Grønneberg, S., & Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics, 41*(2), 436-459 Doi: http://dx.doi.org/10.1111/sjos.12042

THE COPULA INFORMATION CRITERIA

STEFFEN GRØNNEBERG AND NILS LID HJORT

Abstract. We derive two types of AIC-like model selection formulae for the semiparametric pseudo maximum likelihood procedure. We first adapt the arguments leading to the original AIC formula, related to empirical estimation of a certain Kullback–Leibler information distance. This gives a significantly different formula compared to the AIC, which we name the Copula Information Criterion (CIC). However, we show that such a model-selection procedure cannot exist for copula models with densities that grow very fast near the edge of the unit cube. This problem affects most popular copula models. We then derive what we call the Cross-Validation Copula Information Criterion (xv-CIC), which exists under weak conditions and is a first order approximation to exact cross validation. This formula is very similar to the standard AIC formula, but has slightly different motivation. A brief illustration with real data is given.

1. Introduction and summary

A fundamental practical issue in any statistical investigation is the problem of model selection: Suppose several candidate models are available, which model is the best? Many approaches to what "best" means have been suggested in the literature, and the following two are the most common. Firstly, the best model may be the one containing the parameter configuration that minimizes some distance to the postulated true model. Secondly, the best model may be the one giving best predictions for new, and as of yet unobserved cases. Both of these approaches require assumptions on the true data generating mechanism to lead to clear recipes, and in the most famous case – the AIC case of classical parametric statistics – they are connected through an asymptotic equivalence between a certain version of cross-validation and an extended version of the AIC formula called the TIC formula. These basic issues are discussed in Chapter 2 of Claeskens & Hjort (2008).

The AIC formula famously reads

$$\mathrm{AIC} = 2\left(\ell^{\#}_{n,\max} - \mathrm{length}(\theta)\right) \tag{1}$$

where $\ell^{\#}_{n,\max}$ is the maximized likelihood for the model and $\mathrm{length}(\theta)$ is the dimensionality of the parameter set. One computes this AIC score for each candidate model and in the end chooses the model with highest score. This formula is derived under certain rather specific settings, and it is not at all obvious that it is valid outside these conditions. However, it is extremely simple to compute in all likelihood-based investigations, and is heuristically justified as a penalization for complexity. Penalizing for the number of parameters is also attempted for estimation methods which are not purely likelihood based, but usually have names relating to likelihoods, such as pseudo likelihoods. However, this is a very weak justification by itself: it does not give any rational way to prefer the AIC formula compared to, say, the BIC formula

$$\mathrm{BIC} = 2\ell^{\#}_{n,\max} - \log n \times \mathrm{length}(\theta).$$

We believe that there is a genuine need to clarify the use of such formulas in several applied statistical investigations where the classical arguments leading to the AIC formula are invalid,

Date: June 3, 2013.

Key words and phrases. AIC, CIC, copulae, model selection, MPLE, multivariate rank statistics, xv-CIC.

especially when pseudo likelihoods are used. We provide a general description of this problem in Section 2, which are then specialized to the copula case.

When using the unjustified AIC formula, it is implicitly hoped that when an estimation method heuristically resembles the maximum likelihood estimator, there is a continuous relationship between the two technique's model selection behavior, so that the AIC formula is approximately valid. Our current paper investigates the validity of the AIC formula in a semiparametric estimation problem related to copula models. We investigate both the loss-function and prediction perspectives, and the two resulting cases illustrate drastically different levels of continuity regarding model selection behavior. We show that under certain assumptions, the AIC formula is indeed approximately valid from a loss-function perspective. However, outside these rather restrictive conditions, we observe a strong discontinuity in the form of a *non-existence* of such model selection procedures. This discontinuity does not extend to the prediction-perspective of the AIC formula, where a continuous relationship is demonstrated by deriving a generally applicable model-selection formula that approximately equals the classical formula under weak conditions. In this light, our paper motivates further investigation of the AIC formula's use also in other likelihood-like estimation methods.

Our technical setting is as follows. Suppose given independent, identically distributed d-dimensional observations $X_1, X_2, \ldots, X_n$ with density $f^\circ(x)$ and distribution function

$$F^\circ(x) = P(X_{i,1} \leq x_1, X_{i,2} \leq x_2, \ldots X_{i,d} \leq x_d) = C^\circ(F^\circ_\perp(x)).$$

Here, C° is the copula of F° and $F^\circ_\perp$ is the vector of marginal distributions of F°, that is,

$$F^\circ_\perp(x) := (F^\circ_1(x_1), \ldots, F^\circ_d(x_d)), \qquad F_j(x_j) = P(X_{i,j} \leq x_j).$$

We want to fit parametric models to the copula, but leave the marginals unspecified. The copula models are specified through a set of densities $c(u, \theta)$ for $\Theta \subseteq \mathbb{R}^p$ and $u \in [0,1]^d$.

A popular estimator for the copula parameter is the maximum pseudo likelihood estimator $\hat\theta_n$, also called the MPLE. It is defined as the maximizer of the so-called pseudo likelihood

$$\ell_n(\theta) := \sum_{i=1}^n \log c(F_{n,\perp}(X_i), \theta).$$

This estimator sometimes goes by other names, such as the Canonical MLE (Panchenko, 2005). We also note that unrelated estimation techniques are sometimes called the maximum pseudo likelihood estimator in the literature. The pseudo likelihood is expressed in terms of the so-called pseudo observations $F_{n,\perp}(X_i) \in [0,1]^d$, in which $F_{n,\perp}$ is the vector of re-normalized marginal empirical distribution functions

$$F_{n,\perp}(x) := (F_{n,1}(x_1), \ldots, F_{n,d}(x_d)), \text{ where } F_{n,j}(x_j) := \frac{1}{n+1} \sum_{i=1}^n I\{X_{i,j} \leq x_j\}.$$

The non-standard normalization constant $1/(n+1)$ – instead of the classical $1/n$ – is to avoid evaluating $u \mapsto \log c(u, \theta)$ at the boundary $u \in \partial\left([0,1]^d\right)$ where most copula models of interest are infinite. Hence, we consider any size defined in terms of $u \mapsto c(u, \theta)$ as being restricted to $u \in (0,1)^d$.

Many investigations, such as Chen & Fan (2005) and McNeil et al. (2005, Chapter 5), use

$$\text{AIC}^\star = 2\ell_{n,\max} - 2\text{length}(\theta) \tag{2}$$

as a model selection criterion for the MPLE, with $\ell_{n,\max} = \ell_n(\hat\theta_n)$ being the maximum pseudo likelihood. Despite its frequent use, there is no justification for this formula in the literature other than the less than satisfactory heuristic argument mentioned above.

The arguments underlying the derivations of the traditional AIC do not apply here – since $\ell_n(\cdot)$ is not a proper log-likelihood function for a model, but a pseudo likelihood, based on the multivariate rank statistics $F_{n,\perp}$. In other words, the AIC* formula above ignores the noise inherent in the transformation step that takes X_i to $F_{n,\perp}(X_i)$. Such a formula would be appropriate only if we could use $F_k^\circ(X_{i,k})$ – instead of the pseudo observations, or if we would model the marginals $F_1^\circ, \ldots, F_d^\circ$ by parametric models $F_{1,\gamma(1)}, \ldots, F_{d,\gamma(d)}$. This last case would return the estimation problem to a fully parametric one, where the classical AIC formula

$$2\left(\ell_{n,\max}^{\#} - \delta_c - \delta_m\right), \quad \delta_c = \text{length}(\theta), \quad \delta_m = \sum_{k=1}^{d} \text{length}(\gamma(k)) \tag{3}$$

is justified by classical theory. Here $\ell_{n,\max}^{\#}$ is the standard maximized likelihood, δ_c and δ_m each corrects for bias introduced by the estimation of the copula and the marginals respectively.

Note that eq. (3) is only valid when the likelihood is maximized simultaneously in all parameters of the model. The use of multi-stage estimation routines, such as the MPLE or the IFM estimator described in Joe (1997), invalidates the AIC formula. Our paper will focus exclusively on the more complicated MPLE case, but through mimicking the developments of our paper, one could derive a copula model selection procedure based on the IFM, analogous to the AIC formula. Note that because the least false IFM parameter configuration for the copula depends on the marginal misspecification – a deficiency not shared by the MPLE – an IFM-AIC formula would only be valid when the parametric marginal models include the true marginal distributions.

The present paper centres around two contributions. First, we reconsider the steps leading to the original AIC formula in the MPLE setting and derive the appropriate modifications. This leads to two model selection formulae – one valid when the copula model is correctly specified and one valid in general. We will refer to both as the Copula Information Criterion when the context makes it clear which one is meant (or when it does not matter), and will refer to them as the AIC-like and the TIC-like CIC formula when this distinction is needed. These formulae and their derivations are presented in Section 2.1.

The AIC-like CIC formula is of the form

$$2\left(\ell_{n,\max} - \hat{\delta}_c - \hat{\delta}_m\right), \quad \delta_c = \text{length}(\theta) + \text{Tr}\left(\hat{\mathcal{I}}^{-}\hat{W}\right).$$

Again, $\hat{\delta}_c$ and $\hat{\delta}_m$ each takes the estimation of the copula and the marginals into consideration respectively. Now, $\hat{\delta}_c$ has an additional term because we are working with a pseudo likelihood, and $\hat{\delta}_m$ is an estimator of the size $\mathbf{1}^t\Upsilon\mathbf{1}$ where $\Upsilon = (\Upsilon_{a,b})_{1\le a,b\le d}$ is the symmetric matrix with

$$\Upsilon_{a,a} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,a}(u,\theta^\circ) u_a(1-u_a)\,\mathrm{d}C^\circ(u),$$
$$\Upsilon_{a,b} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,b}(u,\theta^\circ)\left[C_{a,b}(u_a,u_b) - u_a u_b\right]\,\mathrm{d}C^\circ(u) \quad (\text{when } a \neq b),$$

and $\zeta''_{a,b}$ is the (a,b)'th element of the matrix function

$$\zeta''(u,\theta^\circ) = \frac{\partial^2}{\partial u^t \partial u}\log c(u,\theta^\circ). \tag{4}$$

and $C_{a,b}$ is the bivariate margin of C corresponding to dimensions a and b.

Section 2.4 includes a simulation illustrating the superiority of the CIC formula to the unmotivated AIC formula for a mixture of Frank and Plackett copulas.

A major difference from the fully parametric case is that $\mathbf{1}^t\Upsilon\mathbf{1}$ may be infinite. The AIC formula provides a certain type of bias-correction, and it turns out that the random variable that causes the systematic deviation we wish to correct for does not even possess a first moment for most popular

copula-models. In a sense made precise in Section 2.1, we show that there does not exist any model selection formula analogous to the AIC for many popular copula models when using the maximum pseudo likelihood estimator. Further differences is that while $\hat{\delta}_c$ is always strictly positive, $\hat{\delta}_m$ may be both positive and negative. Also, in contrast to the penalty term of the classical AIC formula that do not depend on the data, CIC's penalty terms must always be estimated from data.

The second part of the paper pursues the second main path of model selection methodology: prediction. As mentioned above, a certain version of cross-validation and the classical AIC formula are first order equivalent. In Section 4 we show that this is not the case for the CIC and derive a formula that *is* first order equivalent to a version of cross-validation. We name this formula the Cross-Validation Copula Information Criterion, or xv-CIC. This non-equivalence provides a further contrast between MPLE- and MLE-based estimation, and it turns out that the xv-CIC formula is applicable to all common copula models. Thus, the reader who is simply interested in a generally applicable model selection formula for the MPLE can focus on Section 4.

When the parametric copula model is assumed to include the true copula $c^\circ(\cdot)$, the xv-CIC formula is given by

$$2\left(\ell_{n,\max} - \delta_c\right), \quad \delta_c = \text{length}(\theta) + \text{Tr}\left(J_n^{-1}\hat{K}\right),$$

where J_n and $\hat{K}$ are defined below. When $\text{Tr}\left(J_n^{-1}\hat{K}\right)$ is small, this formula provides motivation for the original AIC-formula. A brief illustration of the xv-CIC formula in Section 5 using the Loss-ALAE data. This dataset is used in many papers on copulas, including Frees & Valdez (1998) and Genest et al. (2006).

At the end of our paper, we give some concluding remarks, including some advice on model selection for practitioners in Section 6.2.

We have not conducted a comprehensive simulation study of the small sample performance of the xv-CIC formula, and consider this to be a theme for a separate paper. Because the unmotivated AIC formula has no terms that are estimated from data, it may under certain settings be superior to the xv-CIC formula as an approximation to cross validation. A large scale simulation study would be able to investigate whether or when this is the case.

The paper includes an appendix in the form of a supplementary note, available on the journal web-site. This appendix gathers all but the simplest technical proofs, and includes a script for the R system (as described in R Development Core Team, 2010) to calculate the xv-CIC for certain simple copula models.

We will consistently apply the perpendicular subscript to indicate vectors of marginal distributions, such as $F_{n,\perp}$. Note that we will sometimes use the multivariate empirical distribution function F_n, which is defined with the standard scaling $1/n$ in contrast to our marginal empirical distributions that are scaled according to $1/(n+1)$. We will also use the circle superscript to denote sizes defined in terms of F° and will usually let hats and/or n-subscripts indicate estimators. For example, the Kullback–Leibler least false parameter configuration θ° has a circle superscript, because it is defined in terms of F°, while its estimate is denoted by $\hat{\theta}_n$. We will denote generic elements of $[0,1]^d$ or $[0,1]$ by u or v, while elements of $\mathbb{R}^d$ not constrained to $[0,1]^d$ will be denoted by x or y. For a general introduction to copula models, see Joe (1997), and for a general introduction to the model selection problem, see Claeskens & Hjort (2008). Finally, we will usually let $\mathrm{d}f(x_0)/\mathrm{d}x$ denote $\mathrm{d}f(x)/\mathrm{d}x|_{x=x_0}$.

2. The Copula Information Criterion

Let us take a step back, and consider a fairly abstract summary of the derivation of the AIC formula. Through this discussion, we place the structure of the CIC problem in relation to the

AIC and the so-called Generalized Information Criterion, and indicate the calculations that are required for solving the problem at hand. A detailed derivation of the CIC is then given in Section 2.1.

Maximum likelihood estimation features two statistical functionals – in the sense, say, of Shao (2003) – given by

$$\Phi[F](f(\cdot)) = \int f(x)\,\mathrm{d}F(x)$$

and

$$T[F] = \operatorname*{argmax}_{\theta\in\Theta} \Phi[F](f(\cdot,\theta)). \tag{5}$$

Here F is some cumulative distribution function, and T is defined in terms of a parametric family of densities $\{f(\cdot,\theta) : \theta \in \Theta\}$. The MLE is $\hat{\theta}_{\mathrm{ML}} = T(F_n)$, where F_n is the (multivariate) empirical distribution function.

The Kullback–Leibler divergence between two densities g and h is given by

$$\mathrm{KL}(g(\cdot), h(\cdot)) = \int g(x) \log \frac{g(x)}{h(x)}\,\mathrm{d}x = \int \log g(x)\,\mathrm{d}G(x) - \int \log h(x)\,\mathrm{d}G(x)$$

where $G(x) = \int_{-\infty}^{x} g(y)\,\mathrm{d}y$. Hence,

$$\Phi[F^\circ](\log f(\cdot,\hat{\theta}_n)) = -\mathrm{KL}(f^\circ(\cdot), f(\cdot,\hat{\theta}_n)) + C \tag{6}$$

is the negative attained Kullback–Leibler divergence between $f^\circ(\cdot)$ and $f(\cdot,\theta)$ except for a constant C not depending on $\hat{\theta}_n$. Let us therefore call $\Phi[F^\circ](\log f_\theta)$ the model-relevant part of the KL-divergence (between $f^\circ(\cdot)$ and $f(\cdot,\theta)$). Under regularity conditions, we have

$$\hat{\theta}_n = T(F_n) \xrightarrow[n\to\infty]{\mathscr{P}} T(F^\circ) = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}(f^\circ(\cdot), f(\cdot,\theta)) := \theta^\circ$$

so that $\hat{\theta}_n$ approaches the least false Kullback–Leibler parameter configuration θ°. Also, eq. (6) shows that $\Phi[F^\circ](f(\cdot,\hat{\theta}_n))$ is the attained model-relevant part of KL-divergence. If we are given several candidate models, the AIC-perspective is to use the model with the least attained KL-divergence, or equivalently, the largest attained model-relevant part of KL-divergence. We typically have

$$\Phi[F_n](\log f(\cdot,\hat{\theta}_n)) \xrightarrow[n\to\infty]{\mathscr{P}} \Phi[F^\circ](\log f(\cdot,\theta^\circ)),$$

and we will study the finite sample bias variable

$$\Delta_n := \Phi[F_n](\log f(\cdot,\hat{\theta}_n)) - \Phi[F^\circ](\log f(\cdot,\theta^\circ)) \tag{7}$$

up to a precision level specified shortly. The linearity of Φ in both arguments shows that

$$\begin{aligned}
\Delta_n &= \Phi[F_n - F^\circ](\log f(\cdot,\hat{\theta}_n)) \\
&= \Phi[F_n - F^\circ](\log f(\cdot,\hat{\theta}_n)) - \Phi[F_n - F^\circ](\log f(\cdot,\theta^\circ)) + \Phi[F_n - F^\circ](\log f(\cdot,\theta^\circ)) \\
&= \Phi[F_n - F^\circ](\log f(\cdot,\hat{\theta}_n) - \log f(\cdot,\theta^\circ)) + \Phi[F_n - F^\circ](\log f(\cdot,\theta^\circ)).
\end{aligned} \tag{8}$$

Under regularity conditions (Claeskens & Hjort, 2008) we get

$$\Delta_n = Z_n + \frac{1}{n}\delta_n + o_P(n^{-1}) \tag{9}$$

where

$$Z_n = \Phi[F_n - F^\circ](\log f(\cdot,\theta^\circ)) \tag{10}$$

is a zero mean variable, $\delta_n \xrightarrow[n\to\infty]{\mathscr{W}} \delta$ where $\mathbb{E}\delta \neq 0$. When the model is correct in the sense that $f(\cdot,\theta^\circ) = f^\circ(\cdot)$, we get $\mathbb{E}\delta = \mathrm{length}(\theta)$. The AIC formula of eq. (1) is a sample bias correction for $\Phi[F_n](\log f(\cdot,\hat{\theta}_n))$ as an estimator for $\Phi[F^\circ](\log(\cdot,\hat{\theta}_n))$ based on eq. (9). It is a bias-correction in the weak sense that $o_P(n^{-1})$ is considered low-level noise and is ignored, and that we only try to

approximate the expectation of the weak limit of δ_n, and not the actual attained expectation of δ_n (which may be infinite, see Claeskens & Hjort (2008)).

A generalization of these developments is the Generalized Information Criterion (Konishi & Kitagawa, 2008). Here, the functional T does not have to take on the rather specific form of eq. (5) but may be defined as the maximizer of, say, a *penalized* likelihood function such as for shrinkage estimators. The GIC development concerns reaching an expansion for Δ_n that takes into consideration the general form of T. The functional Φ is, however, maintained as in the original AIC formula.

The CIC exceeds the generality provided by the GIC in the following way. Our model is defined in terms of a parametric copula model $\{c(u,\theta) : \theta \in \Theta\}$. The marginal distributions $f_1^\circ, \ldots, f_d^\circ$ of the observations are unknown and completely unspecified. Concretely, our model is therefore given by

$$\{f(x,\theta) : \theta \in \Theta\} = \left\{ f(x,\theta) = c\left(F_1^\circ(x_1), \ldots, F_d^\circ(x_d), \theta\right) \prod_{k=1}^d f_k^\circ(x_k) : \theta \in \Theta \right\}$$

for $x \in \mathbb{R}^d$. Through a change of variables, we see that the Kullback–Leibler least false parameter configuration is

$$\begin{aligned}
\theta^\circ &= \operatorname*{argmax}_{\theta\in\Theta} \int_{\mathbb{R}^d} \log f(x,\theta)\, \mathrm{d}F^\circ(x) \qquad (11)\\
&= \operatorname*{argmax}_{\theta\in\Theta} \left[\int_{\mathbb{R}^d} \log c\left(F_1^\circ(x_1), \ldots, F_d^\circ(x_d), \theta\right)\, \mathrm{d}F^\circ(x) + \sum_{k=1}^d \int_{\mathbb{R}} \log f_k^\circ(x_k)\, \mathrm{d}F_k^\circ(x) \right] \\
&= \operatorname*{argmax}_{\theta\in\Theta} \int_{\mathbb{R}^d} \log c\left(F_1^\circ(x_1), \ldots, F_d^\circ(x_d), \theta\right)\, \mathrm{d}F^\circ \\
&= \operatorname*{argmax}_{\theta\in\Theta} \int_{[0,1]^d} \log c\left(v_1, \ldots, v_d, \theta\right)\, \mathrm{d}C^\circ(v) \\
&= T(C^\circ).
\end{aligned}$$

Hence, the KL least-false copula parameter only depends on the true copula of the data. Because C° is invariant to monotone transformation of the marginals, empirical estimators of θ° should share this invariance. This point is further discussed in Grønneberg (2010). The rank-based MPLE

$$\hat{\theta}_n = T(C_n),$$

defined in terms of the empirical copula

$$C_n(u) := \frac{1}{n} \sum_{i=1}^n I\{F_{n,\perp}(X_i) \leq u\} = \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^d I\{F_{n,j}(X_{i,j}) \leq u_j\}, \qquad (12)$$

shares this invariance, and consistently estimates θ° under various conditions (Genest et al., 1995). In order to provide a model selection formula for the MPLE, we must again study an analogue of Δ_n defined in eq. (7).

By following the same logic in going from eq. (7) to eq. (8) we get that

$$\Delta_n = \Phi[C_n](\log c(\cdot, \hat{\theta}_n)) - \Phi[C^\circ](\log c(\cdot, \theta^\circ)) = \Delta_{c,n} + \Delta_{m,n} \qquad (13)$$

where

$$\Delta_{c,n} := \Phi[C_n - C^\circ](\log c(\cdot, \hat{\theta}_n) - \log c(\cdot, \theta^\circ))$$

and

$$\Delta_{m,n} := \Phi[C_n - C^\circ](\log c(\cdot, \theta^\circ)).$$

The notation of the two components of Δ_n is chosen as mnemonics to reflect that $\Delta_{c,n}$ is dominated by bias originating from estimating the parametric copula and $\Delta_{m,n}$ is dominated by bias originating from estimating the marginals non-parametrically. This will be shown in Sections 2.1 and 2.2.

The expansion of eq. (13) is seemingly similar to that in eq. (8). However, in the fully parametric case, $\Delta_{m,n}$, which we then denoted by Z_n, had zero mean and could therefore be ignored when providing bias corrections. This zero mean property is a consequence of

$$\mathbb{E}\Phi[F_n](\log f(\cdot,\theta^\circ)) = \Phi[F^\circ](\log f(\cdot,\theta^\circ)),$$

which follows by the definition of the sizes involved. In contrast, we now have $C_n(u) = \Psi[F_n](u)$ for the statistical functional Ψ implicit in eq. (12), which means that

$$\mathbb{E}\Phi[C_n](\log c(\cdot,\theta^\circ)) = \mathbb{E}\Phi \circ \Psi[F_n](\log c(\cdot,\theta^\circ)) \neq \Phi \circ \Psi[F^\circ](\log c(\cdot,\theta^\circ)), \tag{14}$$

and hence, $\mathbb{E}\Delta_{m,n} \neq 0$ due to the presence of the Ψ functional. In order to derive a model selection formula for the MPLE, we need an expansion such as eq. (9) in terms of some (new) zero mean variable Z_n and some δ_n. This computation will be performed in the following subsection, where we will see that

$$\delta_n = \delta_{c,n} + \delta_{m,1,n} + \delta_{m,2,n}.$$

Here, $\delta_{c,n}$ has contributions from $\Delta_{c,n}$ and $\delta_{m,1,n} + \delta_{m,2,n}$ has contributions from $\Delta_{m,n}$. Precisely, we isolate the sizes with non-zero mean that are not $o_P(n^{-1})$. We split up the contributions from $\Delta_{m,n}$ in two, as $\delta_{m,1}$ is zero when the model is correct. The CIC formula consists of correcting the maximized pseudo likelihood with an estimate of the expectation of the weak limit of δ_n.

Finally, we note that the above formulation should apply to many estimation schemes similar to the MPLE. Such a general formula – a generalized GIC – seems to be possible to derive using second order functional expansions. However, this would require a detailed study of the second order functional differentiation of the statistical functional that defines the MPL estimator. This would be technically challenging, as most functional differentiation theory for functionals of interest in statistics (see e.g. van der Vaart & Wellner, 1996; Shao, 2003) focuses on first order differentiation, as this suffices to prove asymptotic Normality – and not the second order differentiation that would be required in order to isolate the terms in Δ_n that are not $o_P(n^{-1})$.

2.1. **Derivation of The Copula Information Criterion.** Like the AIC, the Copula Information Criterion is based on asymptotic (pseudo) likelihood theory. Before we continue our detailed study of Δ_n, we need the following theory for the pseudo likelihood function, some of which generalize previously published results. Central to our investigation is the behavior of the pseudo log-likelihood normalized by sample size

$$A_n(\theta) := \frac{1}{n}\ell_n(\theta) = \int_{[0,1]^d} \log c(u,\theta)\,\mathrm{d}C_n(u).$$

The maximum pseudo likelihood estimator can be written as

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{n}\ell_n(\theta) = \operatorname*{argmax}_{\theta\in\Theta} \int_{u\in[0,1]^d} \log c(u,\theta)\,\mathrm{d}C_n(u)$$

where C_n is the empirical copula of eq. (12). Under conditions such as A1-A5 in Tsukahara (2005), we have

$$\hat{\theta}_n \xrightarrow[n\to\infty]{\mathscr{P}} \operatorname*{argmax}_{\theta\in\Theta} \int_{u\in[0,1]^d} \log c(u,\theta)\,\mathrm{d}C^\circ(u) =: \theta^\circ$$

in which θ° is the least false parameter according to the Kullback–Leibler divergence between the true model $c^\circ(\cdot)$ and $\{c(\cdot,\theta) : \theta \in \Theta\}$. That is,

$$\theta^\circ = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}(c^\circ(\cdot), c(\cdot,\theta)) = \operatorname*{argmin}_{\theta\in\Theta} \int_{u\in[0,1]^d} \log \frac{c^\circ(u)}{c(u,\theta)} c^\circ(u)\,\mathrm{d}u = \operatorname*{argmax}_{\theta\in\Theta} A(\theta) \tag{15}$$

where

$$A(\theta) := \int_{[0,1]^d} c^\circ(u) \log c(u,\theta)\,\mathrm{d}u = \int_{[0,1]^d} \log c(u,\theta)\,\mathrm{d}C^\circ(u). \tag{16}$$

We typically have

$$A_n(\theta) \xrightarrow[n\to\infty]{\mathscr{P}} A(\theta)$$

for each θ, for example under the conditions given in Proposition A1.i in Genest et al. (1995).

Let

$$\phi(u,\theta) = (\partial/\partial\theta) \log c(u,\theta) \tag{17}$$

be the vector of derivatives of $\theta \mapsto \log c(u,\theta)$ and let

$$U_n := \frac{\partial A_n(\theta^\circ)}{\partial\theta} = \frac{1}{n}\frac{\partial \ell_n(\theta^\circ)}{\partial\theta} = \int_{[0,1]^d} \frac{\partial}{\partial\theta} \log c(u,\theta^\circ)\,\mathrm{d}C_n(u) = \int_{[0,1]^d} \phi(u,\theta)\,\mathrm{d}C_n(u) \tag{18}$$

be the normalized pseudo score function, evaluated at θ°. To state the asymptotic distribution of the MPLE, we need the Information matrix

$$\mathcal{I} = \mathbb{E}\phi(\xi,\theta^\circ)\phi(\xi,\theta^\circ)^t \tag{19}$$

and

$$\mathcal{W} = \operatorname{Var} Z, \tag{20}$$

which is the covariance matrix of

$$Z := \sum_{k=1}^d \int_{[0,1]^d} \frac{\partial\phi(u,\theta^\circ)}{\partial u_k} (I\{\xi_k \le u_k\} - u_k)\,\mathrm{d}C^\circ(u) \tag{21}$$

where ξ is a random vector distributed according to C°.

The asymptotic Normality of the normalized score function $\sqrt{n}U_n$ is central to proving the asymptotic Normality of the MPLE. This asymptotic result may be established either through functional weak convergence of the empirical copula process or through the more direct arguments of Ruymgaart et al. (1972) and Ruymgaart (1974). While the direct route is followed in Genest et al. (1995) and Tsukahara (2005), Theorem 6 of Fermanian et al. (2004) shows that the score function is asymptotically normal as a consequence of the process convergence of the empirical copula. Segers (2012) substantially weakened the conditions given in Fermanian et al. (2004) for this process convergence to take place, which in turn implies that using the process convergence of the empirical copula process to prove asymptotic normality of the score function is now of more general applicability. Because we will use this perspective to prove Theorem 1 a bit later, we include the following extension of Theorem 6 of Fermanian et al. (2004). The Lemma features the following condition, which Segers (2012) shows is sufficient for the empirical copula to have a weak limiting distribution. The condition is also necessary for the Gaussian limiting process of the empirical copula to exist and have continuous sample paths.

Condition 1. *For each $j \in \{1,\dots,d\}$, the j'th first-order partial derivative $\dot{C}_j^\circ$ exists and is continuous on the set $V_{d,j} = \{u \in [0,1]^d : 0 < u_j < 1\}$.*

Lemma 1. *Suppose one of the following conditions are fulfilled.*

(1) The score function fulfills condition A1 of Tsukahara (2005).

(2) *That* $u \mapsto \log c(u, \theta)$ *is of bounded Hardy–Krause-variation (defined in A.3 in the supplementary notes) and Condition 1 is fulfilled.*

Then

$$\sqrt{n}U_n \xrightarrow[n\to\infty]{\mathscr{W}} U \sim N_p(0, \Sigma), \quad \Sigma := \mathcal{I} + \mathcal{W} \tag{22}$$

where $\mathcal{I}$ *and* $\mathcal{W}$ *are defined in equations* (19) *and* (20) *respectively.*

Proof. We extend the proof of the bivariate case given in Theorem 6 of Fermanian et al. (2004). Proposition 3.1. of Segers (2012) implies the desired result using the general change of variables formula provided in Section A.3 of the supplementary notes. Other than using the theory of Segers (2012), our only addition to the proof of Fermanian et al. (2004) is to correct their omission of mentioning that their result requires bounded Hardy–Krause-variation and not some other multivariate variational concept. □

Assumptions on the topology of the parameter set Θ analogous to the classical conditions securing asymptotic Normality of the MLE, such as assumptions A1-A5 of Tsukahara (2005), shows

$$\sqrt{n}(\hat{\theta}_n - \theta^\circ) \xrightarrow[n\to\infty]{\mathscr{W}} J^{-1}U \sim N_p(0, J^{-1}\Sigma J^{-1}), \tag{23}$$

where

$$J := -A''(\theta^\circ) = -\int_{[0,1]^d} c^\circ(u) \frac{\partial^2 \log c(u, \theta^\circ)}{\partial\theta\partial\theta^t} \,\mathrm{d}u = -\int_{[0,1]^d} \frac{\partial^2 \log c(u, \theta^\circ)}{\partial\theta\partial\theta^t} \,\mathrm{d}C^\circ \tag{24}$$

is assumed to be of full rank. We will also assume that

$$J_n := -A_n''(\theta^\circ) = -\int_{[0,1]^d} \frac{\partial^2 \log c(u, \theta^\circ)}{\partial\theta\partial\theta^t} \,\mathrm{d}C_n \tag{25}$$

converges in probability to J.

We are now in a position to study the asymptotic behavior of

$$\Delta_n = A_n(\hat{\theta}_n) - A(\hat{\theta}_n). \tag{26}$$

As was the case in the more abstract notation of eq. (13), we get that

$$\begin{aligned}
\Delta_n &= A_n(\hat{\theta}_n) - A(\hat{\theta}_n) \\
&= A_n(\hat{\theta}_n) - A(\hat{\theta}_n) - [A_n(\theta^\circ) - A(\theta^\circ)] + [A_n(\theta^\circ) - A(\theta^\circ)] \\
&= \Big\{ A_n(\hat{\theta}_n) - A(\hat{\theta}_n) - [A_n(\theta^\circ) - A(\theta^\circ)] \Big\} + [A_n(\theta^\circ) - A(\theta^\circ)] \\
&= \Delta_{c,n} + \Delta_{m,n}
\end{aligned}$$

where

$$\Delta_{c,n} = A_n(\hat{\theta}_n) - A(\hat{\theta}_n) - [A_n(\theta^\circ) - A(\theta^\circ)] = \int \log c(u, \hat{\theta}_n) - \log c(u, \theta^\circ) \, d\,[C_n - C^\circ]$$

and

$$\Delta_{m,n} = A_n(\theta^\circ) - A(\theta^\circ) = \int \log c(u, \theta^\circ) \,\mathrm{d}\,[C_n - C^\circ]\,. \tag{27}$$

While $\Delta_{c,n}$ may appear to be more complex than $\Delta_{m,n}$, it is $\Delta_{m,n}$ that causes complications when providing an AIC-like model selection formula for the MPLE. Intuition behind this is that the integrand of $\Delta_{c,n}$ is small, so bias in the integrator $\mathrm{d}[C_n - C]$ turns out not to be as much of a problem as in $\Delta_{m,n}$, where the integrand is constant. Because the mathematical structure behind $\Delta_{c,n}$ is fairly unproblematic when discarding its $o_P(n^{-1})$ components, we only provide a heuristic justification for the condition using classical Taylor-expansions and smoothness conditions. A rigorous proof would basically replicate the expansions performed in Theorem 1 of Tsukahara

(2005) and would involve expansions very similar to but simpler than those of Appendix A.2 in the supplementary notes. In order to maintain brevity and focus, we do not include a formal proof.

Condition 2. *Assume that*

$$\Delta_{c,n} = \frac{1}{n}\delta_{c,n} + o_P(n^{-1}), \quad \delta_{c,n} = \sqrt{n}(\hat{\theta}_n - \theta^\circ)^t \sqrt{n} U_n \xrightarrow[n\to\infty]{\mathscr{W}} D_c,$$

where U_n is the score function of eq (18)*. Also assume that*

$$\delta_c := \mathbb{E} D_c = \operatorname{Tr}\left(J^{-1}\mathcal{I}\right) + \operatorname{Tr}\left(J^{-1}\mathcal{W}\right)$$

where $\mathcal{I}$ and $\mathcal{W}$ are defined in equations (19) *and* (20) *respectively and J is defined in eq.* (24)*.*

Justification 1. *A two-term Taylor-expansion of $\log c(u,\theta)$ around θ° gives*

$$\Delta_{c,n} = (\hat{\theta}_n - \theta^\circ)^t \int \frac{\partial}{\partial\theta} \log c(u,\theta_0)\, \mathrm{d}\left[C_n - C^\circ\right] + \frac{1}{2}(\hat{\theta}_n - \theta^\circ) \int \frac{\partial^2}{\partial\theta\partial\theta^t} \log c(u,\theta_0)\, \mathrm{d}\left[C_n - C^\circ\right] (\hat{\theta}_n - \theta^\circ)^t + o_P(|\hat{\theta}_n - \theta^\circ|^2)$$

As

$$\int \frac{\partial}{\partial\theta} \log c(u,\theta_0)\, \mathrm{d}C^\circ = 0$$

and both

$$\frac{1}{2}(\hat{\theta}_n - \theta^\circ) \int \frac{\partial^2}{\partial\theta\partial\theta^t} \log c(u,\theta_0)\, \mathrm{d}\left[C_n - C^\circ\right] (\hat{\theta}_n - \theta^\circ)^t = o_P(n^{-1}), \quad |\hat{\theta}_n - \theta^\circ|^2 = o_P(n^{-1}),$$

we have that

$$\Delta_n = \frac{1}{n}\delta_{c,n} + \int \log c(u,\theta^\circ)\, \mathrm{d}\left[C_n - C^\circ\right] + o_P(n^{-1}).$$

When these types of expansions are valid, Lemma 1 implies that

$$\sqrt{n}(\hat{\theta}_n - \theta^\circ)^t \sqrt{n} U_n \xrightarrow[n\to\infty]{\mathscr{W}} U^t J^{-1} U = P$$

where

$$\delta_c = \mathbb{E}P = \mathbb{E}U^t J^{-1} U = \operatorname{Tr}(J^{-1}\Sigma) = \operatorname{Tr}\left(J^{-1}\mathcal{I}\right) + \operatorname{Tr}\left(J^{-1}\mathcal{W}\right).$$

by eq. (22)*.* □

Note that similarly to the fully parametric case, we have $\delta_c \geq 0$ since all matrices involved are positive definite, and the trace of positive definite matrices are positive.

Before we study $\Delta_{m,n}$ in detail, let us first give a bound for the stochastic order of the bias $\Delta_{m,n}$ introduces. This bound shows that if we count low-level noise as $o_P(n^{-3/4-\varepsilon})$ for some $0 < \varepsilon < 1/4$ – and not $o_P(n^{-1})$ – we can ignore $\Delta_{m,n}$. However, under Condition 2, the bias originating from $\Delta_{c,n}$ would also be considered low-level noise, and so would the correction terms in the xv-CIC formula derived in Section 4.

In order to state this result, we need the following condition on the copula of the data found in Segers (2012), where the condition is verified for several popular copulas.

Condition 3. *Let $V_{d,j} = \{u \in [0,1]^d : 0 < u_j < 1\}$ for $j \in \{1,\ldots,d\}$ and write $\ddot{C}^\circ_{ij}$ as the second order partial derivative of C° with respect to the i'th and j'th coordinates. Suppose that for every $i, j \in \{1,\ldots,d\}$ the function $\ddot{C}^\circ_{ij}$ is defined and continuous on the set $V_{d,i} \cap V_{d,j}$ and there exists a constant $K > 0$ such that*

$$|\ddot{C}^\circ_{ji}(u)| \leq K \min\left(\frac{1}{u_i(1-u_i)}, \frac{1}{u_j(1-u_j)}\right), \quad u \in V_{d,i} \cap V_{d,j}.$$

Theorem 1. *If Condition 3 is fulfilled and the function $v \mapsto \log c(v, \theta^\circ)$ has finite Hardy-Krause variation, then*

$$\Delta_{m,n} = \int_{\mathbb{R}^d} \log c(F_\perp^\circ(x), \theta^\circ)\, \mathrm{d}[F_n - F^\circ](x) + \breve{Z}_n + O\left(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4}\right)$$

almost surely, where $\breve{Z}_n$ is a random variable with zero mean.

Proof. See Appendix A.3 in the supplementary notes. □

2.2. **The study of $\Delta_{m,n}$.** We now study $\Delta_{m,n}$ defined in eq. (27). If $u \mapsto \log c(u, \theta^\circ)$ is two times continuously differentiable, a two-term Taylor-expansion of each term in $A_n(\theta^\circ)$ around $F_{n,\perp}(X_i) - F_\perp^\circ(X_i)$ gives the fundamental relation

$$\Delta_{m,n} = A_n(\theta^\circ) - A(\theta^\circ) = \int \log c(F_\perp^\circ(x), \theta^\circ)\, \mathrm{d}[F_n - F^\circ](x) + \frac{1}{n}\left(\delta_{m,1,n} + \delta_{m,1,n}\right) + r_n \tag{28}$$

where the m-subscript indicates that the terms originates from the estimation process of the marginals. We have that

$$\delta_{m,1,n}/n = \frac{1}{n}\sum_{i=1}^n \zeta'(F_\perp^\circ(X_i), \theta^\circ)^t (F_{n,\perp}(X_i) - F_\perp^\circ(X_i)),$$

$$\delta_{m,2,n}/n = \frac{1}{2n}\sum_{i=1}^n (F_{n,\perp}(X_i) - F_\perp^\circ(X_i))^t \zeta''(F_\perp^\circ(X_i), \theta^\circ)(F_{n,\perp}(X_i) - F_\perp^\circ(X_i)),$$

in which

$$\zeta'(u,\theta) = \frac{\partial \log c(u,\theta)}{\partial u} \text{ and } \zeta''(u,\theta) = \frac{\partial^2 \log c(u,\theta)}{\partial u \partial u^t} \tag{29}$$

and finally

$$r_n = \frac{1}{2n}\sum_{i=1}^n (F_{n,\perp}(X_i) - F_\perp^\circ(X_i))^t \left[\zeta''(G_n(X_i), \theta^\circ) - \zeta''(F_\perp^\circ(X_i), \theta^\circ)\right](F_{n,\perp}(X_i) - F_\perp^\circ(X_i)), \tag{30}$$

where G_n is a vector function with entries $G_{n,i}(x) = F_i^\circ(x_i) + \tau_{n,i}(x)[F_{n,i}(x_i) - F_i(x_i)]$ for some stochastic vector $\tau_n(x) = (\tau_{n,1}, \ldots, \tau_{n,d}) \in (0,1)^d$.

Theorem 2 will give conditions for when r_n is $o_P(n^{-1})$, and thus considered low-level noise. Clearly, the first term of eq. (28) has zero mean, and it remains to find the expectation of the stochastically significant parts of $\delta_{m,1,n}$ and $\delta_{m,2,n}$. This is described by the following two lemmas, proved in Appendix A.1 of the Supplementary Notes.

Lemma 2. *We have the decomposition $\delta_{m,1,n} = \tilde{\delta}_{m,1,n} + Z_{1,n}$ where $\mathbb{E}Z_{1,n} = 0$ and*

$$\tilde{\delta}_{m,1,n} = \frac{n}{n+1}\int \zeta'(F_\perp^\circ(x), \theta^\circ)^t\, (\mathbf{1} - F_\perp^\circ(x))\, \mathrm{d}F_n(x)$$

and hence

$$\mathbb{E}\delta_{m,1,n} = \frac{n}{n+1}\int_{[0,1]^d} \zeta'(u, \theta^\circ)^t\, (\mathbf{1} - u)\, \mathrm{d}C^\circ(u).$$

Lemma 3. *Let $C_{a,b}$ be the copula of $(X_{1,a}, X_{1,b})$. We have $\mathbb{E}\delta_{m,2,n} \to \mathbf{1}^t\Upsilon\mathbf{1}$ where $\Upsilon = (\Upsilon_{a,b})_{1\le a,b\le d}$ is the symmetric matrix with*

$$\Upsilon_{a,a} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,a}(u,\theta^\circ) u_a(1-u_a)\, \mathrm{d}C^\circ(u),$$

$$\Upsilon_{a,b} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,b}(u,\theta^\circ)\left[C_{a,b}(u_a,u_b) - u_a u_b\right]\, \mathrm{d}C^\circ(u) \quad (\text{when } a \neq b).$$

Here $\zeta''_{a,b}$ are the elements of the matrix function ζ'' defined in eq. (29). Further, $\mathbb{E}\delta_{m,2,n}$ is finite only if $\mathbf{1}^t\Upsilon\mathbf{1}$ is.

This leads to the following result, based on certain growth assumptions of $u \mapsto \log c(u,\theta)$ near $\partial\left([0,1]^d\right)$. A discussion of these assumptions is given at the end of this sub-section.

Theorem 2. *If $u \mapsto \log c(u,\theta)$ is twice continuously differentiable on $(0,1)^d$ and if the conditions of Proposition 1 in Appendix A.2 are met, then*

$$\Delta_{m,n} = A_n(\theta^\circ) - A(\theta^\circ) = \frac{1}{n}(\delta_{m,1,n} + \delta_{m,2,n}) + \tilde{Z}_n + o_P(n^{-1}), \tag{31}$$

in which $\mathbb{E}\tilde{Z}_n = 0$ and

$$\delta_{m,1} := \lim_{n\to\infty} \mathbb{E}\delta_{m,1,n} = \int_{[0,1]^d} \zeta'(u,\theta^\circ)^t\,(\mathbf{1}-u)\;\mathrm{d}C^\circ(u) \tag{32}$$

$$\delta_{m,2} := \lim_{n\to\infty} \mathbb{E}\delta_{m,2,n} = \mathbf{1}^t\Upsilon\mathbf{1} \tag{33}$$

where $\mathbb{E}\delta_{m,1,n}$ and $\mathbb{E}\delta_{m,2,n}$ are infinite only if $\delta_{m,1}$ and $\delta_{m,2}$ respectively are infinite.

Proof. This is a direct consequence of Lemma 2, Lemma 3 and Proposition 1 in Appendix A.2. □

To recapitulate, we are now in the possession of the desired expansion of Δ_n of eq. (26). Under Condition 2 and the assumptions of Theorem 2, we have that

$$\Delta_n = \Delta_{c,n} + \Delta_{m,n} = \hat{Z}_n + \frac{1}{n}\left[\delta_{c,n} + \delta_{m,1,n} + \delta_{m,2,n}\right] + o_P(n^{-1}) \tag{34}$$

where $\mathbb{E}Z_n = 0$ and $\delta_{c,n}$ converges in distribution to a variable with mean $\delta_c = \mathrm{Tr}\left(J^{-1}\mathcal{I}\right) + \mathrm{Tr}\left(J^{-1}\mathcal{W}\right)$ defined in terms of the sizes defined in Section 2.1 and where $\delta_{m,1,n}$ and $\delta_{m,2,n}$ have asymptotic means given by $\delta_{m,1}$ and $\delta_{m,2}$ in equations (32) and (33) respectively.

As announced in the introduction, $\delta_{m,1}$ is usually finite but Υ usually has infinite elements which implies that $\delta_{m,2}$ is infinite. To illustrate this problem, let $d = 2$ and assume that the model is correctly specified, so that $c^\circ(u_1,u_2) = c(u_1,u_2,\theta^\circ)$ for $(u_1,u_2) \in [0,1]^2$. We then have

$$\zeta''_{a,b}(u,\theta^\circ) = \frac{\partial}{\partial u_b}\frac{\partial c^\circ(u)/\partial u_a}{c^\circ(u)} = \frac{\partial^2 c^\circ(u)/\partial u_a \partial u_b}{c^\circ(u)} - \frac{\left[\partial c^\circ(u)/\partial u_a\right]\left[\partial c^\circ(u)/\partial u_b\right]}{c^\circ(u)^2},$$

yielding

$$\Upsilon_{1,2} = \int_{[0,1]^2}\left[c^\circ(u_1,u_2) - \frac{[\partial c^\circ(u_1,u_2)/\partial u_1][\partial c^\circ(u_1,u_2)/\partial u_2]}{c^\circ(u_1,u_2)}\right][C^\circ(u_1,u_2) - u_1u_2]\;\mathrm{d}C^\circ(u_1,u_2),$$

$$\Upsilon_{1,1} = \int_{[0,1]^2}\left[c^\circ(u_1,u_2) - \frac{[\partial c^\circ(u_1,u_2)/\partial u_1][\partial c^\circ(u_1,u_2)/\partial u_1]}{c^\circ(u_1,u_2)}\right]u_1(1-u_1)\,\mathrm{d}C^\circ(u_1,u_2),$$

$$\Upsilon_{2,2} = \int_{[0,1]^2}\left[c^\circ(u_1,u_2) - \frac{[\partial c^\circ(u_1,u_2)/\partial u_2][\partial c^\circ(u_1,u_2)/\partial u_2]}{c^\circ(u_1,u_2)}\right]u_2(1-u_2)\,\mathrm{d}C^\circ(u_1,u_2).$$

Example 1. Consider the bivariate Kimeldorf & Sampson family of copulae with density

$$c(u_1,u_2,\delta) = \frac{1+\delta}{(u_1u_2)^{\delta+1}}\left(1/u_1^\delta + 1/u_2^\delta - 1\right)^{2+1/\delta}, \quad \delta \geq 0$$

which is copula B4 in Joe (1997, p. 141). The B4 density is simply a rational polynomial when $\delta = 1$. This enables us to give closed form expressions for $\Upsilon_{a,b}$ with the help of a computer algebra system, in contrast to most copula densities where numerical integration is needed to compute Υ.

We find that

$$\begin{aligned}\Upsilon_{1,2} &= \int_0^1 \left[\frac{1}{5}u_2^{-1} - \frac{3}{10}u_2 + \frac{1}{10}\right] \mathrm{d}u_2,\\ \Upsilon_{1,1} &= \int_0^1 \left[u_2^{-1} + \frac{1}{2}u_2^{-2} + \frac{3}{2}\right] u_2(1-u_2)\,\mathrm{d}u_2,\\ \Upsilon_{2,2} &= \int_0^1 \frac{1}{2}u_2^{-1}\,\mathrm{d}u_2.\end{aligned}$$

As $\int_0^1 u_2^{-1}\,\mathrm{d}u_2 = \infty$, we get that Υ, and hence also $\mathbb{E}\delta_{m,2,n}$, is infinite.

In fact, the B4 copula is not a pathology. Although it is typical that $\delta_{m,2,n} = O_P(1)$, it is also typical that $\mathbb{E}\delta_{m,2,n}$ is infinite. Almost all of the copula models categorized in Joe (1997) have infinite Υ-values, i.e. the distribution of $\delta_{m,2,n}$ has very heavy tails.

Let us now discuss the assumptions underlying Theorem 2. We see that the central size in the definition of r_n in eq. (30) is ζ'' of eq. (29). Hence, in order to prove that $r_n = o_P(n^{-1})$, we need to impose some growth conditions on ζ'' near the edge of the unit cube to avoid that r_n diverges. The assumptions we use, inspired by Ruymgaart et al. (1972) and Ruymgaart (1974), is that for certain sets of functions $\mathcal{Q}$ and $\mathcal{R}$, there exist functions $q_k \in \mathcal{Q}$ and $r_k, \tilde{r}_{k,l,1}, \tilde{r}_{k,l,2} \in \mathcal{R}$ such that

$$|\zeta''(u,\theta_0)| \le \tilde{r}_{a,b,1}(u_a)\tilde{r}_{a,b,2}(u_b) \prod_{1\le k\le d, k\neq a,b} r_k(u_k) \tag{35}$$

with

$$\int_{[0,1]^d} q_a(u_a)q_b(u_b)\tilde{r}_{a,b,1}(u_a)\tilde{r}_{a,b,2}(u_b) \prod_{1\le k\le d, k\neq a,b} r_k(u_k)\,\mathrm{d}C^\circ(u) < \infty. \tag{36}$$

Typical elements in $\mathcal{Q}$ and $\mathcal{R}$ are

$$q(t) = [t(1-t)]^\zeta, 0 < \zeta < 1/2, \quad r(t) = \rho[t(1-t)]^{-\zeta}, \zeta \ge 0, \rho \ge 0.$$

Hence, for all copula models $c(\cdot,\theta^\circ)$ for which there exists functions in $\mathcal{R}$ to secure eq. (35) – an assumption not depending on the true copula C° – the validity of eq. (36) is quite a lot weaker than the existence of Υ in Lemma 3.

We must, however, admit that similarly to previous investigations on copula models using the quite complicated assumptions of Ruymgaart et al. (1972) and Ruymgaart (1974), we have not conducted a detailed study that proves their validity for a selection of copula models. As Υ is usually infinite, our argument is that we have provided some assumptions securing that the reminder term r_n defined in eq. (30) is $o_P(n^{-1})$, and this conclusion is conjectured to be true also under weaker conditions than ours. Let us also indicate why it should be expected that r_n is $o_P(n^{-1})$. Because

$$2nr_n = \int \mathbb{G}_{n,\perp}(x)^t \left[\zeta''(G_n(X_i),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\right] \mathbb{G}_{n,\perp}(x)\,\mathrm{d}F_n(x),$$

where $\mathbb{G}_{n,\perp}$ is the vector of marginal empirical processes and G_n is defined immediately after eq. (30), it is expected that stochastic process techniques can be used to argue that $2nr_n$ is close to

$$\int W_\perp(x)^t \rho_n(x) W_\perp(x)\,\mathrm{d}F^\circ(x), \quad \rho_n(x) = \zeta''(G_n(x),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0), \tag{37}$$

where $W_\perp(x)$ is defined in terms of an F°-Brownian Bridge W through

$$W_\perp(x) = (W(\pi_1(x_1)), \ldots, W(\pi_d(x_d)))$$

where $\pi_i(x_i)$ maps x_i to $(\infty,\ldots,\infty,x_i,\infty,\ldots,\infty)$ with x_i as the i'th coordinate. Because $\rho_n(x)$ converges to zero uniformly in any compact set contained in $(0,1)^d$, some bounds on $\rho_n(x)$ near the

edge-set $\partial\left([0,1]^d\right)$ would provide the desired $2nr_n = o_P(1)$ based on the approximation indicated in eq. (37). However, if this argument were made precise, Υ would still be infinite for all popular copula models, and the general conclusion of our investigation would still apply.

We will briefly mention a way around these infinite expectation terms in Section 3 by using a weighted version of the MPLE, where the edge of the unit cube is given zero or small weight. The need for such weighting procedures indicates that the MPLE's use of marginal empirical distribution functions blinds the estimation routine from distinguishing between copula densities with different behavior near the edge of the unit cube – at the precision level prescribed by the AIC-programme. This is of practical interest as the MPLE is often used precisely in contexts where the behavior of the copula near the edge of the unit cube is of central interest. Our results can be interpreted as a demarcation for when this use is justified.

Let us finally mention that the finitude of Υ depends on both the least false copula $c(\cdot,\theta_0)$ and the true, unknown copula $c^\circ(\cdot)$. As the true copula is unknown, one cannot know if Υ is finite or not in a given investigation.

2.3. **Empirical estimates.** The CIC formulae now follow from eq. (34) when empirical estimates of the asymptotic expectation of $\delta_{c,n}, \delta_{m,1,n}$ and $\delta_{m,2,n}$ are found. Significant simplifications can be made when the model is assumed correct. This leads to a CIC formula that we call the AIC-like CIC formula, derived in Section 2.3.1. If the model is not assumed correct, nonparametric estimates are required and we get the so-called TIC-like CIC formula, given in Section 2.3.2.

2.3.1. *AIC-like formula.* This section works under the assumption of a correct model, as was the case for the original AIC formula. This assumption leads to several simplifications, as shown by the following result, whose proof is deferred to Appendix A.3.

Proposition 1. *If the parametric model is correctly specified, we have* $\delta_{m,1} = 0$ *and* $\delta_c = \mathrm{length}(\theta) + \mathrm{Tr}(\mathcal{I}^{-1}\mathcal{W})$*, where* $\mathcal{I}$ *and* $\mathcal{W}$ *is defined in equations* (19) *and* (20) *respectively.*

This motivates the AIC-like Copula Information Criterion

$$\mathrm{CIC} = 2\ell_{n,\max} - 2(\hat\delta_c + \hat\delta_{m,2}), \tag{38}$$

where $\hat\delta_c$ and $\hat\delta_{m,2}$ estimates δ_c and $\delta_{m,2}$ respectively.

An obvious estimator of $\delta_{m,2}$ is $\hat\delta_{m,2} = \mathbf{1}^t\hat\Upsilon\mathbf{1}$ where

$$\hat\Upsilon_{a,a} = \frac{1}{2}\int_{[0,1]^d} c(u,\hat\theta_n)\zeta''_{a,a}(u,\hat\theta_n)u_a(1-u_a)\,\mathrm{d}u,$$
$$\hat\Upsilon_{a,b} = \frac{1}{2}\int_{[0,1]^d} c(u,\hat\theta_n)\zeta''_{a,b}(u,\hat\theta_n)\left[C_{a,b}(u_a,u_b,\hat\theta_n) - u_au_b\right]\,\mathrm{d}u$$

in which $C_{a,b}(u_a,u_b,\hat\theta_n)$ is the cumulative copula of (Y_a,Y_b) when $(Y_1,Y_2,\ldots,Y_d)\sim C(u,\hat\theta_n)$. We estimate δ_c by

$$\hat\delta_c = \mathrm{length}(\theta) + \mathrm{Tr}\left(\hat{\mathcal{I}}^-\hat W\right)$$

denoting the generalized inverse of $\hat{\mathcal{I}}$ by $\hat{\mathcal{I}}^-$ and where $\hat{\mathcal{I}}$ is the pseudo empirical information matrix

$$\hat{\mathcal{I}} = \mathbb{E}_{\hat\theta_n}\phi(\tilde\xi,\hat\theta_n)\phi(\tilde\xi,\hat\theta_n)^t \tag{39}$$

estimating the information matrix $\mathcal{I}$ of eq. (19). Here $\phi(u,\theta) = (\partial/\partial\theta)\log c(u,\theta)$ as in eq. (17), and

$$\hat W = \mathrm{Var}_{\hat\theta_n}\left\{\int_{[0,1]^d}\left(\frac{\partial^2}{\partial\theta\partial u^t}\log c(u,\hat\theta_n)\right)^t\left(I\{\xi\le v\}_\perp - u\right)\mathrm{d}C(u,\hat\theta_n)\right\} \tag{40}$$

estimates W of eq. (20). The above covariance matrix is taken with respect to the random vector $\tilde{\xi} \sim C(v, \hat{\theta}_n)$, paralleling the random vector ξ in definition of Z in eq. (21). These integrals can be evaluated in practice through numerical integration routines such as Monte Carlo simulation. We could also use the rank based estimators

$$\hat{\mathcal{I}}^\star = \int_{u\in[0,1]^d} \phi(u,\hat{\theta}_n)\phi(u,\hat{\theta}_n)^t \,\mathrm{d}C_n(u) = \frac{1}{n}\sum_{k=1}^n \phi(\hat{\xi}^{(k)},\hat{\theta}_n)\phi(\hat{\xi}^{(k)},\hat{\theta}_n)^t$$

where $\hat{W}^\star$ as the empirical variance of

$$\int_{[0,1]^d} \left(\frac{\partial^2}{\partial\theta\partial u^t}\log c(u,\hat{\theta}_n)\right)^t (I\{\hat{\xi}^{(k)} \le v\}_\perp - u)\,\mathrm{d}C_n(u)$$

for $\hat{\xi}^{(k)} = F_{n,\perp}(X_k)$ together with analogues for $\hat{\delta}_{m,2}$. While $\hat{\xi}^{(k)}$ is simply the set of pseudo observations, note that it parallels $\tilde{\xi}$ above. An advantage with the rank-based estimators is that they do not require numerical integration. However, numerical integration needs only to be done once for a given copula model, in a grid of θ-values.

2.3.2. *TIC-like formula.* We now have to rely on nonparametric estimators. A natural estimator for $\delta_{m,1}$ is the plug-in estimator

$$\hat{\delta}_{m,1} = \int_{[0,1]^d} \zeta'(u,\hat{\theta}_n)^t\big(\mathbf{1}-u\big)\,\mathrm{d}\hat{C}_n(u)$$

while for $\delta_{m,2}$ we use $\hat{\delta}_{m,2} = \mathbf{1}^t\hat{\Upsilon}\mathbf{1}$, where now

$$\hat{\Upsilon}_{a,a} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,a}(u,\hat{\theta}_n)u_a(1-u_a)\,\mathrm{d}\hat{C}_n(u),$$
$$\hat{\Upsilon}_{a,b} = \frac{1}{2}\int_{[0,1]^d} \zeta''_{a,b}(u,\hat{\theta}_n)\left[\hat{C}_{n,a,b}(u_a,u_b) - u_a u_b\right]\,\mathrm{d}\hat{C}_n(u).$$

Here, $C_{n,a,b}$ is the empirical copula based on $(X_{1,a},X_{1,b}),(X_{2,a},X_{2,b}),\ldots,(X_{n,a},X_{n,b})$. As for the estimation of δ_c, we use $\hat{\delta}_c = \mathrm{Tr}\left(J_n^{-1}\hat{\Sigma}\right)$ where J_n is defined in eq. (25) and

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \left\{\phi(\hat{\xi}^{(i)},\hat{\theta}_n) + \hat{Z}_i\right\}\left\{\phi(\hat{\xi}^{(i)},\hat{\theta}_n) + \hat{Z}_i\right\}^t$$

with

$$\hat{Z}_i = \sum_{j=1}^d \frac{1}{n}\sum_{s=1,s\neq i}^n \left.\frac{\partial\phi(u,\hat{\theta}_n)}{\partial u_j}\right|_{u=\hat{\xi}^{(s)}} \left(I\left\{\hat{\xi}_j^{(i)} \le \hat{\xi}_j^{(s)}\right\} - \hat{\xi}_j^{(s)}\right)$$

using $\hat{\xi}^{(k)} = F_{n,\perp}(X_k)$.

2.4. **Confirmation of the CIC formula through simulation.** This section summarizes a small scale simulation study that confirms the validity of the CIC formula. Some additional numerical illustrations are given in Grønneberg (2010). We will study simulated samples with standard Normal marginals and a mixture copula with CDF $\lambda C_F(u,\theta) + (1-\lambda)C_P(u,\theta)$ with $\lambda = 80\%$. Here,

$$C_F(u,\theta) = C_F(u_1,u_2,\theta) = -\theta^{-1}\log\left([(1-e^{-\theta}) - (1-e^{-\theta u_1})(1-e^{-\theta u_2})]/(1-e^{-\theta})\right)$$

is the CDF of a Frank copula, while

$$C_P(u,\theta) = C_P(u_1,u_2,\theta) = \frac{1}{2}(\theta-1)^{-1}\big\{1+(\theta-1)(u_1+u_2) - \left[(1+(\theta-1)(u_1+u_2))^2 - 4\theta(\theta-1)u_1u_2\right]^{1/2}\big\}$$

is the CDF of a Plackett copula (see chapter 5.1. in Joe, 1997). We will let θ vary, but will use the same parameter value for both copulas.

Because of the Frank and Plackett copula densities' slow growth near $\partial\left([0,1]^2\right)$, the CIC formula exists. We want to use the known (near) unbiasedness of the AIC in the fully parametric case to illustrate that the CIC works as it should, and that the unmotivated AIC formula does not. We can do this by the following procedure.

When fitting parametric models with standard Normal marginals and either a Frank or a Plackett copula, our model is

$$f_i(x,y,\theta) = c_i\left(\Phi^{-1}(x),\Phi^{-1}(y),\theta\right)\phi(x)\phi(y), \quad i \in \{F,P\}$$

where Φ and ϕ is the CDF and density function of the standard Normal distribution. The true copula is known to be a mixture of the two. Denote this density by c°, and let f° be the full data-generating mechanism of (X,Y). We have

$$f^\circ(x,y) = c^\circ\left(\Phi^{-1}(x),\Phi^{-1}(y)\right)\phi(x)\phi(y).$$

By a change of variables, as in eq. (11), the Kullback–Leibler divergence between $f^\circ(\cdot)$ and $f_i(\cdot,\theta)$ is

$$\begin{aligned}\mathrm{KL}(f^\circ(\cdot), f_i(\cdot,\theta)) &= \mathbb{E}\log\frac{f^\circ(X,Y)}{f_i(X,Y,\theta)} = \mathbb{E}\log\frac{c^\circ\left(\Phi^{-1}(X),\Phi^{-1}(Y)\right)}{c_i\left(\Phi^{-1}(X),\Phi^{-1}(Y);\theta\right)}\\ &= \mathrm{KL}(c^\circ(\cdot), c_i(\cdot,\theta)).\end{aligned}$$

implying

$$\begin{aligned}\Delta\mathrm{KL}(f^\circ) &:= \mathrm{KL}(f^\circ, f_F(\cdot,\theta)) - \mathrm{KL}(f^\circ, f_P(\cdot,\theta))\\ &= \mathrm{KL}(c^\circ, c_F(\cdot,\theta)) - \mathrm{KL}(c^\circ, c_P(\cdot,\theta)).\end{aligned}$$

Hence, when estimating a fully specified probability model using the correct marginal models, the difference in AIC values when using the Frank or Plackett copula should be similar on average as to the correct copula-based model selection formula. We let θ vary in a grid of values between 5 and 24 and ran 500 simulations as described above, each with a sample size of 2000. The parameter configurations include weak positive to very strong positive dependence. Figure 1 shows how close the computed CIC-differences and unmotivated AIC-differences are to the correct AIC-differences calculated on the basis of a fully parametric model. The CIC-formula is better than the unmotivated AIC-formula, except for the copula models with weak dependence. When the dependence grows stronger, the unmotivated AIC formula drifts away from the correct value. Note that we used the AIC-like CIC formula, and similarly used the unmotivated AIC formula and not a unmotivated TIC formula (which we have not seen mentioned in the literature).

3. The use of other divergences

The exploding bias correction terms of the CIC are caused by the rapid growth of many copula densities near the edge of the unit cube. One way of reducing the effects of this is to down-weight the sensitivity of the pseudo likelihood near the edge of the unit cube. The standard Kullback-Leibler divergence between $c^\circ(u)$ and $c(u,\theta)$ can be written as

$$\begin{aligned}\mathrm{KL}(c^\circ(\cdot), c(\cdot,\theta)) &= \int_{[0,1]^d} c^\circ(u)\log(c^\circ(u)/c(u,\theta))\,\mathrm{d}u\\ &= \int_{[0,1]^d} c^\circ(u)\log\frac{c^\circ(u)}{c(u,\theta)} - \left[c^\circ(u) - c(u,\theta)\right]\,\mathrm{d}u\end{aligned}$$

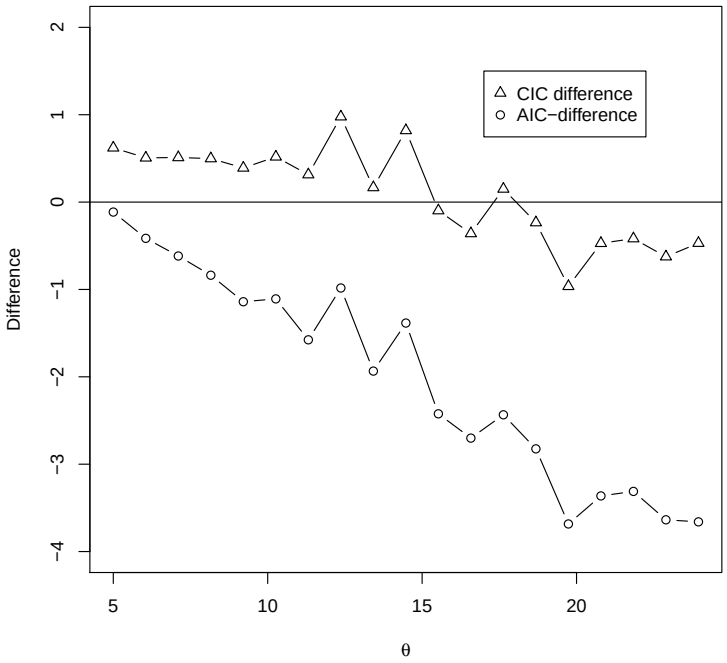

FIGURE 1. Simulated differences between fully parametric AIC and either the CIC or the unmotivated AIC.

as the integral of a density is unity. A natural, weighted generalization of the Kullback–Leibler divergence is

$$\mathrm{KL}_w(c^\circ(\cdot), c(\cdot,\theta)) = \int_{[0,1]^d} w(u) \times \left\{ c^\circ(u) \log \frac{c^\circ(u)}{c(u,\theta)} - [c^\circ(u) - c(u,\theta)] \right\} \mathrm{d}u,$$

where w is some non-negative weighting function. Notice that if $w \equiv 1$, one regains the usual Kullback–Leibler divergence.

This weighted Kullback–Leibler divergence is discussed in Claeskens & Hjort (2008, Section 2.10.2), and can be used to construct minimum divergence estimators and minimum divergence model selection methodologies in the same manner as the MPLE. Generalizing the ideas leading to the MPLE, a natural estimator for the weighted KL-least false parameter

$$\theta_w := \operatorname*{argmin}_{\theta\in\Theta} \mathrm{KL}_w(c^\circ(\cdot), c(\cdot,\theta)) = \operatorname*{argmax}_{\theta\in\Theta} \left[\int_{[0,1]^d} w(u) \log c(u,\theta)\, \mathrm{d}C^\circ(u) - \int_{[0,1]^d} w(u) c(u,\theta)\, \mathrm{d}u \right]$$

is

$$\hat\theta_w := \operatorname*{argmax}_{\theta\in\Theta} \left[\int_{[0,1]^d} w(u) \log c(u,\theta)\, \mathrm{d}C_n(u) - \int_{[0,1]^d} w(u) c(u,\theta)\, \mathrm{d}u \right].$$

One can then work out an AIC-like formula for the least attained *weighted* Kullback–Leibler divergence between the true copula $c^\circ(\cdot)$ and the candidate copula model. The weighting function will then enter the integrals defining the correction terms in a weighted version of the CIC formula, constructed through following the steps leading to Theorem 2. By defining the weighting function as zero, or near zero, near the edge $\partial[0,1]^d$, the weighted CIC formula always exists. It is clear that such a weighted estimation methodology can also be applicable for estimating misspecified copula models, by up-weighting regions of special importance.

In the abstract framework of Section 2, note that the existence problem for the CIC originates from the use of the pseudo likelihood $\Phi[C_n]$, and not merely from estimating $\theta^\circ = T(C^\circ)$ by the MPL estimator $\hat\theta_n = T(C_n)$. For the above sketched program to work, the model selection strategy must use the weighted model relevant KL-divergence, say, Φ_w and derive bias corrections for $\Phi_w[C_n](\log c_{\hat\theta_w})$. The existence problems of the CIC lies fundamentally in the use of the pseudo likelihood $\Phi[C_n]$, as can be seen from the definition of the problematic term $\Delta_{m,n}$.

The need for such a down-weighting scheme indicates that the MPLE is, from the AIC-perspective, unsuited for estimating copula models with log-densities that grow fast near the edge of the unit cube.

4. The Cross-Validation Copula Information Criterion

It is well known that the TIC formula is first order equivalent to a certain version of cross-validation. Indeed, for the ML estimator we have

$$n^{-1}\sum_{i=1}^n \log f(X_i, \hat\theta_{(i)}) = \mathrm{TIC} + o_P(1), \tag{41}$$

where $\hat\theta_{(i)}$ is the ML estimate

$$\hat\theta_{(i)} = \operatorname*{argmax}_\theta \sum_{j\neq i} \log f(X_j, \theta)$$

based on an *iid* sample without the i'th observation.

This section proves that such an equivalence is *not* present for the MPLE. The CIC formula, when it exists, is not first order equivalent to

$$\widehat{\mathrm{xv}}_n = n^{-1}\sum_{i=1}^n \log c(F_{n,\perp,(i)}(X_i), \hat\theta_{(i)}). \tag{42}$$

Here, $F_{n,\perp,(i)}$ is the (rescaled) marginal empirical distribution function and $\hat\theta_{(i)}$ is the MPLE, where both sizes are calculated using the observations $X_1, \ldots, X_n$, excluding X_i.

This non-equivalence is but a curiosity when seen by itself. However, it opens up for the construction of a generally applicable model selection formula. The cross-validation formula $\widehat{\mathrm{xv}}_n$ is well motivated despite not being equivalent to the CIC formula. This leads to what we call the xv-CIC formula, which finds an analytic approximation to $\widehat{\mathrm{xv}}_n$ directly. The xv-CIC is simple to calculate and is of general applicability.

Remark 1. We note a problem with the above definition of $\widehat{\mathrm{xv}}_n$. Recall that the MPLE works with the rescaled empirical distribution function to avoid evaluating $u \mapsto \log c(u, \theta)$ when $\max_{1\leq k\leq d} u_k = 1$. The above $\widehat{\mathrm{xv}}_n$-formula faces an analogous nuisance for the observations $X_{m,k} = \min_{1\leq i\leq n} X_{i,k}$. For these elements, we have that

$$F_{n,k,(m)}(X_{m,k}) = \frac{1}{n}\sum_{i\neq m} I\{X_{i,k} \leq X_{m,k}\} = 0.$$

Most copula log-densities are infinite or undefined at the edge of the unit cube so that each observation with an index in the set

$$I := \left\{1 \leq i \leq n : X_{m,k} = \min_{1\leq i\leq n} X_{i,k} \text{ for all } 1 \leq k \leq d\right\}$$

makes $\widehat{\mathrm{xv}}_n$ infinite. A simple redefinition is

$$\widehat{\mathrm{xv}}_n = n^{-1}\sum_{i=1}^n \log c(\tilde F_{n,\perp,(i)}(X_i), \hat\theta_{(i)})$$

where $\tilde{F}_{n,\perp,(i)}$ is the vector consisting of the functions

$$\tilde{F}_{n,k,(i)}(x_k) = \begin{cases} 1/n, & \text{if } x_k \leq \min_{i\leq j\leq n} X_{j,k}. \\ F_{n,k,(i)}(x_k), & \text{otherwise} \end{cases}$$

Other definitions are possible as well, paralleling the rôle of J_n and K_n in Ruymgaart et al. (1972), using their notation. Our proceeding discussion will be of an asymptotic nature, and the contribution of the at most d elements of I will be insignificant as $n \to \infty$.

The xv-CIC formula is based on the following Theorem, stated in terms of the following condition. The Theorem is proved in Appendix A.4 in the supplementary notes.

Condition 4. *The parametrization of the copula model is such that the statistical functional*

$$Q(F) = \operatorname*{argmax}_{\theta\in\Theta} \int_{[0,1]^d} \phi(u,\theta) + z(u,\theta)\,\mathrm{d}F(u)$$

has an influence function, where $\phi(u,\theta) = (\partial/\partial\theta)\log c(u,\theta)$ *and*

$$z(x,\theta) := \sum_{i=1}^{d} \int \frac{\partial\phi(u,\theta)}{\partial u_i}\left(I\{F^{\circ}_{\perp}(x_i) \leq u_i\}_{\perp} - u_i\right)\,\mathrm{d}C^{\circ}(u).$$

Also, we have the pointwise convergence

$$\hat{z}(x,\theta) := \sum_{k=1}^{d} \int \frac{\partial\phi(u,\theta)}{\partial u_k}\left(I\{x_k \leq u_k\} - u_k\right)\,\mathrm{d}C_n(u) \xrightarrow[n\to\infty]{\mathscr{P}} z(x,\theta)$$

for each x *and each* θ *in an open set around* θ°.

Theorem 3. *Given Conditions 2 and 4, suppose that the score function* U_n *of eq.* (18) *has the expansion*

$$U_n = \int_{[0,1]^d} \phi(u,\theta^{\circ}) + z(u,\theta^{\circ})\,\mathrm{d}[F_n - C^{\circ}](u) + o_P(n^{-1/2}). \tag{43}$$

If $\theta \mapsto \hat{z}(x,\theta)$ *is continuous around* θ°, *we have*

$$\widehat{\mathrm{xv}}_n = n^{-1}\Bigg[\ell_n(\hat{\theta}_n) - n^{-1}\sum_{i=1}^{n}\zeta'(F_{n,\perp}(X_i))^t\left(\mathbf{1}_d - F_{n,\perp}(X_i)\right) + \phi(F_{n,\perp}(X_i),\hat{\theta}_n)^t J^{-1}\phi(F_{n,\perp}(X_i),\hat{\theta}_n) + \phi(F_{n,\perp}(X_i),\hat{\theta}_n)^t J^{-1}\hat{z}(F_{n,\perp}(X_i),\hat{\theta}_n)\Bigg] + o_P(1).$$

Note that assumption A1 of Tsukahara (2005) secures the validity of eq. (43). Also, assuming pointwise convergence of $\hat{z}$ is very non-restrictive, as this convergence must take place to estimate Σ in Lemma 1.

This result motivates the definition of the cross-validation Copula Information Criterion formula

$$\text{xv-CIC} = 2\ell_{n,\max} - 2\left(\hat{\delta}_c + \hat{\delta}_m\right) \tag{44}$$

where

$$\hat{\delta}_c = n^{-1}\sum_{i=1}^{n}\phi(F_{n,\perp}(X_i),\hat{\theta}_n)^t J_n^{-1}\left(\phi(F_{n,\perp}(X_i),\hat{\theta}_n) + \hat{z}(F_{n,\perp}(X_i),\hat{\theta}_n)\right) \tag{45}$$

$$\hat{\delta}_m = n^{-1}\sum_{i=1}^{n}\zeta'(F_{n,\perp}(X_i),\hat{\theta}_n)^t\left(\mathbf{1}_d - F_{n,\perp}(X_i)\right) = \int \zeta'(u,\hat{\theta}_n)^t(\mathbf{1}_d - u)\,\mathrm{d}C_n. \tag{46}$$

More compactly, we may also write

$$\hat{\delta}_c = \mathrm{Tr}\left\{J_n^{-1}\left(\hat{\mathcal{I}} + \hat{K}\right)\right\}, \quad \hat{K} = \int_{[0,1]^d}\phi(u,\hat{\theta}_n)^t z(u,\hat{\theta}_n)\,\mathrm{d}C_n.$$

Note that it is xv-CIC/2 that is first order equivalent with the cross-validation sum of eq. (42). The factor two is included to maintain similarity with the classical AIC formula.

Analogous to the AIC-like CIC formula, simplifications can be made if the model is assumed correct. Indeed, if the model is correct, Proposition 1 implies that $\hat{\delta}_m = o_P(1)$ and

$$\hat{\delta}_c \xrightarrow[n\to\infty]{\mathscr{P}} \text{length}(\theta) + J^{-1}K, \quad K = \int_{[0,1]^d} \phi(u,\theta^\circ)^t z(u,\theta^\circ)\,\mathrm{d}C^\circ.$$

This motivates the definition of the AIC-like cross-validation Copula Information Criterion formula

$$\text{xv-CIC}_{\text{AIC}} = 2\ell_{n,\max} - 2\left(\text{length}(\theta) + J_n^{-1}\hat{K}\right). \tag{47}$$

Note that for $\xi \sim C^\circ$, we have $\mathbb{E}\phi(\xi,\theta^\circ) = 0$ and $\mathbb{E}z(\xi,\theta^\circ) = 0$, so the matrix K is the covariance matrix between $\phi(\xi,\theta^\circ)$ and $Z = z(\xi,\theta^\circ)$. When the model is correct and these random vectors are uncorrelated, the original AIC formula is well-motivated.

Note that if $J^{-1}K$ is small or zero, including the term $J_n^{-1}\hat{K}$ may for small samples only introduce noise into the estimation of the model-relevant KL-divergence. A simulation study could identify the cases where the original AIC-formula is preferred.

Also note that the xv-CIC formulas are motivated by asymptotic approximations of the cross-validation formula eq. (42), and is only valid for large n. Just how large n must be for the formula to become a very good approximation depends on both the data-generating mechanism and the parametric model under consideration. However, for small to medium sized n, one can simply calculate the precise cross-validation expression. As n grows, this becomes intractable. Also, if length(θ) is large, exact cross validation would require solving a very many possibly challenging numerical optimizations. If some of these optimizations do not identify the correct maximum pseudo likelihood solution, this may skew the cross-validation computation. The xv-CIC formula in contrast only requires a single numerical optimization.

5. An empirical example with the xv-CIC formula

We will briefly illustrate the xv-CIC formula on an insurance dataset of losses and allocated loss adjustment expenses (ALAE). The dataset was collected by the US insurance service office, and has been analyzed in several papers. We mention only Frees & Valdez (1998) and Genest et al. (2006), which considered the Clayton, Frank and Gumbel copulae as potential models for the data. All three models have length(θ) = 1. We will follow these authors in ignoring censoring and will further only consider the same copula models as they did.

	Gumbel	Frank	Clayton
$\ell_{n,\max}$	191.4180	161.1961	89.9494
xv	190.4832	160.1406	87.1065
xv-CIC	190.3810	160.1401	86.3736
Difference	0.1023	0.0005	0.7328

Table 1. Calculated cross-validation and xv-CIC scores for the Loss-ALAE data. All three models have length(θ) = 1.

The number of uncensored observations is $n = 1466$. Precise cross-validation is computationally intensive, but still possible to perform. In comparison, the xv-CIC formula is very fast to compute and yields good approximations to the full cross-validation procedure. Table 1 shows the maximized pseudo likelihoods, exact cross-validation scores and xv-CIC scores for the three models. While the match between the full cross-validation score and the xv-CIC is very good for the Frank and

Gumbel copula models, it is slightly less good for the Clayton copula. The final model-ranking agrees with the discussions in the previously cited papers. An R-script included in Section A.4 in the supplementary notes can be used to calculate xv-CIC in settings similar to the one above.

6. Conclusion

6.1. **Concluding remarks.** Our paper has studied the model selection problem for a semiparametric estimation procedure, and shown that it is significantly different compared to its fully parametric counterpart. In the MPLE case, the non-existence of a generally applicable AIC-like criterion is the price to pay for modeling the marginals non-parametrically.

It is well-known that the MPLE is not semiparametrically efficient in the sense of Bickel et al. (1993). However, Grønneberg (2010) argues that while the MPLE's lack of semiparametric efficiency in this sense is not a serious deficiency, its lack of a generally applicable AIC-like criterion is. When the model selection problem is relevant, semiparametric efficiency – which is defined under the assumption of a correct model – does not seem as important as the invariance properties that the MPLE fulfills, as discussed near eq. (11). The lack of a generally applicable AIC formula for the MPLE is not discoverable by the mere root-n normality as derived in e.g. Genest et al. (1995), but is a deeper property of the asymptotic behavior of the MPLE exposed in the present paper. It would be interesting to see if the lack of an AIC-like model selection procedure illustrates a typical feature of estimators based on maximizing a pseudo likelihood.

Of independent interest is the fact that the correction terms of the CIC can be both positive and negative. This means that the "likelihood minus penalty for complexity" interpretation of the AIC formula – often seen as a formalization of Occam's razor – is not a general principle but a consequence of linearity and smoothness properties of the two functionals Φ and T defining the maximum likelihood estimator. This is in contrast to several philosophically oriented discussions that mention the connection between simplicity and statistical model selection, as for example in Section 5 of Baker (2010).

6.2. **Recommendations for practitioners.** Due to the typical non-existence of an AIC-like model selection formula for the MPLE and the non-equivalence between the CIC and the xv-CIC, formally justified model selection procedures for the MPLE is complicated. However, a simple cross validation procedure usually makes sense from a practical point of view. In these cases, exact cross validation should be performed when n is small and the xv-CIC formula should be used when n is large. In most copula investigations, the candidate models are non-nested. In these cases, the TIC-like xv-CIC formula of eq. (44) is usually more appropriate than the AIC-like xv-CIC formula of eq (47).

In the supplementary note accompanying this paper, we have included a script for R (R Development Core Team, 2010) that implements the xv-CIC formula for a modest selection of copula models with $\text{length}(\theta) = 1$. When $\text{length}(\theta) > 1$, the expressions given in eq. (45) and eq. (46) must be found in order to calculate the xv-CIC formula of eq. (44).

Just how large n aught to be before the xv-CIC formula serves as a very good approximation to exact cross validation is still an open question, and is of particular importance for cases when $\text{length}(\theta)$ is large. In thise cases, exact cross validation is practically impossible, while the xv-CIC formula can be computed based on only a single numerical optimization.

Acknowledgements

We are grateful to two anonymous referees, an associate editor and editor Juha Alho for their careful reading of the manuscript and comments that greatly improved the presentation of the manuscript.

References

Baker, A. (2010). Simplicity. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed. Spring 2010 ed.

Bickel, P., Klaassen, A., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive inference in semi-parametric models*. Johns Hopkins University Press, Baltimore.

Chen, X. & Fan, Y. (2005). Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics* **33**, 389–414.

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

Fermanian, J., Radulović, D. & Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* **10**, 847–860.

Frees, E. W. & Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**.

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

Genest, C., Quessy, J.-F. & Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transform. *Scandinavian Journal of Statistics* **33**, 337–366.

Grønneberg, S. (2010). The copula information criterion and its implications for the maximum pseudo likelihood estimator. In *Dependence Modeling: Handbook on Vine Copulae*, D. Kurowicka & H. Joe, eds., chap. 6. World Scientific, pp. 131–163.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall.

Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer.

McNeil, A., Frey, R. & Embrechts, P. (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton Univ Pr.

Panchenko, V. (2005). Goodness-of-fit test for copulas. *Physica A: Statistical Mechanics and its Applications* **355**, 176–182.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics* **2**, 892–910.

Ruymgaart, F. H., Shorack, G. R. & van Zwet, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *The Annals of Mathematical Statistics* **43**, 1122–1135.

Segers, J. (2012). Weak convergence of empirical copula processes under nonrestrictive smoothness assumptions. *Bernoulli* **18**, 764–782.

Shao, J. (2003). *Mathematical Statistics*. Springer Texts in Statistics. Springer.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *The Canadian Journal of Statistics* **33**, 357–375.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer.

Department of Economics, BI Norwegian Business School, 0484 Oslo, Norway
E-mail address: steffeng@gmail.com

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway
E-mail address: nils@math.uio.no

SUPPLEMENTARY NOTES FOR "THE COPULA INFORMATION CRITERIA"

STEFFEN GRØNNEBERG AND NILS LID HJORT

Contents

Appendix A. Supplementary notes for "The Copula Information Criteria"

This appendix gathers technical proofs needed for the results of the main document. We also include a short R-script to calculate the xv-CIC formula when length$(\theta) = 1$. Our notation will follow the main document, but we will also work with the empirical processes

$$\mathbb{G}_{n,k}(x_k) = \sqrt{n}\left[F_{n,k}(x_k) - F_k^\circ(x_k)\right], \qquad \mathbb{G}_{n,\perp}(x) = \sqrt{n}\left[F_{n,\perp}(x) - F_\perp^\circ(x)\right],$$
$$\mathbb{G}_n(x) = \sqrt{n}\left[F_n(x) - F^\circ(x)\right], \qquad \mathbb{C}_n(u) = \sqrt{n}\left[C_n(u) - C^\circ(u)\right].$$

A.1. Proofs for Expectation Structure.

Proof of Lemma 2 in the main document. Define

$$\mathbb{G}_{n,\perp,-i} = \frac{\sqrt{n}}{n+1} \sum_{1\le k\le n, k\neq i} \left[I\{X_k \le x\} - F_\perp^\circ(x)\right]$$

so that $\mathbb{G}_{n,\perp}(x) = \mathbb{G}_{n,\perp,-i}(x) - \sqrt{n}/(n+1)\left[I\{X_i \le x\}_\perp - F_\perp(x)\right]$. This shows

$$\delta_{m,1,n}/n = \frac{1}{\sqrt{n}} \int \zeta'(F_\perp^\circ(x),\theta_0)^t \mathbb{G}_{n,\perp}(x)\,\mathrm{d}F_n(x) = \frac{1}{n\sqrt{n}} \sum_{i=1}^n \zeta'(F_\perp^\circ(X_i),\theta_0)^t \mathbb{G}_{n,\perp,-i}(X_i)$$
$$+ \frac{1}{n^2}\frac{n}{n+1} \sum_{i=1}^n \zeta'(F_\perp^\circ(X_i),\theta_0)^t \left[I\{X_i \le X_i\}_\perp - F_\perp(X_i)\right].$$

The second term is $\tilde{\delta}_{m,1,n}/n$. By independence, we have

$$\mathbb{E}\zeta'(F_\perp^\circ(X_i),\theta_0)^t \mathbb{G}_{n,\perp,-i,+1}(X_i) = \mathbb{E}\mathbb{E}\left[\zeta'(F_\perp^\circ(X_i),\theta_0)^t \mathbb{G}_{n,\perp,-i,+1}(X_i)\middle| X_i\right] = 0.$$

□

Date: June 1, 2013.

Proof of Lemma 3 in the main document. Notice that

$$\begin{aligned}
\delta_{m,2,n}/n =& \frac{1}{2n^2}\sum_{i=1}^n \mathbb{G}_{n,\perp}(X_i)^t \zeta''(F_\perp^\circ(X_i),\theta_0)\mathbb{G}_{n,\perp}(X_i)\\
=& \frac{1}{2n^2}\sum_{i=1}^n \mathbb{G}_{n,\perp,-i}(X_i)^t \zeta''(F_\perp^\circ(X_i),\theta_0)\mathbb{G}_{n,\perp,-i}(X_i)\\
&+ \frac{1}{2n^2}\frac{\sqrt{n}}{n+1}\sum_{i=1}^n \mathbb{G}_{n,\perp,-i}(X_i)^t \zeta''(F_\perp^\circ(X_i),\theta_0)\left[I\{X_i\le X_i\}_\perp - F_\perp(X_i)\right]\\
&+ \frac{1}{2n^2}\frac{\sqrt{n}}{n+1}\sum_{i=1}^n \left[I\{X_i\le X_i\}_\perp - F_\perp(X_i)\right]^t \zeta''(F_\perp^\circ(X_i),\theta_0)\mathbb{G}_{n,\perp,-i}(X_i)\\
&+ \frac{1}{2n^2}\left(\frac{\sqrt{n}}{n+1}\right)^2\sum_{i=1}^n \left[I\{X_i\le X_i\}_\perp - F_\perp(X_i)\right]^t \zeta''(F_\perp^\circ(X_i),\theta_0)\left[I\{X_i\le X_i\}_\perp - F_\perp(X_i)\right].
\end{aligned}$$

After multiplying with n, only the first term will have an effect on the expectation as $n\to\infty$. By independence, its expectation is given by

$$\begin{aligned}
&\frac{1}{2n}\mathbb{E}\int_{\mathbb{R}^d} \mathbb{G}_{n-1,\perp}(x)^t \zeta''(F_\perp^\circ(x),\theta_0)\mathbb{G}_{n-1,\perp}(x)\,\mathrm{d}F^\circ(x)\\
=&\frac{1}{n}\int_{\mathbb{R}^d} \mathbb{E}\left[\mathbb{G}_{n-1,\perp}(x)^t \zeta''(F_\perp^\circ(x),\theta_0)\mathbb{G}_{n-1,\perp}(x)\right]\,\mathrm{d}F^\circ(x)\\
=&\frac{1}{n}\sum_{1\le a,b\le d}\int_{\mathbb{R}^d} \zeta''_{a,b}(F_\perp^\circ(x),\theta_0)\mathbb{E}\left[\mathbb{G}^{(k)}_{n-1,a}(x_a)\mathbb{G}_{n-1,b}(x_b)\right]\,\mathrm{d}F^\circ(x).
\end{aligned}$$

Let $\rho_n = n^2/(n+1)^2$. We have

$$\begin{aligned}
\mathbb{E}\mathbb{G}_{n,a}(x_a)\mathbb{G}_{n,b}(x_b) &= \rho_n\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n I\{X_{i,k}\le x_k\} - F_k^\circ(x_k)\right]\left[\sum_{j=1}^n I\{X_{j,l}\le x_l\} - F_l^\circ(x_l)\right]\\
&= \rho_n\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[I\{X_{i,l}\le x_l\} - F_l^\circ(x_l)\right]\left[I\{X_{i,k}\le x_k\} - F_k^\circ(x_k)\right]\\
&\quad+\rho_n\frac{1}{n}\mathbb{E}\sum_{1\le i,j\le n, i\ne j}\left[I\{X_{i,k}\le x_k\} - F_k^\circ(x_k)\right]\left[I\{X_{j,l}\le x_l\} - F_l^\circ(x_l)\right].
\end{aligned}$$

The second term vanishes by independence, yielding

$$\begin{aligned}
\mathbb{E}\mathbb{G}_{n,a}(x_a)\mathbb{G}_{n,b}(x_b) &= \rho_n\frac{1}{n}\sum_{i=1}^n\Big\{\mathbb{E}\left[I\{X_{i,l}\le x_l\} - F_l^\circ(x_l)\right]I\{X_{i,k}\le x_k\}\\
&\qquad+\mathbb{E}\left[I\{X_{i,l}\le x_l\} - F_l^\circ(x_l)\right]F_k^\circ(x_k)\Big\}\\
&= \rho_n\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[I\{X_{i,l}\le x_l\}I\{X_{i,k}\le x_k\} - F_k^\circ(x_k)F_l^\circ(x_l)\right],
\end{aligned}$$

which is equal to $x_a(1-x_a)$ if $a=b$ and $P\{X_{1,l}\le x_l, X_{1,k}\le x_k\} - F_k^\circ(x_k)F_l^\circ(x_l)$ otherwise. Thus,

$$\begin{aligned}
&\frac{1}{2n}\mathbb{E}\int_{\mathbb{R}^d} \mathbb{G}_{n-1,\perp}(x)^t \zeta''(F_\perp^\circ(x),\theta_0)\mathbb{G}_{n-1,\perp}(x)\,\mathrm{d}F^\circ(x)\\
=&\rho_n\frac{1}{2n}\sum_{1\le a,b\le d, a\ne b}\int_{\mathbb{R}^d} \zeta''_{a,b}(F_\perp^\circ(x),\theta_0)\left[P\{X_{1,a}\le x_a, X_{1,}\le x_b\} - F_a^\circ(x_a)F_b^\circ(x_b)\right]\,\mathrm{d}F^\circ(x)\\
&+\rho_n\frac{1}{2n}\sum_{1\le a\le d}\int_{\mathbb{R}^d} \zeta''_{a,a}(F_\perp^\circ(x),\theta_0)x_a(1-x_a)\,\mathrm{d}F^\circ(x).
\end{aligned}$$

A change of variables shows that this is equal to

$$\rho_n \frac{1}{2n} \sum_{1\le a,b\le d, a\neq b} \int_{[0,1]^d} \zeta''_{a,b}(u,\theta_0)\left[C_{a,b}(u_a,u_b) - u_a u_b\right] \mathrm{d}C^\circ(u) \\ + \rho_n \frac{1}{2n} \sum_{1\le a\le d} \int_{[0,1]^d} \zeta''_{a,a}(u,\theta_0) u_a(1-u_a)\,\mathrm{d}C^\circ(u),$$

which approaches Υ once multiplied by n. □

Proof of Proposition 1 in the main document. The assumption $c^\circ(u) = c(u,\theta_0)$ validates the information matrix equality $J = \mathcal{I}$, which gives the reduced formula for δ_c. As for $\delta_{m,1}$, let us first notice that the fundamental theorem of calculus shows that

$$c(u,\theta_0)\Big|_{u_k=x} = \frac{\mathrm{d}}{\mathrm{d}x}\int_0^x c(u,\theta_0)\,\mathrm{d}u_k = \frac{\mathrm{d}}{\mathrm{d}x}\int_0^1 c(u,\theta_0) I\{0\le u_k\le x\}\,\mathrm{d}u_k.$$

As $c(u,\theta_0)I\{0\le u_k\le x\}$ is dominated by $c(u,\theta_0)$ which is integrable, dominated convergence allows us to move the differential sign in and out of integrals. As $c(u,\theta_0)$ has uniform marginals, this shows

$$\int_0^1\int_0^1\cdots\int_0^1 c(u,\theta_0)\Big|_{v_k=x}\prod_{i\neq k}\mathrm{d}u_i = \frac{\mathrm{d}}{\mathrm{d}x}\int_0^1\int_0^1\cdots\int_0^1\int_0^x c(u,\theta_0)\,\mathrm{d}u_k\prod_{i\neq k}\mathrm{d}u_i = \frac{\mathrm{d}}{\mathrm{d}x}x = 1. \tag{1}$$

We have

$$\begin{aligned}\delta_{m,1} &= \int_{[0,1]^d} \zeta'(u,\theta_0)^t\big(\mathbf{1}-u\big)\,\mathrm{d}C(u,\theta_0)\\ &= \sum_{k=1}^d \int_0^1\int_0^1\cdots\int_0^1 c(u,\theta_0)\frac{\partial\log c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k\prod_{i\neq k}\mathrm{d}u_i\\ &= \sum_{k=1}^d \int_0^1\int_0^1\cdots\int_0^1 \frac{\partial c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k\prod_{i\neq k}\mathrm{d}u_i.\end{aligned}$$

Let $\varepsilon>0$, and write

$$\int_0^1 \frac{\partial c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k = \int_\varepsilon^{1-\varepsilon}\frac{\partial c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k + \int_{[0,1]\setminus(\varepsilon,1-\varepsilon)}\frac{\partial c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k.$$

The first term can be written as

$$\begin{aligned}c(u,\theta_0)(1-u_k)\Big|_{u_k=\varepsilon}^{1-\varepsilon} + \int_\varepsilon^{1-\varepsilon} c(u,\theta_0)\,\mathrm{d}u_k &= c(u,\theta_0)\Big|_{u_k=1-\varepsilon}\varepsilon - c(u,\theta_0)\Big|_{u_k=\varepsilon}(1-\varepsilon)\\ &\quad + \int_\varepsilon^{1-\varepsilon} c(u,\theta_0)\,\mathrm{d}u_k\\ &= c(u,\theta_0)\Big|_{u_k=1-\varepsilon}\varepsilon + c(u,\theta_0)\Big|_{u_k=\varepsilon}\varepsilon - c(u,\theta_0)\Big|_{u_k=\varepsilon}\\ &\quad + \int_\varepsilon^{1-\varepsilon} c(u,\theta_0)\,\mathrm{d}u_k,\end{aligned}$$

through partial integration. By eq. (1), we get

$$\begin{aligned}\delta_{m,1} &= \sum_{k=1}^d \int_0^1\int_0^1\cdots\int_0^1\int_{[0,1]\setminus(\varepsilon,1-\varepsilon)}\frac{\partial c(u,\theta_0)}{\partial u_k}(1-u_k)\,\mathrm{d}u_k\prod_{i\neq k}\mathrm{d}u_i\\ &\quad + 2\varepsilon d - d + \sum_{k=1}^d\int_0^1\int_0^1\cdots\int_0^1\int_\varepsilon^{1-\varepsilon} c(u,\theta_0)\,\mathrm{d}u_k\prod_{i\neq k}\mathrm{d}u_i,\end{aligned}$$

which can be made arbitrarily close to zero by choosing ε sufficiently small. Thus $\delta_{m,1}=0$. □

A.2. **Sufficient conditions for $r_n = o_P(n^{-1})$.** To prove that $r_n = o_P(n^{-1})$, we need to impose some bounds on the edge-behavior of $u \mapsto \zeta''(u, \theta_0)$. We will follow the papers of Genest et al. (1995) and Tsukahara (2005) by using the theory from Ruymgaart et al. (1972) and Ruymgaart (1974).

Definition 1. (1) *Let $\mathcal{Q}$ be the set of continuous functions q on $[0,1]$, which are positive on $(0,1)$, symmetric about $1/2$, decreasing on $[0,1/2]$ and satisfy $\int_0^1 \{q(t)\}^2 \,\mathrm{d}t < \infty$.*

(2) *A function $r : (0,1) \mapsto (0,\infty)$ is called u-shaped if it is symmetric about $1/2$ and decreasing on $(0,1/2]$.*

(3) *For $0 < \beta < 1$ and a u-shaped function r, we define*

$$r_\beta(t) = \begin{cases} r(\beta t), & \text{if } 0 < t \le 1/2; \\ r\left(1 - \beta[1-t]\right), & \text{if } 1/2 < t \le 1 \end{cases}.$$

If for every $\beta > 0$ in a neighborhood of 0, there exists a constant M_β, such that $r_\beta \le M_\beta r$ on $(0,1)$, then r is called a reproducing u-shaped function. We denote by $\mathcal{R}$ the set of reproducing u-shaped functions.

The importance of $\mathcal{Q}$ and $\mathcal{R}$ comes from the following two lemmas, proved in Pyke & Shorack (1968) and Ruymgaart (1974) respectively.

Lemma 1. *Suppose $q_k \in \mathcal{Q}$, then $\|\mathbb{G}_{n,k}/q_k\| = O_P(1)$ where $\mathbb{G}_{n,k}$ is the k'th univariate empirical process.*

Lemma 2. *Suppose $H_{n,k}$ satisfies*

$$\min\left(F_k^\circ(x_k), \frac{1}{n+1}\sum_{i=1}^n I\{X_{i,k} \le x_k\}\right) \le H_{n,k}(x_k) \le \max\left(F_k^\circ(x_k), \frac{1}{n+1}\sum_{i=1}^n I\{X_{i,k} \le x_k\}\right)$$

for all x_k and let $\Lambda_{n,k} = [\min_{1\le i\le n} X_{i,k}, \max_{1\le i\le n} X_{i,k}] \subset \mathbb{R}$. Let $r \in \mathcal{R}$. Then

$$\sup_{x_k \in \Lambda_{n,k}} \frac{r(H_{n,k}(x_k))}{r(F_k^\circ(x_k))} = O_P(1)$$

uniformly in n.

Condition 1. *Suppose $\zeta''(u,\theta_0)$ is continuous, and for each $1 \le k \le d$ and $1 \le a, b \le d$ there exists functions $r_k, \tilde{r}_{k,l,1}, \tilde{r}_{k,l,2} \in \mathcal{R}$, and $q_k \in \mathcal{Q}$ such that*

$$|f(u)| \le \tilde{r}_{a,b,1}(u_a)\tilde{r}_{a,b,2}(u_b) \prod_{1\le k\le d, k\ne a,b} r_k(u_k) \tag{2}$$

where

$$\int_{[0,1]^d} q_a(u_a) q_b(u_b) \tilde{r}_{a,b,1}(u_a)\tilde{r}_{a,b,2}(u_b) \prod_{1\le k\le d, k\ne a,b} r_k(u_k) \,\mathrm{d}C^\circ(u) < \infty. \tag{3}$$

To slightly reduce the complexity of our notation, we here assume that $X_1, X_2, \ldots \sim C^\circ$ so that $F_\perp^\circ(x) = x$. By Lemma 1 of Fermanian et al. (2004) this does not entail any loss of generality.

Proposition 1. *If $u \mapsto \zeta''(u,\theta_0)$ conforms to Condition 1, then $r_n = o_P(n^{-1})$.*

Proof. Note that

$$r_n = \frac{1}{2n^2}\sum_{i=1}^n \mathbb{G}_{n,\perp}(X_i)^t \left[\zeta''(G_n(X_i),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\right] \mathbb{G}_{n,\perp}(X_i).$$

For each $0 < \gamma < 1$, let $S_\gamma = [\gamma, 1-\gamma]^d$ and $S_\gamma^c = [0,1]^d \setminus S_\gamma$. Write

$$\begin{aligned} 2nr_n = &\int_{S_\gamma} \mathbb{G}_{n,\perp}(x)^t \left[\zeta''(G_n(x),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\right] \mathbb{G}_{n,\perp}(x)\, \mathrm{d}F_n(x) \\ &+ \int_{S_\gamma^c} \mathbb{G}_{n,\perp}(x)^t \left[\zeta''(G_n(x),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\right] \mathbb{G}_{n,\perp}(x)\, \mathrm{d}F_n(x), \end{aligned}$$

and denote these integrals by $D_{n,1,\gamma}$ and $D_{n,2,\gamma}$. The absolute value of $D_{n,1,\gamma}$ is bounded by

$$d \sup_{1\le k,l\le d} \left[\|\mathbb{G}_{n,k}\|_{[\gamma,1-\gamma]}\right] \times \|\mathbb{G}_{n,l}\|_{[\gamma,1-\gamma]} \times \|\zeta''(G_n(x),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\|_{S_\gamma},$$

where $\|\cdot\|_E$ is the appropriate sup-norm constrained to the set E. As

$$\|G_n - F_\perp^\circ\| = \|\tau_n[F_{n,\perp} - F_\perp^\circ]\| \le \max_{1\le k\le d} \|\tau_{n,k}\| \|F_{n,\perp} - F_\perp^\circ\| \le \|F_{n,\perp} - F_\perp^\circ\| = o_P(1)$$

by the Glivenko-Cantelli theorem, the assumed continuity of ζ'' on $(0,1)^d$ implies that ζ'' is uniformly continuous on S_γ. Hence, $\|\zeta''(G_n(x),\theta_0) - \zeta''(F_\perp^\circ(x),\theta_0)\| = o_P(1)$. As $\|\mathbb{G}_{n,k}\| = O_P(1)$, this shows $D_{n,1,\gamma} = o_P(1)$. As for $D_{n,2,\gamma}$, its absolute value is bounded by

$$\begin{aligned} \left\|\frac{\mathbb{G}_{n,a}}{q_a}\right\| \left\|\frac{\mathbb{G}_{n,b}}{q_b}\right\| \Bigg[&\int_{S_\gamma^c} \left|q_a(x_a)\zeta''_{a,b}(G_n(x),\theta_0)q_b(x_b)\right| \mathrm{d}F_n(x) \\ &+ \int_{S_\gamma^c} \left|q_a(x_a)\zeta''_{a,b}(F_\perp^\circ(x),\theta_0)q_b(x_b)\right| \mathrm{d}F_n(x)\Bigg], \end{aligned}$$

which by eq. (2) is bounded by

$$\begin{aligned} \left\|\frac{\mathbb{G}_{n,a}}{q_a}\right\| \left\|\frac{\mathbb{G}_{n,b}}{q_b}\right\| \Bigg[&\int_{S_\gamma^c} q_a(x_a)q_b(x_b)\tilde{r}_{a,b,1}(\tilde{x}_a)\tilde{r}_{a,b,2}(\tilde{x}_b) \prod_{1\le k\le d, k\ne a,b} r_k(\tilde{x}_k)\, \mathrm{d}F_n(x) \\ &- \int_{S_\gamma^c} q_a(x_a)q_b(x_b)\tilde{r}_{a,b,1}(x_a)\tilde{r}_{a,b,2}(x_b) \prod_{1\le k\le d, k\ne a,b} r_k(x_k)\, \mathrm{d}F_n(x)\Bigg], \end{aligned}$$

where $\tilde{x}_k = F_{n,\perp}(1,\ldots,1,x_k,1,\ldots,1)$. By Lemma 1, we have $\|\mathbb{G}_{n,a}/q_a\|\|\mathbb{G}_{n,b}/q_b\| = O_P(1)$. It thus suffices to bound

$$\begin{aligned} D_{n,2,\gamma}(a,b,k,l) &:= \int_{S_\gamma^c} q_a(x_a)q_b(x_b)\tilde{r}_{a,b,1}(\tilde{x}_a)\tilde{r}_{a,b,2}(\tilde{x}_b) \prod_{1\le k\le d, k\ne a,b} r_k(\tilde{x}_k)\, \mathrm{d}F_n(x), \\ \tilde{D}_{n,2,\gamma}(a,b,k,l) &:= \int_{S_\gamma^c} q_a(x_a)q_b(x_b)\tilde{r}_{a,b,1}(x_a)\tilde{r}_{a,b,2}(x_b) \prod_{1\le k\le d, k\ne a,b} r_k(x_k)\, \mathrm{d}F_n(x). \end{aligned}$$

By Lemma 2, there exists a constant $M_\varepsilon > 0$ such that

$$\tilde{\Omega}_\varepsilon = \left\{\tilde{r}_{a,b,1}(\tilde{x}_a)\tilde{r}_{a,b,2}(\tilde{x}_b) \prod_{1\le k\le d, k\ne a,b} r_k(\tilde{x}_k) \le M_\varepsilon \tilde{r}_{a,b,1}(x_a)\tilde{r}_{a,b,2}(x_b) \prod_{1\le k\le d, k\ne a,b} r_k(x_k)\right\},$$

with $P(\tilde{\Omega}_\varepsilon) > 1-\varepsilon$ for all n. On $\tilde{\Omega}_\varepsilon$, we have $D_{n,2,\gamma}(a,b,k,l) \le M_\varepsilon \tilde{D}_{n,2,\gamma}(a,b,k,l)$. As ε is arbitrary, it suffices to bound $\tilde{D}_{n,2,\gamma}(a,b,k,l)$. We have

$$\mathbb{E}\left[|\tilde{D}_{n,2,\gamma}|\right] \le \int_{S_\gamma^c} q_a(x_a)q_b(x_b)\tilde{r}_{a,b,1}(x_a)\tilde{r}_{a,b,2}(x_b) \prod_{1\le k\le d, k\ne a,b} r_k(x_k)\, \mathrm{d}F^\circ(x).$$

By the integrability assumption in eq. (3), this expectation converges to zero by the Dominated Convergence Theorem. □

A.3. **Proof of Theorem 3.** The proof of Theorem 3 requires a partial integration result for multivariate Lebesgue-Stieltjes-integrals. Such results does not seem to be well-known in the statistics literature, and therefore we include the following brief description of its main components.

Thanks to the Riesz-representation Theorem and the theory of stochastic integration, functions of finite variation are of fundamental importance in mathematics. Their multivariate counterparts are, however, lesser known. In the multivariate case, several possible definitions of variation is possible. Good references on multivariate variation is Niederreiter (1992) and Owen (2005). We will need the variational concept of Hardy and Krause, which is defined in terms of the Vitali variation.

These variational concepts will be defined in terms of the set $\mathcal{P}$ of all sequences $x_j^0, x_j^1, \ldots, x_j^{m(j)}$ where $1 \leq j \leq d$ and $m(j) \in \mathbb{N}$, such that

$$0 = x_j^0 < x_j^1 < x_j^2 < \ldots < x_j^{m(j)} = 1.$$

In terms of an element of $\mathcal{P}$, define the difference operators

$$\begin{aligned}\Delta_j f(x_1, x_2, \ldots, x_{j-1}, x_j^k, x_{j+1}, \ldots, x_d) \\ = f(x_1, x_2, \ldots, x_{j-1}, x_j^{k+1}, x_{j+1}, \ldots, x_d) - f(x_1, x_2, \ldots, x_{j-1}, x_j^k, x_{j+1}, \ldots, x_d)\end{aligned}$$

and

$$\Delta_j^* f(x_1, x_2, \ldots, x_d) = f(x_1, x_2, \ldots, x_{j-1}, 1, x_{j+1}, \ldots, x_d) - f(x_1, x_2, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_d).$$

Note that this difference operator is clearly unrelated to the difference between the estimated and attained model-relevant KL-divergence of central importance in the main document.

These operators commute, which enables the definition of the composite difference operators

$$\begin{aligned}\Delta_{j(1),j(2),\ldots,j(k)} &= \Delta_{j(1)} \Delta_{j(2)} \cdots \Delta_{j(k)} \\ \Delta^*_{j(1),j(2),\ldots,j(k)} &= \Delta^*_{j(1)} \Delta^*_{j(2)} \cdots \Delta^*_{j(k)}.\end{aligned}$$

Definition 2. *The Vitali-variation of a function $f : [0,1]^d \mapsto \mathbb{R}$ is defined as*

$$V(f) := \sup_{x \in \mathcal{P}} \sum_{j(1)=0}^{m(1)-1} \cdots \sum_{j(d)=0}^{m(d)-1} |\Delta_{1,\ldots,d} f(x_1^{j(1)}, x_2^{j(2)}, \ldots, x_d^{j(d)})|.$$

If $V(f) < \infty$, the function f is said to be of finite Vitali variation.

Theorem 52.2 of McShane (1947) shows that there is an one to one correspondence between regular Borel measures μ on $[0,1]^d$, and functions of bounded Vitali-variation on $[0,1]^d$. This indicates that Vitali-variation is the right variational concept for multivariate integration. However, for a general multivariate Lebesgue–Stieltjes integration by parts formula to be applicable, the following stronger variational concept is required.

Definition 3. *The Hardy–Krause-variation of a function $f : [0,1]^d \mapsto \mathbb{R}$ is defined as the sum of Vitali variation when f is restricted to the various faces of $[0,1]^d$. More precisely, Hardy–Krause-variation of f is given by*

$$\mathbb{V}(f) := \sum_{(i,j) \in S} V(\pi_{i,j} f),$$

in which the index set S is defined y

$$S = \left\{ (i,j) : i \in \{0,1\}^d, j \in \{0,1\}^d, \sum_{k=1}^d i_k - j_k = 0 \right\},$$

and where $V(\pi_{i,k}f)$ is the Vitali-variation of $\pi_{i,j}f$ in the appropriate dimension. Here $\pi_{i,j}$ is the evaluation operator such that $f(x_1, x_2, \ldots, x_d)$ is mapped to

$$f(x_1, x_2, \ldots, x_d)\Big|_{x_{i(1)}=1,\ldots,x_{s(i)}=1,x_{j(1)}=0,\ldots,x_{s(j)}=0},$$

in which $i(k)$ $j(k)$ is the elements which are one, and $s(i), s(j)$ are the number of elements in i and j which are one.

Following Zaremba (1968), we now state the integration by parts formula. This formula makes it apparent why bounded Hardy–Krause-variation is required for its validity. For its formulation, suppose $\phi(r, \ldots, r+k-1; r+k, \ldots, s)$ is an expression which depend only on the partition of the variables $j(r), \ldots, j(s)$ into the sets $\{j(r), \ldots, j(r+k-1)\}$ and $\{j(r+k), \ldots, j(s)\}$. We will let

$$\sum_{1,\ldots,d;k}^{*} \phi(r, \ldots, r+k-1; r+k, \ldots, s)$$

be the sum over all the expressions derived from $\phi(r, \ldots, r+k-1; r+k, \ldots, s)$ by replacing the given partition of $\{j(r), \ldots, j(r+k-1)\}$ successively by all the other partitions of this set into a set of k and a set of $s-r-k-1$, each partition being taken exactly once. This is only meaningful if $0 < k < s-r+1$. If either $k = 1$ or $k = s-r+1$, there is strictly speaking no partition. We then define this sum as being reduced to the single valid term. Finally, let $d_{j(1),\ldots,j(k)}V$ indicate that integration applies only to the variables with subscripts $j(1), \ldots, j(k)$, the other variables being kept constant in the process of integration. Thus, for instance,

$$\Delta_3^* \int_{[0,1]^2} g(x_1, x_2, x_3)\, \mathrm{d}_{1,2}(x_1, x_2, x_3) = \int_{[0,1]^2} g(x_1, x_2, 1)\, \mathrm{d}(x_1, x_2, 1) - \int_{[0,1]^2} g(x_1, x_2, 0)\, \mathrm{d}(x_1, x_2, 0).$$

Lemma 3. *If over $[0,1]^d$, one of the functions $f(x)$ and $g(x)$ is of bounded Hardy–Krause-variation and the other is Lebesgue–Stieltjes integrable with respect to the other, then*

$$\int_{[0,1]^d} f(x)\, \mathrm{d}g(x) = \sum_{k=0}^{d} (-1)^k \sum_{1,2,\ldots,d;k}^{*} \Delta^*_{k+1,\ldots,d} \int_{[0,1]^k} g(x)\, \mathrm{d}_{1,\ldots,k} f(x).$$

Proof. This was first proved in Zaremba (1968) for the Riemann–Stieltjes integral, which gave a straight-forward constructive proof based on iterative use of the univariate summation by parts formula. Lee (2008) gave a nice constructive proof using the Henstock-Kurzweil integral, which implies the formula for the Lebesgue–Stieltjes integral. □

Using the above theory, we can prove Theorem 3.

Proof of Theorem 3. By Lemma 1 of Fermanian et al. (2004), we assume $X_i \sim C^\circ$, so that $F^\circ_\perp(x) = x$. Proposition 4.2 of Segers (2012) reaches

$$\begin{aligned} \mathbb{C}_n(u) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I\{X_i \le u\} - C^\circ(u)] \\ &\quad - \sum_{k=1}^{d} \frac{C^\circ(u)}{\partial u_k} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I\{F_k^\circ(X_{i,k}) \le u_k\} - u_k] + S_n(u). \end{aligned} \tag{4}$$

where $\sup_u |S_n(u)| = O(n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4})$ almost surely, After dividing with $\sqrt{n}$ on both sides of eq (4), the linearity of Stieltjes-integrals shows that

$$\begin{aligned} A_n(\theta_0) - A(\theta_0) = &\int_{[0,1]^d} \log c(u,\theta_0)\,\mathrm{d}[F_n - F^\circ](u) \\ &+ \frac{1}{\sqrt{n}} \sum_{k=1}^d \int_{[0,1]^d} \log c(u,\theta_0)\,\mathrm{d}\left[\frac{C^\circ(u)}{\partial u_k}\{F_{n,k}(u_k) - u_k\}\right] \\ &+ \frac{1}{\sqrt{n}} \int_{[0,1]^d} \log c(u,\theta_0)\,\mathrm{d}S_n(u). \end{aligned}$$

Let the middle integral in the display above be denoted by $\breve{Z}_n$. Note that although the last integral is with respect to S_n, we do not need to assume that S_n itself has finite variation: As S_n is almost surely uniformly bounded for sufficiently large n, it suffices that the variation of $u \mapsto \log c(u,\theta_0)$, that is $\mathbb{V}(\log c(u,\theta_0))$, is finite. This follows from the change of variables formula. We get

$$\frac{1}{\sqrt{n}} \sup_{u\in[0,1]^d} |S_n(u)| \mathbb{V}(\log c(u,\theta_0)) = O\left(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4}\right).$$

The change of variable formula shows

$$\begin{aligned} \mathbb{E}\breve{Z}_n =& \mathbb{E}\frac{1}{\sqrt{n}} \sum_{k=1}^d \int_{[0,1]^d} \log c(u,\theta_0)\,\mathrm{d}\left[\frac{C^\circ(u)}{\partial u_k}\{F_{n,k}(u_k) - u_k)\}\right] \\ =& \mathbb{E}\frac{1}{\sqrt{n}} \sum_{k=1}^d \sum_{l=0}^d (-1)^l \sum_{1,\ldots,d;l}^* \Delta^*_{l+1,\ldots,d} \int_{[0,1]^k} \frac{C^\circ(u)}{\partial u_k} \{F_{n,k}(u_k) - u_k\}\, \mathrm{d}_{1,\ldots,k} \log c(u,\theta_0), \end{aligned}$$

which is equal to

$$\frac{1}{\sqrt{n}} \sum_{k=1}^d \sum_{l=0}^d (-1)^l \sum_{1,\ldots,d;l}^* \Delta^*_{l+1,\ldots,d} \int_{[0,1]^k} \frac{C^\circ(u)}{\partial u_k} \mathbb{E}\{F_{n,k}(u_k) - u_k\}\, \mathrm{d}_{1,\ldots,k} \log c(u,\theta_0)$$

by Fubini. As $F_{n,k}(1) - 1 = F_{n,k}(0) - 0 = 0$, and as $\mathbb{E}F_{n,k}(u_k) - u_k = 0$, the above integral is zero. □

A.4. **Proof of the validity of the xv-CIC.** We now prove Theorem 3 in the main document.

The following Lemma is used to derive the influence function of the MPLE.

Lemma 4. *Given the conditions of Theorem 3, the MPLE has an influence function equal to* $J^{-1}s(y,\theta^\circ)$ *where* $s(x,\theta) = \phi(x,\theta) + z(x,\theta)$.

Proof. Again we apply Lemma 1 of Fermanian et al. (2004) to assume that $F^\circ = C^\circ$, so that $F^\circ_\perp(x) = x$. We have

$$\sqrt{n}(\hat{\theta}_n - \theta^\circ) = \sqrt{n} J_n^{-1} \left(\int_{[0,1]^d} \phi(u,\theta^\circ) + z(u,\theta^\circ)\,\mathrm{d}[F_n - C^\circ](u) \right) + o_P(1). \tag{5}$$

Consider the fundamentally unobservable M-estimator

$$\tilde{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} \int_{[0,1]^d} \phi(u,\theta) + z(u,\theta)\,\mathrm{d}F_n(u).$$

Fubini shows that

$$\begin{aligned} \int_{[0,1]^d} z(u,\theta)\,\mathrm{d}F^\circ(u) &= \int_{[0,1]^d} \int_{[0,1]^d} \frac{\partial\phi(v,\theta)}{\partial v_i} \left(I\{u_i \le v_i\}_\perp - v_i\right)\,\mathrm{d}C^\circ(v)\,\mathrm{d}F^\circ(u) \\ &= \int_{[0,1]^d} \frac{\partial\phi(v,\theta)}{\partial v_i} \int_{[0,1]^d} \left(I\{u_i \le v_i\}_\perp - v_i\right)\mathrm{d}F^\circ(u)\,\mathrm{d}C^\circ(v) = 0 \end{aligned}$$

for any θ. Thus

$$\tilde{\theta}_n \xrightarrow[n\to\infty]{\mathscr{P}} \operatorname*{argmax}_{\theta\in\Theta} \int_{[0,1]^d} \phi(F_\perp^\circ(u),\theta)+z(F_\perp^\circ(u),\theta)\,\mathrm{d}F^\circ(u) = \operatorname*{argmax}_{\theta\in\Theta} \int_{[0,1]^d} \phi(F_\perp^\circ(u),\theta)\,\mathrm{d}F^\circ(u) = \theta^\circ,$$

the same least false parameter as the MPLE. Both estimators are statistical functionals of the empirical distribution function F_n, say $\hat{\theta}_n = Q(F_n)$ and $\tilde{\theta}_n = R(F_n)$, with influence functions

$$\mathrm{infl}_T(F^\circ, y) := \lim_{\varepsilon\to 0} \frac{T((1-\varepsilon)F^\circ + \delta(y)) - T(F^\circ)}{\varepsilon}$$

for T equal to Q or R. We have $\sqrt{n}(\hat{\theta}_n - \theta^\circ) = n^{-1/2}\sum_{i=1}^n \mathrm{infl}(Q, X_i) + o_P(1)$ and $\sqrt{n}(\tilde{\theta}_n - \theta^\circ) = n^{-1/2}\sum_{i=1}^n \mathrm{infl}(R, X_i) + o_P(1)$. The assumed expansion of the score function U_n given in the statement of the Theorem and equation (5) imply that

$$\sqrt{n}(\hat{\theta}_n - \theta^\circ) = \sqrt{n}(\tilde{\theta}_n - \theta^\circ) + o_P(1),$$

which implies that $\hat{\theta}_n$ and $\tilde{\theta}_n$ have the same influence functions. As $\tilde{\theta}_n$ is an ordinary M-estimator, we know that its influence function is $S^{-1}s(y,\theta^\circ)$, where $S = -\mathbb{E}(\partial/\partial\theta^t)s(X,\theta)$ for $X \sim F^\circ$. As $S = J - \mathbb{E}(\partial/\partial\theta^t)z(X,\theta^\circ) = J - (\partial/\partial\theta^t)\mathbb{E}z(X,\theta^\circ) = J$, we get the desired result. □

Proof of Theorem 3 in the main document . We will approximate the cross-validation expression

$$\widehat{\mathrm{xv}}_n = n^{-1}\sum_{i=1}^n \log c(F_{n,\perp,(i)}(X_i), \hat{\theta}_{(i)}).$$

by a Taylor-expansion in both parameters. Let us denote the vector $(\partial/\partial x_1, \ldots, \partial/\partial x_d, \partial/\partial\theta_1, \ldots, \partial\partial\theta_p)$ by $\partial/\partial y$ and let $\Delta y_n(i) = (F_{n,\perp,(i)}(X_i), \hat{\theta}_{(i)})^t - (F_{n,\perp}(X_i)\hat{\theta}_n)^t$. We have

$$\hat{\theta}_{(i)} = \hat{\theta}_n - n^{-1}\mathrm{infl}(F_{n,(i)}, X_i) + o_P(n^{-1}) = \hat{\theta}_n - n^{-1}\mathrm{infl}(F_n, X_i) + o_P(n^{-1}),$$

where $F_{n,(i)}$ is the full empirical distribution function based on $X_1, \ldots, X_n$ except X_i, and is therefore scaled with $1/(n-1)$ and not $1/n$. We have

$$F_{n,\perp,(i)}(X_i) = \frac{1}{n-1}\sum_{k\neq i} I\{X_k \le X_i\}_\perp = \frac{n}{n-1}F_{n,\perp}(X_i) - \frac{1}{n-1}(1,1,\ldots,1).$$

which shows that

$$\Delta y_n(i) = -n^{-1}\begin{pmatrix}\mathbf{1}_d - F_{n,\perp}(X_i)\\ \mathrm{infl}(F_{n,(i)}, X_i)\end{pmatrix} + o_P(n^{-1}) = -n^{-1}\begin{pmatrix}\mathbf{1}_d - F_{n,\perp}(X_i)\\ \mathrm{infl}(F_n, X_i)\end{pmatrix} + o_P(n^{-1}).$$

A Taylor-expansion then shows that

$$\begin{aligned}\widehat{\mathrm{xv}}_n &= n^{-1}\sum_{i=1}^n \log c(F_{n,\perp,(i)}(X_i), \hat{\theta}_{(i)})\\ &= n^{-1}\sum_{i=1}^n \log c(F_{n,\perp}(X_i), \hat{\theta}_n) + \frac{\partial}{\partial y}\log c(F_{n,\perp}(X_i), \hat{\theta}_n)^t \Delta y_n(i) + o_P(1).\end{aligned}$$

Factorizing out n^{-1} and applying eq. (6) shows

$$\widehat{\mathrm{xv}}_n = n^{-1}\left[\ell_n(\hat{\theta}_n) - n^{-1}\sum_{i=1}^n \frac{\partial}{\partial y}\log c(F_{n,\perp}(X_i), \hat{\theta}_n)^t \begin{pmatrix}\mathbf{1}_d - F_{n,\perp}(X_i)\\ \mathrm{infl}(F_n, X_i)\end{pmatrix}\right] + o_P(1).$$

We have

$$\frac{\partial}{\partial y}\log c(F_{n,\perp}(X_i), \hat{\theta}_n) = (\zeta'(F_{n,\perp}(X_i), \hat{\theta}_n), \phi(F_{n,\perp}(X_i), \hat{\theta}_n))^t,$$

which shows that

$$\frac{\partial}{\partial y}\log c(F_{n,\perp}(X_i),\hat\theta_n)^t\begin{pmatrix}\mathbf{1}_d-F_{n,\perp}(X_i)\\ \mathrm{infl}(F_n,X_i)\end{pmatrix}$$
$$=\zeta'(F_{n,\perp}(X_i),\hat\theta_n)^t\,(\mathbf{1}_d-F_{n,\perp}(X_i))+\phi(F_{n,\perp}(X_i),\hat\theta_n)^t\mathrm{infl}(F_n,X_i).$$

The influence function of $\tilde\theta$ is defined in terms of $z(x,\theta)$, which is unobservable. However, we now show that the assumed pointwise convergence of $\hat z$ implies its uniform convergence, so that we can use it as a plug-in estimator. Indeed, we have

$$|\hat z(x,\theta)-z(x,\theta)|\le\sum_{k=1}^d\sum_{*\in\{-,+\}}\left|\int\left[\frac{\partial\phi(u,\theta)}{\partial u_k}\right]^* I\{x_k\le u_k\}\,\mathrm{d}[C_n-C^\circ](u)\right|$$
$$+\left|\int\left[\frac{\partial\phi(u,\theta)}{\partial u_k}\right]^* u_k\,\mathrm{d}[C_n-C^\circ](u)\right|$$

where we write $[x]^*$ for $[x]^+=\max(0,x)$ when $*$ is $+$ and $[x]^-=\min(0,x)$ when $*$ is $-$. The second term does not depend on x and is thus $o_P(1)$. We are left with showing that for a $f\ge 0$ for which

$$I_n(v):=\int_{[0,1]^d} I\{0\le u_i\le 1 \text{ for } i\ne k, v\le u_k\le 1\}\,f(u)\,\mathrm{d}[C_n-C^\circ](u).$$

converges to zero pointwise, we also have $\sup_{0\le u\le 1}|I_n(u)|=o_P(1)$. This follows by the proof of the classical Glivenko-Cantelli Theorem: the known pointwise convergence of $I_n(v)$ implies its uniform consistency as it is a monotone function, see e.g. Chapter 11.4 of Dudley (2003).

This uniform consistency implies that

$$\frac{\partial}{\partial y}c(F_{n,\perp}(X_i),\hat\theta_n)^t\begin{pmatrix}\mathbf{1}_d-F_{n,\perp}(X_i)\\ \mathrm{infl}(F_n,X_i)\end{pmatrix}=\zeta'(F_{n,\perp}(X_i),\hat\theta)^t\,(\mathbf{1}_d-F_{n,\perp}(X_i))$$
$$+\phi(F_{n,\perp}(X_i),\hat\theta_n)^t\hat J^{-1}\phi(F_{n,\perp}(X_i),\hat\theta_n)+\phi(F_{n,\perp}(X_i),\hat\theta_n)^t\hat J^{-1}\hat z(F_{n,\perp}(X_i),\theta^\circ)+o_P(1)$$

which results in the formula stated in the theorem. □

A.5. **Implementation in R for calculating the xv-CIC.** We used the R system as described in R Development Core Team (2010), together with its Copula package, see Yan (2006). The following R-code calculates the xv-CIC through symbolic methods for the case length(θ) = 1. It assumes that R can calculate symbolically with the density of the copula under consideration. This is not the case for e.g. the Gaussian copula. In such cases, one may use numerical differentiation or use other programs capable of more advanced symbolic manipulation than currently R is, such as Mathematica.

The function takes a data-matrix and a copula-object as input, and gives the xv-CIC, the maximized pseudo likelihood and the xv-CIC penalty terms as output.

```
require(copula)

edf <- function (v, adjust = FALSE) { ## This function is from the QRMlib package.
    original <- v
    v <- sort(v)
    vv <- cumsum(!duplicated(v))
    repeats <- tapply(v, v, length)
    add <- rep(cumsum(repeats - 1), repeats)
    df <- (vv + add)/(length(vv) + as.numeric(adjust))
    as.numeric(df[rank(original)])
}

find.xv.cic <- function(X, copula) {
  n <- dim(X)[1]
  d <- dim(X)[2]
```

```
  copula.d <- copula@exprdist$pdf
  log.copula.d <- parse(text=paste(c("log(",deparse(copula.d),")"), sep="", collapse=""))
  for(i in (1:d)) { assign(paste("zeta.d.u", i, sep=""), D(log.copula.d, paste("u", i, sep=""))) }
  phi <- D(log.copula.d, "alpha")
  phi.d.alpha <- D(phi, "alpha")
  for(i in (1:d)) { assign(paste("phi.d.u", i, sep=""), D(phi, paste("u", i, sep=""))) }
  pseudo.obs <- apply(X, 2, edf, adjust=1)
  for(i in (1:d)) { assign(paste("u", i, sep=""), pseudo.obs[,i]) }
  alpha <- fitCopula(copula, pseudo.obs, method="ml")@estimate
  hat.J <- -sum(eval(phi.d.alpha))/n
  hat.J.inv <- 1/hat.J
  r.n <- 0
  for(i in (1:d)) { r.n <- r.n + mean(eval(get(paste("zeta.d.u", i, sep="")))*(rep(1,n) - get(paste("u", i, sep="")))) }
  for(i in (1:d)) { assign(paste("eval.phi.d.u", i, sep=""), eval(get(paste("phi.d.u", i, sep="")))) }
  q.hat <- function(x) {
    res <- 0
    for(i in (1:d)) { res <- res + mean(get(paste("eval.phi.d.u", i, sep="")))*( (x[i] < get(paste("u", i, sep="")))*1 - get(paste("u", i, sep="")))) }
    res
  }
  q.hat.eval <- NULL
  for(i in (1:n)) { q.hat.eval[i] <- q.hat(pseudo.obs[i,]) }
  q.n <- mean(eval(phi)*hat.J.inv*q.hat.eval)
  p.n <- mean(eval(phi)^2*hat.J.inv)
  return(c(fitCopula(copula, pseudo.obs, method="ml")@loglik - (p.n + q.n + r.n), fitCopula(copula, pseudo.obs, method="ml")@loglik, p.n + q.n + r.n))
}
```

References

Dudley, R. (2003). *Real Analysis and Probability*. Cambridge studies in advanced mathematics. Cambridge, 2nd ed.

Fermanian, J., Radulović, D. & Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* **10**, 847–860.

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

Lee, T. (2008). A multidimensional integration by parts formula for the Henstock-Kurzweil integral. *Mathematica Bohemica* **133**, 63–74.

McShane, E. (1947). *Integration*. the Princeton University press.

Niederreiter, H. (1992). *Random number generation and quasi-Monte Carlo methods*. Society for Industrial Mathematics.

Owen, A. (2005). Multidimensional variation for quasi-Monte Carlo. *Contemporary multivariate analysis and design of experiments* , 49.

Pyke, R. & Shorack, G. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems. *The Annals of Mathematical Statistics* **39**, 755–771.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *The Annals of Statistics* **2**, 892–910.

Ruymgaart, F. H., Shorack, G. R. & van Zwet, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *The Annals of Mathematical Statistics* **43**, 1122–1135.

Segers, J. (2012). Weak convergence of empirical copula processes under nonrestrictive smoothness assumptions. *Bernoulli* **18**, 764–782.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *The Canadian Journal of Statistics* **33**, 357–375.

Yan, J. (2006). Enjoy the joy of copulas. *Journal of Statistical Software* **21**, 1–21.

Zaremba, S. K. (1968). Some applications of multidimensional integration by parts. *Annales Polonici Mathematici* **21**.

Department of Economics, BI Norwegian Business School, 0484 Oslo, Norway
E-mail address: steffeng@gmail.com

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway
E-mail address: nils@math.uio.no