The Copula Information Criterion

Based on joint paper with Nils Lid Hjort

17th December 2008 / Delft

S. Grønneberg

1 Based on joint paper with Nils Lid Hjort

Department of Statistics
University of Oslo

1 SFI, and

SFI

Introduction and summary

Basic concepts

Estimation theory (ML)

Model selection with MLE

Model selection for the MPLE

Asymptotics for the MPLE

The MPLE

The CIC

S. Grønneberg

17th December 2008 / Delft
Motivation for The Copula Information Criterion (CIC)

The copula information criterion is a new model selection methodology for the Maximum Pseudo Likelihood Estimator (MPLE)

- What is the MPLE?
  - The MPLE is an estimation paradigm for estimating a parametric copula through multivariate ranks.
  - It is well-used and popular, as its rank based nature makes it very robust to the particularities of the underlying marginal distributions.
  - Estimation through the use of multivariate ranks are equivalent to estimating marginals with the empirical distribution (nonparametrically), and the MPLE can hence be seen as fitting a semiparametric model to data.
Motivation for The Copula Information Criterion (CIC)

- Many copula users have used
  \[ \ell_{n,\text{max}} - \text{length}(\theta) \]  
  as an “AIC model selection formula”. This is incorrect, as the MPLE does not use a real likelihood, but is based on pseudo observations.

- Practitioners are actively using the MPLE for copula fitting, but use either ad-hoc model selection methods or the incorrect AIC-like formula of eq.(1). This provides strong motivation for the investigation of model selection methodologies for the MPLE.

- We went through the proof of the AIC formula, and extended it to the MPLE case. We named the resulting formula the CIC, the Copula Information Criterion.
Motivation for The Copula Information Criterion (CIC)

- We present two new model selection formulae that generalize the AIC for use with the MPLE.
- The first one, which we call the AIC-like CIC, is valid under the assumption of a correctly specified model.
- The second is valid even under a misspecified model (but requires a larger sample size to be useful), and we call it the TIC-like CIC.
- “The Copula Information Criterion” refers to one of these formulae.
Motivation for The Copula Information Criterion (CIC)

However, to our surprise, the CIC often does not exist!

- There does not exist any AIC-like model selection methodology for copulae of central interest in empirical finance (Gumbel, Joe, etc).
- And so not only is using the AIC formula wrong, there is no way to construct a correction in many cases.
- What should we do?

  - We find that the cause of the non-existence is that the rank structure of the MPLE implicitly separates the estimation of marginal and copula structure.
  - This suggests using semiparametric estimators which do not have this separation, such as ML-based sieve estimation.
  - However, alternative estimators lack an important invariance property which motivates the MPLE and that alternatives lack. Further, there does not exist an AIC formula for the competitors.
Basic notation and assumptions

Just to make it clear from the start:

- We will follow standard asymptotic theory of the MPLE by assuming iid-data. That is, we assume that for each \( i = 1, 2, \ldots, n \) the observation

\[
X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d}) \in \mathbb{R}^d
\]

is independent of each other. (However, the elements of \( X_i \) are obviously dependent!)

- Our proofs are (almost) all based on empirical process theory, and so should also be valid under certain non-iid settings such as weak dependence, but we have not done investigated this further.
You all know this, but let’s fix notation

- Let $X = (X_1, \ldots, X_d) \sim F^\circ$ where $X_i \sim F_i^\circ$ and $F^\circ$ is assumed to be continuous.

- Introduce

  $$F_{\perp}^\circ(x) = (F_1(x_1), F_2(x_2), \ldots, F_d(x_d)).$$

- If $F^\circ$ is continuous, there exists an unique CDF $C^\circ$ on $[0, 1]^d$, called the copula of $F^\circ$, that satisfy

  $$F^\circ(x) = C^\circ(F_{\perp}^\circ(x)).$$

- We have that

  $$C^\circ(u) = F^\circ(F_{\perp}^{-1}(u))$$

  where

  $$F_{\perp}^{-1}(u) = \left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_d^{-1}(u_d)\right).$$
ML theory

- Suppose (for the moment) that we wish to fit a fully parametric density

\[ f_{\theta,\gamma}(x) = c_\theta(F_{1,\gamma(1)}(x_1), \ldots, F_{d,\gamma(d)}(x_d)) \]

to observed data \( X_1, \ldots, X_n \sim F^\circ \).

- The MLE paradigm tries to estimate

\[ (\theta^\circ_{\text{ML}}, \gamma^\circ_{\text{ML}}) = \arg\max_{\theta,\gamma} \int \log f_{\theta,\gamma} \, dF^\circ = \arg\max_{\theta,\gamma} A(\theta, \gamma) \]

from empirical data through replacing the unknown \( F^\circ \) with the empirical distribution

\[ \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\} = \frac{1}{n} \sum_{i=1}^{n} \prod_{k=1}^{d} I\{X_{k,i} \leq x_k\}. \]
ML theory

- So an empirical version of

\[(\theta_{\text{ML}}^\circ, \gamma_{\text{ML}}^\circ) = \arg\max_{\theta, \gamma} \int \log f_{\theta, \gamma} \, dF^\circ = \arg\max_{\theta, \gamma} A(\theta, \gamma)\]

is

\[(\hat{\theta}_{\text{ML}}, \hat{\gamma}_{\text{ML}}) = \arg\max_{\theta, \gamma} \int \log f_{\theta, \gamma} \, d\hat{F}_n = \arg\max_{\theta, \gamma} \hat{A}_n(\theta)\]

- Here

\[\hat{A}_n(\theta, \gamma) = \int \log f_{\theta, \gamma} \, d\hat{F}_n = \frac{1}{n} \sum_{i=1}^{n} \log f_{\theta, \gamma}(X_i) = \frac{1}{n} \ell_n(\theta, \gamma)\]

is the empirical version of \(A\), which is \(n^{-1}\) times the log likelihood. This is the most important quantity in the ML paradigm.
The ML-estimator is originally motivated by assuming $f^\circ = f_{\theta_{ML}}^\circ \gamma_{ ML}^\circ$ and then finding the estimator which asymptotically has the least variance for the true parameter.

However, the MLE can be calculated even when $f^\circ \neq f_{\theta_{ML}}^\circ \gamma_{ ML}^\circ$ which consistently maximize $A(\theta)$ through the plug-in principle – but is this desire to maximize $A(\theta)$ really well motivated?
ML theory

- Notice that the relative entropy ("Kullback–Leibler divergence") between \( f^\circ \) and \( f_{\theta,\gamma} \) is

\[
\text{KL}(f^\circ, f_{\theta,\gamma}) = \int f^\circ \log \frac{f^\circ}{f_{\theta,\gamma}} \, dx
= \int f^\circ \log f^\circ \, dx - \int f^\circ \log f_{\theta,\gamma} \, dx
= A(\theta, \gamma)
\]

- Only \( A(\theta, \gamma) \), which we call the model relevant part of Kullback–Leibler divergence, depend on \((\theta, \gamma)\). Hence

\[
\arg\min_{\theta, \gamma} \text{KL}(f^\circ, f_{\theta,\gamma}) = \arg\max_{\theta, \gamma} \int \log f_{\theta,\gamma} \, dF^\circ
= \arg\max_{\theta, \gamma} A(\theta, \gamma) = (\theta^\circ_{\text{ML}}, \gamma^\circ_{\text{ML}}).
\]
Why does this provide a rationale for using the MLE also when $f^\circ \neq f_{\theta_{\text{ML}},\gamma_{\text{ML}}}$ ?

This is since KL-divergence obeys

$$\text{KL}(f, g) = 0$$

if and only if $f = g$ almost surely. This makes it into a “divergence”.

One may also use other argmin-based estimators, but the divergence property is the absolutely minimum demand one must make for these estimators to make sense.

Without this property one can not use argmin-based estimators to distinguish between densities!
Model selection with MLE

- Suppose that we have several models
  
  \[ f_1, \alpha(1), \ldots, f_K, \alpha(K) \]

  and we wish to choose the “best” on the basis of observed iid data \( X_1, \ldots, X_n \).

- A natural generalization of the MLE is then to select, among these candidates, the minimand of Kullback–Leibler divergence to \( f^\circ \).

- This implicitly defines “best” as “KL-best”, which is not always the most useful choice!

- The best model **under this definition** is

  \[ k^\circ = \arg\min_{1 \leq k \leq K} KL(f^\circ, f_k, \alpha(k)^\circ) \]

  where \( \alpha(k)^\circ = \arg\min_{\alpha(k)} KL(f^\circ, f_k, \alpha(k)) = \arg\max_{\alpha(k)} A(k)(\alpha(k)) \).
Model selection with MLE

- \( \hat{A}_n^{(k)}(\theta(k)) \) estimates the maximand of \( A^{(k)}(\alpha(k)) \) (the model relevant part of KL-divergence), suggesting estimating \( k^\circ \) through

\[
\tilde{k}_n = \arg\max_{1 \leq k \leq K} \hat{A}_n^{(k)} \left( \arg\max_{\alpha(k)} \hat{A}_n^{(k)}(\alpha(k)) \right),
\]

That is

\[
\tilde{k}_n = \arg\max_{1 \leq k \leq K} \hat{A}_n^{(k)}(\hat{\alpha}(k)).
\]

where

\[
\hat{\theta}_n(k) = \arg\max_{\alpha(k)} \hat{A}_n^{(k)}(\alpha(k)) = \arg\max_{\alpha(k)} \int \log f_{k,\alpha(k)} \, d\hat{F}_n
\]

- That is, use the model with the largest observed likelihood.
- This is the main philosophical idea residing in the Akaike Information Criterion.
Model selection with MLE

- However, \( \hat{A}_n^{(k)}(\hat{\alpha}(k)) \) is a biased estimator for \( A^{(k)}(\hat{\alpha}(k)) \).
- The AIC specifically tries to find \( p^*(k) \) such that

\[
\int \log f_{k,\alpha(k)} \, d\hat{F}_n - \int \log f_{k,\alpha(k)} \, dF^\circ = \bar{Z}_n + \frac{1}{n} p_n(k) + o_P(n^{-1})
\]

in which \( \mathbb{E} \bar{Z}_n = 0 \) while \( p_n(k) \xrightarrow[n\to\infty]{} p(k) \) with expectation \( p^*(k) \).

- If we know \( p^*(k) \), or can estimate it, this leads to a first order bias free estimate of \( A^{(k)}(\hat{\alpha}(k)) \), the model relevant part of Kullback–Leibler divergence.

- We will define the derivation of asymptotic first order bias corrections to KL-div as the AIC-program, and will extend it to the MPLE in section 7.
Model selection with MLE

The AIC method is then to choose the model with the highest value of

$$\hat{A}_n^{(k)}(\hat{\alpha}(k)) - \frac{1}{n}\hat{p}^*(k)$$

in which $\hat{p}^*(k)$ is an empirical estimate of $p^*(k)$.

Under model conditions $p^*(k)$ is simply length($\alpha(k)$), which is very easy to estimate without any bias.

Under general assumptions, $\hat{p}^*(k)$ is given by the Takeuchi information criterion, which requires more involved estimation techniques that results in more variable estimates.
Model selection with MLE

- The AIC technique is to estimate $\mathbb{E}p(k)$ and not $\mathbb{E}p_n(k)$, which may be infinite.
- However, $\mathbb{E}p(k)$ is always finite!
- There is no a-priori obvious reason for why we estimate $\mathbb{E}p(k)$ instead of $\mathbb{E}p_n(k)$. This might even give better estimates. However, it saves the AIC from possible explosions!
- The explosion of $\mathbb{E}p_n(k)$ happens even in simple models such as the binomial and logistic regression.
- We will see that a similar problem arises with the CIC, but we get $\mathbb{E}r(k) = \mathbb{E}r_n(k)$ for an additional correction term $r_n$ needed for the MPLE case. And so we can not remove the explosion.
- So the above mentioned non-existence of CIC is not as radical as it may seem.
Suppose for the moment that we know the marginal distributions $f_1^\circ, \ldots, f_d^\circ$ and would like to fit a parametric copula $c_\theta$ to data. Hence our model has density

$$f_\theta(x) = c_\theta(F_1^\circ(x_1), \ldots, F_d^\circ(x_d)) \prod_{k=1}^d f_k^\circ(x_k).$$
The MLE technique is to maximize the model relevant part of KL-divergence and use

$$\theta^\circ = \arg \max_{\theta} \int \log f_{\theta} \, dF^\circ$$

$$= \arg \max_{\theta} \left[ \int \log c_{\theta} (F_1^\circ(x_1), \ldots, F_d^\circ(x_d)) \, dF^\circ \right.$$

$$+ \sum_{k=1}^{d} \int \log f_k^\circ(x_k) \, dF^\circ \bigg]$$

$$= \arg \max_{\theta} \int \log c_{\theta} (F_1^\circ(x_1), \ldots, F_d^\circ(x_d)) \, dF^\circ$$

$$= \arg \max_{\theta} \int \log c_{\theta} (v_1, \ldots, v_d) \, dC^\circ(v)$$
So if $F_{\perp}$ is known, the model relevant part of KL-divergence is

$$A(\theta) = \int \log c_\theta(v_1, \ldots, v_d) \ dC^\circ(v).$$

So we need to find a good empirical estimate of $A(\theta)$.

As $F_{\perp}$ is not known, a natural candidate is to use the empirical copula $\hat{C}_n$ as a plug-in for $C^\circ$ and define

$$\hat{A}_n(\theta) = \int \log c_\theta(v_1, \ldots, v_d) \ d\hat{C}_n(v).$$
The empirical copula $\hat{C}_n$ is the multivariate empirical distribution based on $F_{n,\perp}(X_1), \ldots, F_{n,\perp}(X_n)$, where

$$F_{n,\perp} = \left( F_n^{(1)}, \ldots, F_n^{(d)} \right)$$

in which $F_n^{(k)}$ is the $k$'th marginal empirical distribution.

This is motivated from the following: If we were to knew $F_\circ\perp$, we would have $F_\circ\perp(X) \sim C_\circ$. But as we don’t know it, we replace $F_\circ\perp$ with its empirical counterpart $F_{n,\perp}$.

Thus,

$$\hat{C}_n(v) = \frac{1}{n} \sum_{i=1}^{n} I\{F_{n,\perp}(X_i) \leq v\}$$

in index notation.
Notice that $F_{n,\perp}(X_1), \ldots, F_{n,\perp}(X_n)$ is the vectors of marginal ranks of the observations.

Further,

$$R_{nf} = \int f \, d\hat{C}_n$$

is a multivariate rank statistic. Hence, the MPLE is given by

$$\arg\max_{\theta} \int \log c_{\theta}(v_1, \ldots, v_d) \, d\hat{C}_n(v)$$

and is a multivariate rank based statistic.
The MPLE

- Rank based statistics has the nice property of invariance to monotone increasing transformations of the marginals.

- The copula is invariant to monotone increasing transformations of the marginals as well.

- The principle of invariance in classical estimation theory dictates that if a quantity is invariant to some class of transformations, its estimator should also be invariant to the very same transformations.

- This is perhaps the strongest motivation for using the MPLE and not some other estimation scheme which reach the same parameter.
We defined $\hat{C}_n$ through the use of the plug-in estimator $F_{n,\perp}$, and then defined the MPLE through using $\hat{C}_n$ as a plug-in estimator for $C^\circ$.

Hence we have two levels of “replacing the true size with a plug-in estimator”, which makes $\hat{\theta}$ a two-stage estimator.

This will turn out to be a fatal property when constructing an AIC like model selection procedure for the MPLE.
Asymptotics for the MPLE

- Let the pseudo likelihood be defined as
  \[
  \ell_n(\theta) = \sum_{i=1}^{n} \log c_\theta (F_{n,\perp}(X_i)) = n \hat{A}(\theta)
  \]
  and so
  \[
  \hat{\theta} = \arg\max_\theta \ell_n(\theta) = \arg\max_\theta \hat{A}(\theta).
  \]
- This is an M estimator, but with a test function depending on \( n \). Asymptotic normality has been derived despite of this complication.
- Classical, Taylor expansion-based proofs of normality of M-estimators require the asymptotic distribution of the score function
  \[
  U_n = \frac{\partial \hat{A}_n(\theta_0)}{\partial \theta} = n^{-1} \frac{\partial \ell_n(\theta_0)}{\partial \theta}
  \]
However, as
\[ U_n = \int \phi(v, \theta_0) d\hat{C}_n \]
where \( \phi(\cdot, \theta) = \partial/\partial\theta \log c(\cdot, \theta) \), the \( U_n \) is a multivariate rank statistic.

We can apply asymptotic results from the 70’s Annals (Ruymgaart 72 & 74) to get
\[ \sqrt{n}U_n \xrightarrow{\mathcal{W}} \mathcal{N} \quad \xrightarrow{n \to \infty} U \sim N_p(0, \Sigma) \]
in which \( \Sigma \) is an inflated version of the classical ML limit.
Asymptotics for the MPLE

- The covariance matrix $\Sigma$ equals

\[
\mathcal{I} + \text{Var} \left\{ \sum_{i=1}^{d} \int_{[0,1]^d} \frac{\partial \phi(v, \theta_0)}{\partial v_i} (I\{\xi_i \leq v_i\} - v_i) \, dC^\circ(v) \right\}
\]

where $\mathcal{I}$ is the Information matrix

\[
\mathcal{I} = \mathbb{E} \phi(\xi, \theta_0) \phi(\xi, \theta_0)^t,
\]

and $\xi \sim C^\circ$.

- Hence

\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{W}_{n \to \infty}} J^{-1} U \sim N_p(0, J^{-1} \Sigma J^{-1})
\]

where

\[
J = -A''(\theta_0) = -\int_{[0,1]^d} \frac{\partial \log c_{\theta_0}(v)}{\partial \theta \partial \theta^t} \, dC^\circ
\]
Suppose we have $K$ copulae models

\[ c_1, \hat{\theta}(1), \ldots, c_K, \hat{\theta}(K) \]

and wish to choose the best one.

As the MPLE follows the MLE by minimizing KL-div, it is natural to choose the model which somehow minimize KL-div to the truth.

However, should we compare the full semiparametric model with $f^\circ$?

No, the MPLE use the same marginal estimates for all models, and the marginals do not enter the choice of a copula model when using the MPLE.
So we only need to compare the parametric part of the semiparametric model (the copula) with the theoretical analogous part of the true distribution (the true copula).

This, somewhat subtle step, is in fact the most important step of our investigation: it defines what we are trying to estimate.

An appropriate model selection strategy is hence to use model \( k \) which minimize the model relevant part of

\[
\text{KL} \left( c^\circ, c_k, \hat{\theta}(k) \right).
\]

That is, maximize

\[
A^{(k)}(\hat{\theta}(k)) = \int \log c_{k, \hat{\theta}(k)} \, dC^\circ
\]
We can naively use
\[ \tilde{k}_n = \arg\max_{1 \leq k \leq K} \hat{A}_n^{(k)} \left( \arg\max_{\theta(k)} \hat{A}_n^{(k)}(\theta(k)) \right), \]
as \( \hat{A}_n^{(k)}(\cdot) \) is uniformly close to \( A^{(k)}(\cdot) \).

But as in the fully parametric case, \( \hat{A}_n^{(k)}(\hat{\theta}(k)) \) has a first order bias when estimating \( A^{(k)}(\hat{\theta}(k)) \), and the remaining work to reach the CIC is to derive a bias correction.
Model selection for the MPLE

- Suppress the notational dependence on $k$, and let us study the difference between $\hat{A}_n(\hat{\theta})$ and $A(\hat{\theta})$.
- We follow the classical AIC case and use Taylor expansions of $\hat{A}_n(\cdot)$ and $A(\cdot)$ around $\hat{\theta} - \theta^\circ$ to get

$$\hat{A}_n(\hat{\theta}) - A(\hat{\theta}) = \bar{Z}_n + n^{-1} \sqrt{n}(\hat{\theta} - \theta^\circ) J \sqrt{n}(\hat{\theta} - \theta^\circ) + \hat{A}_n(\theta^\circ) - A(\theta^\circ) + o_P(n^{-1}).$$

$:= p_n$

- Here $\mathbb{E}\bar{Z}_n = 0$, so we only need to deal with the remaining terms.
We have

\[ p_n \xrightarrow{\text{w}} (J^{-1} U)^t J (J^{-1} U) = U^t J^{-1} U := p \]

in which \( U \) is the weak limit of the score vector \( \sqrt{n} U_n \).

We have

\[ \mathbb{E} p =: p^* = \mathbb{E} U^t J^{-1} U = \text{Tr}(J^{-1} \Sigma). \]

from that we know \( \text{Cov} \ U = \Sigma \).
Model selection for the MPLE

- Recall

\[ \hat{A}_n(\hat{\theta}) - A(\hat{\theta}) = \bar{Z}_n + n^{-1} \rho_n + \hat{A}_n(\theta^\circ) - A(\theta^\circ) + o_P(n^{-1}). \]

- In the standard ML case with known marginals, we would have

\[
\mathbb{E} \hat{A}_n(\theta^\circ) = \mathbb{E} \int \log c_{\theta^\circ}(\nu) \tilde{C}_n \\
= \frac{1}{n} \sum_{i=1}^{n} \log c_{\theta^\circ}(F^\circ_{\perp}(X_i)) = \int \log c_{\theta^\circ}(F^\circ_{\perp}(x)) \, dF^\circ \\
= \int \log c_{\theta^\circ}(\nu) \, dC^\circ = A(\theta^\circ).
\]

and so \( \mathbb{E} \hat{A}_n(\theta^\circ) - A(\theta^\circ) = 0 \). This finishes the AIC argument, except finding an estimator \( \hat{\rho}^* \) for \( \rho^* \) (easy).
Model selection for the MPLE

- In the MPLE case, however, we have

\[
\mathbb{E} \hat{A}_n(\theta^\circ) = \mathbb{E} \int \log c_{\theta^\circ}(v) \tilde{C}_n \\
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \log c_{\theta^\circ}(F_{n,\perp}(X_i)) \neq A(\theta^\circ).
\]

- Remember that the AIC gives bias-corrections up to the \(o_P(n^{-1})\) precision level. Hence, if we are to provide a genuine extension of the standard AIC, we need to fix this problem.

- Taylor expansion \(\log c_{\theta^\circ}(\cdot)\) around \(F_{\perp}(X_i) - F_{n,\perp}(X_i)\) removes the problematic \(F_{n,\perp}\).
The Taylor expansion gives

\[ \hat{A}_n(\theta^\circ) = O_P(n^{-3/2}) + n^{-1} \sum_{i=1}^{n} \left[ \log c(F_\perp(X_i), \theta^\circ) \\
+ \zeta'(F_\perp(X_i), \theta^\circ)^t (\hat{V}_i - F_\perp(X_i)) \\
+ \frac{1}{2} (\hat{V}_i - F_\perp(X_i))^t \zeta''(F_\perp(X_i), \theta^\circ) (\hat{V}_i - F_\perp(X_i)) \right]. \quad (2) \]

Here

\[ \zeta'(v, \theta) = \frac{\partial \log c(v, \theta)}{\partial v} \quad \text{and} \quad \zeta''(v, \theta) = \frac{\partial^2 \log c(v, \theta)}{\partial v \partial v^t}. \]

The first summation term of eq. (2) has expectation \( A(\theta^\circ) \), as is needed, but we end up with two additional terms whose properties we need to investigate.
Model selection for the MPLE

Through empirical process theory we can show

\[ \hat{A}_n(\hat{\theta}) - A(\hat{\theta}) = \tilde{Z}_n + n^{-1}(p_n + q_n + r_n) + o_P(n^{-1}) \]

in which \( \mathbb{E}\tilde{Z}_n = 0 \), and \( p_n = O_P(1), q_n = O_P(\sqrt{n}) \) and \( r_n = O_P(1) \). Further, \( q_n/\sqrt{n} \xrightarrow{\mathcal{W}} N(0, \sigma^2) \) and

\[
q^*_n = \mathbb{E}q_n = \int_{[0,1]^d} \zeta'(v; \theta_0)^t (1 - v) \, dC^\circ(v) \\
r^*_n = \mathbb{E}r_n = r^* = \mathbb{E}r \\
= \frac{1}{2} \int_{[0,1]^d} 1^t \left\{ \zeta''(v; \theta_0) \otimes \Psi(v) \right\} 1 \, dC^\circ(v)
\]

for symmetric \( \Psi(v) = (\psi_{i,j})_{i,j} \) where

\( \psi_{i,j}(v) = C^\circ_{i,j}(v_i, v_j) - v_i v_j \) when \( i < j \) and

\( \psi_{i,i}(v) = v_i - v_i^2 \)
Model selection for the MPLE

- Empirical estimates of these correction terms can readily be made.
- If one assumes the parametric model is correct, simplifications can be made, and one results in a formula which is visually similar with the classical AIC.
- One gets the “AIC-like” CIC formula

\[
\widehat{\text{CIC}}_{\text{AIC}} = 2\ell_{n,\max} - 2(\hat{p}^* + \hat{r}^*),
\]

- Here

\[
\hat{p}^* = \text{length}(\theta) + \text{Tr} \left( \hat{I} - \hat{W} \right) \quad \text{Nota bene!} \geq 0
\]

and

\[
\hat{r}^* = \frac{1}{2} \int_{[0,1]^d} c(v; \hat{\theta}) 1^t \left\{ \zeta''(v; \hat{\theta}) \otimes \Psi(v) \right\} 1 \, dv
\]
If we do not assume a correctly specified model, one gets the general “TIC-like” CIC formula

\[ \widehat{\text{CIC}}_{\text{TIC}} = 2\ell_{n,\text{max}} - 2(\hat{\rho}^* + \hat{q}^* + \hat{r}^*), \]

Here

\[ \hat{\rho}^* = \text{Tr} \, \hat{J}^{-\hat{\Sigma}} \]
\[ \hat{q}^* = \int_{[0,1]^d} \zeta'(v; \hat{\theta})^t(1 - v) \, d\hat{C}(v), \]
\[ \hat{r}^* = \frac{1}{2} \int_{[0,1]^d} 1^t \left\{ \zeta''(v; \hat{\theta}) \otimes \hat{\Psi}(v) \right\} 1 \, d\hat{C}(v) \]

using appropriate empirical estimates of the matrices involved.
In the fully parametric case, the bias correction terms are strictly positive.

But for the CIC they may be both positive and negative, depending on the dependence structure of the copula one fits.

For the Frank copula, the true bias terms are plotted in the figure below for varying dependence parameters.
These expectations **often do not exist!**.

What goes wrong?

- Interesting copulae have extreme behaviour near the edge of $[0, 1]^d$, such as **extreme tail dependence**. This makes the integrand of the bias-terms non-integrable.
- In the next two slides are plots of $\sum_i \zeta'_i$ and $\sum_{i,j} \zeta''_{i,j}$ when using the AIC-like CIC with the Gumbel copula. Although they are integrable, they diverge to $-\infty$ near the edge of $[0, 1]^d$. 
$\sum_i \zeta_i$ for the Gumbel Copula
\[ \sum_{i,j} \zeta_{i,j}'' \] for the Gumbel Copula
What goes wrong?

Clearly, the need for integrating functions of $\zeta'$ and $\zeta''$ gets us in trouble.

This originates from the $\mathbb{E}\hat{A}_n(\theta^\circ) \neq A(\theta^\circ)$. Is this the problem?

No. Consider using another two-stage estimator (called the IFM). First estimate the marginals, say $F_{\perp,\gamma}$, through MLE. Then use these estimates as plug-in estimates for $F^\circ_{\perp}$ in the copula estimation procedure. We then still have $\mathbb{E}\hat{A}_n(\theta^\circ) \neq A(\theta^\circ)$.

But this escapes the problem: Here we do not need to do a full Taylor-expansion of the $v$ term in $\log c_\theta(v)|_{v=F_{\perp,\hat{\gamma}}(X_i)}$, but only around $\hat{\gamma} - \gamma^\circ$.

This requires no assumption on integrability of functions of $\zeta'$ and $\zeta''$. We would only need very classical conditions on $\sqrt{n}(\hat{\gamma} - \gamma^\circ)$. 
What goes wrong?

- Is the problem that we are using the multivariate empirical distribution as a plug-in estimator for the marginals? Perhaps another choice leads to a fix?
  - No, if we were to use another nonparametric plug-in estimator we would get the very same problem: We would have to make strong integrability assumptions on $\zeta'$ and $\zeta''$.
  - Certain monotonicity assumptions on $F_\perp^\circ$ yields cube root asymptotics for a estimator of $F_\perp^\circ$, say $\tilde{F}_\perp^*$. That is, $\sup_x |F_\perp^\circ(x) - \tilde{F}_\perp^*(x)| = O_P(n^{1/3})$.
  - If we were to use $\tilde{F}_\perp^*$ rather than $F_n,\perp$, we would only need one Taylor-expansion.
  - However, this would still require (too) strong integrability assumptions on $\zeta'$ and hence would not fix the problem.
The problem thus originates from the two-stage nature of the MPLE.

This suggests two things

1. Either: Don’t use a two-stage estimator, but rather use a simultaneous estimator such as the MLE sieve based semiparametric estimator of Chen & Fan 06 in JASA.

2. Or, use other model selection strategies which is not first order bias correction based.

The first possibility has two obvious problems:

- There does not yet exist any model selection procedure for the alternative methods
- The alternatives are not invariant to monotone transformations of the marginals, hence violating the principle of invariance.
A third, rather unsatisfying solution, is to only work with $\ell_{n,\text{max}}$, and not do any bias correction. This is a poor man’s estimator for the KL-best model, but it is clearly consistent. (This is what Chen & Fan does in their Pseudo Likelihood Ratio paper, perhaps without knowing it)
Another method entirely is to follow the strategy of the Focused Information Criterion, which chooses a model which minimize mean squared error of a collection of interest parameters. We think that this would avoid the explosion problems encountered for the CIC.

Semiparametric model selection is somewhat unexplored, and most theory is based on the assumption of a correctly specified finite dimensional part, which is (as we have seen) not a typical situation. There does not yet exist any generally applicable model selection paradigm.