, Figure 1 illustrates that the KL-divergence based model selection criteria that we introduced for two-stage ML estimation (e.g. CIC and $\mathrm{AIC}_{2\mathrm{ML}}$) show the desired behavior in practice.

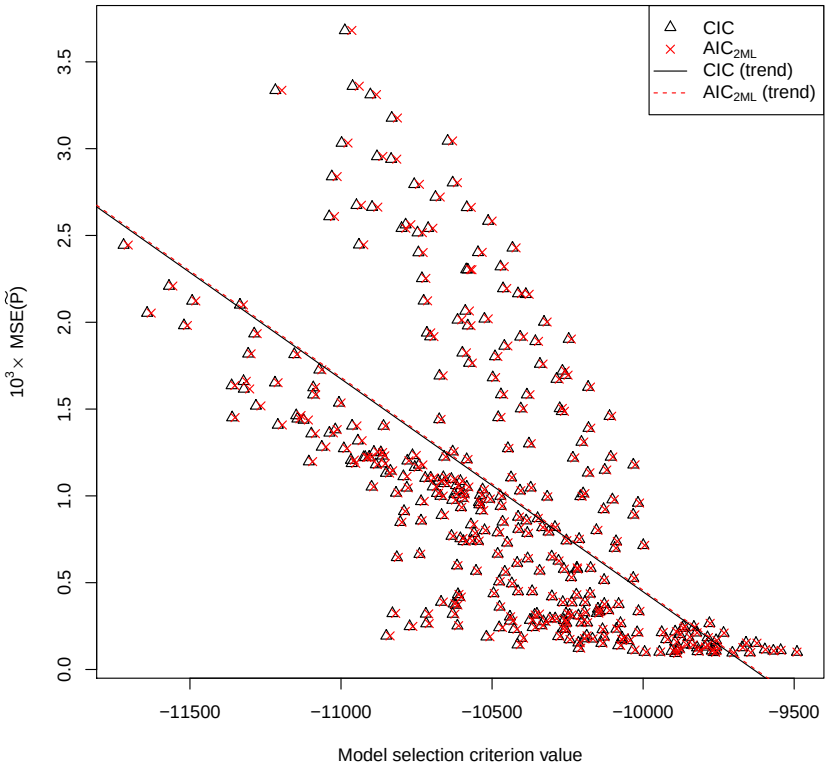

Figure 1: Result from simulation 1. On the x-axis is the value of CIC or $\mathrm{AIC}_{2\mathrm{ML}}$. The 324 different copula models, defined by using the candidate copulas and margins described in Table 3, are fitted to the dataset generated from the data generating model described in Table 2. The y-axis is the mean squared error of $\widetilde{\mathrm{P}}(q_{0.7} < Y)$, which indicates the two-stage ML estimated joint probability that each marginal variable has larger value than its 0.7-quantile value, defined by the data generating model.

Figure 2 uses the same mean squared error as in Figure 1, but the x-axis is now the bias correction term ($\widetilde{p}^*_\eta$ for CIC and p_η for $\mathrm{AIC}_{2\mathrm{ML}}$). The difference between CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ is now more clear. While p_η (dimension of the parameter vector η) for $\mathrm{AIC}_{2\mathrm{ML}}$ is either 9 or 14 depending on the total number of parameters in the model, $\widetilde{p}^*_\eta$ tends to penalize misspecified models more and forms a strong relationship with the mean squared error of estimated $\mathrm{P}(q_{0.7} < Y)$.

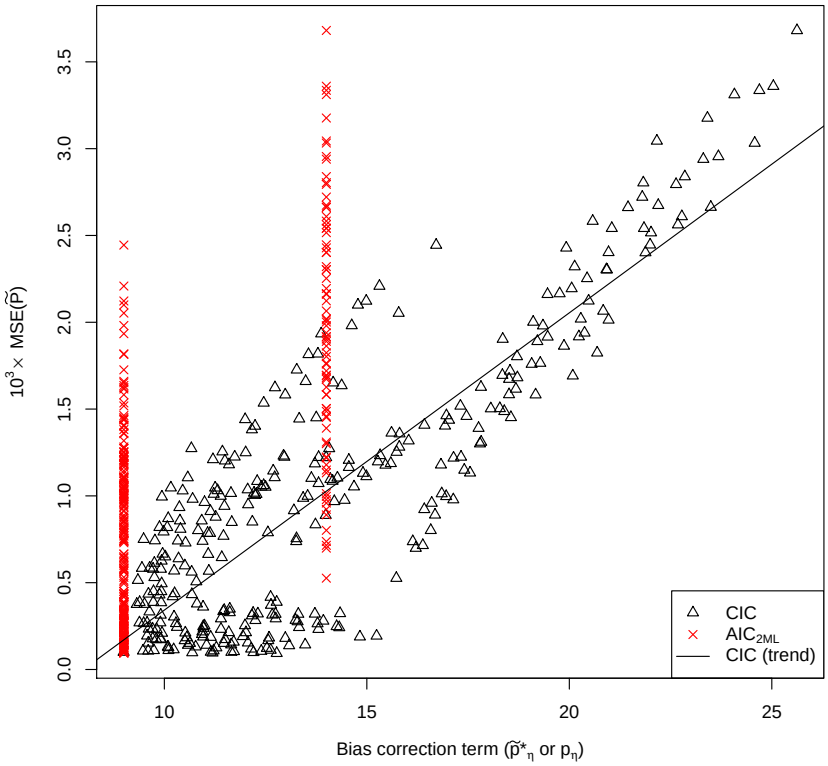

Figure 2: Result from simulation 1. The 324 different copula models, defined by using the candidate copulas and margins described in Table 3, are fitted to the dataset generated from the data generating model described in Table 2. The y-axis is the mean squared error of $\widetilde{\mathrm{P}}(q_{0.7} < Y)$. The x-axis is the value of the bias correction terms in model selection criteria ($\widetilde{p}^*_\eta$ for CIC and p_η for $\mathrm{AIC}_{2\mathrm{ML}}$).

3.2. Simulation 2

In simulation 1, we observed that CIC penalizes misspecified models more in its bias correction part, which is a feature that $\mathrm{AIC}_{2\mathrm{ML}}$ does not have. Does it imply that CIC should outperform $\mathrm{AIC}_{2\mathrm{ML}}$ when they disagree on the model? To find an answer to this question and to study the difference in the behavior of $\mathrm{AIC}_{2\mathrm{ML}}$ and CIC in detail, we performed simulation 2. Although we tried different combinations of copulas and margins, they all led to the same conclusion. Hence, we only display some chosen representative results.

Like we saw in Section 3.1, the difference between CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ is fairly small because of the dominance of the log-likelihood part. This means that there are not that many situations in practice where CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ disagree, especially when the sample size is large. So, we have to look into situations where the difference between two competing models is relatively small. We mimicked such a situation by letting the data come from a mixture of the two models. Another advantage of this simulation design is that it allows us

to study the situation where there is no 'true model' on the candidate model list (when the 'contamination' parameter δ is away from its boundaries).

In simulation 2, we generated datasets of 4 different sizes ($n = 100, 200, 500, 1000$) from the two models described in Table 4. The data generating algorithm generated $(1-\delta)\cdot 100\%$ of data points from model 1 and $\delta \cdot 100\%$ of data points from model 2. So, δ can be seen as the degree of 'contamination' to model 1 by model 2. Then, we fitted both model 1 and model 2 to these data by using two-stage ML. This process was repeated 10000 times. We used $\delta = 0, 0.05, \cdots, 0.95, 1$.

Note that the copulas used in model 1 and 2 are easily distinguishable with reasonable sample size (e.g. 200), since Clayton copula has stronger lower tail dependence than Frank copula. This enables us to observe different degrees of misspecification with varying δ values.

Table 4: Description of the models used in simulation 2.

	Dependence measure	Copula	Margin 1	Margin 2
Model 1	Kendall's $\tau = 0.56$ Spearman's $\rho = 0.76$	Frank $\theta = 7$	Log-normal $\alpha_1 = (0.9, 0.8)^{\mathrm{T}}$ (mean, SD)	Log-normal $\alpha_2 = (0.3, 0.8)^{\mathrm{T}}$ (mean, SD)
Model 2	Kendall's $\tau = 0.60$ Spearman's $\rho = 0.79$	Clayton $\theta = 3$	Weibull $\alpha_1 = (1.5, 4)^{\mathrm{T}}$ (shape, scale)	Gamma $\alpha_2 = (2, 1)^{\mathrm{T}}$ (shape, rate)

From Figure 3, we can first confirm our theoretical finding from Section 2.3 that CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ become equal when the model is correctly specified. (For model 1, it is when $\delta = 0$ and for model 2, it is when $\delta = 1$.) As the degree of model misspecification grows (for model 1, it is when δ moves towards 1 and for model 2, it is when δ moves towards 0), the two model selection criteria diverge from each other. To be more specific, CIC always has smaller value than $\mathrm{AIC}_{2\mathrm{ML}}$. Since the only difference between the two model selection criteria is the bias correction part and the bias correction part of $\mathrm{AIC}_{2\mathrm{ML}}$ is constant for all values of δ, this can only imply that $\widetilde{p}^*_\eta$ gets larger as the degree of misspecification grows. The divergence between CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ gets smaller and smaller as the sample size (n) grows. This is because the size of the log-likelihood part of grows linearly with the sample size while the bias correction part does not.

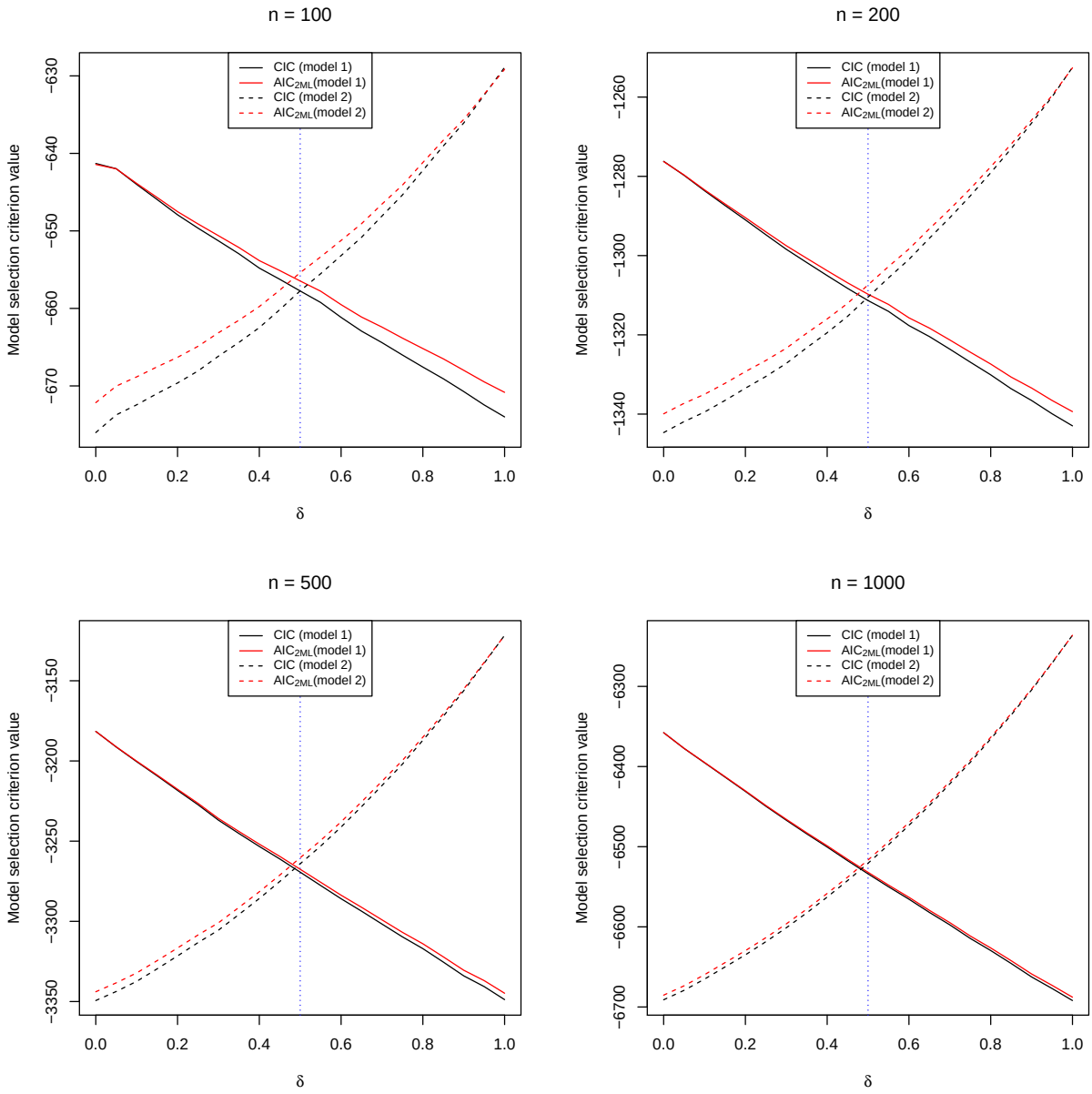

Figure 3: Result of simulation 2. The y-axis is the value of CIC or $\mathrm{AIC}_{2\mathrm{ML}}$. The x-axis is δ, the proportion of data points generated by model 2 described in Table 4. The remaining $(1-\delta)\cdot 100\%$ data points were generated by model 1. The simulation was repeated 10000 times and the results were averaged. The blue dashed vertical line is $\delta = 0.5$.

Figure 4 shows how often each model is chosen as the best model by each model selection criterion. One could intuitively think that CIC should choose the 'more correct' model (which is model 1 for $\delta < 0.5$ and model 2 for $\delta > 0.5$) more often than $\mathrm{AIC}_{2\mathrm{ML}}$ since CIC is a model robust model selection criterion and the misspecified model gets penalized by $\widetilde{p}^*_\eta$. However, the result shows that this is not the case. Figure 4 shows

that CIC always prefers model 1 above model 2, both when $\delta < 0.5$ and $\delta > 0.5$. We tried many different copulas and margins and CIC always preferred a certain model across all range of δ. Which model was preferred, depended on the combinations of copula and margins. Figure 5, where we plot the bias correction term against δ, explains this phenomenon. At a first glance, we can observe that $\widetilde{p}^*_\eta$ and p_η are equal when the model is correctly specified and that $\widetilde{p}^*_\eta$ grows as the degree of model misspecification increases. The key thing is that this growth happens in both models and that each model has its own speed of growth for $\widetilde{p}^*_\eta$. Back to $\mathrm{AIC}_{2\mathrm{ML}}$, since both model 1 and 2 have same number of parameters, the model with higher log-likelihood value becomes the 'winning model'. In other words, whether $\mathrm{AIC}_{2\mathrm{ML}}$ choose model 1 or 2 as winning model, depends on

$$\mathrm{sgn}\left(\widetilde{\ell}_{n,\mathrm{M1}} - \widetilde{\ell}_{n,\mathrm{M2}}\right),$$

where $\widetilde{\ell}_{n,\mathrm{M}j}$ indicates the under two-stage ML maximized log-likelihood value of model j. So, CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ disagree on the winning model if and only if

$$\mathrm{sgn}\left(\widetilde{\ell}_{n,\mathrm{M1}} - \widetilde{\ell}_{n,\mathrm{M2}}\right) \neq \mathrm{sgn}\left(\widetilde{\ell}_{n,\mathrm{M1}} - \widetilde{\ell}_{n,\mathrm{M2}} - (\widetilde{p}^*_{\eta,\mathrm{M1}} - \widetilde{p}^*_{\eta,\mathrm{M2}})\right), \tag{5}$$

where $\widetilde{p}^*_{\eta,\mathrm{M}j}$ indicates the bias correction term of model j, as in (3). Thus, if $\widetilde{p}^*_{\eta,\mathrm{M1}} < \widetilde{p}^*_{\eta,\mathrm{M2}}$, CIC chooses model 2 more often as winning model (relatively to $\mathrm{AIC}_{2\mathrm{ML}}$). If the inequality holds in opposite direction, CIC chooses model 1 more often as winning model (again, relatively to $\mathrm{AIC}_{2\mathrm{ML}}$). From Figure 5, we see that both $\widetilde{p}^*_{\eta,\mathrm{M1}}$ and $\widetilde{p}^*_{\eta,\mathrm{M2}}$ have value close to 5, when the model is correctly specified, but as the misspecification appears, $\widetilde{p}^*_{\eta,\mathrm{M2}}$ grows faster than $\widetilde{p}^*_{\eta,\mathrm{M1}}$. (This is also visible by the fact that the point where $\widetilde{p}^*_{\eta,\mathrm{M1}}$-curve and $\widetilde{p}^*_{\eta,\mathrm{M2}}$-curve cross each other lies far more right to $\delta = 0.5$.) Combining this growth speed difference with (5), it is logical that CIC chooses model 1 more often as winning model compared to $\mathrm{AIC}_{2\mathrm{ML}}$.

To sum up, CIC, compared to $\mathrm{AIC}_{2\mathrm{ML}}$, chooses the model with smaller value of $\widetilde{p}^*_\eta$ more often as winning model, which in practice is not necessarily more correct model as we noticed in Figure 4.

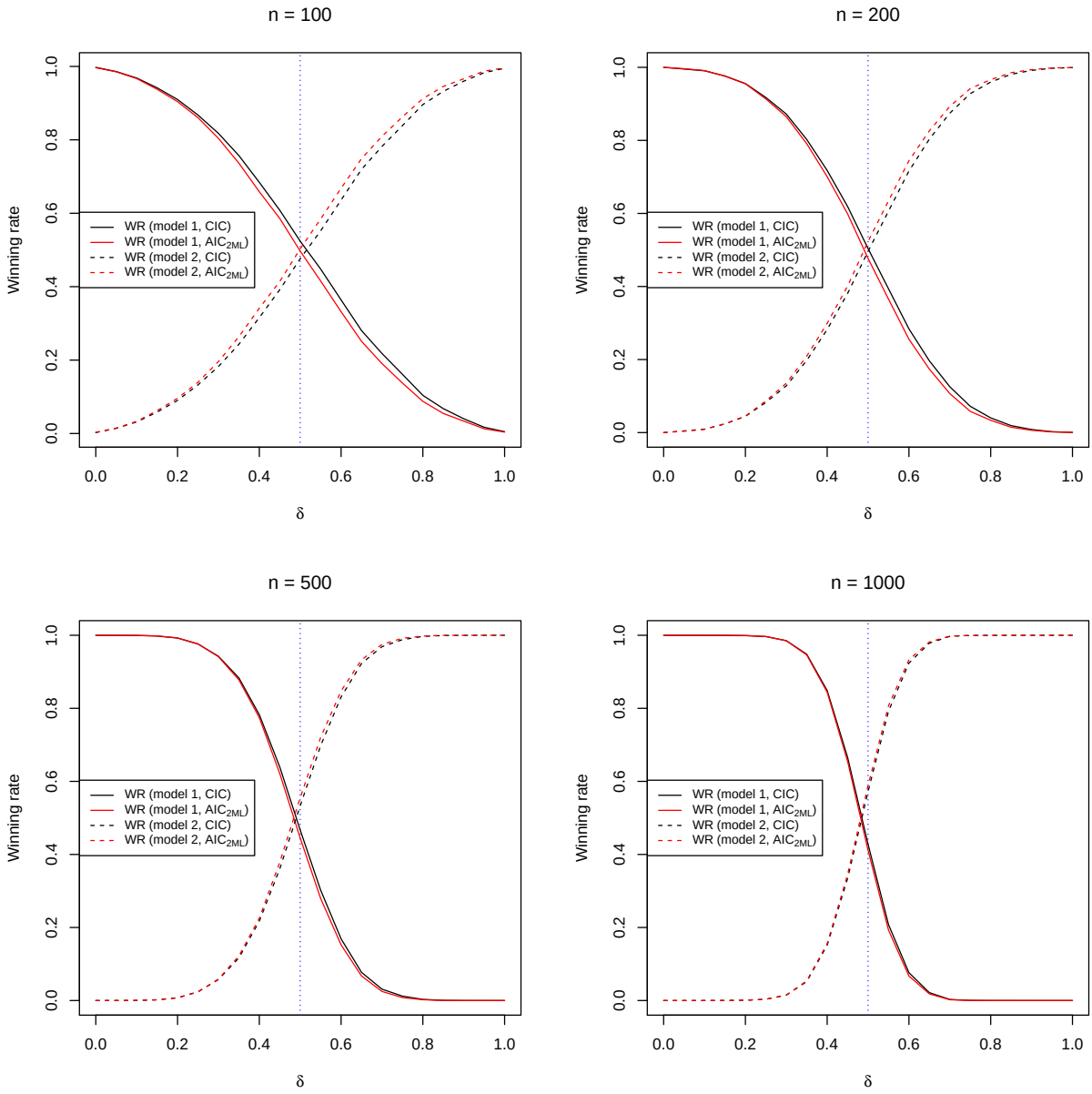

Figure 4: Result of simulation 2. The y-axis is what we call 'winning rate (WR)', the proportion among 10000 repetitions that the concerning model (model 1 or model 2) is picked as the best model according to the model selection criterion of choice (CIC or $\mathrm{AIC}_{\mathrm{2ML}}$). The x-axis is δ, the proportion of data points generated by model 2 described in Table 4. The remaining $(1-\delta)\cdot 100\%$ data points were generated by model 1. The blue dashed vertical line is $\delta = 0.5$.

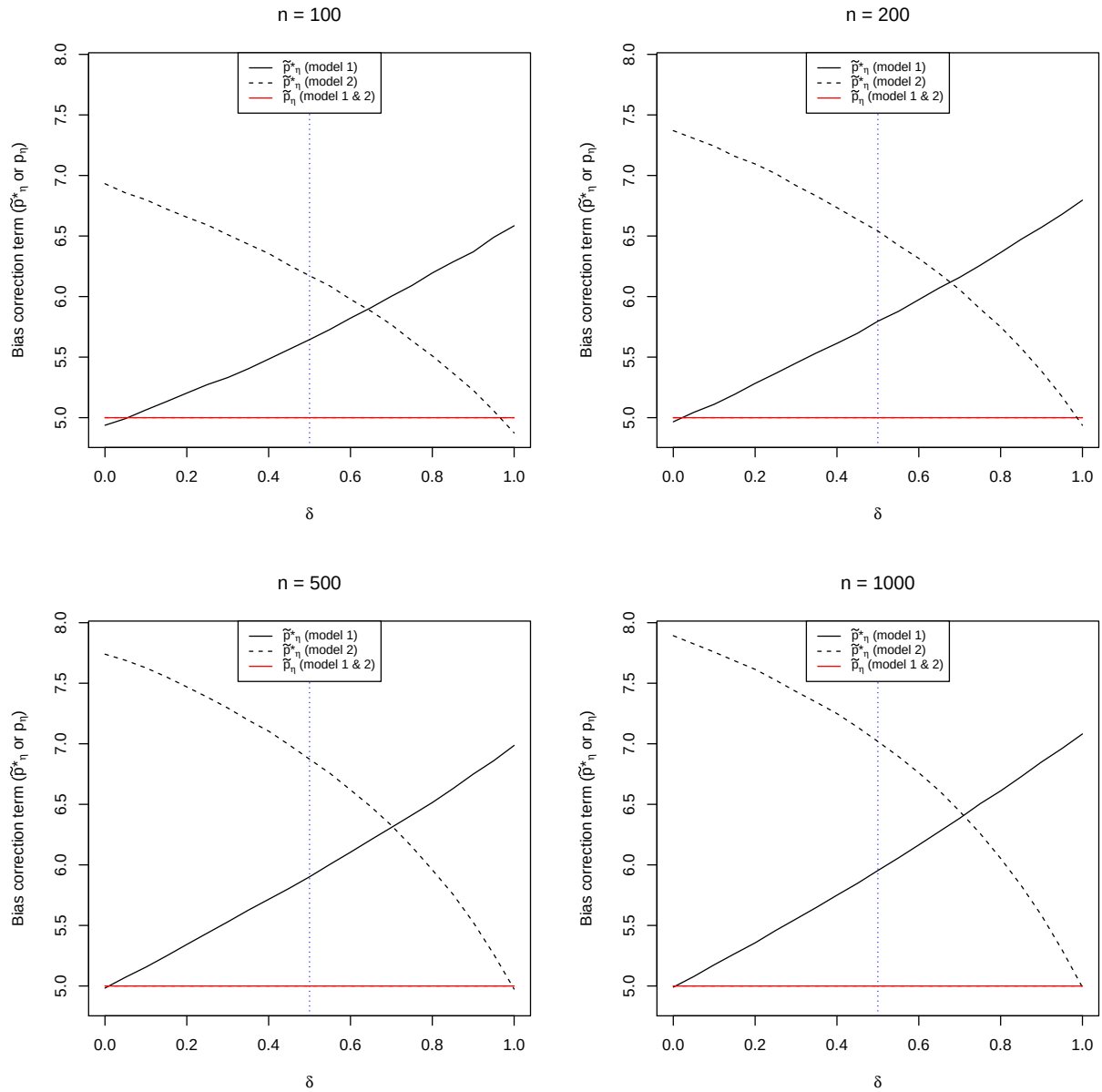

Figure 5: Result of simulation 2. The y-axis is the value of bias correction term ($\widetilde{p}^*_\eta$ for CIC and p_η for $\mathrm{AIC}_{2\mathrm{ML}}$). The x-axis is δ, the proportion of data points generated by model 2 described in Table 4. The remaining $(1-\delta)\cdot 100\%$ data points were generated by model 1. The simulation was repeated 10000 times and the results were averaged. The blue dashed vertical line is $\delta = 0.5$.

Interestingly, there does exist a situation where, theoretically at least, the difference in $\widetilde{p}^*_\eta$ between two competing models always favors the correct model: When one of the two model is the 'true model'. i.e. $\delta = 0$ or 1. In this case, as one can see in Figure 5, $\widetilde{p}^*_\eta$ of the correct model will be (almost) equal to p_η, while $\widetilde{p}^*_\eta$ of the misspecified model will outgrow p_η. So, there is no issue of growth speed difference. However,

in this situation, the log-likelihood of the 'true model' will be much higher than that of the misspecified model. So, the $\mathrm{AIC}_{2\mathrm{ML}}$ would also be able to point the true model as winning model. (e.g. Winning rate close to 1 in Figure 4 when $\delta = 0$ or 1.) So, in practice, there will be no noticeable advantage of CIC in this situation.

Furthermore, the fact that $\widetilde{p}^*_\eta$ gets lower than p_η in lower sample sizes is because of the estimation errors involved in estimating matrices as mentioned in Section 2.3 of Burnham and Anderson (2003).

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 copulas. 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 set of simulation studies. In the first simulation study, we observed that CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ have strong agreement in how they rank the model and that the 'better models' according to CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ show better prediction performance (measured in MSE). In addition, $\widetilde{p}^*_\eta$ alone has a strong correlation with the prediction performance. So, one could consider to utilize $\widetilde{p}^*_\eta$ as an extra tool to evaluate the model. (We showed earlier that $\widetilde{p}^*_\eta$ can be used to measure the divergence of a candidate model from the true model.)

In the second simulation study, we showed that CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ are indeed identical when the model is correctly specified. The two model selection criteria diverge from each other as the degree of misspecification gets larger. The difference between CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ however becomes smaller and smaller as sample size increases, because the log-likelihood part grows much faster than the bias correction part.

Further, we showed that when CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ disagree on the winning model, CIC does not necessarily

outperforms $\mathrm{AIC}_{2\mathrm{ML}}$. CIC has tendency to prefer a model that shows slow growth $\widetilde{p}_\eta^*$ of when the degree of model misspecification increases. Further, CIC should theoretically outperform $\mathrm{AIC}_{2\mathrm{ML}}$ in the situation when there is true model on the candidate model list. However, in practice, this effect is shadowed by the log-likelihood term which dominates the criterion because the fit of the true model is much better than the misspecified model. In this regard, $\mathrm{AIC}_{2\mathrm{ML}}$ already picks the true model as the best model and there is almost no space for CIC to outperform $\mathrm{AIC}_{2\mathrm{ML}}$. Hence, no real practical advantage of CIC over $\mathrm{AIC}_{2\mathrm{ML}}$. Moreover, CIC requires large sample sizes to estimate its bias correction elements accurately and when sample size is large, the dominance of log-likelihood term gets even larger. In this sense, $\mathrm{AIC}_{2\mathrm{ML}}$ is a parsimonious approach to CIC. These practical issues are also the reason why TIC is rarely applied in practice for ML estimated models (Burnham and Anderson, 2003).

Because of the large number of possible models in high dimensional settings, the number of situations that we could examine was limited. (In case of a 4-dimensional copula model with 5 candidate copulas and 3 candidate margins for each variable, there are 324 models that we have to test.) Another problem was that we could not try all copulas 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 copulas in variety of situations would be fruitful.

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

Theoretically, CIC can be directly applied to pair-copula constructions as long as one uses two-stage ML for estimation. A pair-copula construction is after all just a special case of high-dimensional copula (Aas et al., 2009). However, in practice, it can be difficult to estimate all the necessary components due to the large number of parameters.

Although CIC and $\mathrm{AIC}_{2\mathrm{ML}}$ perform 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. The authors of this study are currently developing a model selection criterion that can take this into account, for copula models under the two-stage ML scheme (Ko et al., 2019).

Acknowledgments

The authors would like to thank Steffen Grønneberg and Ingrid Hobæk Haff 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. We are also grateful to the editor and reviewers for constructive comments which helped improve our presentation.

References

Aas, K., Czado, C., Frigessi, A., Bakken, H., 2009. Pair-copula constructions of multiple dependence. Insurance: Mathematics and Economics 44, 182–198.

Acar, E.F., Craiu, R.V., Yao, F., 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.

Burnham, K.P., Anderson, D.R., 2003. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. 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 27, 951–981.

Ko, V., Hjort, N.L., 2019. Model robust inference with two-stage maximum likelihood estimation for copulas. Journal of Multivariate Analysis (in press).

Ko, V., Hjort, N.L., Hobæk Haff, I., 2019. Focused information criterion for copulas. Scandinavian Journal of Statistics (in press).

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.