Semiparametric model selection for copulas

Master's Thesis in Statistics

Financial Theory and Insurance Mathematics

Lars Arne Jordanger

UNIVERSITY OF BERGEN

Faculty of Mathematics and Natural Sciences

Department of Mathematics

June 2013

Abstract

This thesis will consider the performance of the cross-validation copula information criterion, xv-CIC, in the realm of finite samples.

The theory leading to the xv-CIC will be outlined, and an analysis will be conducted on an assorted collection of bivariate one-parameter copula models. The restriction to the bivariate case is not a grave one, since more complex d-variate samples can be broken down into a study of conditioned bivariate samples, by the methodology of regular vine-copulas, the pair copula construction and stepwise-semiparametric estimation of parameters.

As a by-product of our analysis, we can give an advice with regard to the selection of model selection method in the semiparametric realm.

Acknowledgements

First and foremost, I would like to thank Prof. Dag Tjøstheim, Ph.D., for his excellent supervision during the writing of this thesis.

Furthermore, I would also like to thank both my fellow Master's Degree students and the staff at the Department of Mathematics, for making the last two years an enjoyable period of my life.

Finally, I would like to express my gratitude toward all those participating in the creation of open source programs like LaTeX, `R`, `Sweave`, `kntir` and so on, since the writing of this thesis hardly would have been manageable without these programs.

Contents

List of Figures

List of Tables

Notation and Abbreviations

$\boldsymbol{\mathcal{X}}_n$	independent observations in $\mathbb{R}^d$
${}^u\boldsymbol{\mathcal{X}}_n$	independent idealized observations in $[0,1]^d$
${}^p\boldsymbol{\mathcal{X}}_n$	dependent pseudo-observations in $[0,1]^d$
ℓ	log-likelihood function (on $\boldsymbol{\mathcal{X}}_n$ or ${}^u\boldsymbol{\mathcal{X}}_n$)
${}^p\ell$	pseudo-log-likelihood (on ${}^p\boldsymbol{\mathcal{X}}_n$)
$\widehat{\boldsymbol{\theta}}$	ml-estimate of copula-parameter(s) (on $\boldsymbol{\mathcal{X}}_n$ or ${}^u\boldsymbol{\mathcal{X}}_n$)
${}^p\widehat{\boldsymbol{\theta}}$	mpl-estimate of copula-parameter(s) (on ${}^p\boldsymbol{\mathcal{X}}_n$)
AIC	Akaike Information Criterion (on $\boldsymbol{\mathcal{X}}_n$ or ${}^u\boldsymbol{\mathcal{X}}_n$)
^{p}AIC	pseudo-Akaike Information Criterion (on ${}^p\boldsymbol{\mathcal{X}}_n$)
BIC	Bayesian Information Criterion
CIC	Copula Information Criterion (on ${}^p\boldsymbol{\mathcal{X}}_n$)
CIC^{AIC}	Copula Information Criterion, AIC-like assumptions
CIC^{TIC}	Copula Information Criterion, TIC-like assumptions
IFM	Inference Functions for Margins
KLIC	Kullback-Leibler Information Criterion
ml	maximum likelihood
MLE	Maximum Likelihood Estimator
mpl	maximum pseudo-likelihood
MPLE	Maximum Pseudo-Likelihood Estimator
PCC	Pairwise Copula Construction
SSP	Stepwise semiparametric
TIC	Takeuchi Information Criterion (on $\boldsymbol{\mathcal{X}}_n$ or ${}^u\boldsymbol{\mathcal{X}}_n$)
xv	leave-one-out cross-validation (on $\boldsymbol{\mathcal{X}}_n$ or ${}^u\boldsymbol{\mathcal{X}}_n$)
^{p}xv	pseudo leave-one-out cross-validation (on ${}^p\boldsymbol{\mathcal{X}}_n$)
xv-CIC	cross validation Copula Information Criterion (on ${}^p\boldsymbol{\mathcal{X}}_n$)

Chapter 1

Introduction

Every day vast amounts of data are encountered in science and commerce, and there is a high demand for the construction of statistical models that for instance can be used to attain reasonable predictions with regard to what the future might bring, be it stock values, levels of precipitation, losses due to natural disasters or whatnot.

In the construction of such models, one need to make distributional assumptions for the available d-variate observations $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_i\}_{i=1}^n$. However, if the family of distributions used does not allow for more extreme cases, this might e.g. result in severely underestimated risk-rates for ruin in a financial setting.

It is not a trivial task to find a distributional model that can describe a set of d-variate observations $\boldsymbol{\mathcal{X}}_n$ in a faithful way. Even though theoretically sound algorithms exist for the estimation of the parameters, these algorithms might be rendered useless if the number of parameters used to describe the structures inherent in $\boldsymbol{\mathcal{X}}_n$ leads to a computational time that spans eons.

Moreover, in addition to fitting models to our observations $\boldsymbol{\mathcal{X}}_n$, we will also like to rank the models in order to select the one which seems to give the best description of the data-generating process.

This thesis will investigate the finte sample performance of such a model selection method in a semiparametric setting, and this first chapter will now proceed by presenting an outline of the theory motivating the *cross-validation copula information criterion*.

1.1 Classical multivariate theory

Some data are easier to describe than others. For instance, one could have data originating from a biological process, where most of the observations will be clustered in the proximity of the mean, and where we due to the biological nature of the data-generating process will find rather few extreme deviations.

In the univariate case, the distribution that often describes the features of such observations have in fact received its name from the multitude of cases where it popped up, i.e. the *normal* distribution.

Moreover, as we can learn from a textbook like e.g. Johnson and Wichern [1], the univariate normal distribution can be extended to a multivariate distribution, uniquely determined by its mean $\boldsymbol{\mu}$ and its covariance matrix $\boldsymbol{\Sigma}$.

As in the univariate case where we might need to consider the t-distribution when we want to do an inference based on a sample, we might need to use the multivariate t-distribution instead of the multivariate normal-distribution. The multivariate normal distribution and multivariate t-distribution are commonly known as elliptical distributions, due to the elliptical contour-lines they feature. If our data feature such symmetrical properties, these distributions surely are worth to consider in our quest for a model.

However, our data might not fit into this fold, e.g. a daily measure of temperature, precipitation and wind (direction + strength) would after a while probably contain a much higher amount of extreme observations (for one or more of the covariates) than what we would have expected to see if the data-generating process were of an elliptical nature.

Another example of data that might be erroneous to model by an elliptical distribution is the daily values of stocks. In this case we might expect there to be less correlation between the prices in a bull market than in a bear market, i.e. the tail-dependencies will be different. In particular, when some stocks in a bull market start to rise, we do not expect all other stocks to rise as well – but in a bear market, many stocks will simultaneously plummet into the abyss.

Whereas we rather easily can produce an elliptical model for a set of observations that exhibits symmetrical features, the situation is tremendously more difficult in the general case – and we will be forced to assume quite a few simplifying assumptions in order to produce any model at all.

It is thus important to keep strictly in mind that the model we obtain after our simplifying assumptions not should be confused with the unknown data-generating process that our observations stemmed from. The following quote from G. E. P. Box is a nice reminder with regard to how we should look upon our models:

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> *Empirical Model-Building by Box and Draper, p. 74.*

1.2 Sklar's theorem

The theory of copulas, introduced in Sklar [2] tells us that there is a nice connection between general d-variate distributions with support on $\mathbb{R}^d$ and a particular class of d-variate distributions, called copulas. To be precise, a d-variate copula is a function $C : [0,1]^d \to [0,1]$ satisfying the following two properties

$$(i) \quad \text{For every } \boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d, \tag{1.1a}$$
$$C(\boldsymbol{u}) = 0 \text{ if at least one coordinate of } \boldsymbol{u} \text{ is } 0,$$
$$\text{and for all } j = 1, \ldots, d$$
$$C(1, \ldots, 1, u_j, 1 \ldots, 1) = u_j.$$
$$(ii) \quad C \text{ is } d\text{-increasing (see Nelsen [3])}. \tag{1.1b}$$

If we make the simplifying assumption that the cumulative distribution function $F(\boldsymbol{x})$ is continuous, Sklar's theorem tells us that there exists an unique copula $C(\boldsymbol{u})$ on $[0,1]^d$ such that

$$F(\boldsymbol{x}) = C(\boldsymbol{F}_{\perp}(\boldsymbol{x})), \tag{1.2}$$

where $\boldsymbol{F}_{\perp}(\boldsymbol{x}) \stackrel{\text{def}}{=} (F_1(x_1), \ldots, F_d(x_d))$, with $F_j(x_j) \stackrel{\text{def}}{=} \mathrm{P}(X_j \leq x_j)$ the vector of cdf's for the marginal distributions.

This implies that the copula C encodes all the multivariate dependencies of the distribution F, and by this connection we can separate the study of the dependency structure of $F(\boldsymbol{x})$ from the structure of its marginal distributions.

Furthermore, this enables us to construct new multivariate distributions, since the copula C can be given any set of cdfs as inputs – and thereby give us new d-variate distributions G with the same dependency structure as the F we started with.

A consequence of this is that we "easily" can produce a plethora of different d-variate models by using different combinations of d-variate copula models C and different collections of marginal distributions $\{F_i\}_{i=1}^d$.

However, with regard to the computational costs required for the fitting of a model to the observations $\boldsymbol{\mathcal{X}}_n$, it is recommendable to restrict our attention to models that we initially think might have a decent match with the data.

When we consider this, there are two parts to deal with – the marginal models $\{F_i\}_{i=1}^d$ and the d-variate copula model C – and in both of these parts there will in general be necessary to make simplifying assumptions in order to get a manageable estimation problem. The discussion of estimation related to the copula will be postponed to the next section.

Models for the marginal distributions could of course be investigated by considering the relevant subset from $\boldsymbol{\mathcal{X}}_n$, in order to restrict our attention to models that at least have a decent fit to the corresponding empirical data. However, an option (that will be further discussed later on) is to get rid of the marginal distributions completely, by using the empirical marginals to create a set of pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ in $[0,1]^d$.

There are several reasons that justifies this simplification. One reason is that we in general will be more interested in the parameters describing the interdependencies of $\boldsymbol{\mathcal{X}}_n$, and that we thus can sacrifice all those other parameters that the marginals would have introduced. Furthermore, if we want the estimation-algorithm to finish within a reasonable timespan, it is a necessity to keep the number of parameters as low as possible. Thirdly, even though the transformation $\boldsymbol{\mathcal{X}}_n \to {}^p\boldsymbol{\mathcal{X}}_n$ introduces some extra bias into our computation, the effect of this dwindles when we have larger samples, cf. appendix B.

1.3 Bivariate copulas and vine copulas

When it comes to the copula-models, there is a tremendous difference between the bivariate case and the general case – which we now will briefly comment upon.

In the bivariate case, one or two copula-parameters is sufficient to reasonably model features like symmetry and tail-dependence. In addition to the copulas corresponding to the symmetric elliptical distributions (i.e. the `normal`-copula[1] and the `t`-copula), a major part of the toolbox in the bivariate case of copula-modeling is the Archimedean copulas – which can be constructed by the help of a generating function (see chapter 2 for the formal definition, and cf. Joe [4] or Nelsen [3] for lengthy lists of such copulas).

[1] The `normal`-copula is also known as the Gaussian-copula.

When the dimension d increases, there is a rapid increase in the number of possible internal symmetries and the number of corners/edges in $[0,1]^d$ where some kind of tail-dependence can occur. In particular, this implies that the copula-models need a higher number of parameters in order to faithfully model d-variate observations.

However, the problem of finding suitable d-variate copula models is rather complicated. We can construct Archimedean copulas in any dimension, but they are stuck with the one or two parameters that they have in the bivariate case – and they are thus to rigid to be useful. The other easy available source of d-variate copula-models is the copula models corresponding to the elliptical distributions, but these are not suitable when the observations $\boldsymbol{\mathcal{X}}_n$ depart from symmetry, e.g. by having a prominent one-sided tail-dependency.

The lack of flexible d-variate copula models made the usefulness of Sklar's theorem rather limited in the general d-variate setting, but this changed when the pair-copula construction (PCC) and vine-copulas entered the stage, in particular when this was successfully applied in a modeling problem, see Aas and Berg [5] and Aas et al. [6].

1.3.1 The pair copula construction

This section will briefly sketch the pair copula construction (PCC) and only mention the concept of vine copulas. These concepts are needed in order to give at least a partly justification for restricting the analysis in chapter 4 to the bivariate case, but otherwise we will not mention them again. The interested reader can find a rigorous introduction to this theory in Brechmann [7].

For the discussion below, recall that we are working under the simplifying assumption that we have a continuous d-variate distribution $F(\boldsymbol{x}) = C(\boldsymbol{F}_{\perp}(\boldsymbol{x}))$, with notation as given in eq. (1.2).

We will henceforth denote by respectively f and c the pdfs of F and C, and we will furthermore denote by f_i the pdfs of the marginal distributions F_i.

The following connection between the pdfs follows directly from eq. (1.2):

$$f(\boldsymbol{x}) = \left(\prod_{k=1}^{d} f_k\left(x_k\right) \right) \cdot c(\boldsymbol{F}_{\perp}(\boldsymbol{x})). \tag{1.3}$$

In addition to this expression, the density $f(\boldsymbol{x})$ can also be expressed by the help of conditional distributions like

$$f(\boldsymbol{x}) = f_1\left(x_1\right) f_{2|1}\left(x_2|x_1\right) \cdots f_{d|12\ldots(d-1)}\left(x_d|x_1,\ldots,x_{d-1}\right), \tag{1.4}$$

or we might equally well have used any permutation of the indices $(1, \ldots, d)$.

Joe [8] showed how the conditional densities in eq. (1.4) could be expressed by conditional densities derived from the copula C, and how an iterative process then gives a formula which decomposes the d-variate density f in terms of d univariate marginal densities and $d(d-1)/2$ bivariate copula densities. We will here briefly sketch this result.

With notation borrowed from Haff [9], where a more detailed argument can be found, the following rule tells us how a conditioning variable x_j in $f_{i|j \cup v}$ can be "removed" by using the bivariate conditional copula $c_{ij|v}(\cdot, \cdot|x_v)$ instead, i.e.

$$f_{i|j \cup v}(x_i|x_j, x_v) = c_{ij|v}\left(F_{i|v}(x_i|x_v), F_{j|v}(x_j|x_v)|x_v\right) \cdot f_{i|v}(x_i|x_v). \tag{1.5}$$

Note that v here is a nonempty subset of $\{1, \ldots, d\} \setminus \{i, j\}$, and that x_v thus is to be interpreted as the collection of variables indexed by v.

An iteration of eq. (1.5) leads to an expression of the following form for the density f,

$$f(\boldsymbol{x}) = \left(\prod_{k=1}^{d} f_k(x_k)\right) \cdot \left(\prod_{\ell=1}^{d-1} \prod_{(i,j,v)} c_{ij|v}\left(F_{i|v}(x_i|x_v), F_{j|v}(x_j|x_v)|x_v\right)\right), \tag{1.6}$$

in which the triplet (i, j, v) is such that v is a subset of size $\ell - 1$ from $\{1, \ldots, d\} \setminus \{i, j\}$.

If we in eq. (1.6) make the simplifying assumption that all the bivariate conditional copula densities $c_{ij|v}(\cdot, \cdot|x_v)$ are constant with regard to x_v, i.e. that we assume that the random pair $\left(F_{i|v}(X_i|X_v), F_{j|v}(X_j|X_v)\right)$ is independent of the random vector X_v, then we arrive at the following simpler expression for the density $f(\boldsymbol{x})$

$$f(\boldsymbol{x}) = \left(\prod_{k=1}^{d} f_k(x_k)\right) \cdot \left(\prod_{\ell=1}^{d-1} \prod_{(i,j,v)} c_{ij|v}\left(F_{i|v}(x_i|x_v), F_{j|v}(x_j|x_v)\right)\right). \tag{1.7}$$

According to Haff et al. [10], the nonparametric shape constraint assumption of independence is satisfied by copulas like the `t`-copula and the `clayton` copula. Moreover, even though this assumption does not hold in general, it provides in many cases an acceptable approximation to the true distribution. The interested reader is urged to check out Haff [9] for further details.

An important consequence of eq. (1.7) with regard to the study of a d-variate sample $\boldsymbol{\mathcal{X}}_n$ (from some unknown data-generating process), is that we can draw on all the knowledge we have of bivariate copulas. This gives a tremendous increase in flexibility, which is clearly preferable if we want the resulting model to have a god fit to the data.

1.3.2 Regular vines and vine copulas

The combinatorial rules that the triplets (i, j, v) from eq. (1.7) must satisfy, was investigated in Bedford and Cooke [11, 12], and lead to the definition of regular vines and vine copulas.

For our purposes the following informal description of a vine copula is sufficient: A *vine copula* is a collection of bivariate copula models that labels all the edges of a *regular vine*, where the latter can be described as a nested set of trees where the edges of tree i are the nodes of tree $i+1$, and where two edges in tree i are joined by an edge in tree $i+1$ only if they share a common node.[2]

With the vine copula as a part of our toolkit, the problem of finding a copula model that can describe the interdependencies in a d-variate set $\boldsymbol{\mathcal{X}}_n$ is much simplified – and the models we get have plenty of parameters that can be tweaked in order to fit the models to our observations.

However, with an increased number of parameters comes a formidable increase in computing time could result. Remember that we can factorize the d-variate pdf f into d univariate marginals and $d(d-1)/2$ bivariate copula densities, and if all of these distributions then have one or two parameters, any work on the resulting structure would demand a large computational cost

Moreover, we would typically like to compare several fitted models against each other, in order to see if some of them seems to fit $\boldsymbol{\mathcal{X}}_n$ better than the others. This also heightens the total computational load. Even for rather low values of d will it be infeasible to try to fit all possible regular vine structures, and all the different vine copulas they can be labeled with, when searching for a model for some $\boldsymbol{\mathcal{X}}_n$. In short, we need techniques that can delimit our attention to the most promising candidates.

1.3.3 Delimiting techniques

As mentioned above, we need some way to restrict our attention to the most promising models when we want to model a d-variate set of observations $\boldsymbol{\mathcal{X}}_n$. And in addition we need some way to decide if one of the models have a better fit to the data, and thus should be the preferred one if we want to make some predictions, say.

Like in the previous sections, we will once more only give a sketch of the arguments here, the interested reader can find a complete treatment of these topics in Brechmann [7].

[2] The rigorous formal definitions of regular vines and vine copulas will not be given in this thesis, but the interested reader can find an exquisite introduction in Brechmann [7, Chap 2.4].

First of all, we need to pick a regular vine to be used as the frame for the vine copula models. The standard strategy with regard to this is to employ an algorithm that picks a vine where as much as possible of the dependence structure is captured in the first levels.

When a regular vine has been chosen, we can use that as a frame for a vine copula model, i.e. we need a collection of $d(d-1)/2$ bivariate copula models to put on the edges of the vine. It is trivial to produce a lot of such vine copula models, but the time needed for the estimation of the parameters implies that we still need to apply simplifying assumptions if we want an answer within a reasonable timespan.

When the focal point of interest is the multivariate dependencies between the covariates, a standard procedure is to "sacrifice" the parameters related to the marginal distributions. In particular, we avoid the specification of models for the marginals by replacing $\boldsymbol{F}_{\boldsymbol{\gamma},\perp}$ with $\boldsymbol{F}_{n,\perp}(\boldsymbol{x}) \stackrel{\text{def}}{=} (F_{n,1}(x_1), \ldots, F_{n,d}(x_d))$, where $F_{n,j}(x_j)$ stands for $\frac{n}{n+1}$ times the empirical marginal, i.e. $F_{n,j}(x_j) \stackrel{\text{def}}{=} \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{X_j \leq x_j\}$. [3]

We thus replace the independent observations $\boldsymbol{\mathcal{X}}_n$ from $\mathbb{R}^d$, with dependent pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ in $[0,1]^d$. This strategy simplifies the estimation process of the remaining parameters, but there is a price to pay since the transformation $\boldsymbol{\mathcal{X}}_n \to {}^p\boldsymbol{\mathcal{X}}_n$ introduces extra noise that must be accounted for when we consider the bias of our estimated parameters, and the corresponding fitted models.

This semiparametric approach – that solely focus on the parameters of the copula – might still not reduce our problem to a manageable one, since we even for quite few covariates might reach an intractable number of parameters to estimate.

Stepwise semiparametric, reduction to the bivariate case: When we use a *vine-copula* as a model for our pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$, we can counter the problem of too many parameters by using the *stepwise semiparametric* (SSP) strategy – which at the price of some further simplifying assumptions enables us to estimate our parameters in smaller batches. The strategy is to focus on the regular vine, and start out with the tree at its lowest level. At each edge of this tree there will be a corresponding conditioned bivariate set, and we will fit our bivariate copulas to these sets. When all the trees in a level is accounted for, the process goes on to the next level. Haff [9] can be consulted for further details with regard to this procedure.

To emphasize: The SSP strategy for parameter estimation implies that we even for d-variate observations $\boldsymbol{\mathcal{X}}_n$ will restrict our attention to the bivariate case when we want to find a copula that models the interdependencies of our observations.

[3] The rescaling with the factor $\frac{n}{n+1}$ ensures that we avoid points on the edge of the unit cube in $\mathbb{R}^d$. This is important since many copula models of interest are heavy tailed, and points on the boundary could then introduce infinities into our calculations.

A consequence: The results from the approach used in chapter 4, where we restrict our attention to the bivariate case when we investigate the performance of our semiparametric selection methods, will be useful in a more general setting too.

Some additional comments: The consequence mentioned above is the one of interest for this thesis, but there are two additional concepts that deserves to be mentioned before we leave this subject.

The point is, that even though the SSP strategy splits the estimation of the parameters into manageable chunks, i.e. one bivariate copula at a time, we might still have that the number of such chunks results in a total computational time beyond acceptable limits.

A strategy that can be used to face this problem, is to stop the SSP-process by either using *simplification* or *truncation* when the conditioned observations at a level is sufficiently close to a normal copula or to the independence copula, such that the remaining layers then become approximated by the chosen structure. See Brechmann [7] and Haff [9] for the definitions of simplification and truncation, and note that [7] in addition gives criteria and algorithms for determining at which level of the vine (if any) it might be advisable to use simplification or truncation in order to reduce the computational load even further.

Simplifications and truncations give a simpler model than the one we otherwise would have obtained, and quite probably a much simpler model than the one that generated the data. But as we anyway never can expect to find the true data-generating process from a finite sample, the important point is whether or not our model can enlighten us to the workings of the data-generating process. Caution must anyway be applied with regard to how much we can deduce from our model, and the following warning from Whitehead should be observed.

> The aim of science is to seek the simplest explanations of complex facts. We are apt to fall into the error of thinking that the facts are simple because simplicity is the goal of our quest. The guiding motto in the life of every natural philosopher should be, "Seek simplicity and distrust it."
>
> The Concept of Nature (1926), Alfred North Whitehead.

1.4 Model selection, bias and the need for simulations

The previous section mentioned challenges that we encounter when we search for a model which in an adequate way can describe a set of d-variate observations $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_i\}_{i=1}^n$, and we saw how techniques based on the vine-copulas and the pair-copula construction could enable us to produce an abundance of models with enough parameters to capture interesting structures in $\boldsymbol{\mathcal{X}}_n$, like symmetry and tail-dependence.

Moreover, we mentioned how the technique of stepwise semiparametric estimation could enable us to compute the estimates of the parameters in a vine copula within a reasonable timespan, by replacing the task of estimating all the parameters at the same time with the more feasible task of estimating the parameters of the bivariate copulas in the vine copula one level at the time.

But in general the fitting of a model to the observations is only a part of the story. We will typically attempt to fit an assorted collection of models to our observations $\boldsymbol{\mathcal{X}}_n$, and when all of these have been fitted to the data – we would like to rank the fitted models[4] in order to see which model to select, i.e. to find the model that based on the observations $\boldsymbol{\mathcal{X}}_n$ (or pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$) seems to give the the best approximation to the process that generated our observations.

The important thing to keep in mind is that we need enough parameters to capture the interesting structures in $\boldsymbol{\mathcal{X}}_n$, but not so many that we over-fit our model to the specific set of observations. We should also take into account the number of observations that will be necessary in the estimation process in order to get estimates with an acceptable level of error.

In a setting where we use a fully parametric approach to the modeling of $\boldsymbol{\mathcal{X}}_n$, i.e. where we use parametric models for the univariate marginals F_i in eq. (1.2), we could use a model selection method like AIC, introduced in Akaike [13], in order to rank different proposed models against each other.

The validity of the AIC-formula is due to a likelihood-based argument on the independent observations in $\boldsymbol{\mathcal{X}}_n$. But when we in a practical application use the empirical marginals in order to reduce the amount of parameters to be estimated – we end out with an analysis that must be based on the dependent pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$. And in this framework the likelihood is replaced with the pseudo-likelihood introduced by Besag [14].

Even though AIC as a selection method is based on a maximum likelihood (ml) argument, it has been common practice to "tweak" it slightly and apply it without further ado in the semiparametric setting where a maximum pseudo likelihood (mpl) is at play.

[4] By "a fitted model" we mean the model we get when we use the estimated parameters.

This difference is a central one in this thesis, and we will henceforth introduce the notation ${}^{p}\text{AIC}$ when the AIC-like selection method is applied on models based on a set of dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$. The formulas that is used for their respective computation is as follows,

$$\text{AIC}(F) = 2 \cdot \ell_{\max}(F) - 2 \cdot \dim(F) \tag{1.8}$$

and

$${}^{p}\text{AIC}(C) = 2 \cdot {}^{p}\ell_{\max}(C) - 2 \cdot \dim(C), \tag{1.9}$$

in which $\ell(F)$ represents the log-likelihood-function of a model F on the independent observations $\boldsymbol{\mathcal{X}}_n$ from $\mathbb{R}^d$, while ${}^{p}\ell(C)$ represents the pseudo-log-likelihood-function of a copula model on the dependent pseudo observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ in $[0,1]^d$. The subindex max is to indicate that we are considering the maximum of these functions, i.e. that they are evaluated at the estimated values of the parameters in the models. Further, $\dim(\cdot)$ simply refers to the number of parameters that the two models contain, and it is present in order to give us a bias-correcting term when different models to the observations (or pseudo-observations) are to be ranked against each other.

The formal validity of eq. (1.9) was investigated in Grønneberg and Hjort [15], and it turned out that the simple bias-correction term did not properly account for the noise in the transformation $\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$. An analysis was conducted in [15] in order to find theoretically valid model selection methods for the semiparametric case, but the two resulting formulas turned out to have the rather unappealing property of not being generally applicable (see latter chapters for details).

However, in Grønneberg [16, Part III] a generally applicable model selection method for the semiparametric realm was created as an analytical approximation to the semiparametric leave-one-out cross-validation technique.

The result of this approach, named the cross-validation copula information criterion (xv-CIC), is the main theme of this thesis. An outline for its theoretical foundation is given in chapter 3.

Bias-correction and the need for simulations. An essential detail to be aware of for model selection methods, is that their bias-correcting terms often are based on the asymptotic behavior of the finite-sample bias-correcting distributions, i.e. that the bias-correction is given as the expectation of this limiting distribution.

The use of limiting distributions in order to simplify the computations should introduce a negligible error when the number of observations is large enough, since the deviations

from the true (but unobtainable) finite-sample bias-correcting terms then should be so small that the ranking of the considered models should be unaffected.

But how large must the number of observations be before we can trust the conclusion of the selection method? The answer to that question will in fact depend on which models we are considering, as the number of observations needed for a good approximation of the bias-correcting term for one model might turn out to give an awful approximation for another one.

It seems that the only way we can probe how much faith we should put in the conclusion of a selection method (for a given size of observations and a given list of candidate models) is to perform simulations in order to see how often the method actually hits the mark.

This thesis gives in chapter 4 the performance-results of the xv-CIC as a model selection method, when tested on bivariate samples of sizes $N \in \{100, 250, 500, 1000\}$.

Note that we do not make a severe restriction when we only test the performance in the bivariate case, since the methodology of the stepwise semiparamteric estimation procedure breaks the estimation process into chunks that handles one bivariate copula at the time.

A note of caution. When we investigate a collection of models, that for some reason has been proposed as the ones we want to fit to some observations $\boldsymbol{\mathcal{X}}_n$ (or pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$), it is paramount that we are aware of the fact that most selection methods does not inform us if any of these models are any good.

To emphasize: The selection method will rank the fitted models and tell us which of these that gives the best description of our data, but if all of our initially considered models are useless – then even the best of them will share that deficiency.

Unless we are quite certain that one of the fitted models should give a good description of our data, we should apply some goodness of fit test to investigate whether or not the chosen model is any good at all. If none of the models seems to be worth their salt, we should consider other models instead – confer the discussion in section 4.4 for a concrete example.

Chapter 2

Some background theory

This chapter contains a collection of assorted background theory.

We will in section 2.1 discuss the copula models that will be used when we test the small-sample behavior of xv-CIC in chapter 4, and we will moreover consider how the use of the empirical marginals in the transformation $\boldsymbol{\mathcal{X}}_n \to {}^p\boldsymbol{\mathcal{X}}_n$ creates noise in our estimated parameters and bias-correcting terms.

Section 2.2 will mention different strategies used to estimate parameters in parametric and semiparametric models, and i.e. why we might like to replace a parametric approximation $F_{(\boldsymbol{\theta},\gamma)}$ on $\boldsymbol{\mathcal{X}}_n$ with a semiparametric approximation $C_{\boldsymbol{\theta}}$ on ${}^p\boldsymbol{\mathcal{X}}_n$.

Finally, in section 2.3 comes a list of things we might wish to consider when we want to do a model selection.

2.1 Seven copula models

This section presents the seven bivariate copula-families that will be used when we in chapter 4 investigate the small-sample performance of xv-CIC, i.e. we will first use them to create simulations $\boldsymbol{\mathcal{X}}_n$, and then they will be used as proposed models $\boldsymbol{\mathcal{C}}$ that will be fitted to the corresponding pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$.

We will in this thesis use names that matches those in the `copula`-package, and the copula models will henceforth be denoted by `clayton`, `frank`, `galambos`, `gumbel`, `huslerReiss`, `normal` and `t`.

The first five of these copula models are all one-parametric Archimedean copulas, while the two last are elliptical copula models. Note that we in the bivariate case only have

one parameter in the correlation matrix, so the bivariate `normal` copula is thus also one-parametric. The `t` copula have an additional parameter in its degrees of freedom, but we will in our analysis fix this value to four, i.e. we should strictly speaking rather write `t (df=4)` than `t`.

In chapter 1 we skipped the discussion of how the different copula models where obtained, so let us mention that before we look closer upon the models mentioned above.

One way to produce copula models is to start out with Sklar's theorem, see eq. (1.2) and use the inverses of the marginals to obtain the following expression for the copula C corresponding to F,

$$C(u_1 \ldots, u_d) = F({F_1}^{-1}(u_1), \ldots, {F_d}^{-1}(u_d)). \tag{2.1}$$

This method is only useful in those cases where we do have a well defined d-variate cdf F to start with, and this restricts the possibilities. But we can at least obtain d-variate copula models from the d-variate elliptical distributions by this strategy.

Another class of d-variate copula models is the Archimedean copulas:
A d-variate copula $C(\boldsymbol{u})$ is Archimedean if there exists a $\psi(t) : [0, \infty] \to [0, 1]$ such that

$$C(\boldsymbol{u}) = \psi\left(\sum_{i=1}^{d} \psi^{-1}(u_i)\right), \tag{2.2}$$

where the generator ψ is subject to the following restrictions

1. ψ is continuous and decreasing, with $\psi(0) = 1$ and $\psi(\infty) \stackrel{\text{def}}{=} \lim_{t\to\infty} \psi(t) = 0$,
2. ψ is strictly decreasing on $[0, \psi^{-1}(0)]$,
3. $\psi \in C^d(0, \infty)$ (i.e. at least d times differentiable) and for all k in $\{1, \ldots, d\}$ we have $(-1)^k \psi^{(k)}(t) \geq 0$, and
4. the inverse $\psi^{-1}(\tau) : [0, 1] \to [0, \infty]$ has $\psi^{-1}(0) \stackrel{\text{def}}{=} \inf\{t : \psi(t) = 0\}$.

Note: This definition of Archimedean copula is in line with the one used for the bivariate case in Joe [4, p 86], and the multivariate formulation here is based on Hofert [17, p 52]. This definition is the most convenient to work with when we want to find tractable expressions for the derivatives, otherwise we could just as well have followed the receipt given in Nelsen [3, Chapter 4]. In the latter case we would have used a strictly decreasing function $\phi(\tau) : [0, 1] \to [0, \infty]$, with $\phi(1) = 0$, and then defined an Archimedean copula to be a copula expressible as $C(\boldsymbol{u}) = \phi^{[-1]}(\phi(u_1) + \cdots + \phi(u_d))$, where $\phi^{[-1]}(t)$ (the generalized inverse) is defined as $\phi^{-1}(t)$ on $[0, \phi(0))$ and as 0 otherwise, and where moreover $\phi^{[-1]}(t)$ must satisfy $(-1)^k \left(\phi^{[-1]}(t)\right)^{(k)} \geq 0$ for $k = 1, \ldots, d$.

Remember from chapter 1 that the copula models mentioned above in general can be too rigid to be useful when we want to model a d-variate set of observations, but that

we by the help of vine copulas can model $\boldsymbol{\mathcal{X}}_n$ by using bivariate copula models on the conditioned bivariate subsets corresponding to the edges of a regular vine. Note: Another approach to the modeling of d-variate observations is to use nested Archimedean copulas – interested readers can consult Hofert [17] for further details.

Some cdfs: With regard to the testing performed in chapter 4, the explicit functions/generators describing our seven copula models are of no interest whatsoever, since everything is taken care of by the functions in the `copula`-package – but for the sake of completeness we nevertheless include the cdfs for the seven bivariate copulas we consider.

1. cdf for the bivariate `normal` copula, $\rho \in (-1, 1)$

$$C(u_1, u_2) = \Phi_\rho \left(\Phi^{-1}(u_1), \Phi^{-1}(u_2) \right),$$

   where Φ_ρ denotes the bivariate standard normald cdf with correlation ρ.

2. cdf for the bivariate `t` copula, $\rho \in (-1, 1)$, $\nu > 0$ degrees of freedom

$$C(u_1, u_2) = t_{\rho,\nu} \left({t_\nu}^{-1}(u_1), {t_\nu}^{-1}(u_2) \right),$$

   where $t_{\rho,\nu}$ denotes the cdf of the bivariate standard t distribution with correlation parameter ρ and ν degrees of freedom.

3. cdf for the bivariate `clayton` copula, $\theta > 0$

$$C(u_1, u_2) = \left({u_1}^{-\theta} + {u_2}^{-\theta} - 1 \right)^{-\frac{1}{\theta}}.$$

4. cdf for the bivariate `frank` copula, $\theta \in \mathbb{R} \setminus \{0\}$

$$C(u_1, u_2) = -\frac{1}{\theta} \log \left[1 + \frac{\left(e^{-\theta u_1} - 1 \right) \left(e^{-\theta u_2} - 1 \right)}{e^{-\theta} - 1} \right].$$

5. cdf for the bivariate `gumbel` copula, $\theta \geq 1$

$$C(u_1, u_2) = \exp \left[- \left((-\log(u_1))^\theta + (-\log(u_2))^\theta \right)^{\frac{1}{\theta}} \right].$$

6. cdf for the bivariate `galambos` copula, $\theta > 0$

$$C(u_1, u_2) = \exp \left[\log(u_1 u_2) \cdot \left(1 - \left(\left(\frac{\log(u_2)}{\log(u_1 u_2)} \right)^{-\theta} + \left(1 - \left(\frac{\log(u_2)}{\log(u_1 u_2)} \right) \right)^{-\theta} \right)^{-\frac{1}{\theta}} \right) \right]$$

7. cdf for the bivariate `huslerReiss` copula, $\theta > 0$

$$C(u_1, u_2) = \exp\left[\log(u_1 u_2) \cdot \left(\frac{\log(u_2)}{\log(u_1 u_2)} \cdot \Phi\left(\frac{1}{\theta} + \frac{\theta}{2} \cdot \log\left(\frac{\frac{\log(u_2)}{\log(u_1 u_2)}}{1 - \frac{\log(u_2)}{\log(u_1 u_2)}}\right)\right)\right.$$
$$\left. + \left(1 - \frac{\log(u_2)}{\log(u_1 u_2)}\right) \cdot \Phi\left(\frac{1}{\theta} - \frac{\theta}{2} \cdot \log\left(\frac{\frac{\log(u_2)}{\log(u_1 u_2)}}{1 - \frac{\log(u_2)}{\log(u_1 u_2)}}\right)\right)\right)\Bigg]$$

Note: The first five cdfs are taken from Brechmann [7, p. 9-11], in which more details can be found. The two latest cdfs are as given in the `copula`-package. It is a trivial task to simplify the cdf for the `huslerReiss`, but then it would not be given exactly like it is presented in the `huslerReissCopula`-function.

Some plots: Instead of the dreary cdfs mentioned above, what we really should consider are the plots that show us the properties of our copula models, i.e. how they are suited for the different structures that a bivariate set of observations can contain. We will briefly discuss this in the following paragraphs and plots, without going deeply into the technical details.

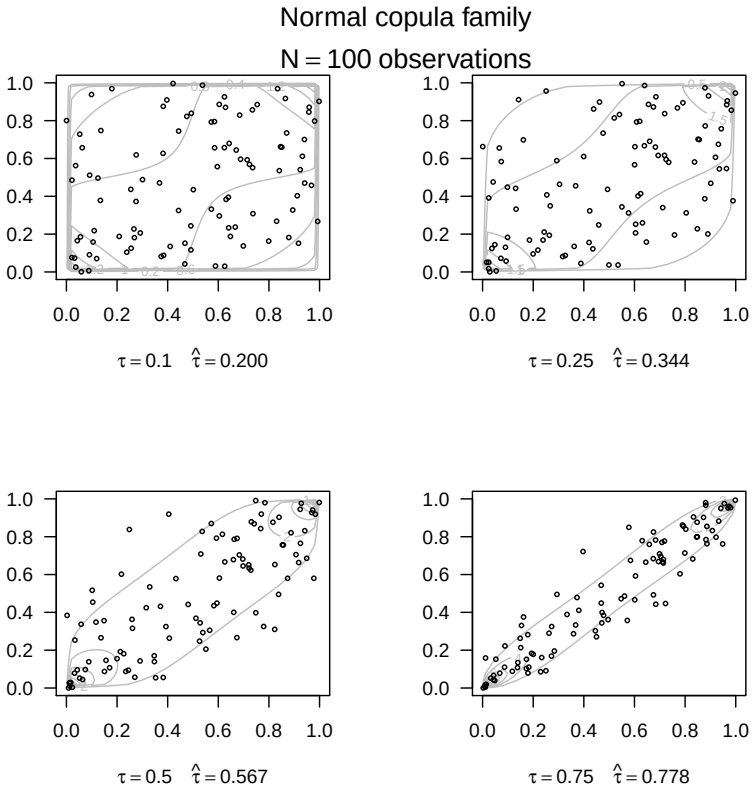
We will give four plots in each figure, in order to show how the contour-lines of our copulas change with the value of Kendall's τ. In addition to the three values used later on in chapter 4, i.e. $\tau \in \{0.25, 0.50, 0.75\}$, we will also include an example with $\tau = 0.10$. The lowest value of τ is included in order to see how all these copula models approaches the independence copula when τ becomes small

In addition to the contour lines, the plots also contain some observations generated from the model in order to see realizations of a concrete sample of size $N = 100$. Moreover, the empirical value of Kendall's τ is also presented – but note that this might deviate somewhat from the true value since the sample-size is rather small.

Elliptical copula models: Lets start out by considering figs. 2.1 and 2.2, which gives us the plots corresponding to the `normal` copula and the `t (df=4)` copula. The symmetry in these plots are easy to spot – and if we compare the subplots corresponding to the different values of Kendall's τ, we can see that the `t (df=4)` have some more weight at the lower left and upper right corner. This difference will diminish when the degrees of freedom increase, and it is customary to use the `normal`-copula as an approximation if the degrees of freedom in the `t`-copula exceed thirty.

Extreme value models: From Nelsen [3, p. 97] we learn that a copula C_* is an extreme value copula if there exists a copula C such that $C_*(u, v) = \lim_{n\to\infty} C^n(u^{1/n}, v^{1/n})$. Three of our five Archimedean copula models belongs to this class, i.e. `galambos`, `gumbel` and `huslerReiss`. Plots corresponding to these three are given in figs. 2.3 to 2.5. The

FIGURE 2.1: `normal` copula - four values of tau



contour-lines in these plots shows that there are a higher chance of finding observations where the two covariates both have large values or small values, and that the former of these cases is the most frequent one. These three copula models is hard to distinguish from each other, and their closeness makes it difficult for the model selection methods to correctly identify the data-generating process based on small samples.

The `frank` copula In fig. 2.6 the contour lines and sample-examples for the `frank` copula is presented. This is a symmetric copula, which (for higher values of τ) has prominent tails and otherwise has a lot of its density concentrated along the diagonal from the lower left to the upper right corner.

The `clayton` copula The plot of the last copula model, the `clayton`, is given in fig. 2.7. This copula is asymmetric, with greater dependence in the negative tail. It is thus easier for the selection methods to correctly identify data generated by this copula.

FIGURE 2.2: `t` copula - four values of tau

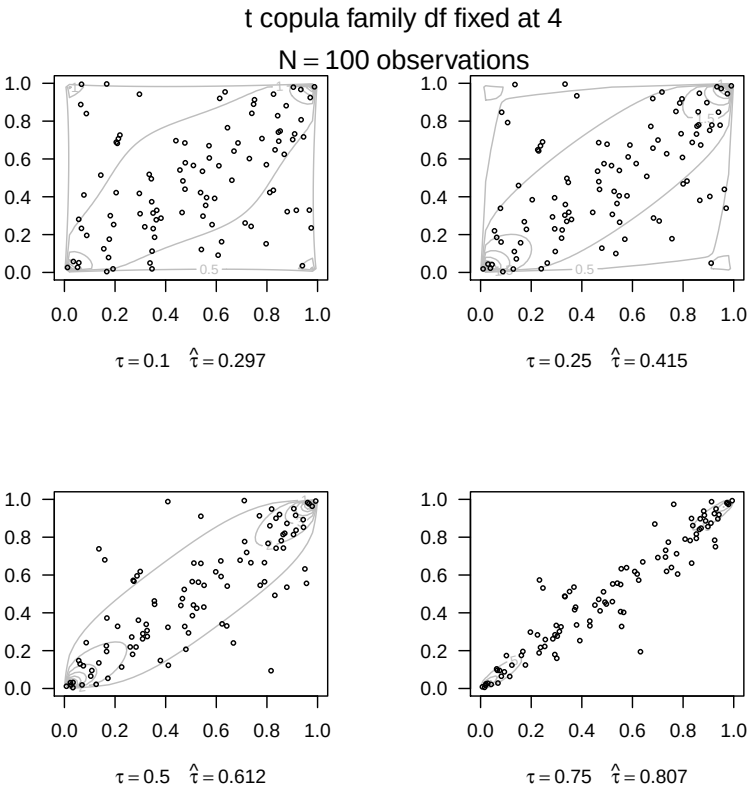


FIGURE 2.3: `galambos` copula - four values of tau

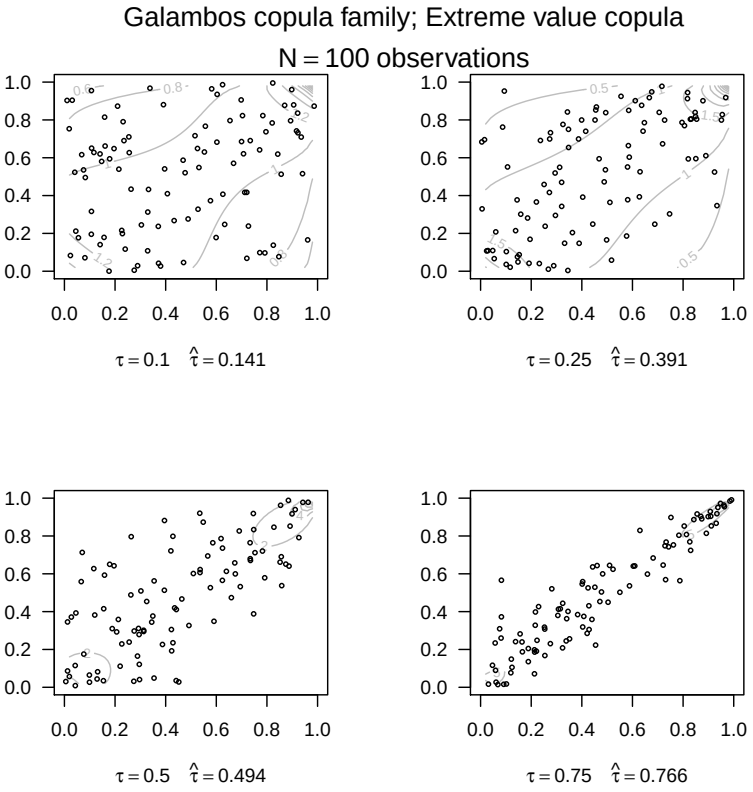


FIGURE 2.4: `gumbel` copula - four values of tau

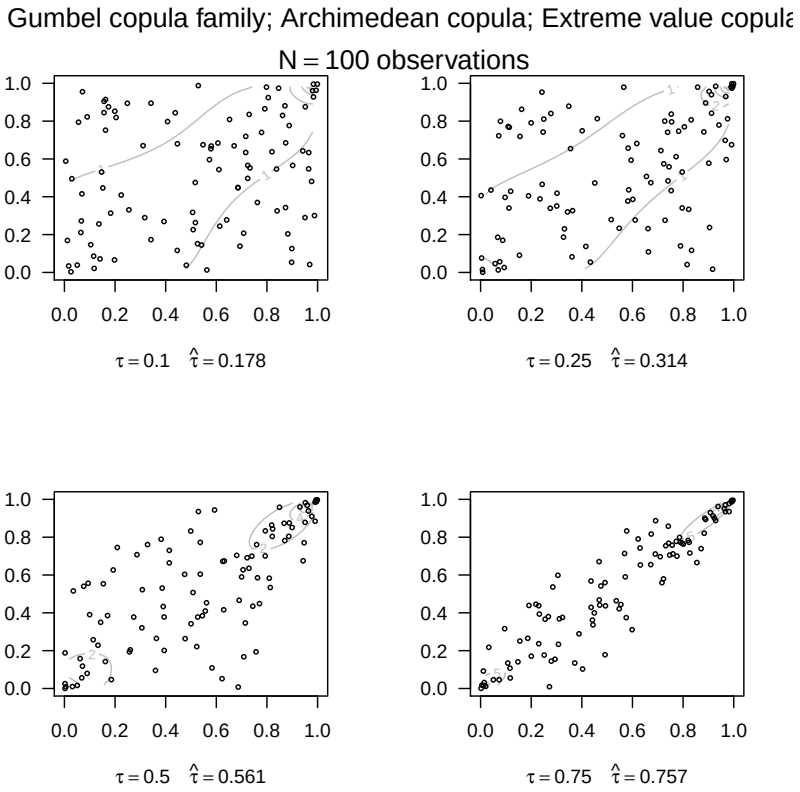


FIGURE 2.5: `huslerReiss` copula - four values of tau

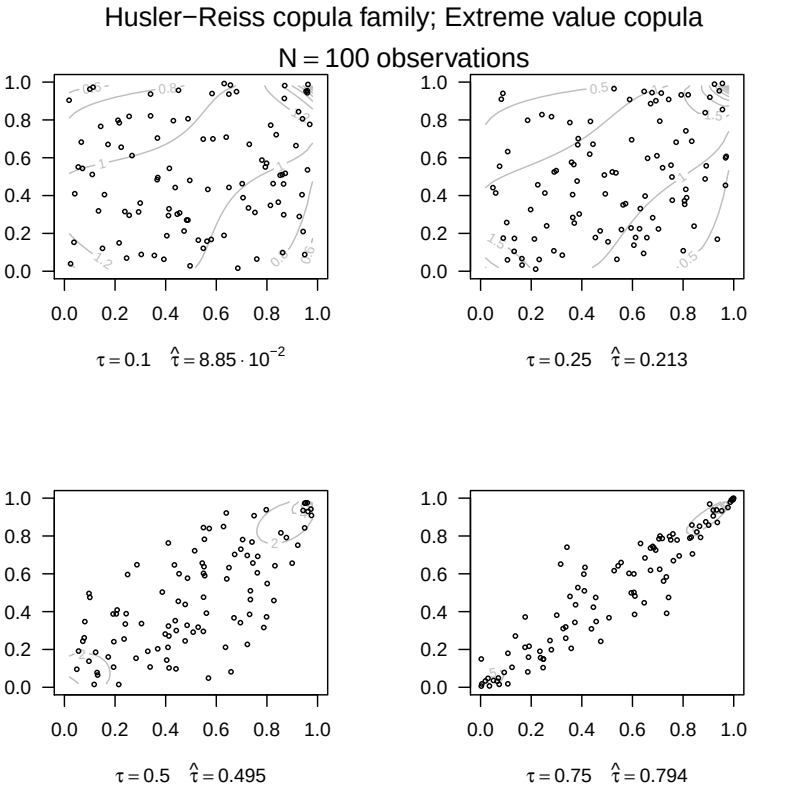


FIGURE 2.6: `frank` copula - four values of tau

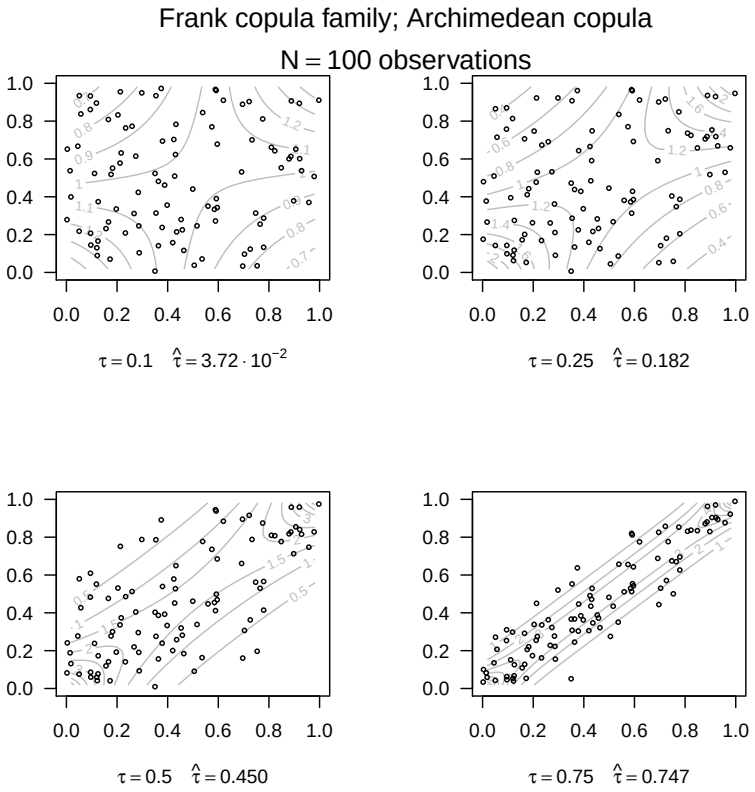
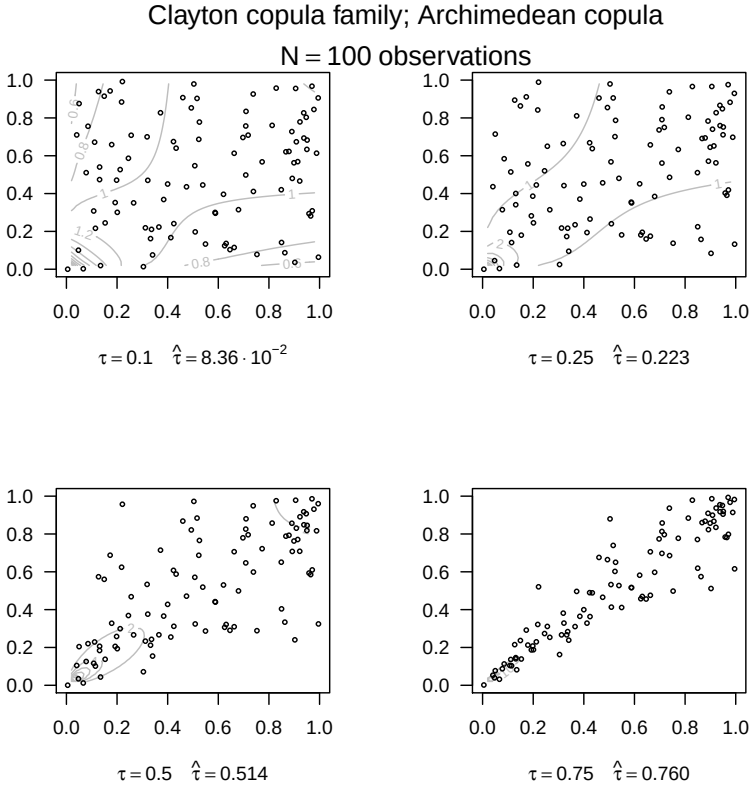


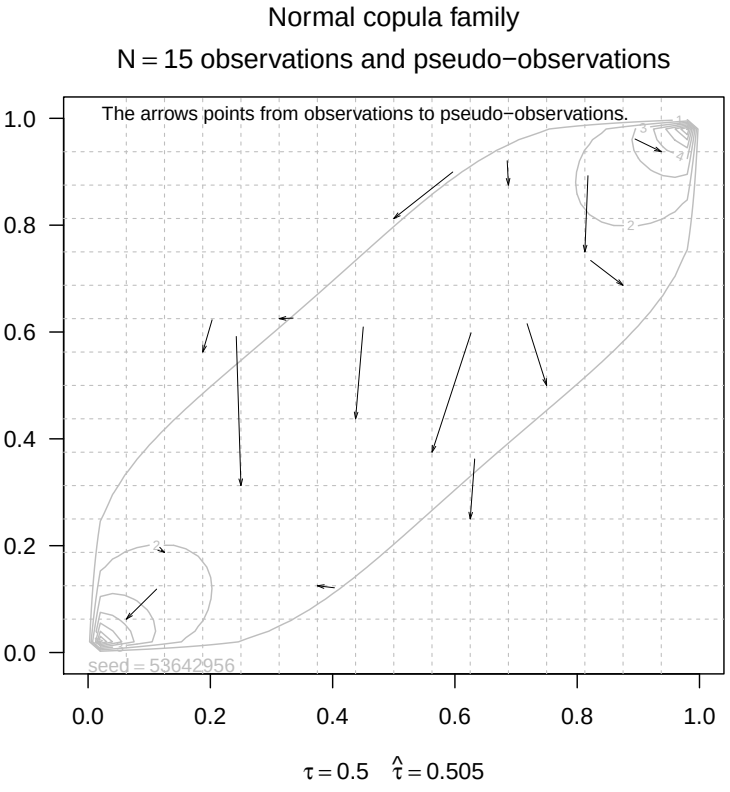
FIGURE 2.7: `clayton` copula - four values of tau



Noise in transformation $\boldsymbol{\mathcal{X}}_n \to {}^p\boldsymbol{\mathcal{X}}_n$ Throughout this thesis, we often refer to the *noise* that originates when the empirical marginals are used to replace a set of independent observations $\boldsymbol{\mathcal{X}}_n$ in $\mathbb{R}^d$ with dependent pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ in $[0,1]^d$.

When this phrase is used, we have in mind the effect which is illustrated in fig. 2.8, where we have used a ridiculous small sample of size $N = 15$ to stress the fact that the observations $\boldsymbol{\mathcal{X}}_n$ (the starting point of the arrows) will be shuffled around quite a bit before we have the dependent pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ (the end points of the arrows).

FIGURE 2.8: Noise in transformation $\boldsymbol{\mathcal{X}}_n \to {}^p\boldsymbol{\mathcal{X}}_n$



The noise that this shuffling incurs on the mpl-based estimate of the parameter, and the corresponding maximum of the pseudo-log-likelihood function, is investigated in section 4.3 – and not surprisingly will the effect diminish when the number of observations grows, cf. figs. 4.1 to 4.4.

Another detail that can be commented upon in view of fig. 2.8, is that the pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ always will be positioned at the gridpoints shown in the figure, i.e. that a bivariate ${}^p\boldsymbol{\mathcal{X}}_n$ based on a sample of size n will be restricted to a subset of n^2 points in $[0,1]^2$.

Moreover, due to the way ${}^p\boldsymbol{\mathcal{X}}_n$ is created by the help of the empirical marginals – we will have that there only can be one pseudo-observations on each horizontal/vertical line

in the grid.[1] This latest statement is under the assumption that there is no ties in our observations $\boldsymbol{\mathcal{X}}_n$, which we can considered to be "automatically" fulfilled whenever we have samples from a continuous data-generating process.

In a practical situation ties can of course occur due to rounding of observed values or grouping of data. In such cases the data can be jiggled slightly to get rid of the ties, and then we consider the pseudo-observation ${}^p\boldsymbol{\mathcal{X}}_n$ corresponding to these jiggled data.

In section 4.4 there is a concrete example where we get rid of ties this way. As discussed there, it seems to be no reason at all to fear that another jiggled version of the data should lead to a different conclusion.

2.2 Estimation of parameters.

This section will comment upon different approaches that can be used when we want to estimate the parameters $\boldsymbol{\alpha}$ of a model $F_{\boldsymbol{\alpha}}(\boldsymbol{x})$, in order to get an optimal fit to some d-variate observations $\boldsymbol{\mathcal{X}}_n$.

We know from Sklar's theorem, cf. eq. (1.2), that we can express the cdf $F_{\boldsymbol{\alpha}}$ by a copula C that takes care of all the dependencies between the covariates, and the d cdfs of the marginal distribution. It is thus possible to separate the content of $\boldsymbol{\alpha}$ into a couple of vectors $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, which respectively represent the parameters of the copula C and the collection of parameters that are needed in the d marginal distributions $(F_1, \ldots, F_d)$.

To make this explicit in our notation, we will write $\boldsymbol{\mathcal{F}}_{(\boldsymbol{\theta},\boldsymbol{\gamma})}$ instead of $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$ and we will write $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x}) = C_{\boldsymbol{\theta}}\left(\boldsymbol{F}_{\boldsymbol{\gamma},\perp}(\boldsymbol{x})\right)$, with $\boldsymbol{F}_{\boldsymbol{\gamma},\perp}(\boldsymbol{x}) = \left(F_{\boldsymbol{\gamma}(1)}(x_1), \ldots, F_{\boldsymbol{\gamma}(d)}(x_d)\right)$. Note that there is no a priori reason to assume that the parameters of the marginal distributions should be disjunct. The notation $\boldsymbol{\gamma}(i)$ thus represents the sub-vector of $\boldsymbol{\gamma}$ needed for the parametrization of the i'th marginal cdf, i.e. $F_i(x_i) = F_{\boldsymbol{\gamma}(i)}(x_i)$.

We will consider different approaches with regard to estimating the parameters for a model that aims to describe some d-variate observations $\boldsymbol{\mathcal{X}}_n$. If the observations stem from some (unknown) distribution F°, and we want to find the parameters $(\boldsymbol{\theta},\boldsymbol{\gamma})$ that bests fit a model $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x})$ to $\boldsymbol{\mathcal{X}}_n$ – then we would like some kind of "measure of distance" in order to decide the optimal parameters given the data we have available.

We will in section 3.2 discuss in detail one way to estimate the parameters $(\boldsymbol{\theta},\boldsymbol{\gamma})$ that ensures that our fitted model $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x})$ is the "closest one" to the true model F°, and

[1] This implies that the pseudo-observations are dependent.

that is the Kullback-Leibler information criterion – defined relatively the pdfs f° and f,

$$\begin{aligned}\mathrm{KLIC}\left(f^\circ, f\right) &\stackrel{\text{def}}{=} \int \log \frac{f^\circ}{f}\,\mathrm{d}F^\circ \\ &= \int \log f^\circ\,\mathrm{d}F^\circ - \int \log f\,\mathrm{d}F^\circ \\ &= \mathrm{E}_{f^\circ}\left[\log f^\circ\right] - \mathrm{E}_{f^\circ}\left[\log f\right]. \end{aligned} \tag{2.3}$$

For the purpose of this section, the result of interest (see section 3.3.1) is that the parameters that minimize the KLIC-function will be the same that maximize the likelihood-function.

Thus, if we insert the maximum likelihood estimate of the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ into our function $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x})$, we have the model that based on our sample $\boldsymbol{\mathcal{X}}_n$ is the closest one to F°. This implies that the following will be our strategy in a fully parametric setting.

First approach, MLE: We make assumptions regarding which parametric families the d marginals $F_{\boldsymbol{\gamma}(i)}$ and the copula $C_{\boldsymbol{\theta}}$ stem from, and then we use a *maximum likelihood estimation* (MLE) to find an estimate $(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{MLE}})$ of the optimal parameter configuration $(\boldsymbol{\theta}^\circ, \boldsymbol{\gamma}^\circ) \stackrel{\text{def}}{=} \operatorname{argmin}_{(\boldsymbol{\theta},\boldsymbol{\gamma})} \mathrm{KLIC}\left(f^\circ, f_{(\boldsymbol{\theta},\boldsymbol{\gamma})}\right)$.

A problem with this approach is that the estimation of all the parameters at the same time can imply that the estimation-algorithm literally will take for ever to finish. It might thus be prudent to consider the following approach instead.

Second approach, IFM: We make the same assumptions as in the first approach, but instead of MLE we use the two-step procedure named *inference functions for margins* (IFM) method, that was introduced in Joe and Xu [18]. In the first step we find an estimate $\widehat{\boldsymbol{\gamma}}_{\mathrm{IFM}}$ of the marginal parameters $\boldsymbol{\gamma}$,

$$\widehat{\boldsymbol{\gamma}}_{\mathrm{IFM}} = \operatorname*{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{d}\sum_{j=1}^{n} \log f_{\boldsymbol{\gamma}(i)}\left(X_{ij}\right), \tag{2.4}$$

and then we proceed in the next step by replacing $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x}) = C_{\boldsymbol{\theta}}\left(\boldsymbol{F}_{\boldsymbol{\gamma},\perp}(\boldsymbol{x})\right)$ with $F_{(\boldsymbol{\theta},\mathrm{IFM})}(\boldsymbol{x}) = C_{\boldsymbol{\theta}}(\boldsymbol{F}_{\widehat{\boldsymbol{\gamma}}_{\mathrm{IFM}},\perp}(\boldsymbol{x}))$, and finally we use MLE to find an estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{IFM}}$ of ${\boldsymbol{\theta}^\circ}_{\mathrm{IFM}} \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\theta}} \mathrm{KLIC}\left(f^\circ, f_{(\boldsymbol{\theta},\,\mathrm{IFM})}\right)$.

Note that this procedure simplifies if the marginal distributions F_i do not have any parameters in common, since we then can use MLE on each individual marginal in the first step. Yan [19] contains further details regarding this approach (and the next one), together with a numerical computation that supports the use of the resulting estimate $(\widehat{\boldsymbol{\theta}}_{\mathrm{IFM}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{IFM}})$ as an approximation to $(\boldsymbol{\theta}^\circ, \boldsymbol{\gamma}^\circ)$.

Third approach, MPLE: When the goal is an estimate of the multivariate dependencies of the distribution F°, we can assume that the copula belongs to some parametric family and avoid the specification of models for the marginals by replacing $\boldsymbol{F}_{\boldsymbol{\gamma},\perp}$ with $\boldsymbol{F}_{n,\perp}(\boldsymbol{x})$, as discussed on page 8.

Since the marginals now are unspecified, we need to use *maximum pseudo-log-likelihood estimation*(MPLE) instead of MLE when we are searching for our parameters.

In this approach we thus replace the problem of fitting $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}(\boldsymbol{x}) = C_{\boldsymbol{\theta}}\left(\boldsymbol{F}_{\boldsymbol{\gamma},\perp}(\boldsymbol{x})\right)$ to the observations $\boldsymbol{\mathcal{X}}_n$ with the problem of fitting $C_{\boldsymbol{\theta}}(\boldsymbol{u})$ to the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ – in which the latter part has to be done by the help of a mpl-technique.

Fourth approach, SSP: The goal of this approach is the same as in the previous one, i.e. our quest is the interdependence parameters. This method is used when $d \geq 3$ and we have used vines and the pair-copula construction in order to create our copula C. A direct application of the third method is then exchanged with a *stepwise semiparametric* (SSP) estimation, in which the strategy resembles the one from IFM, i.e. we estimate the parameters one level at a time. The resulting estimate $\widehat{\boldsymbol{\theta}}_{\text{SSP}}$ obtained by this method will then be used in our model instead of $\widehat{\boldsymbol{\theta}}_{\text{MPLE}}$, which would have been the one we would have obtained if we had considered all the levels simultaneous.

Some comments upon these approaches: The four approaches mentioned above have different drawbacks. The first approach will, unless the number of parameters is small, require quite a demanding computational cost. The computational cost is significantly reduced in the other approaches, but then there is a question regarding how much the estimates have gone astray from the true target $\boldsymbol{\theta}^\circ$.

In addition, the estimates of the interdependency parameters $\boldsymbol{\theta}$ in the two first approaches will be sensitive to the marginal models that has been used. To emphasize, even if we have found the correct model for the copula – if the parametric families we have used for the marginals does not hit the mark, the resulting estimate of the $\boldsymbol{\theta}$-parameters can be severely affected.

The two last approaches do not suffer from any risk of misspecified marginals, as they instead use the empirical marginals to replace the estimation problem to one concerning the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$. This transformation will in addition reduce the computational costs involved in our estimate. And moreover, for higher values of d, limitations on the available computational resources might make SSP our only realistic alternative.

There is however a drawback with the use of the unspecified marginals in the third and fourth approach, and that is that the transformation process from observations to

pseudo-observations introduces some noise into our estimation process. But as we will see later on, this is hardly an issue at all if we have a decent-sized sample to work upon.

2.3 Some model selection issues

The following list is based on a similar list in Claeskens and Hjort [20, chapter 1], and it presents some of the issues that might influence the model selection process.

1. **Models are approximations.** The "true" model that generated the data will in general be unknown. Furthermore, if the true model is very complex, it might be useful to consider a simpler model (e.g. if we have limitations with regard to computational power).

2. **The bias-variance trade-off.** Simplicity vs. complexity: A simple model with few parameters to estimate will have a lower variability at the cost of introducing modeling bias, while complex models with many parameters give a small bias at the cost at higher variability. Statistical model selection methods must seek a proper balance between overfitting (a model with too may parameters, more than actually needed) and underfitting (a model with too few parameters, not capturing the right signal).

3. **Parsimony.** Keep it as simple as possible. Only include more parameters if this improves the predictional quality of the model.

4. **The context.** All modeling is rooted in an appropriate scientific context and is for a certain purpose. "The purpose of models is not to fit the data but to sharpen the questions", S. Karlin at the 11th R. A. Fisher Memorial Lecture, Royal Society 20, April 1983.

5. **The focus.** Some quantities or functions of parameters might be more interesting than others, and model building and model selection could emphasize good performance precisely for those quantities that are more important. Different aims might thus introduce different loss functions for the same data, and thus result in the preference of different models.

6. **Conflicting recommendations.** Different model selection strategies might end up offering different advise for the same data and the same list of candidate models. It is thus important to know how different selection schemes are constructed and what their aims and properties are.

7. **Model averaging.** If a selection strategy does not assign a clear winner, it might be advantageous to combine inference output across the best models.

Chapter 3

The Copula Information Criteria

This chapter will sketch the theory leading to the three different strands of the copula information criterion (CIC). The first two, CIC^{AIC} and CIC^{TIC} were introduced in Grønneberg and Hjort [15], and also discussed in Grønnebergs contribution to Kurowicka and Joe [21],[1] whereas the third one, xv-CIC were presented in Grønneberg [16, Part III].

3.1 An overview

The CIC is a machinery aimed at selecting the best available copula-model c from a collection $\boldsymbol{\mathcal{C}}$ of proposed dependency-models for a set of pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$. It is a semiparametric analogue to selection methods like the Akaike Information Criterion (AIC), introduced in Akaike [13], the Bayesian Information Criterion (BIC) introduced in Schwarz [22], and the Takeuchi Information Criterion (TIC), introduced in Takeuchi [23]. In particular, the main point of CIC is to assign a numerical value to each of the models from $\boldsymbol{\mathcal{C}}$, and then pick the one with the best result, and this assignment must balance bias versus variance by introducing a "penalty" for the most complex models.

Note that several issues might influence the model selection process,[2] and thus our decision of which model that is the "best".[3] Grønneberg [16] states it like this:

> 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

[1] His contribution to [21] is also included in his dissertation, Grønneberg [16, Part II]

[2] See section 2.3 for a list of model selection issues taken from Claeskens and Hjort [20].

[3] There is no guarantee that the "best" of the proposed models is a good model for our observations $\boldsymbol{\mathcal{X}}_n$, so it might be recommendable to apply a goodness-of-fit test to see if the chosen model can be trusted in further applications.

> 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.

The first of these approaches can be referred to as the "loss-function perspective", whereas the second can be refereed to as the "prediction perspective".

The AIC-method was developed in a setting where the models to be ranked are fully parametric, but it has nevertheless been a standard procedure to use a tweaked version of it in the realm of semiparametric models. As mentioned in chapter 1, the formula for the AIC is then modified by replacing the maximum of the log-likelihood function with the maximum of the pseudo-log-likelihood function, cf. eqs. (1.8) and (1.9). In order to emphasize the difference between these two, we will in this thesis use ^{p}AIC as the name of the mpl-based pseudo-variant of AIC.

Note that the AIC formula requires a proper log-likelihood, cf. section 3.3.2, and the ^{p}AIC is thus not formally valid since it uses a pseudo-log-likelihood instead. The use of ^{p}AIC has, according to Grønneberg and Hjort [15], been justified by the belief that in the limit there could be a continuous connection between AIC and ^{p}AIC. However, when the extra noise from the transformation-step from observations to pseudo-observations is accounted for, we end up with the $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ selection formulas, whose bias-correcting terms does not behave as nicely as those encountered in the fully parametric setting. In contrast to the case of AIC, the bias-terms of $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ can actually attain infinite values for copulas with heavy tail-dependence.

The $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ can of course be used to rank a list of semiparametric models for some pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, but all the time we might doubt the conclusion (due to the lack of general applicability), we probably should do something else instead (like using ^{p}AIC).

Even though these two CIC-variants lacks the desired property of being generally applicable, we will include the arguments leading to them since these are useful when giving the formula for the better behaved xv-CIC– which where introduced in Grønneberg [16, Part III].

As discussed in Grønneberg [16, Part III], when the "loss-function perspective" is used instead of the "prediction perspective", we can construct the cross-validation copula information criterion, xv-CIC. The xv-CIC-formula use another threshold for what it consider as low-level noise in the (pseudo)observations, and the resulting formula turns out to be of a generally applicable nature.

The arguments presented in Grønneberg [16] for the semiparametric case is motivated by the machinery that in the parametric case leads to the AIC-formula, and the first sections

of this chapter will thus consider the theory connecting the AIC with the Kullback-Leibler information criterion, and how this is connected with TIC and cross-validation.[4] The latter sections then sketch how CIC^{AIC}, CIC^{TIC} and xv-CIC are developed based on a similar approach with respect to the semiparametric situation.

3.2 Kullback-Leibler, $\text{KLIC}\left(f^\circ, f\right)$

We will in this section look closer upon the Kullback-Leibler information criteria, KLIC, which were defined in eq. (3.1) on page 23. The KLIC is closely related to the loss-function perspective of model selection, since it gives us a way to gauge how much a postulated model f from $\boldsymbol{\mathcal{F}}$ deviates from the true data generating-model f°.

For the sake of completeness of the present section, we restate the definition of KLIC once more.

$$\begin{aligned}\text{KLIC}\left(f^\circ, f\right) &\stackrel{\text{def}}{=} \int \log \frac{f^\circ}{f}\, \mathrm{d}F^\circ \\ &= \int \log f^\circ \, \mathrm{d}F^\circ - \int \log f \, \mathrm{d}F^\circ \\ &= \mathrm{E}_{f^\circ}\left[\log f^\circ\right] - \mathrm{E}_{f^\circ}\left[\log f\right]. \end{aligned} \tag{3.1}$$

By the use of e.g. the Jensen inequality, it can be shown that $\text{KLIC}\left(f^\circ, f\right)$ always will be non-negative, and that it is equal to zero if and only if $f = f^\circ$ almost everywhere. This implies that the KLIC-values can be used to rank the models in $\boldsymbol{\mathcal{F}} = \{f_i\}_{i \in I}$ according to how well they approximate our distribution f°, whereupon we can find the best available approximation by choosing the one which minimizes $\text{KLIC}\left(f^\circ, f_i\right)$.

Note that the KLIC, as defined in eq. (3.1), gives a non-symmetric measure of the difference between two probability distributions, and thus does not give a metric on the space of probability distributions.[5]

In a practical setting with real data $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we will have a situation where the models in $\boldsymbol{\mathcal{F}}$ are those we think might fit our data in the best possible fashion, while the true model f° that we want to approximate will be unknown. In this case, with an unknown f°, we must settle for the empirical distribution and use the observations in $\boldsymbol{\mathcal{X}}_n$ to get an estimate of the Kullback-Leibler divergence.

[4] Kullback-Leibler information criterion is also known as information divergence, information gain, relative entropy or Kullback-Leibler divergence, see e.g. Claeskens and Hjort [20] for details regarding the connection to information and entropy).

[5] A symmetric version of KLIC does exist, but eq. (3.1) is the one used in Grønneberg [16].

If our goal is to find the model in $\boldsymbol{\mathcal{F}}$ which is the best approximation to f°, an inspection of eq. (3.1) shows that it will be sufficient to estimate $\mathrm{E}_{f^\circ}[\log f]$. This follows since the estimate of $\mathrm{E}_{f^\circ}[\log f^\circ]$ will be present in all the estimated Kullback-Leibler divergences, and as we seek the model which minimizes this divergence it will thus be sufficient to find the model which maximizes the part $\mathrm{E}_{f^\circ}[\log f]$.

A note of warning: When we simplify our quest for an optimal model from $\boldsymbol{\mathcal{F}}$, by ignoring the estimation of $\mathrm{E}_{f^\circ}[\log f^\circ]$, we also lose the information telling us whether or not our chosen model $\widetilde{f}$ is an adequate approximation to (the empirical estimate of) f°. In particular: If all of the proposed models in $\boldsymbol{\mathcal{F}}$ are awful as approximations of f°, the one picked by this simplified method might turn out to be rather ill-suited for our intended purposes.

Unless we have some good a priori reason to expect that (at least one of) the proposed models in $\boldsymbol{\mathcal{F}}$ are in the KLIC-vicinity of the true model f°, it is highly recommendable to employ a goodness-of-fit (GoF) test to check the adequacy of $\widetilde{f}$. Brechmann [7] contains (among other things) a very good overview with respect to available GoF-tests for the copula setting, and the interested reader should take a look there.

3.3 Parametric model selection

We will in this section consider the realm of model selection between fully parametric models, before we in section 3.4 presents the adjustments from [16] that is required to deal with the nonparametric setting.

3.3.1 Kullback-Leibler and MLE

We will follow the argumentation from Grønneberg [16], and first consider the justification of the maximum likelihood estimator (MLE) as a method of selecting the member from a fully parametric family $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}} = \{f_{\boldsymbol{\alpha}}\}$ that in a best possible fashion fit a set of independent identically distributed d-variate observations $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. The estimated p-dimensional parameter configuration that identifies this member will be denoted by $\widehat{\boldsymbol{\alpha}}$.

Selection by the use of KLIC. As outlined in section 3.2, the model in $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$ that minimizes the KLIC-value is the same that maximizes the value of $\mathrm{E}_{f^\circ}[\log f_{\boldsymbol{\alpha}}]$. In

particular, the optimal parameter-configuration will be given by $\boldsymbol{\alpha}^\circ$ satisfying

$$\begin{aligned}\boldsymbol{\alpha}^\circ &\overset{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\alpha}} \operatorname{KLIC}(f^\circ, f_{\boldsymbol{\alpha}}) \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \mathrm{E}_{f^\circ}[\log f_{\boldsymbol{\alpha}}] \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \int \log f_{\boldsymbol{\alpha}} \, \mathrm{d}F^\circ. \end{aligned} \tag{3.2}$$

If we have a number of observations $\boldsymbol{\mathcal{X}}_n$ stemming from f°, but f° itself is unknown, we can exchange $\mathrm{d}F^\circ$ with $\mathrm{d}\widehat{F}_n$ in eq. (3.2) – where the d-variate empirical distribution $\widehat{F}_n$ is defined by

$$\begin{aligned}\widehat{F}_n(\boldsymbol{x}) &\overset{\text{def}}{=} n^{-1} \sum_{i=i}^{n} \prod_{j=1}^{d} \mathbb{1}\{X_{j,\,i} \leq x_j\} \\ &= n^{-1} \sum_{i=1}^{n} \mathbb{1}\{\boldsymbol{x}_i \leq \boldsymbol{x}\}. \end{aligned} \tag{3.3}$$

This implies that we based on our observations $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ can compute an estimate $\widehat{\boldsymbol{\alpha}}_n$ of $\boldsymbol{\alpha}^\circ$ by the following expression

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_n &\overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\alpha}} \int \log f_{\boldsymbol{\alpha}} \, \mathrm{d}\widehat{F}_n \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} n^{-1} \sum_{i=1}^{n} \log f_{\boldsymbol{\alpha}}(\boldsymbol{x}_i). \end{aligned} \tag{3.4}$$

The integral in eq. (3.4) is a Lebesgue-Stieltjes integral, and as such its continuity properties implies that given uniform convergence of $\widehat{F}_n$, i.e.

$$\lim_{n \to \infty} \sup_{\boldsymbol{x} \in \mathbb{R}^d} \left| \widehat{F}_n(\boldsymbol{x}) - F^\circ(\boldsymbol{x}) \right| = 0 \qquad \text{almost surely,}$$

we will have that $\widehat{\boldsymbol{\alpha}}_n$ will converge almost surely to $\boldsymbol{\alpha}^\circ$ as $n \to \infty$.

Selection by the use of MLE. When we want to find the optimal parameter configuration by the use of the maximum likelihood estimator, our goal will be to find the parameter value $\widehat{\boldsymbol{\alpha}}$ that maximizes the likelihood function with respect to the available

i.i.d. observations $\boldsymbol{\mathcal{X}}_n$, i.e.

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &\overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\alpha}} L\left(\boldsymbol{\alpha}; \boldsymbol{\mathcal{X}}_n\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \prod_{i=1}^{n} f_{\boldsymbol{\alpha}}\left(\boldsymbol{x}_i\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} e^{\log\left(\prod_{i=1}^{n} f_{\boldsymbol{\alpha}}(\boldsymbol{x}_i)\right)} \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log f_{\boldsymbol{\alpha}}\left(\boldsymbol{x}_i\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}; \boldsymbol{\mathcal{X}}_n),\end{aligned} \tag{3.5}$$

where ℓ is the log-likelihood corresponding to our proposed model $f_{\boldsymbol{\alpha}}$.

The factor n^{-1} in the first of eqs. (3.4) and (3.5) does not affect the attained argmax with respect to $\boldsymbol{\alpha}$, and the two approaches thus gives the same result.

This states that the procedure used to find the maximum likelihood estimate, i.e. that we solve the system of p equations given by setting the score vector $\boldsymbol{u}(\boldsymbol{\alpha}; \boldsymbol{\mathcal{X}}_n) = \frac{\partial}{\partial \boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}; \boldsymbol{\mathcal{X}}_n)$ equal to the zero-vector, gives the same answer as the one we find by using the strategy based on the Kullback-Leibler divergence. In particular, under ordinary regularity conditions, we have

$$\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_n \xrightarrow{\text{a.s}} \boldsymbol{\alpha}^\circ \qquad \text{as } n \to \infty. \tag{3.6}$$

If the true model f° is contained in the parametric family $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$, we will have $f_{\boldsymbol{\alpha}^\circ} = f^\circ$, and in this case we can expect our approximation by $f_{\widehat{\boldsymbol{\alpha}}}$ to be working quite well (at least if the number of observations in $\boldsymbol{\mathcal{X}}_n$ is sufficiently large).

3.3.2 Kullback-Leibler and AIC/TIC

The previous section looked upon the situation where the approximations to f° all stemmed from one parametric family $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$, and we saw that the selection strategy based on the minimization of the KLIC-values, i.e.

$$\widetilde{f} = f_{\boldsymbol{\alpha}^\circ} \qquad \text{where} \qquad \boldsymbol{\alpha}^\circ = \operatorname*{argmin}_{\boldsymbol{\alpha}} \operatorname{KLIC}\left(f^\circ, f_{\boldsymbol{\alpha}}\right), \tag{3.7}$$

gave us the same result we would have obtained by using an approach based on the maximum likelihood estimator.

If the approximating model is to be chosen from a set of N parametric families,[6] i.e. $\boldsymbol{\mathcal{F}} = \cup_{k=1}^{N} \boldsymbol{\mathcal{F}}_{k,\boldsymbol{\alpha}(k)}$, it is clear that the task of finding the best approximation to f° from $\boldsymbol{\mathcal{F}}$ boils down to finding the best approximation from the set that constitutes the best approximations *within* each parametric family. This implies that it is enough to consider the set $\boldsymbol{\mathcal{F}}^\circ = \left\{ f_{1,\boldsymbol{\alpha}(1)^\circ}, \ldots, f_{K,\boldsymbol{\alpha}(N)^\circ} \right\}$ and find the value k° that minimizes the KLIC-values of these models with respect to the model f°. This gives us the following expression for the best approximation to f°:

$$\widetilde{f} = f_{k^\circ,\boldsymbol{\alpha}(k^\circ)^\circ} \qquad \text{where} \qquad \begin{aligned} k^\circ &= \operatorname*{argmin}_{1\leq k\leq N} \operatorname{KLIC}\left(f^\circ, f_{k,\boldsymbol{\alpha}(k)^\circ}\right) \\ &= \operatorname*{argmax}_{1\leq k\leq N} \int \log f_{k,\boldsymbol{\alpha}(k)^\circ} \, \mathrm{d}F^\circ. \end{aligned} \tag{3.8}$$

In a practical situation where we have observations $\boldsymbol{\mathcal{X}}_n = \{\boldsymbol{x}_1 \ldots, \boldsymbol{x}_n\}$ from an unknown model f°, we will need a strategy to select the optimal approximation from $\boldsymbol{\mathcal{F}}$ with respect to the available information. To do this we will need estimates $\widehat{\boldsymbol{\alpha}(k)}_n$ for $\boldsymbol{\alpha}(k)^\circ$, for $k = 1, \ldots, N$ and an estimate $\widetilde{k}_n$ for k.

With the same reasoning as in section 3.3.1, we see that within each parametric family $\boldsymbol{\mathcal{F}}_{k,\boldsymbol{\alpha}(k)}$, the expression for $\widehat{\boldsymbol{\alpha}(k)}_n$ will be given as in eq. (3.4)

$$\widehat{\boldsymbol{\alpha}(k)}_n \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\alpha}(k)} \int \log f_{k,\boldsymbol{\alpha}(k)} \, \mathrm{d}\widehat{F}_n. \tag{3.9}$$

We know that $\widehat{\boldsymbol{\alpha}(k)}_n$ converges almost surely to $\boldsymbol{\alpha}(k)^\circ$ as $n \to \infty$, but in a situation with a finite sample $\boldsymbol{\mathcal{X}}_n$ there will be a non-negligible bias. This bias can be ignored when we seek the best model within a parametric family $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}(k)}$, but we can not ignore it when we want to select a model from a set of several parametric families.

It is thus necessary to establish a bias-correcting term Bias_k, such that the estimator of k° becomes

$$\widetilde{k}_n \overset{\text{def}}{=} \operatorname*{argmax}_{1\leq k\leq N} \left[\int \log f_{k,\widehat{\boldsymbol{\alpha}(k)}_n} \, \mathrm{d}\widehat{F}_n - \operatorname{Bias}_k \right]. \tag{3.10}$$

Bias correction Following the arguments in Grønneberg [16], we will now present a bias correcting term for the estimators of $\int \log f_{k,\boldsymbol{\alpha}^\circ} \, \mathrm{d}F^\circ$. Since the situation is identical

[6] For example if we think that one of the parameters in a distribution is connected with the observed covariates through some polynomial relation, and want to find out whether a linear or quadratic relation gives the best model.

for all the K parametric families $\boldsymbol{\mathcal{F}}_{k,\boldsymbol{\alpha}(k)}$ in $\boldsymbol{\mathcal{F}}$, we can simplify the notation and consider the integral $R(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \int \log f_{\boldsymbol{\alpha}} \, \mathrm{d}F^\circ$.

We will now consider the following objects:

$$R(\boldsymbol{\alpha}^\circ) = \int \log f_{\boldsymbol{\alpha}^\circ} \, \mathrm{d}F^\circ, \tag{3.11}$$

$$\widehat{R}_n = R(\widehat{\boldsymbol{\alpha}}_n) = \int \log f_{\widehat{\boldsymbol{\alpha}}_n} \, \mathrm{d}F^\circ, \tag{3.12}$$

$$\widehat{Q}_n = n^{-1} \sum_{i=1}^{n} \log f_{\widehat{\boldsymbol{\alpha}}_n}(\boldsymbol{x}_i) = n^{-1} \ell\left(\widehat{\boldsymbol{\alpha}}_n; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right). \tag{3.13}$$

Note: $R(\boldsymbol{\alpha}^\circ)$, as given in eq. (3.11), is the true value we want to estimate. In eq. (3.12) the estimator $\widehat{\boldsymbol{\alpha}}_n$ for $\boldsymbol{\alpha}^\circ$ has been inserted into $R(\boldsymbol{\alpha})$, such that we get $\widehat{R}_n$ as an estimator for $R(\boldsymbol{\alpha}^\circ)$. However, $\widehat{R}_n$ can only be computed when we know the data generating model f°, and thus we need to modify our estimator by exchanging $\mathrm{d}F^\circ$ with $\mathrm{d}\widehat{F}_n$ (the differential of the estimator for the empirical distribution). This leads us to the estimator $\widehat{Q}_n$, as given in eq. (3.13).

The problem, at least for small values of n, is that the estimator $\widehat{Q}_n$ gives biased estimates of $R(\boldsymbol{\alpha}^\circ)$, and thus we need to find a bias correcting term. We will here follow the exposition given in Claeskens and Hjort [20, chapter 2] and arrive at an expression for this bias by the help of p-variate[7] Taylor expansions around $\boldsymbol{\alpha}^\circ$ for the two functions $R(\boldsymbol{\alpha})$ and $Q(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} n^{-1}\ell(\boldsymbol{\alpha}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

The Taylor expansion of $R(\boldsymbol{\alpha})$. Our quest for a bias-correcting term for $\widehat{Q}_n$ starts with a first order Taylor expansion (with second order error term) for the function $R(\boldsymbol{\alpha})$ around the optimal value $\boldsymbol{\alpha}^\circ$.

To be specific, we want to write $R(\boldsymbol{\alpha})$ as the sum of a linearization around $\boldsymbol{\alpha}^\circ$ and a second order error term. This gives us

$$\begin{aligned} R(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha}^\circ) &+ \left[\left(\frac{\partial}{\partial \boldsymbol{\gamma}} R(\boldsymbol{\gamma}) \right) \Big|_{\boldsymbol{\gamma}=\boldsymbol{\alpha}^\circ} \right]^T \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \\ &+ \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)^T \cdot \left[\left(\frac{\partial^2}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} R(\boldsymbol{\gamma}) \right) \Big|_{\boldsymbol{\gamma}=\boldsymbol{\alpha}^\circ + h_1 \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)} \right] \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ), \end{aligned} \tag{3.14}$$

for some $h_1 \in [0, 1]$.

Under the assumption that $\log f_{\boldsymbol{\gamma}}$ behaves nicely enough to allow us to interchange integration with respect to $\mathrm{d}F^\circ$ and partial derivation with respect to $\boldsymbol{\gamma}$, we could just

[7] Here p is the length of the vector $\boldsymbol{\alpha}$.

as well have started out with a Taylor expansion of the integrand $\log f_{\boldsymbol{\gamma}}$. In this case we readily see that eq. (3.14) can be rewritten as

$$\begin{aligned} R(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha}^\circ) &+ \left(\int \boldsymbol{u}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\alpha}^\circ} \, \mathrm{d}F^\circ \right)^T \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \\ &+ \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)^T \cdot \int \boldsymbol{I}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\alpha}^\circ + h_1 \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)} \, \mathrm{d}F^\circ \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \,, \end{aligned} \tag{3.15}$$

where

$$\boldsymbol{u}(\boldsymbol{\gamma})^T = \left[\frac{\partial}{\partial \gamma_1} \log f_{\boldsymbol{\gamma}}, \ldots, \frac{\partial}{\partial \gamma_p} \log f_{\boldsymbol{\gamma}} \right], \qquad \text{and} \tag{3.16}$$

$$\boldsymbol{I}(\boldsymbol{\gamma}) = \begin{bmatrix} \frac{\partial^2}{\partial \gamma_1^2} \log f_{\boldsymbol{\gamma}} & \cdots & \frac{\partial^2}{\partial \gamma_1 \partial \gamma_p} \log f_{\boldsymbol{\gamma}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \gamma_p \partial \gamma_1} \log f_{\boldsymbol{\gamma}} & \cdots & \frac{\partial^2}{\partial \gamma_p^2} \log f_{\boldsymbol{\gamma}} \end{bmatrix}. \tag{3.17}$$

The expressions given in eqs. (3.16) and (3.17) are respectively the score vector and the information matrix of the model $f_{\boldsymbol{\gamma}}$. Since the integrals involved in eq. (3.15) represents expectations, we can rewrite them as

$$\begin{aligned} R(\boldsymbol{\alpha}) &= R(\boldsymbol{\alpha}^\circ) + \left(\mathrm{E}_{f^\circ} \left[\boldsymbol{u}(\boldsymbol{\alpha}^\circ) \right] \right)^T \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \\ &\quad + \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)^T \cdot \mathrm{E}_{f^\circ} \left[\boldsymbol{I}(\boldsymbol{\alpha}^\circ + h_1 \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)) \right] \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \\ &= R(\boldsymbol{\alpha}^\circ) - \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)^T \cdot \boldsymbol{J}(\boldsymbol{\alpha}^\circ + h_1 \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ)) \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}^\circ) \end{aligned} \tag{3.18}$$

Note that the expectation of the score vector at the optimal parameter value $\boldsymbol{\alpha}^\circ$ equals zero, and that term has thus been removed from eq. (3.18). Furthermore, the notation $\boldsymbol{J}(\boldsymbol{\gamma}) = -\mathrm{E}_{f^\circ} \left[\boldsymbol{I}(\boldsymbol{\gamma}) \right]$ has been introduced to be in accordance with the notation in Claeskens and Hjort [20, chapter 2]. $\boldsymbol{J}(\boldsymbol{\gamma})$ evaluated at $\boldsymbol{\alpha}^\circ$ gives the *Fisher information matrix* of our model, and this will be denoted by $\boldsymbol{J}$.

With the Taylor expansion from eq. (3.18), we can consider the behavior of the stochastic function $\widehat{R}_n = R(\widehat{\boldsymbol{\alpha}}_n)$. First of all, remember from eq. (3.6) that our estimator $\widehat{\boldsymbol{\alpha}}_n$ is the same as the maximum likelihood estimator, and thus we have the following result

$$\widehat{\boldsymbol{\alpha}}_n = \boldsymbol{\alpha}^\circ + \boldsymbol{J}^{-1} \cdot \bar{\boldsymbol{u}}_n + o_{_P}\left(n^{-1/2} \right), \tag{3.19}$$

where $\bar{\boldsymbol{u}}_n = n^{-1} \sum_{i=1}^n u(\boldsymbol{\alpha}^\circ, \boldsymbol{x}_i)$.

When n grows, and under regularity conditions like those found in Hjort and Pollard [24], the central limit theorem tells us that there is a convergence in distribution

$$\sqrt{n}\bar{\boldsymbol{u}}_n \xrightarrow{d} \boldsymbol{u}' \sim \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{K}\right), \tag{3.20}$$

where $\boldsymbol{K} = \mathrm{Var}_{f^\circ}\left(\boldsymbol{u}(\boldsymbol{\alpha}^\circ)\right)$ is the covariance matrix of the score vector evaluated at the optimal parametric value $\boldsymbol{\alpha}^\circ$.

Note that $\boldsymbol{x}_n = o_{\mathrm{P}}(1)$ means that for all $\epsilon > 0$ we have $\lim_{n\to\infty} \mathrm{P}\left(|\boldsymbol{x}_n| > \epsilon\right) = 0$, and furthermore, $\boldsymbol{x}_n = o_{\mathrm{P}}(g(n))$ means that $\boldsymbol{x}_n/g(n) = o_{\mathrm{P}}(1)$. We can thus rewrite eq. (3.19) as

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ\right) = \boldsymbol{J}^{-1} \cdot \left(\sqrt{n}\bar{\boldsymbol{u}}_n\right) + o_{\mathrm{P}}(1)\,. \tag{3.21}$$

This, together with (3.20), gives the following convergence in distribution

$$\boldsymbol{v}_n \overset{\text{def}}{=} \sqrt{n}\left(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ\right) \xrightarrow{d} \boldsymbol{J}^{-1} \cdot \boldsymbol{u}' \sim \mathrm{N}_p\left(\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}\right). \tag{3.22}$$

To abbreviate our notation we will henceforth use $\boldsymbol{v}_n$ as defined in eq. (3.22) and moreover we will introduce $\widetilde{\boldsymbol{J}}_{h_1,n} \overset{\text{def}}{=} \boldsymbol{J}(\boldsymbol{\alpha}^\circ + h_1 \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ))$. With these conventions we can rewrite $\widehat{R}_n$, as given by inserting $\widehat{\boldsymbol{\alpha}}_n$ into eq. (3.18), as

$$\widehat{R}_n = R(\boldsymbol{\alpha}^\circ) - \frac{1}{2} n^{-1} \boldsymbol{v}_n^T \cdot \widetilde{\boldsymbol{J}}_{h_1,n} \cdot \boldsymbol{v}_n. \tag{3.23}$$

The Taylor expansion of $\widehat{Q}_n$ When used on $Q(\boldsymbol{\alpha}) = n^{-1}\ell(\boldsymbol{\alpha}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, the Taylor expansion argument gives the following expression for $\widehat{Q}_n = n^{-1}\sum_{i=1}^n \log f_{\widehat{\boldsymbol{\alpha}}_n}(\boldsymbol{x}_i)$, the estimator from eq. (3.13),

$$\widehat{Q}_n = n^{-1}\sum_{i=1}^n \log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) \tag{3.24a}$$

$$+ n^{-1}\sum_{i=1}^n \boldsymbol{u}(\boldsymbol{\alpha}^\circ; \boldsymbol{x}_i)^T \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ) \tag{3.24b}$$

$$+ n^{-1}\sum_{i=1}^n \frac{1}{2}\left(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ\right)^T \cdot \boldsymbol{I}\left(\boldsymbol{\alpha}^\circ + h_2 \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ)\,; \boldsymbol{x}_i\right) \cdot \left(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ\right), \tag{3.24c}$$

for some $h_2 \in [0, 1]$

Before we investigate the difference between the two estimators $\widehat{Q}_n$ and $\widehat{R}_n$, we will in accordance with Claeskens and Hjort [20] rewrite/rename the parts of eq. (3.24) to get a more compact notation.

For the first part, eq. (3.24a), we introduce the variables $Z_i = \log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) - R(\boldsymbol{\alpha}^\circ)$, $i = 1, \ldots, n$ and $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$, such that we can write

$$n^{-1} \sum_{i=1}^{n} \log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) = n^{-1} \sum_{i=1}^{n} \{\log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) - R(\boldsymbol{\alpha}^\circ) + R(\boldsymbol{\alpha}^\circ)\} \tag{3.25a}$$

$$= n^{-1} \sum_{i=1}^{n} Z_i + R(\boldsymbol{\alpha}^\circ) \tag{3.25b}$$

$$= \bar{Z}_n + R(\boldsymbol{\alpha}^\circ). \tag{3.25c}$$

For the second part we will use $\bar{\boldsymbol{u}}_n = n^{-1} \sum_{i=1}^n u(\boldsymbol{\alpha}^\circ, \boldsymbol{x}_i)$ and $\boldsymbol{v}_n = \sqrt{n}\,(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ)$ to rewrite eq. (3.24b) like

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{\alpha}^\circ; \boldsymbol{x}_i)^T \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ) = \bar{\boldsymbol{u}}_n^T \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ) \tag{3.26a}$$

$$= n^{-1} \left(\sqrt{n}\bar{\boldsymbol{u}}_n^T\right) \cdot \left(\sqrt{n}\,(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ)\right) \tag{3.26b}$$

$$= n^{-1} \left(\sqrt{n}\bar{\boldsymbol{u}}_n^T\right) \cdot \boldsymbol{v}_n. \tag{3.26c}$$

The inclusion of $\sqrt{n}$ with the factor $\bar{\boldsymbol{u}}_n$ is done so that we can take advantage of (3.20) when we go to the limit.

In the error term we introduce $\boldsymbol{J}_{h_2,n} \overset{\text{def}}{=} -n^{-1} \sum_{i=1}^n \boldsymbol{I}\left(\boldsymbol{\alpha}^\circ + h_2 \cdot (\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ); \boldsymbol{x}_i\right)$, which together with $\boldsymbol{v}_n$ enables us to express eq. (3.24c) as $-\frac{1}{2} n^{-1} \boldsymbol{v}_n^T \cdot \boldsymbol{J}_{h_2,n} \cdot \boldsymbol{v}_n$. Note that $\boldsymbol{J}_{h_2,n}$ converges to $\boldsymbol{J}$ in probability, and this is an important element of our analysis of the behavior of $\widehat{Q}_n - \widehat{R}_n$.

All together this implies that we can rewrite eq. (3.24) like

$$Q_n = \bar{Z}_n + R(\boldsymbol{\alpha}^\circ) + n^{-1} \left(\sqrt{n}\bar{\boldsymbol{u}}_n^T\right) \cdot \boldsymbol{v}_n - \frac{1}{2} n^{-1} \boldsymbol{v}_n^T \cdot \boldsymbol{J}_{h_2,n} \cdot \boldsymbol{v}_n, \tag{3.27}$$

which together with eq. (3.23) gives us the following expression for $\widehat{Q}_n - \widehat{R}_n$:

$$\widehat{Q}_n - \widehat{R}_n = \bar{Z}_n + n^{-1} \left(\sqrt{n}\bar{\boldsymbol{u}}_n^T\right) \cdot \boldsymbol{v}_n - \frac{1}{2} n^{-1} \boldsymbol{v}_n^T \cdot \left(\boldsymbol{J}_{h_2,n} - \widetilde{\boldsymbol{J}}_{h_1,n}\right) \cdot \boldsymbol{v}_n. \tag{3.28}$$

Due to the fact that $\boldsymbol{J}_{h_2,n}$ and $\widetilde{\boldsymbol{J}}_{h_1,n}$ both converge in probability to $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\alpha}^\circ)$, their difference in the last part of eq. (3.28) tends to the zero-matrix. In particular, this

implies that we have

$$n \cdot \left[-\frac{1}{2} n^{-1} \boldsymbol{v}_n^T \cdot \left(\boldsymbol{J}_{h_2,n} - \widetilde{\boldsymbol{J}}_{h_1,n} \right) \cdot \boldsymbol{v}_n \right] \xrightarrow[n\to\infty]{P} -\frac{1}{2} \left(\boldsymbol{J}^{-1} \cdot \boldsymbol{u}' \right)^T \cdot \mathbf{0} \cdot \left(\boldsymbol{J}^{-1} \cdot \boldsymbol{u}' \right) = 0, \quad (3.29)$$

i.e. the expression for the error term is simply of type $o_P\left(n^{-1}\right)$. With this observation we can rewrite eq. (3.28) as

$$\widehat{Q}_n - \widehat{R}_n = \bar{Z}_n + n^{-1} p_n + o_P\left(n^{-1}\right), \quad (3.30)$$

where $p_n \overset{\text{def}}{=} \sqrt{n}\bar{\boldsymbol{u}}_n^T \cdot \boldsymbol{v}_n = \sqrt{n}\left(n^{-1}\sum_{i=1}^n u(\boldsymbol{\alpha}^\circ, \boldsymbol{x}_i)\right)^T \cdot \left(\sqrt{n}\left(\widehat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^\circ\right)\right)$.

Note that by eq. (3.19) we can rewrite p_n as $p_n = \sqrt{n}\bar{\boldsymbol{u}}_n^T \cdot \left(\boldsymbol{J}^{-1} \cdot \sqrt{n}\bar{\boldsymbol{u}}_n + o_P(1)\right)$, which by eq. (3.20) will converge in distribution to

$$p_\infty \overset{\text{def}}{=} \left(\boldsymbol{u}'\right)^t \cdot \boldsymbol{J}^{-1} \cdot \boldsymbol{u}'. \quad (3.31)$$

In accordance with the convention from [20] and [16], we will denote the expectation of p_∞ with p^*, and thus we have

$$p^* \overset{\text{def}}{=} \mathrm{E}_{\mathcal{L}}\left[p_\infty\right] = \mathrm{Tr}\left(\boldsymbol{J}^{-1}\boldsymbol{K}\right), \quad (3.32)$$

where $\boldsymbol{J}$ is the Fisher information matrix, and $\boldsymbol{K}$ is the covariance matrix from eq. (3.20).

Note that we, since we deal with n-variate stochastic variables of the i.i.d. variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (all distributed like $\boldsymbol{X}$), need to take the expectations with respect to their joint distribution function in eq. (3.32), i.e. the expectations are with respect to $\mathcal{L} \overset{\text{def}}{=} \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^n f^\circ(\boldsymbol{x}_i)$.

We now have at our disposal all the ingredients, and we can thus go through the required steps leading to an expression for the bias-correcting term Bias_k from eq. (3.10),

The expectation of $\widehat{Q}_n - \widehat{R}_n$. We can compute the expectation of one part of eq. (3.30) even without knowledge of the data-generating distribution f°: The term $\bar{Z}_n = n^{-1}\sum_{i=1}^n \left[\log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) - R(\boldsymbol{\alpha}^\circ)\right]$ has a very simple structure since it does not contain $\widehat{\boldsymbol{\alpha}}_n$. Due to the definition of $R(\boldsymbol{\alpha}^\circ)$ from eq. (3.11), the following simple computation shows that its expectation equals zero.

$$
\begin{aligned}
\mathrm{E}_{\mathcal{L}}\left[\bar{Z}_n\right] &= n^{-1}\sum_{i=1}^{n}\mathrm{E}_{\mathcal{L}}\left[\log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i) - R(\boldsymbol{\alpha}^\circ)\right] \\
&= n^{-1}\sum_{i=1}^{n}\left[\int\cdots\int\prod_{j=1}^{n} f^\circ(\boldsymbol{x}_j)\log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_1\cdots\mathrm{d}\boldsymbol{x}_n - R(\boldsymbol{\alpha}^\circ)\right] \\
&= n^{-1}\sum_{i=1}^{n}\left[\int f^\circ(\boldsymbol{x}_i)\log f_{\boldsymbol{\alpha}^\circ}(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_i - R(\boldsymbol{\alpha}^\circ)\right] \\
&= n^{-1}\sum_{i=1}^{n}\left[R(\boldsymbol{\alpha}^\circ) - R(\boldsymbol{\alpha}^\circ)\right] \\
&= 0.
\end{aligned}
$$

With this observation, we see that the expectation of the difference $\widehat{Q}_n - \widehat{R}_n$ becomes

$$
\mathrm{E}_{\mathcal{L}}\left[\widehat{Q}_n - \widehat{R}_n\right] = n^{-1}\,\mathrm{E}_{\mathcal{L}}\left[p_n\right] + \mathrm{E}_{\mathcal{L}}\left[o_P\left(n^{-1}\right)\right]. \tag{3.33}
$$

Note that the expectations of $o_P\left(n^{-1}\right)$ in eq. (3.33) might be nigh on impossible to compute, and it is by no means certain that the expectation of the $o_P\left(n^{-1}\right)$-part or p_n should be finite.[8]

If we want to estimate the two terms of eq. (3.33), or to be more precise the expectation of the last two terms of eq. (3.28), then this can of course be estimated by the plug-in of $\prod_{i=1}^{n}\widehat{f^\circ(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_i}$ instead of $\prod_{i=1}^{n} f^\circ(\boldsymbol{x}_i)\,\mathrm{d}\boldsymbol{x}_i$, which – in the case of the joint distribution where we only have one observation $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ – leads to the estimate given by $n^{-1}\left(\sqrt{n}\bar{\boldsymbol{u}}_n^T\right)\cdot\boldsymbol{v}_n - \frac{1}{2}n^{-1}\boldsymbol{v}_n^T\cdot\left(\boldsymbol{J}_{h_2,n} - \widetilde{\boldsymbol{J}}_{h_1,n}\right)\cdot\boldsymbol{v}_n$, where the only information we have regarding h_1 and h_2 in the error-term is that they both belong to $[0,1]$. This implies that we instead of an estimated value of the expectation get an intractable bivariate function of (h_1,h_2).

Under the assumptions needed for the Taylor expansions etc. to be true, there will exist h_1 and h_2 that ensures that this function gives ean equality in eq. (3.33). In lack of knowledge regarding the values to use for h_1 and h_2, we could in theory attempt to compute upper and lower limits of the desired expectation. Such an attempt would of course be a rather tedious affair, and probably quite computational expensive too. It is thus desirable to find our estimate by other means.

[8] Note: The expectation of p_n does not exist for many parametric families (e.g. the binomial model, see Claeskens and Hjort [20]), but this does not constitute a problem since the limiting distribution p_∞ have a finite expectation. The breakdown of $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ in the case of semiparametric models is due to the fact that the first-order bias correcting terms then can have limiting distributions with an infinite expectation.

When we have n large, we may take advantage of the negligible behavior of the $o_P\left(n^{-1}\right)$-term and replace that part with zero, and furthermore we can then also replace p_n with the limiting distribution p_∞ from eq. (3.31).

With the expression for the expectation from eq. (3.32), we then end up with the following bias-correcting term,

$$\mathrm{E}_{\mathcal{L}}\left[\widehat{Q}_n - \widehat{R}_n\right] \sim n^{-1}\,\mathrm{E}_{\mathcal{L}}\left[p_\infty\right] = n^{-1}p^* = n^{-1}\,\mathrm{Tr}\left(\boldsymbol{J}^{-1}\boldsymbol{K}\right). \tag{3.34}$$

The quality of the estimate in eq. (3.34) will depend both on the true distribution f° and the number of observations n. For sufficiently large values of n, we would expect this to be a nice estimate, but that does not imply that we get an acceptable result based on the dataset $\boldsymbol{\mathcal{X}}_n$ that we have at our disposal. In order to check this we need to do some simulations to gauge how good this approximation turns out to be for different sizes of $\boldsymbol{\mathcal{X}}_n$ from known distributions.

The AIC- and TIC-connection Since the two matrices $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\alpha}^\circ) = -\,\mathrm{E}_{f^\circ}\left[\boldsymbol{I}(\boldsymbol{\alpha}^\circ)\right]$ and $\boldsymbol{K} = \mathrm{Var}_{f^\circ}\left(\boldsymbol{u}(\boldsymbol{\alpha}^\circ)\right)$ are defined with respect to the true model f° and the optimal parameter configuration $\boldsymbol{\alpha}^\circ$, which both are unknown, we will need to find an estimate $\widehat{p^*}$ to replace the value p^*.

An estimator of p^* can e.g. be obtained by computing estimates of the two matrices $\boldsymbol{J}$ and $\boldsymbol{K}$, but we can simplify this drastically if we assume that we have a correctly specified model, i.e. that $f^\circ = f_{\boldsymbol{\alpha}^\circ}$ for some parameter $\boldsymbol{\alpha}^\circ$. We will then have an equality $\boldsymbol{J} = \boldsymbol{K}$, which implies that $p^* = \mathrm{Tr}\left(\boldsymbol{I}\right) = p$, i.e. we do not need to use our observations to compute the bias-correction, since the assumptions imply that p^* will be identical to $p = \mathrm{length}\left(\boldsymbol{\alpha}\right)$, the number of parameters in our model. (This is the estimate used by AIC, see e.g. Claeskens and Hjort [20, chapter 2] for details.)

If we do not know if our model is correctly specified, we should instead use $\widehat{p^*} = \mathrm{Tr}\left(\widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{K}}\right)$ as our estimate of p^*. (This estimate is used by TIC,[9] [20, chapter 2.3].)

We will now summarize and see how our choice of p^* as our bias-correcting term in our KLIC-based selection criterion connects with AIC and TIC.

[9] In many practical situations, AIC is used on all the models, even though it is obvious that the assumption of correctly specified model can not be valid for all of them – and the rankings attained by AIC and TIC are then almost identical, cf. the discussion in appendix B. The gain from the computationally more expensive TIC-estimate of $\widehat{p^*}$ might thus be of minor significance.

We start out without a specification of whether we assume our model to be correctly specified or not, and thus considers the following bias-corrected estimator of $R(\boldsymbol{\alpha}^\circ)$

$$\begin{aligned}\widetilde{Q}_n &\overset{\text{def}}{=} \widehat{Q}_n - n^{-1}\,\mathrm{E}_{\mathcal{L}}\left[p_\infty\right] \\ &= n^{-1}\sum_{i=1}^{n}\log f_{\widehat{\boldsymbol{\alpha}}_n}(\boldsymbol{x}_i) - n^{-1}\widehat{p^*} \\ &= n^{-1}\left(\ell(\widehat{\boldsymbol{\alpha}}_n;\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) - \widehat{p^*}\right),\end{aligned} \tag{3.35}$$

where the last equality is under the assumption that the observations are independent and identically distributed.

We thus arrive at the following estimator of the k° in eq. (3.8)

$$\begin{aligned}\widetilde{k}_n^* &\overset{\text{def}}{=} \underset{1\le k\le K}{\operatorname{argmax}}\,\widetilde{Q}(k)_n \\ &= \underset{1\le k\le N}{\operatorname{argmax}}\left[\int \log f_{k,\widehat{\boldsymbol{\alpha}(k)}_n}\,\mathrm{d}\widehat{F}_n - n^{-1}p(k)^*\right] \\ &= \underset{1\le k\le N}{\operatorname{argmax}}\left[n^{-1}\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n;\boldsymbol{\mathcal{X}}_n) - n^{-1}p(k)^*\right]\end{aligned} \tag{3.36}$$

Note that the number n of observations will influence how good $n^{-1}p_\infty$ will work as an approximation to the two terms $n^{-1}p_n + o_p\left(n^{-1}\right)$ in eq. (3.30), and thus the quality of the bias correcting part of the estimator $\widetilde{Q}_n$ will be affected if the sample is small.[10] On the other hand, as we readily see from eq. (3.36), the factor n^{-1} will be a common factor in all the calculations, and therefore it does not affect the selection process.

If the models under consideration are nested,[11] $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}(1)} \subset \cdots \subset \boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}(N)}$, and if we make the assumption that the true model for f° belongs to the family of models under consideration, then our selection criterion can be written like

$$\widetilde{k}_n^* = \underset{1\le k\le N}{\operatorname{argmax}}\left[n^{-1}\left(\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n;\boldsymbol{\mathcal{X}}_n) - p\right)\right]. \tag{3.37}$$

If we compare this with the AIC criterion,

$$k_n^{\text{AIC}} \overset{\text{def}}{=} \underset{1\le k\le N}{\operatorname{argmax}}\left[2\left(\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n;\boldsymbol{\mathcal{X}}_n) - p\right)\right], \tag{3.38}$$

we immediately see that they give the same result.

[10] According to Claeskens and Hjort [20, chapter 3], in the presentation of the Bayesian Information Criterion $\text{BIC} = 2\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n;\boldsymbol{\mathcal{X}}_n) - \log(n)p$, the AIC will not succeed in detecting "the true model" with probability tending to 1 when the sample size increases. The reason for this is that an increase in the sample-size will increase the maximized log-likelihood-value, and then the AIC-formula does not sufficiently penalize the model for its number of parameters p.

[11] For example a setting where we want to check if one of the parameters have a polynomial relation to the observed covariates.

If we do not have a collection of nested models, or if we do not want to assume that the true model for f° is included in $\boldsymbol{\mathcal{F}} = \cup_{i=1}^{N} \left\{\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}(i)}\right\}$, we should use the estimator $\widehat{p^*} = \mathrm{Tr}\left(\widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{K}}\right)$. This leads to the following selection criterion

$$\widetilde{k}_n^* = \underset{1\leq k\leq N}{\mathrm{argmax}} \left[n^{-1}\left(\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n; \boldsymbol{\mathcal{X}}_n) - \mathrm{Tr}\left(\widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{K}}\right)\right)\right], \tag{3.39}$$

which a comparison with the TIC criterion,

$$k_n^{\mathrm{TIC}} \overset{\mathrm{def}}{=} \underset{1\leq k\leq N}{\mathrm{argmax}} \left[2\left(\ell_k(\widehat{\boldsymbol{\alpha}(k)}_n; \boldsymbol{\mathcal{X}}_n) - \mathrm{Tr}\left(\widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{K}}\right)\right)\right], \tag{3.40}$$

tells us that we have arrived at the same selection criterion.

The above arguments shows that the KLIC-based selection criterion and the AIC and the TIC, are the same. In particular, we can compare the models by computing the maximum of their respective log-likelihood-functions and then subtract the (generalized) dimension of the model - and then we select the model which attains the highest value.

3.3.3 TIC vs. cross-validation

The connection between TIC and "leave-one-out cross-validation" from the parametric setting gives the heart of the argument in Grønneberg [16] with regard to the introduction of the cross-validation copula information criterion xv-CIC.

In this section we will briefly present the relevant result, while avoiding the technicalities. The interested reader can e.g. check out Claeskens and Hjort [20, chapter 2.9] for the formal argument that is required to prove this connection.

The main point of interest is that we have a "convergence" between the "loss-function perspective" and the "prediction perspective" when we search for the best model to describe our observations $\boldsymbol{\mathcal{X}}_n$. In particular, when the number of observations increases, we can expect an increased chance for the two selection strategies to propose the same model as the best one.

As previously observed, in eq. (3.1), we only need to compute the following part of the KLIC-value,

$$\mathrm{E}_{f^\circ}\left[\log f_{\boldsymbol{\alpha}}\right] = \int \log f_{\boldsymbol{\alpha}} \, \mathrm{d}F^\circ, \tag{3.41}$$

when we want to investigate the closeness of a model $f_{\boldsymbol{\alpha}}$ from some parametric family $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$ to a specified model f°, and the optimal member from $\boldsymbol{\mathcal{F}}_{\boldsymbol{\alpha}}$ is the one whose parameter $\boldsymbol{\alpha}^\circ$ maximizes eq. (3.41).

When the true value $\boldsymbol{\alpha}^\circ$ is unknown, we can use as an approximation $\widehat{\boldsymbol{\alpha}}$, the maximum likelihood estimator based on the observations $\boldsymbol{\mathcal{X}}_n$.

In the realm of i.i.d. observations, we find that the estimate we then are looking for equals the expectation of a new observation $\boldsymbol{X}_{\text{new}}$, i.e. we have

$$\int \log f_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{x})\, \mathrm{d}F^\circ(\boldsymbol{x}) = \mathrm{E}_{f^\circ}\left[\log f_{\widehat{\boldsymbol{\alpha}}}(\boldsymbol{X}_{\text{new}})\right], \tag{3.42}$$

see [20, p. 51] for further details.

Based on eq. (3.42) we introduce the following "prediction perspective" selection method:

$$\widehat{\mathrm{xv}}_n = n^{-1} \sum_{k=1}^{n} \log f_{\widehat{\boldsymbol{\alpha}}_{n(k)}}(\boldsymbol{x}_k), \tag{3.43}$$

in which $\widehat{\boldsymbol{\alpha}}_n(i)$ is the ml-estimate based on the sample without the i'th observation, i.e.

$$\widehat{\boldsymbol{\alpha}}_n(i) = \underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \int \log f_{\boldsymbol{\alpha}}(\boldsymbol{x})\, \mathrm{d}\widehat{F}_{n\setminus i}(\boldsymbol{x}). \tag{3.44}$$

By the help of influence functions, it is proved in [20] that $\widehat{\mathrm{xv}}_n$ with probability one tends toward $n^{-1}\left\{\ell_n(\widehat{\boldsymbol{\alpha}}) - \mathrm{Tr}\left(\widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{K}}\right)\right\}$, which implies the following relation between $\widehat{\mathrm{xv}}_n$ and TIC_n

$$\mathrm{TIC}_n = 2n\widehat{\mathrm{xv}}_n + o_p(1). \tag{3.45}$$

This is used in Grønneberg [16] to construct the selection method xv-CIC for the semiparametric realm, and we will in the following sections give an outline of the analysis leading to xv-CIC.

3.4 Semiparametric model selection, CIC

As discussed in chapters 1 and 2, we will in many cases restrict our attention to semiparametric models, i.e. that we instead of a quest for a model $F_{(\boldsymbol{\theta},\boldsymbol{\gamma})}$ for some independent observations $\boldsymbol{\mathcal{X}}_n$ will look for a copula model $C_{\boldsymbol{\theta}}$ for the corresponding dependent pseudo-observations ${}^p\boldsymbol{\mathcal{X}}_n$ (obtained from $\boldsymbol{\mathcal{X}}_n$ by the help of the empirical marginal distributions).

We will now consider how the machinery from section 3.3 must be tweaked to deal with model selection in the semiparametric situation.

3.4.1 Semiparametric models and MPLE.

In the semiparametric approach, we avoid the specification of models for the marginals by transforming our observations $\boldsymbol{\mathcal{X}}_n$ into a set of pseudo-independent observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ – the latter constituted of the pseudo-observations ${}^{p}\boldsymbol{x}_j \overset{\text{def}}{=} \boldsymbol{F}_{n,\perp}(\boldsymbol{x}_j)$ originating from $\boldsymbol{x}_j \in \boldsymbol{\mathcal{X}}_n$ with the help of the empirical marginals $\boldsymbol{F}_{n,\perp}$.

When we then want to find the parametric copula $c_{\boldsymbol{\theta}}$, from a copula-family $\boldsymbol{\mathcal{C}}_{\boldsymbol{\theta}}$, that is closest to the copula c° corresponding to the data-generating model f°, we need to use pseudo-likelihoods instead of likelihoods in order to take into account the noise from the transformation $\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$. [12]

Furthermore, when there are several different copula-candidates available, we would like to rank them in a consistent way such that we can pick out the one that fits the pseudo-data ${}^{p}\boldsymbol{\mathcal{X}}_n$ in the best possible fashion.

As the copulas themselves are d-variate distributions, this selection procedure might be expected to be fairly close to the one treated in the fully parametric case – and it has been normal practice to use an adjusted AIC-like formula without further ado.[13]

However, the requirements for the machinery of the AIC-formula is missing in the semiparametric context, since the pseudo-observations requires the machinery of maximum pseudo-log-likelihood introduced by Besag [14].

In particular, this implies that our model selection formula must take into account that we will use the maximizer of the pseudo-log-likelihood for $C_{\boldsymbol{\theta}}\left({}^{p}\boldsymbol{x}\right)$ in our analysis

$$ {}^{p}\ell_n\left(\boldsymbol{\theta}; {}^{p}\boldsymbol{\mathcal{X}}\right) = \sum_{i=1}^{n} \log c_{\boldsymbol{\theta}}\left({}^{p}\boldsymbol{x}_i\right), \tag{3.46} $$

which, as discussed in detail in Grønneberg and Hjort [15], Grønneberg [16], implies quite a few modifications of the machinery mentioned in the previous discussions connecting MLE and KLIC with AIC/TIC.

3.4.2 MPLE and the score-function.

We will now look closer upon the changes that follows when MLE is exchanged with MPLE. The asymptotic behavior of the MPLE is in particular of interest, and as for the MLE-case we can use the relation between the score-function and the likelihood-function, i.e. $\boldsymbol{u}_n \overset{\text{def}}{=} n^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} {}^{p}\ell_n(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\circ}}$, to extract this information. This is summarized

[12] See the discussion around fig. 2.8 for a reminder of what we mean by noise in this context.

[13] The adjustment being that the maximum of the pseudo likelihood function is used instead of the maximum of the likelihood function.

in Lemma 1 from Grønneberg [16, Part III], which states that under the necessary regularity conditions on $\boldsymbol{u}_n$, we have a weak convergence[14]

$$\sqrt{n}\bar{\boldsymbol{u}}_n \xrightarrow[n\to\infty]{W} \boldsymbol{u} \sim \mathrm{N}_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), \tag{3.47}$$

where $\boldsymbol{\Sigma} = \boldsymbol{I} + \boldsymbol{W}$, with

$$\boldsymbol{I} \overset{\text{def}}{=} \mathrm{E}\left[\boldsymbol{\phi}_{\boldsymbol{\theta}^\circ}(\boldsymbol{\xi})\, \boldsymbol{\phi}_{\boldsymbol{\theta}^\circ}(\boldsymbol{\xi})^T\right], \tag{3.48}$$

$$\boldsymbol{W} \overset{\text{def}}{=} \mathrm{Var}\left(\boldsymbol{Z}\right), \tag{3.49}$$

with

$$\boldsymbol{Z} \overset{\text{def}}{=} \sum_{k=1}^{d} \int \frac{\partial \boldsymbol{\phi}_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})}{\partial v_k} \cdot \left(\mathbb{1}\{\xi_k \leq v_k\} - v_k\right) \mathrm{d}C^\circ\left(\boldsymbol{v}\right), \tag{3.50}$$

in which $\boldsymbol{\xi}$ is a random vector distributed according to C° and

$$\boldsymbol{\phi}_{\boldsymbol{\theta}^\circ}(\boldsymbol{v}) \overset{\text{def}}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \log c_{\boldsymbol{\theta}}\left(\boldsymbol{v}\right). \tag{3.51}$$

This lemma implies that it is possible to give conditions on the parametrization that ensures a weak convergence $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\circ\right) \xrightarrow[n\to\infty]{W} \boldsymbol{J}^{-1}\boldsymbol{u} \sim \mathrm{N}_p\left(\boldsymbol{0}, \boldsymbol{J^{-1}\Sigma J^{-1}}\right)$, where the limit is defined in terms of a full-rank matrix

$$\boldsymbol{J} \overset{\text{def}}{=} -\int_{[0,1]^d} \frac{\partial^2 \log c_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{v}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mathrm{d}C^\circ\left(\boldsymbol{v}\right). \tag{3.52}$$

3.4.3 The arguments leading to $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$.

The argumentation in Grønneberg [16] leading to the formulas for $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ is motivated by the connection between MLE and KLIC and AIC/TIC These arguments therefore closely follows the discussion of the previous part of this chapter, and we will thus not delve deeply into the details of the arguments in this section.

The difference between the $\mathrm{CIC}^{\mathrm{AIC}}$ and the $\mathrm{CIC}^{\mathrm{TIC}}$ is analogous to the difference between AIC and TIC, i.e. their bias-correcting terms are molded with regard to whether or not we find it reasonable to assume that we have a correct specified model for our pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$. Thus the arguments leading to these two formulas are in the same

[14] The use of the empirical marginals introduces a bunch of discontinuities that must be accounted for (see e.g. Ruymgaart [25]), and an effect of this is that the covariance matrix $\boldsymbol{\Sigma}$ in addition to the information matrix $\boldsymbol{I}$ gains an additional matrix $\boldsymbol{W}$ (cf. Genest et al. [26] for details). According to Grønneberg [16], this $\boldsymbol{W}$ can be seen as accounting for the fact that we are dealing with a pseudo likelihood and not a proper likelihood.

vein, with the bifurcation point appearing at the moment we need to decide what kind of estimation regime we will use in our computation.

The argument regarding the connection between the maximum pseudo-log-likelihood estimator ${}^{p}\ell_n$ and the Kullback-Leibler divergence between the copula $C^\circ(\boldsymbol{v})$ of the true model $F^\circ(\boldsymbol{x}) = C^\circ(\boldsymbol{F}_\perp(\boldsymbol{x}))$ and the copula $C_{\boldsymbol{\theta}}(\boldsymbol{v})$ of our semiparametric model $F_{\boldsymbol{\theta}}(\boldsymbol{x}) = C_{\boldsymbol{\theta}}\left(\boldsymbol{F}_{n,\perp}(\boldsymbol{x})\right)$, goes as summarized below.

Similar to the case of the MLE, cf. eqs. (3.4) and (3.5), we first consider the maximizer of the pseudo-log-likelihood of a copula-model $C_{\boldsymbol{\theta}}(\boldsymbol{v})$

$$ {}^{p}\widehat{\boldsymbol{\theta}}_n \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} n^{-1}{}^{p}\ell_n(\boldsymbol{\theta}; {}^{p}\boldsymbol{\mathcal{X}}_n) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} n^{-1} \int_{[0,1]^d} \log c_{\boldsymbol{\theta}}(\boldsymbol{v})\, \mathrm{d}\widehat{C}_n(\boldsymbol{v}), \tag{3.53} $$

where $\widehat{C}_n(\boldsymbol{v})$ is the empirical copula, defined by

$$ \widehat{C}_n(\boldsymbol{v}) \overset{\text{def}}{=} n^{-1} \sum_{j=1}^{n} \mathbb{1}\{\boldsymbol{F}_{n,\perp}(\boldsymbol{x}_j) \leq \boldsymbol{v}\} = n^{-1} \sum_{j=1}^{n} \prod_{i=1}^{d} \mathbb{1}\{F_{n,i}(x_{j,i}) \leq v_i\}. \tag{3.54} $$

Under suitable regularity conditions, e.g. like those in Genest et al. [26], this maximizer turns out to behave just like eqs. (3.2) and (3.6). We thus have the following convergence in probability:

$$ {}^{p}\widehat{\boldsymbol{\theta}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}^\circ \overset{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{[0,1]^d} \log c_{\boldsymbol{\theta}}\, \mathrm{d}C^\circ \tag{3.55} $$

$$ = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{[0,1]^d} \log \frac{c^\circ}{c_{\boldsymbol{\theta}}}\, \mathrm{d}C^\circ \tag{3.56} $$

$$ = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{KLIC}(c^\circ, c_{\boldsymbol{\theta}}). \tag{3.57} $$

If we have a family of copula-models $\boldsymbol{\mathcal{C}}_{\boldsymbol{\theta}} = \{c_{\boldsymbol{\theta}}\}$ parametrized by $\boldsymbol{\theta}$,[15] and if we know the true model c°, then the best approximation from the family $\boldsymbol{\mathcal{C}}$ would be $\widetilde{c} = c_{\boldsymbol{\theta}^\circ}$. In a practical situation with an unknown true model, we will replace $\boldsymbol{\theta}^\circ$ with ${}^{p}\widehat{\boldsymbol{\theta}}_n$, and thus use $\widehat{c} = c_{\boldsymbol{\theta}}\big|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_n}$ as our estimate.

If we do not have any competing models, we could let it rest here – but when we want to use this as basis for a model-selection criterion we need to recap the arguments that took us from eq. (3.8) to eq. (3.10):

If the best approximation to an unknown model c° is to be found within a collection of different parametric models, i.e. $\boldsymbol{\mathcal{C}} = \cup_{k=1}^{N} \boldsymbol{\mathcal{C}}_{k,\boldsymbol{\theta}(k)}$, it is enough to search for the best

[15] The elements of $\boldsymbol{\mathcal{C}}_{\boldsymbol{\theta}}$ could in simple cases e.g. be from an elliptical or Archimedean copula-family. In a multivariate setting with different kind of tail-dependencies, it might be more natural that $\boldsymbol{\mathcal{C}}_{\boldsymbol{\theta}}$ instead consists of conglomerated structures described by some vine copula or nested Archimedean copula.

approximation from the set $\boldsymbol{\mathcal{C}}^\circ = \left\{c_{1,\,\boldsymbol{\theta}(1)^\circ}, \ldots, c_{N,\,\boldsymbol{\theta}(N)^\circ}\right\}$ that constitutes the best approximations *within* each parametric family. From this set we then pick the index $\widetilde{k}$ that minimizes the Kullback-Leibler distance to our empirical estimate of the true model.

Unless we have an exceptional huge number of observations, we must expect our estimates ${}^{p}\widehat{\boldsymbol{\theta}(k)}_n$ to have a non-negligible bias, which will induce a non-negligible bias in $\int \log c_{k,\,{}^{p}\widehat{\boldsymbol{\theta}(k)}_n}\,\mathrm{d}\widehat{C}_n$ – and this implies that we will need to adjust our estimates of the Kullback-Leibler values before we pick our $\widetilde{k}$, in particular we get an expression similar to eq. (3.10)

$$\widetilde{k}_n \overset{\text{def}}{=} \underset{1\le k\le N}{\operatorname{argmax}} \left[\int \log c_{k,\,{}^{p}\widehat{\boldsymbol{\theta}(k)}_n}\,\mathrm{d}\widehat{C}_n - \text{Bias}_k\right]. \tag{3.58}$$

The ingredients needed in our recipe for the bias-correcting terms is based on Taylor-series expansion around $\boldsymbol{\theta}^\circ$ of the two functions

$$A(\boldsymbol{\theta}) \overset{\text{def}}{=} \int_{[0,1]^d} \log c_{\boldsymbol{\theta}}\left(\boldsymbol{v}\right)\mathrm{d}C^\circ\left(\boldsymbol{v}\right), \tag{3.59}$$

$$\begin{aligned} A_n(\boldsymbol{\theta}) &\overset{\text{def}}{=} \int_{[0,1]^d} \log c_{\boldsymbol{\theta}}\left({}^{p}\boldsymbol{x}\right)\mathrm{d}\widehat{C}_n({}^{p}\boldsymbol{x}) \\ &= n^{-1}{}^{p}\ell_n(\boldsymbol{\theta};\,{}^{p}\boldsymbol{\mathcal{X}}_n), \end{aligned} \tag{3.60}$$

where $A(\boldsymbol{\theta})$ represents the ideal situation where we actually know the copula of the data-generating model $F^\circ\left(\boldsymbol{x}\right) = C^\circ\left(\boldsymbol{F}^\circ_{n,\,\perp}(\boldsymbol{x})\right)$ – whereas $A_n(\boldsymbol{\theta})$ is the one we must settle for when the true model is unknown and the best we can do is to make an estimate based on the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ (obtained from the observations $\boldsymbol{\mathcal{X}}_n$ by the help of the empirical marginals).

The connection between $A_n(\boldsymbol{\theta})$ and the pseudo log-likelihood function grants us information about its asymptotic behavior, and under suitable regularity conditions (e.g. Ruymgaart [25]) we have the following convergence in probability

$$A_n(\boldsymbol{\theta}) \xrightarrow[n\to\infty]{P} A(\boldsymbol{\theta}). \tag{3.61}$$

In accordance with eqs. (3.11) to (3.13), we can consider

$A(\boldsymbol{\theta}^\circ)$, the true value/target we want to estimate,

$A({}^{p}\widehat{\boldsymbol{\theta}}_n)$, an estimate of $A(\boldsymbol{\theta}^\circ)$ based on the MPLE ${}^{p}\widehat{\boldsymbol{\theta}}_n$,

$A_n({}^{p}\widehat{\boldsymbol{\theta}}_n)$, the estimate we can obtain when the true copula is unknown.

The argument in Grønneberg [16] is to use $A_n({}^{p}\widehat{\boldsymbol{\theta}}_n)$ to approximate $A({}^{p}\widehat{\boldsymbol{\theta}}_n)$, since the difference $A_n({}^{p}\widehat{\boldsymbol{\theta}}_n) - A({}^{p}\widehat{\boldsymbol{\theta}}_n)$ then can be used to make small-sample corrections to the

estimator $A_n({}^p\widehat{\boldsymbol{\theta}}_n)$. In particular, the goal is to find a decomposition of the form

$$A_n({}^p\widehat{\boldsymbol{\theta}}_n) - A({}^p\widehat{\boldsymbol{\theta}}_n) = \zeta_n + n^{-1}\alpha_n + n^{-1}\beta_n, \tag{3.62}$$

where $\mathrm{E}[\zeta_n] = 0$, and where α_n and β_n are bias terms such that α_n is $\mathcal{O}_P(1)$, but not $o_P(1)$ and β_n is $o_P(1)$. In particular: the $n^{-1}\beta_n$-term can be considered as low-level noise when n is large enough.

Similarly to the strategy used in the fully parametric case, we can use asymptotic theory to find a limiting distribution α of α_n — and when we want an estimate of the bias, we can use the estimate of $\mathrm{E}[\alpha]$ as our estimate.

Note: It turns out, in contrast to the fully parametric case, that we might have limiting distributions with an infinite expectation – and thus we do not obtain a generally applicable selection model.

When we linearize $A(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^\circ$, we get the exact same expressions as we found in eqs. (3.14) to (3.18):

$$\begin{aligned} A(\boldsymbol{\theta}) = A(\boldsymbol{\theta}^\circ) &+ \left(\int \boldsymbol{u}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\theta}^\circ} \, \mathrm{d}c^\circ \right)^T \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \\ &+ \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^T \cdot \int \boldsymbol{I}(\boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\theta}^\circ + h_1 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)} \, \mathrm{d}c^\circ \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ), \end{aligned} \tag{3.63}$$

where

$$\boldsymbol{u}(\boldsymbol{\gamma})^T = \left[\frac{\partial}{\partial \gamma_1} \log c_{\boldsymbol{\gamma}}, \ldots, \frac{\partial}{\partial \gamma_p} \log c_{\boldsymbol{\gamma}} \right], \qquad \text{and} \tag{3.64}$$

$$\boldsymbol{I}(\boldsymbol{\gamma}) = \begin{bmatrix} \frac{\partial^2}{\partial \gamma_1^2} \log c_{\boldsymbol{\gamma}} & \cdots & \frac{\partial^2}{\partial \gamma_1 \partial \gamma_p} \log c_{\boldsymbol{\gamma}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \gamma_p \partial \gamma_1} \log c_{\boldsymbol{\gamma}} & \cdots & \frac{\partial^2}{\partial \gamma_p^2} \log c_{\boldsymbol{\gamma}} \end{bmatrix}. \tag{3.65}$$

The expressions given in eqs. (3.64) and (3.65) is respectively the score vector and the information matrix of the model $c_{\boldsymbol{\theta}}$. Since the integrals involved in eq. (3.63) represents expectations, we can rewrite them as

$$\begin{aligned} A(\boldsymbol{\theta}) &= A(\boldsymbol{\theta}^\circ) + \left(\mathrm{E}_{f^\circ}[\boldsymbol{u}(\boldsymbol{\theta}^\circ)]\right)^T \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \\ &\quad + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^T \cdot \mathrm{E}_{f^\circ}[\boldsymbol{I}(\boldsymbol{\theta}^\circ + h_1 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ))] \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ) \\ &= A(\boldsymbol{\theta}^\circ) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^T \cdot \boldsymbol{J}(\boldsymbol{\theta}^\circ + h_1 \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)), \end{aligned} \tag{3.66}$$

where the score vector term disappears since its expectation always equals zero, and as usual the notation $\boldsymbol{J}(\boldsymbol{\gamma}) = -\operatorname{E}_{f^\circ}\left[\boldsymbol{I}(\boldsymbol{\gamma})\right]$ is used since the *Fisher information matrix*, i.e. $\boldsymbol{J}(\boldsymbol{\gamma})$ evaluated at $\boldsymbol{\theta}^\circ$, is denoted by $\boldsymbol{J}$.

However, when we want to consider the linearization of $A_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta} = \boldsymbol{\theta}^\circ$, and then want to evaluate it at $\boldsymbol{\theta} = {}^p\widehat{\boldsymbol{\theta}}_n$, we find that although it is equal in form to eq. (3.24), it does differ in the important aspect that it is to be evaluated with respect to the pseudo-observations:

$$A_n({}^p\widehat{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^{n}\log c_{\boldsymbol{\theta}^\circ}\left({}^p\boldsymbol{x}_i\right) \tag{3.67a}$$

$$+ n^{-1}\sum_{i=1}^{n}\boldsymbol{u}(\boldsymbol{\theta}^\circ; {}^p\boldsymbol{x}_i)^T\cdot\left({}^p\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\circ\right) \tag{3.67b}$$

$$+ n^{-1}\sum_{i=1}^{n}\frac{1}{2}\left({}^p\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\circ\right)^T\cdot\boldsymbol{I}\left(\boldsymbol{\theta}^\circ + h_2\cdot\left({}^p\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\circ\right); {}^p\boldsymbol{x}_i\right)\cdot\left({}^p\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\circ\right), \tag{3.67c}$$

for some $h_2 \in [0,1]$.

In particular, this implies that we must account for the noise from the transformation from $\boldsymbol{\mathcal{X}}_n$ to ${}^p\boldsymbol{\mathcal{X}}_n$ in this expression, when we want to find the bias-correcting term at the $\mathcal{O}_P\left(n^{-1}\right)$-level. As done in Grønneberg [16], this requires a closer look upon the difference $A_n(\boldsymbol{\theta}^\circ) - A(\boldsymbol{\theta}^\circ)$.

In order to adjust for the deviations from the true values that the transformation step generates, we need to do a Taylor expansion (two terms plus remainder) of the expression for $A_n(\boldsymbol{\theta})$. In particular, we will expand $\log c_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})$ around $\boldsymbol{v} = F^\circ_\perp(\boldsymbol{x})$.

To summarize, in order to find the bias-correcting terms at the $\mathcal{O}_P\left(n^{-1}\right)$-level, we need to consider the following difference

$$A_n(\boldsymbol{\theta}^\circ) - A(\boldsymbol{\theta}^\circ) = n^{-1}\sum_{j=1}^{n}\log c_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j)\right) - \int_{\mathbb{R}^d}\log c_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}^\circ_\perp(\boldsymbol{x})\right)\mathrm{d}F^\circ(\boldsymbol{x}), \tag{3.68}$$

which is what we get from the definitions in eqs. (3.59) and (3.60) when we want to emphasize that the copulas we consider is related to observations $\boldsymbol{\mathcal{X}}_n$ of some d-variate distribution $F^\circ(\boldsymbol{x}) = C^\circ\left(\boldsymbol{F}^\circ_\perp(\boldsymbol{x})\right)$.

The argument needed in order to arrive at the expression from Grønneberg and Hjort [15], given in eq. (3.69) below, is to first restate the sum in eq. (3.68) as a Lebesgue-Stieltjes integral and then do a second order Taylor-expansion of $\log c_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})$ around $\boldsymbol{v} = F^\circ_\perp(\boldsymbol{x})$. In order to get the desired result we have to do a minor reshuffling of the terms, and then go back to sums instead of Lebesgue-Stieltjes integrals for some of the terms.

Under the condition that the function $\log c_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})$ can be differentiated twice, we get

$$A_n(\boldsymbol{\theta}^\circ) - A(\boldsymbol{\theta}^\circ) = \int_{\mathbb{R}^d} \log c_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_\perp^\circ(\boldsymbol{x})\right) \mathrm{d}[F_n(\boldsymbol{x}) - F^\circ(\boldsymbol{x})] + Q_n + R_n + B_n, \tag{3.69}$$

where Q_n and R_n denotes the first and second order terms of the Taylor expansion, i.e.

$$Q_n = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{\zeta}'_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right)^T \left[\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j) - \boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right],$$
$$R_n = \frac{1}{2n}\sum_{j=1}^{n} \left[\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j) - \boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right]^T \boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right) \left[\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j) - \boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right],$$

in which

$$\boldsymbol{\zeta}'_{\boldsymbol{\theta}}(\boldsymbol{v}) = \frac{\partial \log c_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{v}}, \tag{3.70}$$
$$\boldsymbol{\zeta}''_{\boldsymbol{\theta}}(\boldsymbol{v}) = \frac{\partial^2 \log c_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial \boldsymbol{v}\partial \boldsymbol{v}^T} \tag{3.71}$$

and where B_n represents the remainder term, given by

$$B_n = \frac{1}{2n}\sum_{j=1}^{n} \left[\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j) - \boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right]^T \left[\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{G}_n(\boldsymbol{X}_j)\right) - \boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right)\right] \left[\boldsymbol{F}_{n,\perp}(\boldsymbol{X}_j) - \boldsymbol{F}_\perp^\circ(\boldsymbol{X}_j)\right],$$

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

We need (estimates of) the expectation of eq. (3.69) in order to find the bias-correcting terms. The expectation of the integral is zero, so it is the other three parts on the right hand side that we need to work upon. The line of argument mimics the one encountered in the parametric setting, i.e. we will use expectations of the limiting distributions as bias-correcting terms in the final selection formulas $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$.

Grønneberg [16, Part III] gives arguments in Lemma 2 and Lemma 3 that looks closer upon the two terms Q_n and R_n, and presents conditions that ensures that we have $B_n = o_P\left(n^{-1}\right)$,[16] such that this annoying term disappears when we go to the limit.

The first result is that Q_n can be decomposed as $n^{-1}q_n + Z_{Q,n}$, with

$$\mathrm{E}\left[Z_{Q,n}\right] = 0, \tag{3.72}$$
$$q_n = \frac{n}{n+1}\int_{\mathbb{R}^d} \boldsymbol{\zeta}'_{\boldsymbol{\theta}^\circ}\left(\boldsymbol{F}_\perp^\circ(\boldsymbol{x})\right)^T \left(\mathbf{1} - \boldsymbol{F}_\perp^\circ(\boldsymbol{x})\right) \mathrm{d}F_n(\boldsymbol{x}) = \mathcal{O}_P(1), \tag{3.73}$$
$$\mathrm{E}\left[q_n\right] = \frac{n}{n+1}\int_{[0,1]^d} \boldsymbol{\zeta}'_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})^T (\mathbf{1} - \boldsymbol{v})\, \mathrm{d}C^\circ(\boldsymbol{v}). \tag{3.74}$$

[16] The conditions are given in Grønneberg [16, Part III], in Proposition 1 from Appendix A.2

Note that $Z_{Q,n}$ is defined in a similar way as $\bar{Z}_n$ on page 36, but its explicit form is not of any importance since its expectation is identical to zero.

Moreover, with $C_{a,b}$ the cumulative copula of $\left(\boldsymbol{X}_{1,a}, \boldsymbol{X}_{1,b}\right)$,[17] and $r_n \stackrel{\text{def}}{=} nR_n$, we have $\mathrm{E}\left[r_n\right] \to \mathbf{1}^T \boldsymbol{\Upsilon} \mathbf{1}$, where $\boldsymbol{\Upsilon}$ is given by

$$\Upsilon_{a,a} = \frac{1}{2} \int_{[0,1]^d} \left(\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})\right)_{a,a} v_a(1-v_a)\, \mathrm{d}C^\circ(\boldsymbol{v}), \tag{3.75}$$

$$\Upsilon_{a,b} = \frac{1}{2} \int_{[0,1]^d} \left(\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})\right)_{a,b} \left[C^\circ_{a,b}(v_a, v_b) - v_a v_b\right] \mathrm{d}C^\circ(\boldsymbol{v}), \quad \text{(when } a \neq b\text{)}, \tag{3.76}$$

in which $\left(\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})\right)_{a,b}$ are the elements of the matrix function $\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})$ from eq. (3.71).

Grønneberg [16, Part III, Theorem 1] gives the desired expression for $A_n(\boldsymbol{\theta}^\circ) - A(\boldsymbol{\theta}^\circ)$ as

$$A_n(\boldsymbol{\theta}^\circ) - A(\boldsymbol{\theta}^\circ) = n^{-1}(q_n + r_n) + \tilde{Z}_n + o_P\left(n^{-1}\right), \tag{3.77}$$

in which $\mathrm{E}\left[\tilde{Z}_n\right] = 0$, and introduces

$$q^* \stackrel{\text{def}}{=} \lim_{n\to\infty} \mathrm{E}\left[q_n\right] = \int_{[0,1]^d} \left(\boldsymbol{\zeta}'_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})\right)^T \cdot (\mathbf{1} - \boldsymbol{v})\, \mathrm{d}C^\circ(\boldsymbol{v}), \tag{3.78}$$

$$r^* \stackrel{\text{def}}{=} \lim_{n\to\infty} \mathrm{E}\left[r_n\right] = \mathbf{1}^T \boldsymbol{\Upsilon} \mathbf{1}. \tag{3.79}$$

As for $Z_{Q,n}$, the exact form of $\tilde{Z}_n$ can for our purposes be ignored since its expectation is identical to zero.

The next section will present the estimates of q^* and r^* that give the two selection models $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$, but before we venture into those details it might be worth to note that we alas can have situations where r^*, the limit of the expectation of r_n, might become infinite.

The reason for this is that copula models $c_\theta(\boldsymbol{v})$ that increases rapidly close to the boundary of $[0,1]^d$, can have a derived matrix-function $\boldsymbol{\zeta}''_{\boldsymbol{\theta}^\circ}(\boldsymbol{v})$ whose elements in the vicinity of the boundary grows fast enough for the integrals in $\boldsymbol{\Upsilon}$ to become infinite.

A concrete computation showing this phenomenon is given in [16], with the bivariate Archimedean B4-copula (see e.g. Joe [4] for the definition) as an example. The generator of the B4-copula enables us to compute the defining integrals of $\boldsymbol{\Upsilon}$ analytically, and the conclusion is that this gives infinite elements in $\boldsymbol{\Upsilon}$, and thus an infinite r^*.

So even though the term we are considering is $\mathcal{O}_P(1)$, we will for copula models with some tail-dependence find that the expectation of the limiting distribution is infinite –

[17] This cumulative copula is the copula associated with $\boldsymbol{F}_{a,b}(X_a, X_b)$, defined from $F(X_1, \ldots, X_d)$ by letting the superfluous observators tend to infinity.

in contrast to the situation known from the parametric case where the expectation of the limiting distribution does not have such an undesirable behavior.

3.4.4 The estimators needed in the computation of $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$.

Based on the results mentioned in the previous section, we can now state Proposition 1 from Grønneberg [16, Part III] that motivates the empirical estimates to be used under the assumption of a correctly specified model (AIC-like variant) or a more general situation (TIC-like variant):

> If the parametric model is correctly specified, we have $q^* = 0$ and $p^* = \mathrm{length}\,(\boldsymbol{\theta}) + \mathrm{Tr}\left(\boldsymbol{I}^{-1}\boldsymbol{W}\right)$.

Reminder: The matrices $\boldsymbol{I}$ and $\boldsymbol{W}$ are those defined in eq. (3.48) and eq. (3.49), i.e. the components that constitute the covariance matrix $\boldsymbol{\Sigma}$ of the limiting normal distribution when we must use the pseudo-likelihood instead of the likelihood in our analysis.

The case of $\mathrm{CIC}^{\mathrm{AIC}}$: The above result motivates the AIC-like Copula Information Criterion

$$\mathrm{CIC}^{\mathrm{AIC}} \stackrel{\mathrm{def}}{=} 2 \cdot {}^{p}\ell_{n,\max} - 2 \cdot \left(\widehat{p^*} + \widehat{r^*}\right), \tag{3.80}$$

where ${}^{p}\ell_{n,\max}$ is the maximum of the pseudo-log-likelihood, and where $\widehat{p^*}$ and $\widehat{r^*}$ are the estimates of p^* and r^* defined below.

A natural estimator of r^* in this case, where we assume that we have $c_{\boldsymbol{\theta}^\circ} = c^\circ$, is $\widehat{r^*} = \mathbf{1}^T\widehat{\boldsymbol{\Upsilon}}\mathbf{1}$, defined in terms of the plug-in estimators

$$\widehat{\Upsilon}_{a,a} = \frac{1}{2}\int_{[0,1]^d}\left(\boldsymbol{\zeta}''_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{v})\right)_{a,a} v_a(1-v_a)\,\mathrm{d}C_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{v}), \tag{3.81}$$

$$\widehat{\Upsilon}_{a,b} = \frac{1}{2}\int_{[0,1]^d}\left(\boldsymbol{\zeta}''_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{v})\right)_{a,b}\left[C_{{}^{p}\widehat{\boldsymbol{\theta}}_n;\,a,\,b}(v_a, v_b) - v_a v_b\right]\mathrm{d}C_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{v}), \tag{3.82}$$

A natural estimator for p^* is, with regard to the result of the proposition, to use

$$\widehat{p^*} = \mathrm{length}\,(\boldsymbol{\theta}) + \mathrm{Tr}\left(\widehat{\boldsymbol{I}}^{-}\widehat{\boldsymbol{W}}\right), \tag{3.83}$$

where $\widehat{\boldsymbol{I}}^{-}$ is the generalized inverse of $\widehat{I}$, the pseudo empirical information matrix

$$\widehat{\boldsymbol{I}} = \mathrm{E}\left[\boldsymbol{\phi}_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{\xi})\,\boldsymbol{\phi}_{{}^{p}\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{\xi})^T\right] \tag{3.84}$$

and

$$\widehat{W} = \operatorname{Var}\left(\int_{[0,1]^d}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{u}^T}\log c_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{u})\right)^T(\mathbb{1}\{\boldsymbol{\xi}\leq v\}_\perp - \boldsymbol{u})\,\mathrm{d}C_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{u})\right), \tag{3.85}$$

where $\boldsymbol{\xi}\sim C_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{v})$ and where $\boldsymbol{\phi}_{\boldsymbol{\theta}} = \frac{\partial}{\partial\boldsymbol{\theta}}\log c_{\boldsymbol{\theta}}(\boldsymbol{v})$.

The case of $\mathrm{CIC}^{\mathrm{TIC}}$: When we consider a situation where we do not want to impose an assumption about a correctly specified model, we need nonparametric estimators that gives us a TIC-like formula.

$$\mathrm{CIC}^{\mathrm{TIC}} \stackrel{\mathrm{def}}{=} 2\cdot {}^{p}\ell_{n,\max} - 2\cdot\left(\widehat{p^*}+\widehat{q^*}+\widehat{r^*}\right), \tag{3.86}$$

with ${}^{p}\ell_{n,\max}$ the maximum of the pseudo-log-likelihood, and $\widehat{p^*}$, $\widehat{q^*}$ and $\widehat{r^*}$ as given in the estimates below.

The natural estimator for q^* is the plug-in estimator

$$\widehat{q^*} = \int_{[0,1]^d}\left(\boldsymbol{\zeta}'_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{v})\right)^T\cdot(\mathbf{1}-\boldsymbol{v})\,\mathrm{d}\widehat{C}_n(\boldsymbol{v}), \tag{3.87}$$

while the estimator for r^*, $\widehat{r^*} = \mathbf{1}^T\widehat{\boldsymbol{\Upsilon}}\mathbf{1}$, in this case must use the empirical copula

$$\widehat{\Upsilon}_{a,a} = \frac{1}{2}\int_{[0,1]^d}\left(\boldsymbol{\zeta}''_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{v})\right)_{a,a} v_a(1-v_a)\,\mathrm{d}\widehat{C}_n(\boldsymbol{v}), \tag{3.88}$$

$$\widehat{\Upsilon}_{a,b} = \frac{1}{2}\int_{[0,1]^d}\left(\boldsymbol{\zeta}''_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{v})\right)_{a,b}\left[\widehat{C}_{n;a,b}(v_a,v_b) - v_a v_b\right]\mathrm{d}\widehat{C}_n(\boldsymbol{v}), \tag{3.89}$$

where $\widehat{C}_{n,a,b}(v_a,v_b)$ is the empirical bivariate copula corresponding to the pairs of covariates indexed by a and b, i.e. $(x_{1,a},x_{1,b}),(x_{2,a},x_{2,b}),\ldots,(x_{n,a},x_{n,b})$.

For the estimation of p^*, we use $\widehat{p^*} = \operatorname{Tr}\left(\widehat{\boldsymbol{J}}_n^{-1}\widehat{\boldsymbol{\Sigma}}\right)$, based on estimates of $\boldsymbol{\Sigma} = \boldsymbol{I}+\boldsymbol{W}$ and of $\boldsymbol{J}$ defined in eq. (3.52).

In this case Grønneberg [16] gives us

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{\phi}_{p_{\widehat{\boldsymbol{\theta}}_n}}\left(\widehat{\boldsymbol{\xi}}^{(i)}\right)+\widehat{\boldsymbol{Z}}_i\right\}\left\{\boldsymbol{\phi}_{p_{\widehat{\boldsymbol{\theta}}_n}}\left(\widehat{\boldsymbol{\xi}}^{(i)}\right)+\widehat{\boldsymbol{Z}}_i\right\}^T, \tag{3.90}$$

with

$$\widehat{\boldsymbol{Z}}_i = \sum_{j=1}^{d} n^{-1}\sum_{s=1,s\neq i}^{n}\left.\frac{\partial\boldsymbol{\phi}_{p_{\widehat{\boldsymbol{\theta}}_n}}(\boldsymbol{v})}{\partial v_j}\right|_{\boldsymbol{v}=\widehat{\boldsymbol{\xi}}^{(s)}}\left(\mathbb{1}\left\{\widehat{\xi}_j^{(i)}\leq\widehat{\xi}_j^{(s)}\right\}-\widehat{\xi}_j^{(s)}\right) \tag{3.91}$$

using $\widehat{\boldsymbol{\xi}}^{(k)} = \boldsymbol{F}_{n,\perp}(\boldsymbol{x}_k)$.

A note of warning: The estimators given above for the two selection methods $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ can be computed for any set of pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, but the estimate of r^* will only be reasonable in those cases where the data-generating process c° does not give an infinite $\boldsymbol{\Upsilon}$-matrix. If we do not know what kind of process the data originated from, the use of these estimators could lead us astray if we end up using them on a situation for which they are not applicable.

In this authors opinion: Even though the theory leading to $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$ is of interest, the lack of general applicability implies that they should be shunned in practical applications.

3.4.5 The arguments leading to xv-CIC.

In section 3.3.3 we mentioned the parametric case connection between the "loss-function perspective", and the "prediction perspective", i.e. eq. (3.45) which states that we for a n-sized set of observations $\boldsymbol{\mathcal{X}}_n$ have the following correspondence

$$\mathrm{TIC}_n = 2n\widehat{\mathrm{xv}}_n + o_p(1)\,, \tag{3.92}$$

with $\widehat{\mathrm{xv}}_n$ as defined in eq. (3.43).

We will in this section consider the adjustments that must be applied when the empirical marginals replaces observations $\boldsymbol{\mathcal{X}}_n$ with pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$.

We already now from the previous sections that the semiparametric replacement for the TIC, the $\mathrm{CIC}^{\mathrm{TIC}}$ from section 3.3.2, does suffer from the rather undesirable property of not being applicable for observations stemming from copula models with tail-dependence. As previously mentioned, this implies that we probably are better off if we do not use $\mathrm{CIC}^{\mathrm{TIC}}$ as a selection model.

In contrast, the generalization of the leave-one-out cross-validation xv_n, the ${}^{p}\mathrm{xv}_n$ defined below in eq. (3.93), gives a selection method that is well behaved for all cases.

The ${}^{p}\mathrm{xv}_n$ is computationally expensive to apply directly, so Grønneberg [16, Part III] sets out to produce an analytical approximation to it, and the end result of this is the xv-CIC-formula given in eq. (3.97).

Since xv-CIC is first order equivalent to the generally applicable ${}^{p}\mathrm{xv}_n$, we then have at our hand a generally applicable model selection tool that can be used when we want to find the best available copula-model c from a collection $\boldsymbol{\mathcal{C}}$ that describes the interdependencies of a set of pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$.

The definition of ${}^{p}\mathrm{xv}_n$. If we make the obvious modifications to eq. (3.92), we obtain the following semiparametric version of the leave-one-out cross-validation formula.[18]

$$ {}^{p}\mathrm{xv}_n \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \log c_{\boldsymbol{\theta}} \left(\boldsymbol{F}_{n,\perp,(i)}(\boldsymbol{X}_i) \right) \big|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{(i)}}, \tag{3.93} $$

where ${}^{p}\widehat{\boldsymbol{\theta}}_{(i)}$ is the maximum pseudo-likelihood estimate (MPLE)

$$ {}^{p}\widehat{\boldsymbol{\theta}}_{(i)} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{j\neq i} \log c_{\boldsymbol{\theta}} \left(\boldsymbol{F}_{n,\perp,(i)}(\boldsymbol{X}_j) \right), $$

and where $\boldsymbol{F}_{n,\perp,(i)}$ is the ($\frac{n-1}{n}$-rescaled) marginal empirical distribution function based on all observations $\boldsymbol{X}_1, \dots \boldsymbol{X}_n$ except $\boldsymbol{X}_i$.

There is no reference to the data-generating model c° in the definition of ${}^{p}\mathrm{xv}_n$, which implies that it does not experience the effect that incurs the $\mathrm{CIC}^{\mathrm{TIC}}$-formula to become inapplicable. The ${}^{p}\mathrm{xv}_n$ thus gives a well defined semiparametric model selection strategy, but the requirement that we need to compute a total of n MPLEs ${}^{p}\widehat{\boldsymbol{\theta}}_{(i)}$ implies that the computational cost becomes formidable.

Motivated by the general applicability of the ${}^{p}\mathrm{xv}_n$, Grønneberg [16], sets out to find an expression similar to eq. (3.45) for the semiparametric case, i.e. the goal is to find an asymptotically equivalent analytical approximation to ${}^{p}\mathrm{xv}_n$.

The main step toward the desired xv-CIC-formula is Theorem 2 from Grønneberg [16, Part III], whose proof both requires the use of influence functions like those needed in the proof of eq. (3.92), and the use of a Taylor series expansion of $\log c_{\boldsymbol{\theta}}(\boldsymbol{v})$ in both $\boldsymbol{v}$ and $\boldsymbol{\theta}$, as sketched in section 3.4.3 for the case of $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$.

Note: The similarities with the line of argument used for the $\mathrm{CIC}^{\mathrm{AIC}}$ and $\mathrm{CIC}^{\mathrm{TIC}}$, implies that we in the notation below have $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{v})$ and $\boldsymbol{\zeta}'_{\boldsymbol{\theta}}(\boldsymbol{v})$ as defined in eqs. (3.51) and (3.70) respectively.

The result we need in order to motivate the definition of xv-CIC goes as follows: With regularity conditions like those in Genest et al. [26], wich for each $\boldsymbol{\theta}$ around $\boldsymbol{\theta}^\circ$ secures the point-wise convergence of

$$ \widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x}) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \int \frac{\partial \boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{v})}{\partial v_k} \left(\mathbb{1}\{x_k \leq v_k\} - v_k \right) \mathrm{d}\widehat{C}_n(\boldsymbol{u}) \xrightarrow[n\to\infty]{P} \boldsymbol{z}_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{3.94} $$

[18] Strictly speaking, $\boldsymbol{F}_{n,\perp,(i)}$ in eq. (3.93) should have been replaced with a slightly modified version in order to ensure that we do not have to evaluate it at the border of $[0,1]^d$, since many copula models might then be undefined (or attain infinte values). However, as discussed in detail in Remark 2 in Grønneberg [16, Part III], since the arguments later on is of an asymptotic nature, the few potential problematic values will be insignificant as $n \to \infty$.

and given that the function $\widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}(\boldsymbol{x})$ is continuous around $\boldsymbol{\theta}^{\circ}$, we have

$$
\begin{aligned}
{}^{p}\mathrm{xv}_{n} = n^{-1} \Bigg[& {}^{p}\ell_{n}(\boldsymbol{\theta}) - n^{-1} \sum_{i=1}^{n} \boldsymbol{\zeta}'_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} (\mathbf{1}_{d} - {}^{p}\boldsymbol{X}_{i}) \\
& + \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} \widehat{\boldsymbol{J}}^{-1} \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i}) + \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} \widehat{\boldsymbol{J}}^{-1} \widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i}) \Bigg] \Bigg|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{n}} + o_{P}(1) \,,
\end{aligned}
\tag{3.95}
$$

in which ${}^{p}\widehat{\boldsymbol{\theta}}_{n}$ is the MPLE for $\boldsymbol{\theta}$ in $c_{\boldsymbol{\theta}}$ (the model we are considering) with regard to the pseudo-observators in ${}^{p}\boldsymbol{\mathcal{X}}_{n}$. The matrix $\widehat{\boldsymbol{J}}$ is given by the Lebesgue-Stieltjes integral (i.e. a sum over the pseudo-observators in ${}^{p}\boldsymbol{\mathcal{X}}_{n}$)

$$
\widehat{\boldsymbol{J}} \stackrel{\text{def}}{=} - \int_{0,1]^{d}} \frac{\partial^{2} \log c_{\boldsymbol{\theta}}\left({}^{p}\boldsymbol{X}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{T}} \Bigg|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{n}} \mathrm{d}\widehat{C}_{n}\left({}^{p}\boldsymbol{X}\right) , \tag{3.96}
$$

wich under the regularity conditions of e.g. Ruymgaart [25] will converge in probability towards the matrix $\boldsymbol{J}$ defined in eq. (3.52).

Motivated by this result, the cross-validation Copula Information Criterion is defined by[19]

$$
\text{xv-CIC} \stackrel{\text{def}}{=} 2 \cdot {}^{p}\ell_{n,\max} - 2 \cdot (\widehat{p}_{n} + \widehat{q}_{n} + \widehat{r}_{n}) \,, \tag{3.97}
$$

where ${}^{p}\ell_{n,\max}$ as usual is the maximum of the pseudo-log-likelihood, and where

$$
\widehat{p}_{n} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} \widehat{\boldsymbol{J}}^{-1} \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i}) \Big|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{n}} \,, \tag{3.98a}
$$

$$
\widehat{q}_{n} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \boldsymbol{\phi}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} \widehat{\boldsymbol{J}}^{-1} \widehat{\boldsymbol{z}}_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i}) \Big|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{n}} \,, \tag{3.98b}
$$

$$
\widehat{r}_{n} \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \boldsymbol{\zeta}'_{\boldsymbol{\theta}}({}^{p}\boldsymbol{X}_{i})^{T} \left(\mathbf{1}_{d} - {}^{p}\boldsymbol{X}_{i}\right) \Big|_{\boldsymbol{\theta}={}^{p}\widehat{\boldsymbol{\theta}}_{n}} \,. \tag{3.98c}
$$

Grønneberg [16] also considers how the terms in xv-CIC would look like if we assume that we have picked the correct copula model, i.e. that we have $c^{\circ} = c_{\boldsymbol{\theta}^{\circ}}$. In this case it is possible to proove that we under standard coniditons have

$$
{}^{p}\mathrm{xv}_{n} = n^{-1} \left[{}^{p}\ell_{n}({}^{p}\widehat{\boldsymbol{\theta}}_{n}) - \text{length}\left(\boldsymbol{\theta}\right) - \boldsymbol{J}^{-1} \widehat{\boldsymbol{W}}_{n} \right] + o_{P}(1) \,, \tag{3.99}
$$

which then motivates the following AIC-like cross-validation Copula Information Criterion

$$
\text{xv-CIC}_{\text{AIC}} \stackrel{\text{def}}{=} 2\, {}^{p}\ell_{n,\max} - 2 \left(\text{length}\left(\boldsymbol{\theta}\right) + \boldsymbol{J}^{-1} \widehat{\boldsymbol{W}}_{n} \right) . \tag{3.100}
$$

[19] It is actually xv-CIC/2 that is first order equivalent with the cross-validation sum, but the resulting model ranking is not affected by this scaling, and the modification is made in order to maintain similarity with the classical AIC formula.

As mentioned in [16], the matrix $\widehat{\boldsymbol{W}}_n$ estimates $\boldsymbol{W}$ from eq. (3.49), which can be seen to account for the fact that we are dealing with a pseudo likelihood and not a proper likelihood. If it turns out that the term $\boldsymbol{J}^{-1}\boldsymbol{W}$ is small, it might introduce more variance than it bias-corrects. In such cases the classical AIC approach (i.e. the approach named ^{p}AIC in this thesis) might be preferable to use.

Note that the xv-CIC formula are motivated by an asymptotic approximation of the cross-validation formula, and it might thus only be reasonable to apply when n is large enough to make the $o_P(1)$ term in eq. (3.95) negligible. The size of n needed for a very good approximation will depend on both the data-generating mechanism and the parametric model $c_{\boldsymbol{\theta}}$ under consideration. If n is small to medium sized, we could always resort to the computation of the full cross-validation in eq. (3.93). When n is large enough to make the use of eq. (3.93) intractable, we could probably safely apply xv-CIC.

Grønneberg [16] comments that a large simulation study would be necessary to gauge the small-sample behavior of the xv-CIC for various families of copulas, and the result from such simulations for an assorted collection of copulas is presented in chapter 4.

Chapter 4

Results from simulations

This chapter presents the result from simulations executed in order to see how the selection method xv-CIC fares in the realm of finite samples $\boldsymbol{\mathcal{X}}_n$. In addition we will combine this analysis with a comparison to the results from the selection method ^{p}AIC.

Remember from the previous chapters that Grønneberg and Hjort [15] shows that ^{p}AIC, i.e. the selection method where we use an AIC-strategy on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ as if they actually were true independent observations $\boldsymbol{\mathcal{X}}_n$, does not rest on a theoretically sound framework. [15] tells us that we, when the noise from the transformation to pseudo-observations is taken into account, either should use CIC^{AIC} or CIC^{TIC} (depending on whether or not we assume a correctly specified copula model).

But the CIC^{AIC} and CIC^{TIC} share the unfortunate fate that their bias-correcting terms can attain infinite values, which implies that they are not generally applicable. However, the xv-CIC from Grønneberg [16, Part III] gives us a valid selection method for the semi-parametric settings we are interested in.

But even though the model selection method xv-CIC rests on a theoretical sound framework, it is important to remember that the precision of its bias correcting term will depend on how many observations we have in $\boldsymbol{\mathcal{X}}_n$. This is directly connected to how negligible the $o_P(1)$ term in the approximation is, and this is in addition connected to what kind of copulas we wish to rank against each other as possible models for the interdependencies of $\boldsymbol{\mathcal{X}}_n$.

In the following sections we will first present the framework that has been used to generate the data, and then we present the analysis based on them.

4.1 The setup

In order to find out how xv-CIC fares as a selection method in the finite realm, we need to test it on data we know the origin of. This section presents a sketch of the setup that has been used in order to accomplish this.

First of all, the program that has been used is `R` (version 2.15.2) and it is in particular the `R`-package `copula` that is at the heart of the computations. Quite a few other packages have also been necessary in order to properly store and work with the results from the different steps of the algorithm, and the interested reader can consult appendix C for further details.

The copula models: As mentioned in section 2.1, we will use the following bivariate copula models in our analysis: `clayton`, `frank`, `galambos`, `gumbel`, `huslerReiss`, `normal` and `t`. Of these, the first five are all Archimedean copulas, and the third to fifth of them are in addition extreme-value copulas. These five models are all included due to the fact that the code could deal with all Archimedean copulas in a unified and cost-efficient manner.[1]

The two copula models `normal` and `t`, which respectively refers to the copulas *corresponding* to the bivariate normal distribution and the bivariate student's t-distribution (which in our case has had degrees of freedom fixed to the value four), have been included since they are mandatory in any analysis of the interdependence structure of a set of bivariate data.

The data: The generated samples have been of size $N \in \{100, 250, 500, 1000\}$, and the parameters in the models have been picked in order to correspond to values of Kendall's τ in the set $\{0.25, 0.5, 0.75\}$.

A total of $R = 5000$ replicates were created for each combination of N and τ, by performing the steps described below.

1. For all combinations of copula models, sample-sizes N and parameter-values (as introduced above), the `rCopula`-function was used in order to generate our "idealized" samples ${}^{u}\boldsymbol{\mathcal{X}}_n$, please see the discussion at the the end of this section for the connection between the original samples $\boldsymbol{\mathcal{X}}_n$, the idealized samples ${}^{u}\boldsymbol{\mathcal{X}}_n$ and the pseudo-samples ${}^{p}\boldsymbol{\mathcal{X}}_n$.

2. The `pobs`-function was then used to create the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, and then we used the `fitCopula`-function on these to find the maximum pseudo likelihood

[1] See the discussion in appendix C for further details.

estimate ${}^{p}\widehat{\boldsymbol{\theta}}$ and the corresponding maximum of the pseudo log-likelihood function ${}^{p}\ell$ for all the seven copula models.

3. Based on the value of ${}^{p}\widehat{\boldsymbol{\theta}}$ (and on N) the bias-correcting terms of the xv-CIC formula was computed for each pseudo-observation ${}^{p}\boldsymbol{\mathcal{X}}_n$ and each model copula.

4. For every replicate, the values of ^{p}AIC and xv-CIC was computed, and the seven copula models was then ranked according to the result.

The analysis of the data: If we restrict our attention to the model with the highest rankings, we can easily see how well the xv-CIC has worked with regard to identifying the correct model for the different combinations of data-generating model, size of sample and value of parameter.

In addition to a measure of how often the correct model is identified, we can also find a measure of how often the selected model actually was the model that generated the data. This will be investigated in the next section, where we also will see how good/bad xv-CIC is as a model selection in these cases when compared to ^{p}AIC.

The actual values of the parameters: For those that prefer to have more detailed knowledge of the values the parameters in the different models must have in order to give the desired values of Kendall's τ, the relevant values (with a precision of six digits) are to be found in table 4.1. These parameters where computed by `iTau` from the `copula`-package.

TABLE 4.1: The exact parameter values corresponding to τ in $\{0.25, 0.50, 0.75\}$.

copula	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
`clayton`	0.666667	2.000000	6.000000
`frank`	2.371930	5.736283	14.138504
`galambos`	0.597775	1.284823	3.290396
`gumbel`	1.333333	2.000000	4.000000
`huslerReiss`	0.987050	1.803681	4.099378
`normal`	0.382683	0.707107	0.923880
`t (df=4)`	0.382683	0.707107	0.923880

Note on $\boldsymbol{\mathcal{X}}_n$, ${}^{u}\boldsymbol{\mathcal{X}}_n$ and ${}^{p}\boldsymbol{\mathcal{X}}_n$. In the description of how the data was created, the term "idealized" observations ${}^{u}\boldsymbol{\mathcal{X}}_n$ was introduced, and this requires an explanation.

First of all it might be prudent to explain why it for our analysis of the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ is sufficient to create a sample from a copula model $C(u_1, u_2)$ instead of a sample from a general bivariate model $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$.

Let us start out by noting that the `mvdc`-function from the `Copula` package, which is the function we would use to create a sample from a bivariate function (specified by its

copula and its marginal models), would start out by creating a sample like ${}^{u}\boldsymbol{\mathcal{X}}_n$ from the copula-model, and then it would use the inverse cdfs ${F_1}^{-1}$ and ${F_2}^{-1}$ in order to create the actual realization $\boldsymbol{\mathcal{X}}_n$ of our sample.

Since the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ is created by applying the empirical marginals to our observations $\boldsymbol{\mathcal{X}}_n$, and since the marginal cdfs (and their inverses) are strictly increasing functions, there is no need to invest computational resources in order to create an "original" sample $\boldsymbol{\mathcal{X}}_n$ from the "idealized" sample ${}^{u}\boldsymbol{\mathcal{X}}_n$, since the resulting pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ will be unaffected by this extra step.

This clarifies why it for our analysis is sufficient to sample from the copula models by the help of `rCopula`, but the reason for the term "idealized" on the sample ${}^{u}\boldsymbol{\mathcal{X}}_n$ still needs an explanation.

For the purposes of section 4.2, where all the analysis is based on ${}^{p}\boldsymbol{\mathcal{X}}_n$, we do not need to worry about the distinction between $\boldsymbol{\mathcal{X}}_n$ and ${}^{u}\boldsymbol{\mathcal{X}}_n$. But when we in section 4.3 and appendix B investigate the performance of the parametric model selections methods AIC and TIC when used on ${}^{u}\boldsymbol{\mathcal{X}}_n$, it is paramount to stress that we have an exceptionally rare situation.

In an ordinary situation with some bivariate observations $\boldsymbol{\mathcal{X}}_n$, we would need to specify marginal distributions in addition to the copula models before we fitted our models to $\boldsymbol{\mathcal{X}}_n$ and applied AIC to rank them. When we in our analysis use ${}^{u}\boldsymbol{\mathcal{X}}_n$ instead, we have the luxurious knowledge that the marginal distributions are uniform on $[0, 1]$. In a practical setting this would imply that we from $\boldsymbol{\mathcal{X}}_n$ could construct ${}^{u}\boldsymbol{\mathcal{X}}_n$, which only could happen under the idealized condition that we had exact knowledge of the marginal distributions.

The "cheating" we do when we use ${}^{u}\boldsymbol{\mathcal{X}}_n$ instead of $\boldsymbol{\mathcal{X}}_n$ makes the discussion in section 4.3 and appendix B less general in nature. However, as discussed in appendix B it might still be some insight to be gained from this approach.

4.2 The results

In this section we will present some tables that informs us how good the model selection method xv-CIC has worked in the cases mentioned in the previous section. And furthermore, we will see how this selection method compares to the ^{p}AIC in these cases.

The comparison of xv-CIC vs. ^{p}AIC is of interest since the latter has been widely applied in practical settings,[2] and it is thus of interest to see if the well established practice of using ^{p}AIC is superior, inferior or equivalent to the xv-CIC, the latter criterion having a formally correct theoretical basis.

As it turns out, at last for the cases considered in this thesis, it really does not seem to matter which one of the two selection methods that are applied. They appear to work just as good/bad in all the cases investigated, a result which in this authors opinion is as expected. After all, it resembles the situation from the parametric case where AIC formally only should be applied under the assumption that we have a correctly specified model, and otherwise we should fall back on the TIC. However, when we encounter finite sized samples we will without further ado happily use the computational much simpler AIC, since the end result in the long run has turned out to be more or less identical, cf. appendix B for further details.

Some tables: In the following paragraphs, we will present a few tables that summarize some information from the generated data. Most of the tables are postponed to appendix A, because this chapter otherwise would become quite cluttered.

Let us start with a table which shows how many times the different selection methods ranked the models first, where the data is based on a total of 5000 replicates, where each replicate was created with parameters corresponding to a value of Kendall's τ of 0.5, and a sample-size of $N = 250$.

TABLE 4.2: xv-CIC vs. ^{p}AIC, $N = 250$ and $\tau = 0.5$ – counting.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	4992	2	0	0	0	5	1
clayton	xv-CIC	4974	9	0	0	0	11	6
frank	^{p}AIC	3	4661	6	23	3	268	36
frank	xv-CIC	0	4738	8	27	5	198	24
galambos	^{p}AIC	0	20	1341	1755	1595	170	119
galambos	xv-CIC	0	24	1307	2011	1477	108	73
gumbel	^{p}AIC	0	28	1078	2745	813	133	203
gumbel	xv-CIC	0	36	1025	3023	726	83	107
huslerReiss	^{p}AIC	0	7	909	467	3347	245	25
huslerReiss	xv-CIC	0	9	1014	591	3209	161	16
normal	^{p}AIC	8	133	72	53	265	4174	295
normal	xv-CIC	1	193	114	89	350	4021	232
t	^{p}AIC	8	27	42	228	15	184	4496
t	xv-CIC	4	42	54	370	17	208	4305

[2] Remember that the notation ^{p}AIC in this thesis is to stress that the computations are based on the dependent pseudo-observations $^{p}\boldsymbol{\mathcal{X}}_n$ instead of the original independent observations $\boldsymbol{\mathcal{X}}_n$, and that the customary notation in the literature thus simply is AIC.

Table 4.2 can be used to deduce that the two selection methods follow each other rather closely, and we furthermore see that ^{p}AIC has more hits on the correct model for the five cases `clayton`, `galambos`, `huslerReiss`, `normal` and `t`, whereas xv-CIC only "wins" the two models `frank` and `gumbel`. Note that the scores of the three extreme-value copulas `galambos`, `gumbel` and `huslerReiss` is as expected since they do have a high degree of resemblance - which makes them hard to distinguish based on small samples. If we blur their distinction and collects them into one "extreme-value copula folder", we see that the selection methods then have a decent level of prediction (compared to the others) with regard to telling us that an extreme value copula is at play.

Furthermore, table 4.2 shows that there is quite a difference between the number of erroneous predictions for the models that are considered. The `clayton` copula is almost never ranked as number one when the data is from another model, while the `normal` and `t` copulas in comparison has been wrongly proposed as the correct model in quite a few cases.

To investigate to what extent this can be an effect of the transformation to pseudo-observations, section 4.3 will compare the result of a ranking based on AIC on the original independent observations $\boldsymbol{\mathcal{X}}_n$ (before the transformation), with those rankings we achieved by using ^{p}AIC on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$.

But before this we need to properly look upon the results from all the other combinations of τ and N that we have simulated data from. Tables showing the percentages for the twelve combinations we get from the values of N and τ are given on the following pages. Note in particular that table 4.7 is the one corresponding to table 4.2 above.

In order to emphasize the most interesting features, the following enhancements has been applied to the tables.

1. To distinguish ^{p}AIC from xv-CIC, the rows corresponding to the former is given with a light-gray tone.
2. For each row, a function has been applied that inserted a $*$ in front of the highest value. This makes it easier to find the model that the selection method in most cases proposed as the one generating the observations at hand.
3. For those cells corresponding to a match between the data generating model and the proposed model, the cell containing the highest score has been changed to **boldface**. In those cases were a tie occurs, both cells are emphasized in this way.

TABLE 4.3: xv-CIC v.s. ^{p}AIC $N = 100$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**81.28**	4.98	0	0.32	0.58	5.90	6.94
clayton	xv-CIC	*73.34	7.10	0.04	0.52	0.86	9.02	9.12
frank	^{p}AIC	11.30	*46.66	1.50	6.52	8.88	16.02	9.12
frank	xv-CIC	7.18	∗**49.92**	1.68	7.78	9.72	15.86	7.86
galambos	^{p}AIC	2.06	6.86	**8.00**	26.94	*37.42	9.00	9.72
galambos	xv-CIC	1.38	7.52	7.06	33.10	*35.78	7.82	7.34
gumbel	^{p}AIC	2.20	7.42	6.52	*34.70	27.74	7.34	14.08
gumbel	xv-CIC	1.46	8.16	6.04	∗**40.92**	26.10	6.66	10.66
huslerReiss	^{p}AIC	1.78	5.92	6.54	21.32	∗**47.08**	9.76	7.60
huslerReiss	xv-CIC	1.00	6.34	7.10	26.12	*45.28	8.50	5.66
normal	^{p}AIC	15.82	17.14	2.42	6.20	18.30	∗**30.80**	9.32
normal	xv-CIC	10.74	18.96	2.72	8.18	20.04	*30.74	8.62
t	^{p}AIC	10.50	4.72	1.58	10.60	4.72	4.44	∗**63.44**
t	xv-CIC	7.34	5.68	1.58	14.98	4.86	4.80	*60.76

TABLE 4.4: xv-CIC v.s. ^{p}AIC $N = 100$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**95.74**	1.30	0	0	0	1.28	1.68
clayton	xv-CIC	*91.48	2.36	0	0	0	2.96	3.20
frank	^{p}AIC	2.34	*70.76	0.98	2.80	2.92	14.02	6.18
frank	xv-CIC	0.94	∗**75.78**	1.00	3.66	3.40	10.56	4.66
galambos	^{p}AIC	0.08	3.18	11.62	30.82	*38.60	7.86	7.84
galambos	xv-CIC	0.04	3.92	**11.90**	*38.38	35.52	5.30	4.94
gumbel	^{p}AIC	0	3.06	11.18	*40.64	27.74	6.28	11.10
gumbel	xv-CIC	0	3.60	9.60	∗**49.18**	25.48	4.46	7.68
huslerReiss	^{p}AIC	0.06	2.08	10.24	17.66	∗**55.60**	10.32	4.04
huslerReiss	xv-CIC	0	2.84	11.06	22.56	*54.14	6.56	2.84
normal	^{p}AIC	4.02	8.54	2.40	2.94	13.78	∗**55.60**	12.72
normal	xv-CIC	2.08	11.32	3.10	4.72	17.36	*50.20	11.22
t	^{p}AIC	4.38	3.46	1.96	9.92	3.66	9.66	∗**66.96**
t	xv-CIC	2.48	4.66	2.06	15.60	4.14	9.38	*61.68

TABLE 4.5: xv-CIC v.s. ^{p}AIC $N = 100$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**98.46**	0.84	0	0	0	0.22	0.48
clayton	xv-CIC	*95.58	2.56	0	0	0	0.46	1.40
frank	^{p}AIC	0.24	*89.14	0.34	1.32	0.60	5.94	2.42
frank	xv-CIC	0	∗**92.22**	0.28	1.68	0.62	3.52	1.68
galambos	^{p}AIC	0	1.66	**15.18**	*37.26	28.28	7.76	9.86
galambos	xv-CIC	0	2.48	14.14	*46.22	25.98	4.80	6.38
gumbel	^{p}AIC	0	1.68	14.30	*44.12	22.76	6.30	10.84
gumbel	xv-CIC	0	2.48	12.94	∗**53.18**	20.50	4.04	6.86
huslerReiss	^{p}AIC	0	1.40	13.46	16.96	∗**52.40**	11.16	4.62
huslerReiss	xv-CIC	0	1.96	14.52	23.26	*50.18	7.42	2.66
normal	^{p}AIC	0.64	4.34	1.94	2.28	8.78	∗**64.38**	17.64
normal	xv-CIC	0.30	6.54	3.22	5.00	11.80	*57.02	16.12
t	^{p}AIC	1.16	2.28	1.64	8.44	2.62	13.54	∗**70.32**
t	xv-CIC	0.64	3.22	2.02	13.96	3.06	12.70	*64.40

TABLE 4.6: xv-CIC v.s. ^{p}AIC $N = 250$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**95.30**	1.20	0	0.02	0	2.12	1.36
clayton	xv-CIC	*92.04	1.96	0	0.02	0	3.94	2.04
frank	^{p}AIC	3.60	*70.90	0.88	3.00	1.94	17.42	2.26
frank	xv-CIC	2.24	∗**73.12**	1.02	3.54	2.30	15.82	1.96
galambos	^{p}AIC	0.02	2.34	15.40	30.70	*41.18	7.34	3.02
galambos	xv-CIC	0.02	2.70	**16.16**	34.30	*38.66	5.82	2.34
gumbel	^{p}AIC	0.12	3.20	13.64	*49.40	21.80	5.46	6.38
gumbel	xv-CIC	0.08	3.44	12.72	∗**54.88**	20.12	4.36	4.40
huslerReiss	^{p}AIC	0.08	1.70	13.88	19.50	∗**54.96**	7.86	2.02
huslerReiss	xv-CIC	0.04	1.98	14.92	22.70	*52.62	6.42	1.32
normal	^{p}AIC	5.32	14.52	2.64	4.28	10.74	∗**58.56**	3.94
normal	xv-CIC	3.00	16.12	3.18	5.68	11.88	*57.16	2.98
t	^{p}AIC	2.76	1.36	0.96	5.92	0.86	2.38	∗**85.76**
t	xv-CIC	1.88	1.74	1.04	8.60	0.66	2.68	*83.40

TABLE 4.7: xv-CIC v.s. ^{p}AIC $N = 250$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**99.84**	0.04	0	0	0	0.10	0.02
clayton	xv-CIC	*99.48	0.18	0	0	0	0.22	0.12
frank	^{p}AIC	0.06	*93.22	0.12	0.46	0.06	5.36	0.72
frank	xv-CIC	0	∗**94.76**	0.16	0.54	0.10	3.96	0.48
galambos	^{p}AIC	0	0.40	**26.82**	*35.10	31.90	3.40	2.38
galambos	xv-CIC	0	0.48	26.14	*40.22	29.54	2.16	1.46
gumbel	^{p}AIC	0	0.56	21.56	*54.90	16.26	2.66	4.06
gumbel	xv-CIC	0	0.72	20.50	∗**60.46**	14.52	1.66	2.14
huslerReiss	^{p}AIC	0	0.14	18.18	9.34	∗**66.94**	4.90	0.50
huslerReiss	xv-CIC	0	0.18	20.28	11.82	*64.18	3.22	0.32
normal	^{p}AIC	0.16	2.66	1.44	1.06	5.30	∗**83.48**	5.90
normal	xv-CIC	0.02	3.86	2.28	1.78	7.00	*80.42	4.64
t	^{p}AIC	0.16	0.54	0.84	4.56	0.30	3.68	∗**89.92**
t	xv-CIC	0.08	0.84	1.08	7.40	0.34	4.16	*86.10

TABLE 4.8: xv-CIC v.s. ^{p}AIC $N = 250$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*100	0	0	0	0	0	0
clayton	xv-CIC	∗**99.92**	0.08	0	0	0	0	0
frank	^{p}AIC	0	*99.54	0	0.02	0	0.26	0.18
frank	xv-CIC	0	∗**99.82**	0	0.02	0	0.06	0.10
galambos	^{p}AIC	0	0.08	**29.96**	*45.28	19.98	1.68	3.02
galambos	xv-CIC	0	0.10	28.42	*51.08	17.98	0.84	1.58
gumbel	^{p}AIC	0	0.12	26.82	*55.48	12.32	1.46	3.80
gumbel	xv-CIC	0	0.14	24.82	∗**61.22**	10.92	0.88	2.02
huslerReiss	^{p}AIC	0	0.08	20.14	6.40	∗**69.02**	3.60	0.76
huslerReiss	xv-CIC	0	0.10	22.98	8.30	*66.38	1.86	0.38
normal	^{p}AIC	0	0.48	0.64	0.50	1.92	∗**88.50**	7.96
normal	xv-CIC	0	0.86	1.66	1.12	3.14	*86.36	6.86
t	^{p}AIC	0	0.10	0.56	2.46	0.10	4.22	∗**92.56**
t	xv-CIC	0	0.20	0.66	5.02	0.16	4.88	*89.08

TABLE 4.9: xv-CIC v.s. ^{p}AIC $N = 500$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**99.32**	0.16	0	0	0	0.40	0.12
clayton	xv-CIC	*98.98	0.20	0	0	0	0.60	0.22
frank	^{p}AIC	0.20	*86.86	0.14	1.10	0.24	11.08	0.38
frank	xv-CIC	0.12	∗**87.86**	0.24	1.18	0.28	10.04	0.28
galambos	^{p}AIC	0	0.46	26.72	30.78	*38.00	3.40	0.64
galambos	xv-CIC	0	0.52	**26.92**	33.62	*35.78	2.88	0.28
gumbel	^{p}AIC	0	0.80	19.22	*62.68	13.14	2.38	1.78
gumbel	xv-CIC	0	0.82	18.06	∗**65.86**	12.16	1.92	1.18
huslerReiss	^{p}AIC	0	0.24	19.82	12.38	∗**63.56**	3.80	0.20
huslerReiss	xv-CIC	0	0.30	21.20	14.46	*60.90	3.00	0.14
normal	^{p}AIC	1.34	8.62	1.54	1.52	4.10	∗**81.96**	0.92
normal	xv-CIC	0.94	9.50	1.82	1.82	4.92	*80.20	0.80
t	^{p}AIC	0.26	0.22	0.22	2.38	0.04	0.72	∗**96.16**
t	xv-CIC	0.22	0.32	0.22	3.24	0.04	0.78	*95.18

TABLE 4.10: xv-CIC v.s. ^{p}AIC $N = 500$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	xv-CIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	*99.06	0	0	0	0.86	0.08
frank	xv-CIC	0	∗**99.40**	0	0	0	0.54	0.06
galambos	^{p}AIC	0	0	∗**41.76**	32.02	24.88	0.92	0.42
galambos	xv-CIC	0	0.02	*41.42	35.30	22.44	0.58	0.24
gumbel	^{p}AIC	0	0	27.84	*63.42	7.30	0.52	0.92
gumbel	xv-CIC	0	0	25.54	∗**67.26**	6.50	0.34	0.36
huslerReiss	^{p}AIC	0	0.02	20.42	3.20	∗**74.98**	1.34	0.04
huslerReiss	xv-CIC	0	0.02	22.22	3.86	*73.16	0.72	0.02
normal	^{p}AIC	0.02	0.48	0.28	0.10	0.70	∗**97.18**	1.24
normal	xv-CIC	0.02	0.66	0.50	0.16	1.10	*96.62	0.94
t	^{p}AIC	0	0	0.10	1.24	0.02	0.74	∗**97.90**
t	xv-CIC	0	0.04	0.12	2.02	0.02	0.88	*96.92

TABLE 4.11: xv-CIC v.s. ^{p}AIC $N = 500$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	xv-CIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	xv-CIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	*44.86	43.20	11.38	0.24	0.32
galambos	xv-CIC	0	0	**43.38**	*46.04	10.28	0.14	0.16
gumbel	^{p}AIC	0	0	32.68	*61.72	5.04	0.04	0.52
gumbel	xv-CIC	0	0	29.68	∗**65.50**	4.46	0.04	0.32
huslerReiss	^{p}AIC	0	0	17.20	1.50	∗**80.84**	0.44	0.02
huslerReiss	xv-CIC	0	0	18.84	1.78	*79.10	0.26	0.02
normal	^{p}AIC	0	0.02	0.06	0.02	0.16	∗**98.22**	1.52
normal	xv-CIC	0	0.08	0.16	0.02	0.28	*98.08	1.38
t	^{p}AIC	0	0	0.06	0.44	0	0.74	∗**98.76**
t	xv-CIC	0	0	0.08	0.98	0.04	0.92	*97.98

TABLE 4.12: xv-CIC v.s. ^{p}AIC $N = 1000$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**99.94**	0	0	0	0	0.06	0
clayton	xv-CIC	∗**99.94**	0	0	0	0	0.06	0
frank	^{p}AIC	0	*95.02	0.04	0.04	0.02	4.88	0
frank	xv-CIC	0	∗**95.54**	0.04	0.04	0.02	4.36	0
galambos	^{p}AIC	0	0.02	*41.02	24.46	33.76	0.74	0
galambos	xv-CIC	0	0.02	∗**41.62**	26.08	31.66	0.62	0
gumbel	^{p}AIC	0	0.10	20.44	*73.42	5.26	0.48	0.30
gumbel	xv-CIC	0	0.12	19.34	∗**75.14**	4.94	0.32	0.14
huslerReiss	^{p}AIC	0	0.02	23.54	5.28	∗**70.32**	0.84	0
huslerReiss	xv-CIC	0	0.02	25.12	5.78	*68.42	0.66	0
normal	^{p}AIC	0.06	3.32	0.28	0.34	0.66	∗**95.32**	0.02
normal	xv-CIC	0.04	3.84	0.38	0.42	0.90	*94.42	0
t	^{p}AIC	0.02	0	0	0.24	0	0	∗**99.74**
t	xv-CIC	0.02	0	0	0.40	0	0.02	*99.56

TABLE 4.13: xv-CIC v.s. ^{p}AIC $N = 1000$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	xv-CIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	xv-CIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	∗**56.42**	28.04	15.42	0.08	0.04
galambos	xv-CIC	0	0	*55.80	30.04	14.12	0.04	0
gumbel	^{p}AIC	0	0	25.46	*72.82	1.60	0.02	0.10
gumbel	xv-CIC	0	0	23.60	∗**74.88**	1.50	0	0.02
huslerReiss	^{p}AIC	0	0	14.42	0.36	∗**85.18**	0.04	0
huslerReiss	xv-CIC	0	0	15.90	0.40	*83.68	0.02	0
normal	^{p}AIC	0	0.02	0.04	0	0.06	∗**99.82**	0.06
normal	xv-CIC	0	0.04	0.04	0	0.08	*99.80	0.04
t	^{p}AIC	0	0	0	0.12	0	0.02	∗**99.86**
t	xv-CIC	0	0	0	0.18	0	0.02	*99.80

TABLE 4.14: xv-CIC v.s. ^{p}AIC $N = 1000$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	xv-CIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	xv-CIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	∗**56.54**	39.36	4.10	0	0
galambos	xv-CIC	0	0	*55.18	41.24	3.58	0	0
gumbel	^{p}AIC	0	0	31.90	*67.24	0.86	0	0
gumbel	xv-CIC	0	0	29.86	∗**69.40**	0.74	0	0
huslerReiss	^{p}AIC	0	0	8.52	0.04	∗**91.44**	0	0
huslerReiss	xv-CIC	0	0	9.54	0.04	*90.42	0	0
normal	^{p}AIC	0	0	0	0	0	*99.88	0.12
normal	xv-CIC	0	0	0	0	0	∗**99.94**	0.06
t	^{p}AIC	0	0	0	0	0	0.04	∗**99.96**
t	xv-CIC	0	0	0	0.02	0.04	0.08	*99.86

Identifying the correct model — confident conclusions?

The twelve tables from table 4.3 to table 4.14 are nice to consider when we want to see how the two selection methods fared with regard to proposing different models as the source of the data that was inspected, and they roughly show the same trend as in table 4.2. But to examine this closer we look at the data from a slightly different perspective.

To be specific, the main point of interest is to consider the following two measures

1. *Hit rate for selection*, i.e. how often do the selection method propose the correct model as its candidate for the data-generating model.

2. *Confidence in conclusion*, i.e. how much faith can we put in the result from the selection method. We compute this column-wise by dividing the number of correctly proposed models with the total number of proposed models. For example, if we from table 4.2 would like to find the confidence in conclusion for xv-CIC and the `gumbel`-copula, we find that it was proposed correctly 3023 times, but altogether it was proposed 6111 times, leading to a meager confidence in conclusion of 49.47 percent.

To elaborate: A selection method with a high hitting rate for one model might be rather lousy if it frequently err by proposing the same model as candidate for data from the other models as well. Conversely, a selection method with a low hitting rate for one of the models might still be worth to consider if it almost never propose that model as a candidate when data from other models are considered.

The sequence of tables that is presented from table 4.15 to table 4.28 gives us information, sorted by the copula models, regarding how the selection method xv-CIC scores against ^{p}AIC when it comes to hit rate for selection and confidence in conclusion.

TABLE 4.15: xv-CIC v.s. ^{p}AIC — hit rate for selection — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**81.28**	**95.74**	***98.46**
100	xv-CIC	73.34	91.48	*95.58
250	^{p}AIC	**95.30**	**99.84**	***100**
250	xv-CIC	92.04	99.48	*99.92
500	^{p}AIC	**99.32**	***100**	***100**
500	xv-CIC	98.98	***100**	***100**
1000	^{p}AIC	**99.94**	***100**	***100**
1000	xv-CIC	**99.94**	***100**	***100**

TABLE 4.16: xv-CIC v.s. ^{p}AIC — confidence in conclusion — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	65.05	89.79	*97.97
100	xv-CIC	**71.59**	**94.28**	***99.02**
250	^{p}AIC	88.89	99.62	***100**
250	xv-CIC	**92.68**	**99.89**	***100**
500	^{p}AIC	98.21	**99.98**	***100**
500	xv-CIC	**98.72**	**99.98**	***100**
1000	^{p}AIC	99.92	***100**	***100**
1000	xv-CIC	**99.94**	***100**	***100**

TABLE 4.17: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	46.66	70.76	*89.14
100	xv-CIC	**49.92**	**75.78**	∗**92.22**
250	^{p}AIC	70.90	93.22	*99.54
250	xv-CIC	**73.12**	**94.76**	∗**99.82**
500	^{p}AIC	86.86	99.06	∗**100**
500	xv-CIC	**87.86**	**99.40**	∗**100**
1000	^{p}AIC	95.02	∗**100**	∗**100**
1000	xv-CIC	**95.54**	∗**100**	∗**100**

TABLE 4.18: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**49.79**	**76.59**	∗**87.96**
100	xv-CIC	48.14	72.53	*82.73
250	^{p}AIC	**74.45**	**95.55**	∗**99.14**
250	xv-CIC	72.35	93.80	*98.53
500	^{p}AIC	**89.21**	**99.49**	∗**99.98**
500	xv-CIC	88.28	99.26	*99.92
1000	^{p}AIC	**96.48**	**99.98**	∗**100**
1000	xv-CIC	95.98	99.96	∗**100**

TABLE 4.19: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**8.00**	11.62	∗**15.18**
100	xv-CIC	7.06	**11.90**	*14.14
250	^{p}AIC	15.40	**26.82**	∗**29.96**
250	xv-CIC	**16.16**	26.14	*28.42
500	^{p}AIC	26.72	**41.76**	∗**44.86**
500	xv-CIC	**26.92**	41.42	*43.38
1000	^{p}AIC	41.02	**56.42**	∗**56.54**
1000	xv-CIC	**41.62**	*55.80	55.18

TABLE 4.20: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**30.12**	30.27	∗**32.39**
100	xv-CIC	26.92	∗**30.73**	30.00
250	^{p}AIC	32.48	∗**38.89**	**38.35**
250	xv-CIC	**32.95**	*37.10	36.18
500	^{p}AIC	**39.49**	**46.19**	∗**47.29**
500	xv-CIC	39.32	46.12	*47.08
1000	^{p}AIC	48.07	∗**58.56**	58.31
1000	xv-CIC	**48.11**	*58.52	**58.34**

Comparing xv-CIC **vs.** ^{p}AIC**:** An inspection of the tables from table 4.15 to table 4.28 tells us the following.

1. Both xv-CIC and ^{p}AIC fares better when the value of τ increases.
2. Both xv-CIC and ^{p}AIC fares better when the value of N increases.
3. ^{p}AIC has a better hit rate for data generated by the five copula models `clayton`, `galambos`, `huslerReiss`, `normal` and `t`.
4. xv-CIC has a higher confidence in the conclusion for data generated by the five copula models `clayton`, `galambos`, `huslerReiss`, `normal` and `t`.
5. The difference in performance between xv-CIC and ^{p}AIC is altogether rather small, i.e. when the size of the samples increases. They either both perform good or they are equally bad.

Note that the interchange we see in the scores of the hit rate and confidence in conclusion is a consequence of the affinity/aversion the two selection methods show towards the different copula models. In particular, xv-CIC does not only give a good hit rate for

TABLE 4.21: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	34.70	40.64	*44.12
100	xv-CIC	**40.92**	**49.18**	∗**53.18**
250	^{p}AIC	49.40	54.90	*55.48
250	xv-CIC	**54.88**	**60.46**	∗**61.22**
500	^{p}AIC	62.68	*63.42	61.72
500	xv-CIC	**65.86**	∗**67.26**	**65.50**
1000	^{p}AIC	*73.42	72.82	67.24
1000	xv-CIC	∗**75.14**	**74.88**	**69.40**

TABLE 4.22: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**32.55**	**38.78**	∗**39.97**
100	xv-CIC	31.09	36.67	*37.11
250	^{p}AIC	**43.78**	∗**52.07**	**50.37**
250	xv-CIC	42.30	*49.46	48.29
500	^{p}AIC	**56.54**	∗**63.43**	**57.74**
500	xv-CIC	54.80	*61.93	57.29
1000	^{p}AIC	**70.74**	∗**71.85**	**63.05**
1000	xv-CIC	69.66	*70.97	62.69

TABLE 4.23: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**47.08**	∗**55.60**	**52.40**
100	xv-CIC	45.28	*54.14	50.18
250	^{p}AIC	**54.96**	**66.94**	∗**69.02**
250	xv-CIC	52.62	64.18	*66.38
500	^{p}AIC	**63.56**	**74.98**	∗**80.84**
500	xv-CIC	60.90	73.16	*79.10
1000	^{p}AIC	**70.32**	**85.18**	∗**91.44**
1000	xv-CIC	68.42	83.68	*90.42

TABLE 4.24: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**32.53**	**39.07**	∗**45.39**
100	xv-CIC	31.74	38.66	*44.74
250	^{p}AIC	**41.80**	55.43	*66.78
250	xv-CIC	41.68	**55.48**	∗**67.33**
500	^{p}AIC	53.37	69.50	*82.98
500	xv-CIC	**53.38**	**70.87**	∗**84.00**
1000	^{p}AIC	63.91	83.29	*94.85
1000	xv-CIC	**64.58**	**84.20**	∗**95.39**

data generated from the `gumbel` copula, it also frequently propose the `gumbel` for data from other models too – which naturally leads to a lower score on the confidence in conclusion.

Furthermore, note that the deterioration in performance due to lower values of τ is as expected – since the copula models approaches the independence copula when τ is small.

The obvious observation that an increase in the sample size N gives a better performance of the selection methods does not require any further discussion. However, it is worth mentioning that the differences in performance between the two selection methods does become quite small when N increases – and depending on the copula under consideration they can actual be rather close for small samples too.

TABLE 4.25: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**30.80**	**55.60**	∗**64.38**
100	xv-CIC	30.74	50.20	*57.02
250	^{p}AIC	**58.56**	**83.48**	∗**88.50**
250	xv-CIC	57.16	80.42	*86.36
500	^{p}AIC	**81.96**	**97.18**	∗**98.22**
500	xv-CIC	80.20	96.62	*98.08
1000	^{p}AIC	**95.32**	**99.82**	*99.88
1000	xv-CIC	94.42	99.80	∗**99.94**

TABLE 4.26: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**36.99**	52.94	*58.90
100	xv-CIC	36.85	**56.13**	∗**63.38**
250	^{p}AIC	57.89	80.59	*88.74
250	xv-CIC	**59.41**	**83.94**	∗**91.02**
500	^{p}AIC	79.00	95.68	*98.53
500	xv-CIC	**80.66**	**96.93**	∗**98.63**
1000	^{p}AIC	93.15	99.83	∗**99.95**
1000	xv-CIC	**93.98**	**99.91**	*99.92

TABLE 4.27: xv-CIC v.s. ^{p}AIC
— hit rate for selection —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	**63.44**	**66.96**	∗**70.32**
100	xv-CIC	60.76	61.68	*64.40
250	^{p}AIC	**85.76**	**89.92**	∗**92.56**
250	xv-CIC	83.40	86.10	*89.08
500	^{p}AIC	**96.16**	**97.90**	∗**98.76**
500	xv-CIC	95.18	96.92	*97.98
1000	^{p}AIC	**99.74**	**99.86**	∗**99.96**
1000	xv-CIC	99.56	99.80	*99.86

TABLE 4.28: xv-CIC v.s. ^{p}AIC
— confidence in conclusion —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	^{p}AIC	52.76	*60.58	60.52
100	xv-CIC	**55.22**	**64.10**	∗**64.72**
250	^{p}AIC	81.87	*86.87	85.48
250	xv-CIC	**84.72**	∗**90.38**	**89.06**
500	^{p}AIC	95.96	97.31	*97.64
500	xv-CIC	**97.04**	∗**98.35**	**98.11**
1000	^{p}AIC	99.68	99.80	*99.88
1000	xv-CIC	**99.85**	**99.93**	∗**99.93**

Conclusion: For those cases treated in this chapter, the two semiparametric selection methods xv-CIC and ^{p}AIC seems to perform good and bad at the same cases. When they do perform good, their difference is small enough to make it tempting to consider them to be interchangeable. Tables 4.29 and 4.30 gives a summary of the *differences* in hit rates and in confidence in conclusion for the sample-size $N = 1000$. i.e. a positive number represents a case in which ^{p}AIC fares better than xv-CIC.

TABLE 4.29: Difference in hit rates for $N = 1000$: ^{p}AIC − xv-CIC

$N = 1000$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	0.00	0.00	0.00
frank	-0.52	0.00	0.00
galambos	-0.60	0.62	1.36
gumbel	-1.72	-2.06	-2.16
huslerReiss	1.90	1.50	1.02
normal	0.90	0.02	-0.06
t	0.18	0.06	0.10

TABLE 4.30: Difference in confidence in conclusion for $N = 1000$: ^{p}AIC − xv-CIC

$N = 1000$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	-0.02	0.00	0.00
frank	0.51	0.02	0.00
galambos	-0.04	0.04	-0.03
gumbel	1.08	0.88	0.36
huslerReiss	-0.67	-0.90	-0.55
normal	-0.83	-0.08	0.04
t	-0.18	-0.14	-0.06

An inspection of table 4.29 tells us that ^{p}AIC have an edge over xv-CIC with regard to the *hit rate* (it is best in two thirds of the situations), and the difference in *confidence in conclusion* given in table 4.30 seems to be negligible. This authors impression is thus that there has been no payoff for all the extra computational investment that was made in the production of the xv-CIC-values.

With regard to this, it seems reasonable to propose the same practice here as the one used in the fully parametric setting, cf. appendix B, where the costly computation of TIC is avoided in favor of the inexpensive AIC. Even though AIC in cases like this at most can be correct for one of the models, we still use it for all of them since the resulting selection tool for most practical uses can be considered to be just as good.

A remark: Neither of the two selection methods performs any good when N is small, and it might be tempting to wonder if we for those cases should use the leave-on-out-cross-validation ^{p}xv directly. The ^{p}xv, however, incurs a tremendous increase in the computational cost, and additional simulations are needed in order to test if this would result in a selection method with a satisfying hit rate and an acceptable level of confidence in the conclusion.

Another remark: The `fitCopula`, used to find the mpl-estimates ${}^{p}\widehat{\boldsymbol{\theta}}$, can sometimes fail in its optimization process. The initial small-samples tests used for this analysis indicated that this might be a problem, but tests on larger sized samples gave the impression that the effect on the hit-rates should be negligible – and the problem were thus ignored for the time being. To justify this strategy, a more substantial foundation

is provided in table 4.31, which gives the number of `NA` that occurred when the fitting process tried to fit the correct data-generating copula to the data. As we can see, only 2 cases occurred (of a total of 420000), and both of these were in the smallest case $N = 100$ – which implies that we for this collection of copulas rarely need to fear the occurrence of this problem.

In particular, the bad performance of the two selection methods in the case of samples of size $N = 100$ is not due to any failure of the estimation of the parameters ${}^{p}\widehat{\boldsymbol{\theta}}$. Moreover, since a computation of ${}^{p}\text{xv}_N$ requires N mpl-estimates to be carried out, this might indicate that the use of ${}^{p}\text{xv}$ on small samples might encounter this problem more frequently.

TABLE 4.31: `NA`-number, negligible effect on hit rates.

N	τ	clayton	frank	galambos	gumbel	huslerReiss	normal	t
100	0.25	0	0	0	1	1	0	0
100	0.50	0	0	0	0	0	0	0
100	0.75	0	0	0	0	0	0	0
250	0.25	0	0	0	0	0	0	0
250	0.50	0	0	0	0	0	0	0
250	0.75	0	0	0	0	0	0	0
500	0.25	0	0	0	0	0	0	0
500	0.50	0	0	0	0	0	0	0
500	0.75	0	0	0	0	0	0	0
1000	0.25	0	0	0	0	0	0	0
1000	0.50	0	0	0	0	0	0	0
1000	0.75	0	0	0	0	0	0	0

4.3 The noise in the transformation from $\boldsymbol{\mathcal{X}}_n$ to ${}^{p}\boldsymbol{\mathcal{X}}_n$

This section will consider how the noise from the "empirical marginals transformation", cf. page 21, that take independent observations $\boldsymbol{\mathcal{X}}_n$ and shuffle them around to dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, affects the values of the estimated parameters and the corresponding maximum of the log-likelihood.

We will also discuss tables akin to those used in the comparison of xv-CIC vs. ${}^{p}\text{AIC}$, in order to see how much the noise from the transformation "mess things up" with regard to the performance of our semiparametric model selection tool ${}^{p}\text{AIC}$.

Remember from section 4.1 that we really do not have at our disposition simulations of "original" independent observations $\boldsymbol{\mathcal{X}}_n$, and that we thus will have to do with the less general "idealized" samples ${}^{u}\boldsymbol{\mathcal{X}}_n$. This implies that the discussion at the end of this section, where the performance of AIC used on ${}^{u}\boldsymbol{\mathcal{X}}_n$ is compared to ${}^{p}\text{AIC}$ on ${}^{p}\boldsymbol{\mathcal{X}}_n$, gives AIC an advantage it would not have had in a general setting. Nevertheless, as discussed

in appendix B, it might still be something to be gained from this approach, since we at least can get an impression of how the noise in the transformation ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$ affects our results.

${}^{u}\boldsymbol{\mathcal{X}}_n$ vs. ${}^{p}\boldsymbol{\mathcal{X}}_n$ — estimates of parameters and likelihood. In order to see how the transformation from ${}^{u}\boldsymbol{\mathcal{X}}_n$ to ${}^{p}\boldsymbol{\mathcal{X}}_n$ can affect the values we are interested in, let us start out with an example where we see how this process influences which points we will be using as basis for our fitting process. For this purpose, consider figs. 4.1 to 4.4 – where we have plotted arrows pointing from observations in ${}^{u}\boldsymbol{\mathcal{X}}_n$ to their corresponding pseudo-observations in ${}^{p}\boldsymbol{\mathcal{X}}_n$, for four sample-sizes from the `normal` copula. Note that the observations in the first plots are subsets of those later on. Another detail to mention is that we have added the value of $\hat{\tau}$, the empirical estimate of Kendall's τ, to these plots.[3]

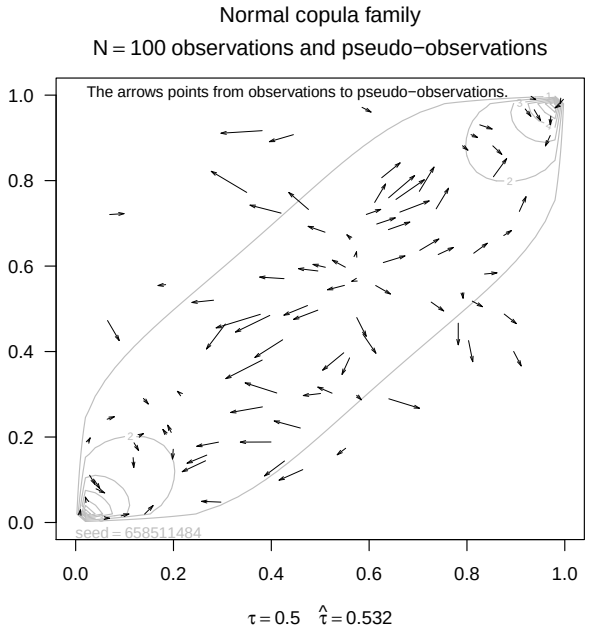


FIGURE 4.1: Visualization of noise in ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$, sample of size $N = 100$

We observe from these plots that the difference between ${}^{u}\boldsymbol{\mathcal{X}}_n$ and ${}^{p}\boldsymbol{\mathcal{X}}_n$ diminishes when the size of the sample increases from $N = 100$ to $N = 1000$, and it is thus reasonable to expect that we for large samples will have that the noise from the transformation $\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$ will incur a less severe effect on our computations.

Let us consider this in more detail by comparing the estimates obtained by maximum likelihood (ml) on ${}^{u}\boldsymbol{\mathcal{X}}_n$ with those found by the help of maximum pseudo-likelihood (mpl) on ${}^{p}\boldsymbol{\mathcal{X}}_n$. We will restrict our attention to the case of the `normal` copula, since the situation is similar for the other copula models.

[3] Since $\widehat{\tau}$ is a measure based on concordance, its value is unaffected by the shuffling made by the transformation $\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$.

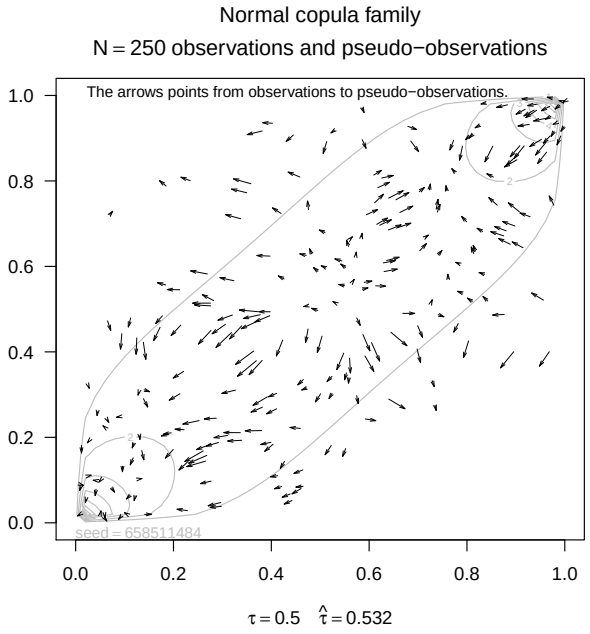


FIGURE 4.2: Visualization of noise in ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$, sample of size $N = 250$

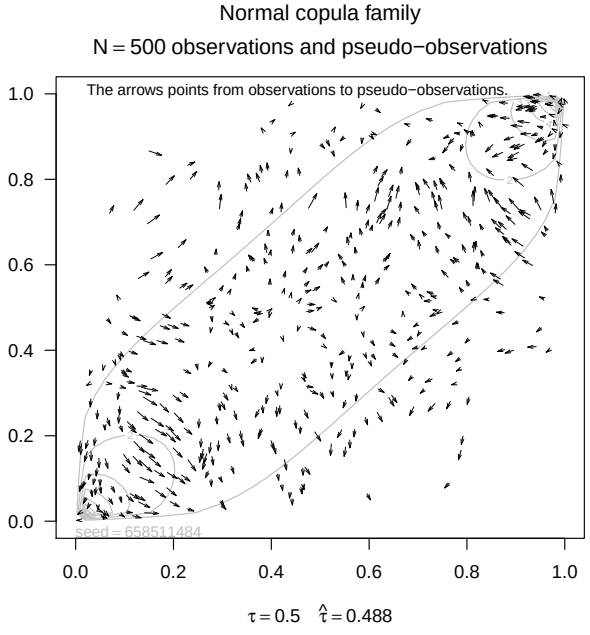


FIGURE 4.3: Visualization of noise in ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$, sample of size $N = 500$

Before we consider the tables based on the ml- and mpl-estimates of our parameters, it might be nice to refresh from table 4.1 on page 59, that the values they are attempting to estimate is $\rho_{0.25} = 0.382684$, $\rho_{0.50} = 0.707107$ and $\rho_{0.75} = 0.923880$.

When we inspect table 4.32, which gives us the mean from the parameters estimated on all or 5000 replicates, we observe the expected tendency that both the ml- and mpl-based estimates approaches the true values. From table 4.33 we furthermore observe that the

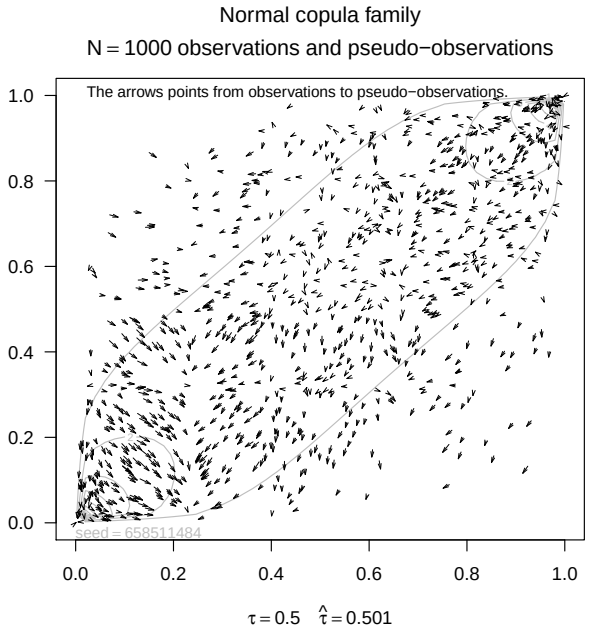


FIGURE 4.4: Visualization of noise in ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$, sample of size $N = 1000$

mean of the variance of these parameter-estimates decreases nicely when the sample-size increases, and we also see how the noise from the transformation $\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$ incurs a larger variance on the estimates.

TABLE 4.32: Mean of estimated parameters, copula = `normal`.

N	method	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	ml	0.38213	0.70719	0.92394
100	mpl	0.39840	0.71594	0.92244
250	ml	0.38155	0.70656	0.92374
250	mpl	0.38991	0.71137	0.92322
500	ml	0.38271	0.70715	0.92389
500	mpl	0.38730	0.70984	0.92364
1000	ml	0.38271	0.70720	0.92391
1000	mpl	0.38508	0.70860	0.92374

TABLE 4.33: Mean of estimated var.est, copula = `normal`.

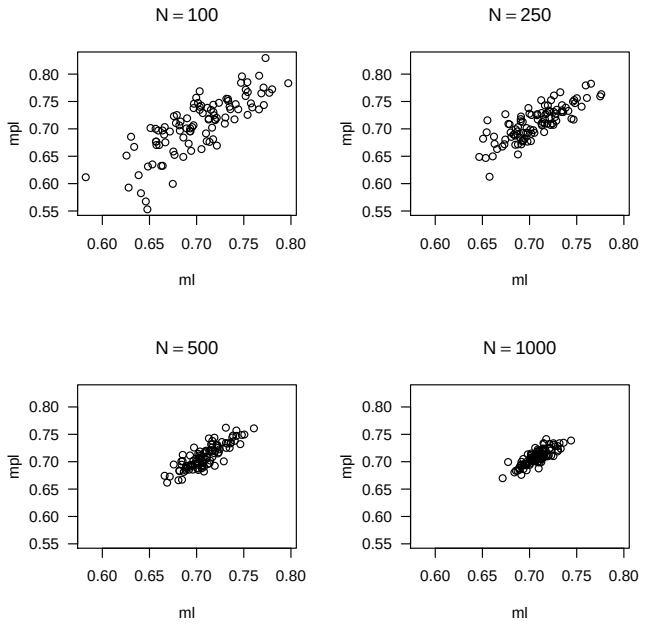
N	method	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	ml	0.006491	0.001701	0.000118
100	mpl	0.007676	0.002494	0.000295
250	ml	0.002564	0.000674	0.000047
250	mpl	0.002975	0.000990	0.000101
500	ml	0.001275	0.000335	0.000023
500	mpl	0.001475	0.000495	0.000047
1000	ml	0.000637	0.000167	0.000012
1000	mpl	0.000732	0.000248	0.000022

Some plots of estimated parameters and log-likelihood values. Tables 4.32 and 4.33 only gives us information with regard to the mean of the estimated parameters and the mean of the corresponding variance of the estimate, but we need to consider a plot if we want to see how the transformation ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$ affects the individual cases.

To investigate this effect, we have created four series of 100 replicates from the `normal` copula, with Kendall's τ equal to 0.5, in which the smaller samples are subsets of the larger ones. In figs. 4.5 and 4.6 we can respectively see the effect on the quality of the estimated parameters, and their error-limits, when we increase the size of our samples.

These plot shows us that the shuffling due to the transformation step can have some effect on the estimates of the parameters and the size of their corresponding variance, but this effect dwindles when the sample grows – in accordance with what we observed in figs. 4.1 to 4.4.

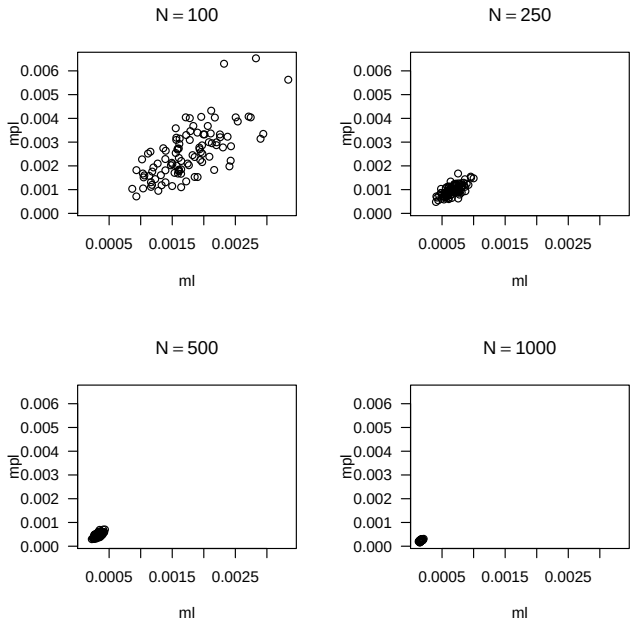
FIGURE 4.5: The `normal` copula, $\tau = 0.50$, 100 replicates.
Effect of sample size on the estimated parameters



When we consider the maximums of the ml-based log-likelihood ℓ versus the mpl-based pseudo-log-likelihood ${}^{p}\ell$, we can not use the same setup as in figs. 4.5 and 4.6 – since the scales for the different sample-sizes will differ due to the growth of the log-likelihood values. One way to resolve this is to consider a histogram of their ratios instead, which is done in fig. 4.7.

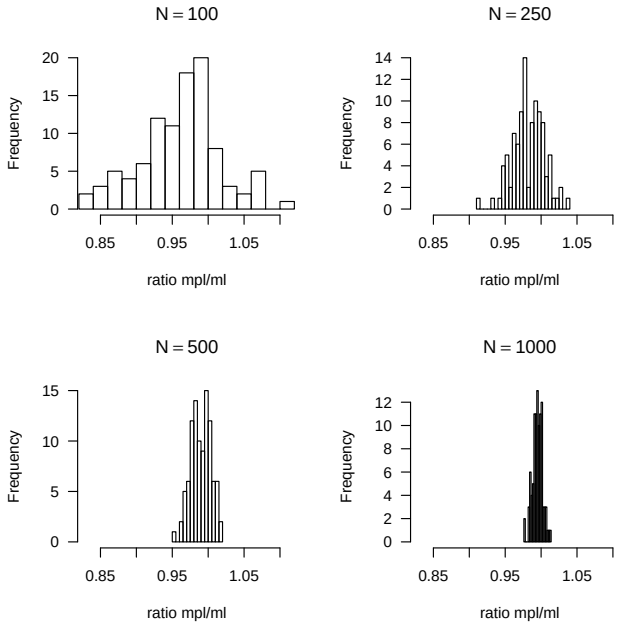
These histograms tell us that there can be some difference in the attained values of

FIGURE 4.6: The `normal` copula, $\tau = 0.50$, 100 replicates.
Effect of sample size on the size of the error



the maximum of the ${}^{p}\ell$ and the corresponding maximum of ℓ, and it is thus reasonable to expect that AIC used on the observations ${}^{u}\boldsymbol{\mathcal{X}}_n$ and ^{p}AIC used on the pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ can differ with regard to which model they rank first. But once more we can see that the spread declines for larger samples, making it more plausible that the two selection methods will propose the same model.

FIGURE 4.7: The `normal` copula, $\tau = 0.50$, 100 replicates.
Effect of sample size on the ratio of ${}^{p}\ell/\ell$



Parametric vs. semiparametric — AIC vs. ^{p}AIC. In the previous section we used ^{p}AIC and xv-CIC on pseudo-observations $^{p}\boldsymbol{\mathcal{X}}_n$, in order to figure out how these fared as pseudo-parametric model selection methods.

Remember from section 4.1 that the pseudo-observations $^{p}\boldsymbol{\mathcal{X}}_n$ were created by using the `pobs`-function from the `copula`-package on samples generated by the `rCopula`-function on the seven copula models `clayton`, `frank`, `galambos`, `gumbel` `huslerReiss`, `normal` and `t`, which implies that we here have the luxury of actually knowing $^{u}\boldsymbol{\mathcal{X}}_n$ too.

In particular, since we know that the independent observations $^{u}\boldsymbol{\mathcal{X}}_n$ were created using uniform marginals in the copula-models, we can perform a maximum likelihood based fitting of the copula models to our observations – and we can then use AIC to rank these models against each other.

As mentioned at the end of section 4.1, please note that we kind of "cheat" in this approach. We do know much more than we normally would when we instead of "original" observations $\boldsymbol{\mathcal{X}}_n$ have access to the "idealized" observations $^{u}\boldsymbol{\mathcal{X}}_n$ – where we know that the marginal distributions are uniform on $[0, 1]$. Nevertheless, even though this approach is not as general as we might wish – we should still be able to gain some insight from this line of argument, as discussed in appendix B.

Let us keep in mind the lack of generality of our observations $^{u}\boldsymbol{\mathcal{X}}_n$, and perform an analysis similar to the one conducted for ^{p}AIC vs. xv-CIC in the semiparametric case. To be precise, we want to compare the parametric AIC based on $^{u}\boldsymbol{\mathcal{X}}_n$ with the semiparametric ^{p}AIC on $^{p}\boldsymbol{\mathcal{X}}_n$.

The tables corresponding to most of those presented earlier in this chapter has been collected in appendix A, and if we we inspect the tables from table A.1 to table A.26 (pages 88 to 94) – we see that we for small samples will have that a selection method based on the observations $^{u}\boldsymbol{\mathcal{X}}_n$ is superior to a selection method that only are given the pseudo-observations $^{p}\boldsymbol{\mathcal{X}}_n$. As expected: The AIC wins over ^{p}AIC both with regard to the hit rate and the level of confidence we can have in the conclusion.[4]

Note on the normal copula: Tables A.23 and A.24, which respectively gives the hit rates and confidence in conclusion for the `normal` copula, implies that there for small samples seems to be a major difference between the result before and after the transformation $^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$, i.e. that the effect of replacing observations $^{u}\boldsymbol{\mathcal{X}}_n$ with pseudo-observations $^{p}\boldsymbol{\mathcal{X}}_n$ seems to be particularly severe for this case.

[4] Caution: Remember that we did "cheat" here, and that we in a general parametric situation first would need to guess on models for the marginal distributions, and then our AIC-approach would need to estimate parameters for both the copula and the marginals.

However, in a setting with multivariate original observations $\boldsymbol{\mathcal{X}}_n$, where the bivariate data we are considering has originated from a SSP analysis of a vine copula, cf. chapter 2 – we might expect this not be much of an issue, since an attempt at finding a multidimensional structure based on as few as $N = 100$ observations probably not would be conducted.

Note on the performance of ^{p}AIC when N grows. In a practical setting, we will not have exact knowledge of the models that the data originated from, so we do not have the option to choose between AIC or ^{p}AIC as our model selection method.

It is however nice to know that the semiparametric model selection strategy of using ^{p}AIC on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, does gives a decent performance with regard to identifying the correct model.

Tables 4.34 and 4.35 gives us the same kind of information as tables 4.29 and 4.30, i.e. they give us the difference between ^{p}AIC and AIC with regard to hit rates and confidence in conclusion. (Negative numbers thus means that AIC gave the best result.)

TABLE 4.34: Difference in hit rates for $N = 1000$: ^{p}AIC − AIC

$N = 1000$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	-0.06	0.00	0.00
frank	-0.64	0.00	0.00
galambos	-1.18	-0.62	-4.02
gumbel	0.18	0.94	4.10
huslerReiss	-2.78	-2.78	-4.46
normal	-0.66	-0.16	-0.12
t	-0.10	-0.12	-0.02

TABLE 4.35: Difference in confidence in conclusion for $N = 1000$: ^{p}AIC − AIC

$N = 1000$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	-0.06	0.00	0.00
frank	-0.06	-0.02	0.00
galambos	-1.26	-1.38	-2.09
gumbel	-1.88	-1.28	-1.73
huslerReiss	-0.72	0.75	1.09
normal	-1.59	-0.14	-0.02
t	-0.04	-0.18	-0.12

We see, as expected, from tables 4.34 and 4.35 that AIC based on the independent observations ${}^{u}\boldsymbol{\mathcal{X}}_n$ gives better results than ^{p}AIC on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, but the differences in performance is of a minor magnitude.

If we take for granted that any attempts at finding a multivariate copula model for a d-variate set of observations $\boldsymbol{\mathcal{X}}_n$, will be based on sets which contain a decent amount of data, we can from this see that ^{p}AIC used on the corresponding pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ does deserves its role as a much used model selection method.

4.4 Danish Fire Loss Data

We will in this section consider an example used in McNeil [27], i.e. we will consider the danish fire loss data, from Copenhagen Reinsurance, which is available at `http://www.ma.hw.ac.uk/~mcneil/data.html`.

These data contain the losses, in millions of Danish Krone, from 2167 fires over the period 1980 to 1990. The losses have been adjusted for inflation to reflect their 1985 values.

A few of the initial rows from the dataset is given in table 4.36, in order to introduce the headings. We wish to investigate the interdependency between "loss on contents" and "loss on profits", when we restrict our attention to the subset of 604 observations where both of them are nonzero.

TABLE 4.36: Danish Fire Loss Data

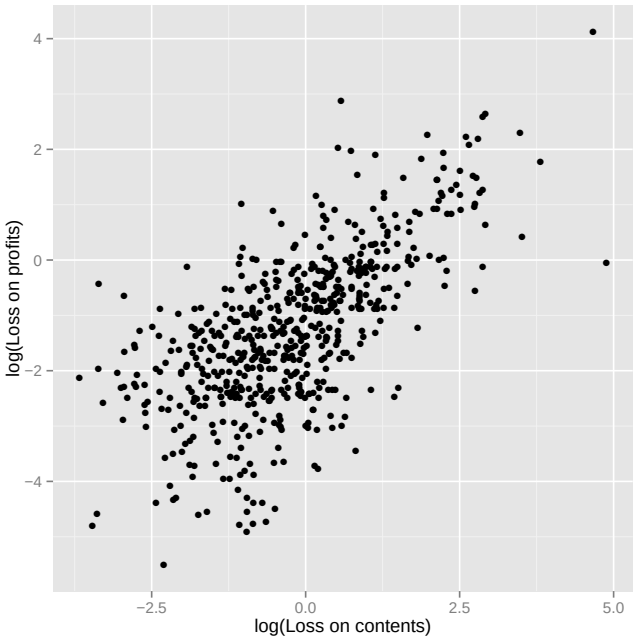
Positions	building	contents	profits	total
01/03/1980	1.098097	0.585651	0.000000	1.683748
01/04/1980	1.756955	0.336750	0.000000	2.093704
01/05/1980	1.732581	0.000000	0.000000	1.732581
01/07/1980	0.000000	1.305376	0.474378	1.779754
01/07/1980	1.244510	3.367496	0.000000	4.612006
01/10/1980	4.452040	4.273234	0.000000	8.725274
01/10/1980	2.494876	3.543192	1.860908	7.898975
01/16/1980	0.775690	0.993117	0.439239	2.208045

Since the subset of observations we are interested in have values in the span from 0.004084 to 132.0132 millions of Danish Krone, and most of the observations have low values (the median is 0.4599, whereas the mean is 1.605), we will apply the logarithm to them in order to make them more tractable for analysis.

The logarithmically transformed observations are plotted in fig. 4.8. We can from this figure see that the majority of the observations are at low values, but there are quite a few extreme cases too. What is not evident from this plot is the amount of ties that are present in our data. Along the first axis we have 514 unique values, while we along the second axis only have 381 unique values. With a total of 604 observations, this implies that we do have a lot of ties.

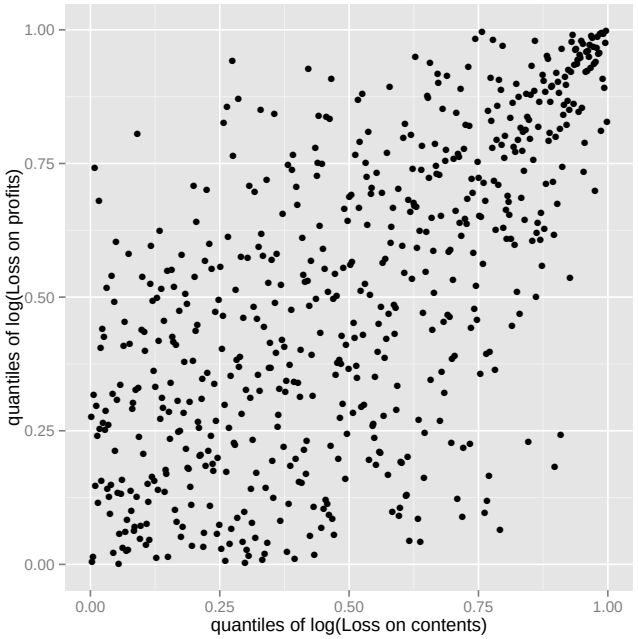
In our quest for a model that can describe the interdependencies of our data, we will use the empirical marginals in order to convert them into a setting where we can use ^{p}AIC and xv-CIC to rank different copula models against each other. The function `pobs` from the `copula`-package has been used with `ties.method="random"` in order to get the

FIGURE 4.8: Danish Fire Loss Data, logarithmic plot



points given in fig. 4.9, i.e. before the empirical marginals was applied the observations where slightly jiggled in order to get rid of the ties.

FIGURE 4.9: Danish Fire Loss Data, pseudo-observations



From fig. 4.9 it seems evident that a copula model for these data ought to have some tail dependency, but we will nevertheless in our further analysis try to fit all the seven copula models that we have considered in chapter 4. Table 4.37 presents the mpl-estimates ${}^{p}\widehat{\boldsymbol{\theta}}$

obtained by using `fitCopula` on the pseudo-observations, including the variance of these estimates and the corresponding maximum of the pseudo-log-likelihood ${}^{p}\ell$.

Table 4.38 gives the estimates of the bias-correcting terms that we need in order to compute the xv-CIC-values of these models, cf. eq. (3.98) on page 55, and in table 4.39 we find the resulting values of ^{p}AIC and xv-CIC.

TABLE 4.37: Danish Fire Loss Data, fitting of copula models

	estimate	var.est	loglik
`clayton`	0.81120	0.00369	80.08194
`frank`	5.14198	0.12561	162.67866
`galambos`	1.16026	0.00468	191.21300
`gumbel`	1.87268	0.00506	192.00155
`huslerReiss`	1.61860	0.00607	187.15355
`normal`	0.65482	0.00047	165.81343
`t (df=4)`	0.63753	0.00087	160.37309

TABLE 4.38: Danish Fire Loss Data, pqr-values for copula models

	$\widehat{p}_N$	$\widehat{q}_N$	$\widehat{r}_N$
`clayton`	1.3954996	-0.5246320	3.3954385
`frank`	0.9765355	0.0488948	0.0951288
`galambos`	1.0661042	-0.0413715	0.4080203
`gumbel`	0.9756528	-0.0253286	0.3815788
`huslerReiss`	1.4380186	-0.1135916	0.3640977
`normal`	1.1689508	-0.1326271	1.8292126
`t (df=4)`	0.9188460	-0.0472993	2.4046977

TABLE 4.39: Danish Fire Loss Data, IC-values for copula models

	^{p}AIC	xv-CIC
`clayton`	158.16	151.63
`frank`	323.36	323.12
`galambos`	380.43	379.56
`gumbel`	382.00	381.34
`huslerReiss`	372.31	370.93
`normal`	329.63	325.90
`t (df=4)`	318.75	314.19

From table 4.39 it is evident that both ^{p}AIC and xv-CIC gives the following ranking of the models with regard to the pseudo-observations plotted in fig. 4.9: `gumbel`, `galambos`, `huslerReiss`, `normal`, `frank`, `t`, and `clayton`.

In view of the size of our sample (604 observations) and the discussion in section 4.2, it is hardly a surprise that the two selection methods did agree in this case.

Furthermore, an inspection of the xv-CIC and ^{p}AIC-values of the three copula models with the highest rating, i.e. `gumbel`, `galambos` and `huslerReiss`, shows that these are

clustered close together. This fits well with the fact that these three extreme-value copulas are hard to distinguish, and it might thus be prudent to investigate some more before we decide which copula model to settle for in this case.

We can apply a *Goodness of Fit test* to our models in order to get a better idea with regard to their suitability as explanations for the interdependencies observed in fig. 4.9. Table 4.40 gives the result we obtained when we used `gofCopula` with the default setting. i.e. `Parametric bootstrap based GOF test with 'method'="Sn", 'estim.method'="mpl"`, see the documentation in the `copula`-package for further details.

TABLE 4.40: Danish Fire Loss Data, GoF-values for copula models

copula	p-value
`clayton`	0.0004995005
`frank`	0.0004995005
`galambos`	0.0074925075
`gumbel`	0.0064935065
`huslerReiss`	0.0024975025
`normal`	0.0004995005
`t (df=4)`	0.0004995005

From table 4.40, we see that none of the p-values are exceptionally high, but there are still some differences between them. The p-values for the three extreme-value copulas are the highest, which imply that they should be considered with least suspicion. If we from the values of table 4.39 are convinced that an extreme-value copula is the most decent way to explain our pseudo-observations, then in view of table 4.40 it might be more or less equally good/bad to pick either `gumbel` or `galambos` as our copula-model.

Note that the low p-values in table 4.40 could be taken as an indicator that none of our models are any good, and that we should look for other options in our quest for a describing model. A possibility to consider in this context is to see if the models might fit better on a *transformed* version of our pseudo-observations.

This strategy was applied in Berentsen et al. [28], in which the copula models where fitted to the "flipped" version of the pseudo-observations,[5] which resulted in the conclusion that the "flipped" `clayton` copula were chosen as the best candidate for the description of the interdependencies of or observations. To be precise, the copula model $C_{\text{clayton}}(1-u_1, 1-u_2)$ were picked as the best candidate for the interdependencies between U_1 (corresponding to the logarithm of the loss on contents) and U_2 (corresponding to the logarithm of the loss on profits).

A note with regard to the use of a GoF-test: In this setting, where the data we are analyzing is obtained by empirical means, we should considered it as a standard

[5] Whit "flipped" we mean that we want to fit a copula model $C(u_1, u_2)$ to the set of observations obtained by sending (u_1, u_2) to the point $(1-u_1, 1-u_2)$.

procedure to compute the *Goodness of Fit*-values, since we do not know anything certain about the properties of the process that our data originated from.

Regarding the ties in our sample: As mentioned above, there is quite a few ties in the original data we started out with. In order to apply the methodology of ^{p}AIC and xv-CIC we needed to jiggle them slightly in order to get the points separated from each other, such that we got distinct pseudo-observations. With regard to the high number of ties in our data, one might wonder if other randomizations than the one we happened to use above, would have influenced the conclusion.

To investigate the effect of a different randomization to get rid of the ties, a total of 100 realizations of the pseudo-observations were computed, and then `fitCopula` where used to find their estimates of ${}^{p}\widehat{\boldsymbol{\theta}}$ and ${}^{p}\ell$.

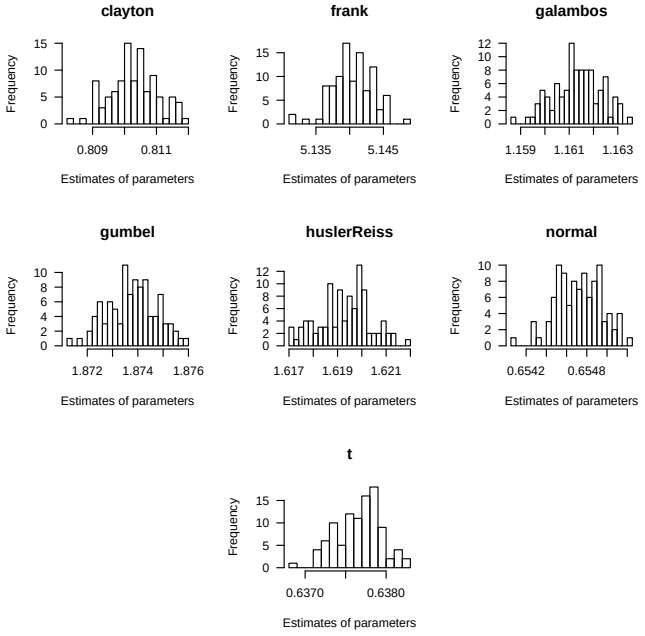
The conclusion from this little experiment with different realizations of the pseudo-observations, is that there does not seem to be any effect at all with regard to the corresponding ranking of the seven copula models. Not only where the same model ranked as number one for all the cases, but the ordering of the remaining once where also identical.

The estimates of the parameters from the different realizations of our pseudo-observations are presented in table 4.41 and fig. 4.10, and it might not be that surprising that they are so close to each other. If we keep in mind how many significant digits we should use from our estimates when they are based on a collection of 604 observations, it hardly matters at all which realization of the pseudo-observations we use.

TABLE 4.41: Danish Fire Loss Data, estimate of parameters

copula	min	mean	max
`clayton`	0.80837	0.81029	0.81184
`frank`	5.13157	5.14045	5.14881
`galambos`	1.15879	1.16135	1.16359
`gumbel`	1.87121	1.87378	1.87592
`huslerReiss`	1.61708	1.61930	1.62194
`normal`	0.65406	0.65473	0.65523
`t (df=4)`	0.63688	0.63767	0.63827

FIGURE 4.10: Danish Fire Loss Data, estimate of parameters



Chapter 5

Conclusion

In this thesis, the cross-validation copula information criterion (xv-CIC), introduced in Grønneberg [16, Part III], has been tested with regard to its performance as a semi-parametric model selection method. The basis for this testing is the combination of copula-models, parameters and sample-sizes, that are described in section 4.1.

As discussed in section 4.2, the xv-CIC does fare reasonable well when the size of the sample grows. Moreover, for the smaller samples the selection method can still be considered trustworthy if we do not have too small values of Kendall's τ.[1] However, the performance of the semiparametric selection method ^{p}AIC is just as good as xv-CIC – but it comes at a fraction of the computational cost.[2]

The xv-CIC is interesting from a theoretical perspective, but it can not be recommended as a practical tool for semiparametric model selection - at least not for the one-parametric bivariate cases considered in this thesis.

To conclude: If you want to investigate which copula model from a collection $\boldsymbol{\mathcal{C}}$ that best describes the interdependencies in a collection of bivariate data $\boldsymbol{\mathcal{X}}_n$, use the empirical marginals to create pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, fit your models to this set by using the maximum pseudo-likelihood strategy, and use ^{p}AIC to rank the models. It might furthermore be advisable to employ a *Goodness of Fit*-test in order to check that at least some of the models in $\boldsymbol{\mathcal{C}}$ gives a decent description of ${}^{p}\boldsymbol{\mathcal{X}}_n$. The same advice is valid if the bivariate data to be fitted originate from a SSP-strategy used on a regular vine-copula: Use ^{p}AIC.

[1] For small values of τ the copula models tend toward the independence copula, and as such it becomes harder to distinguish the data-generating models based on small samples.

[2] The notation ^{p}AIC, is used in this thesis to stress that we apply AIC on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$ as if they where proper independent observations.

Appendix A

Tables and plots for section 4.3

Tables related to AIC **vs.** ^{p}AIC**.** In order to avoid to clutter up chapter 4 to much, this appendix collects tables from section 4.3 that compares the rankings of AIC used on the independent observations ${}^{u}\boldsymbol{\mathcal{X}}_{n}$ versus the rankings of ^{p}AIC used on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_{n}$.

Note that AIC used on ${}^{u}\boldsymbol{\mathcal{X}}_{n}$ does not represent a completely general situation, cf. the discussion in the beginning of appendix B, but it might nevertheless tell us something interesting with regard to the trustworthiness of ^{p}AIC as a selection method in the semi-parametric case.

The most interesting information is given in tables 4.34 and 4.35, while the tables in this appendix mostly is included for the sake of completeness.

TABLE A.1: ^{p}AIC v.s. AIC $N = 100$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*81.28	4.98	0	0.32	0.58	5.90	6.94
clayton	AIC	∗**82.24**	5.02	0	0.14	0.34	6.88	5.38
frank	^{p}AIC	11.30	*46.66	1.50	6.52	8.88	16.02	9.12
frank	AIC	8.52	∗**56.62**	0.82	5.28	6.38	16.88	5.50
galambos	^{p}AIC	2.06	6.86	**8.00**	26.94	*37.42	9.00	9.72
galambos	AIC	1.14	7.44	7.70	24.32	*41.12	10.92	7.36
gumbel	^{p}AIC	2.20	7.42	6.52	∗**34.70**	27.74	7.34	14.08
gumbel	AIC	1.20	8.54	7.16	*32.84	28.78	9.54	11.94
huslerReiss	^{p}AIC	1.78	5.92	6.54	21.32	*47.08	9.76	7.60
huslerReiss	AIC	0.96	6.50	7.12	16.72	∗**51.70**	11.68	5.32
normal	^{p}AIC	15.82	17.14	2.42	6.20	18.30	*30.80	9.32
normal	AIC	12.22	20.78	1.62	5.12	13.04	∗**41.48**	5.74
t	^{p}AIC	10.50	4.72	1.58	10.60	4.72	4.44	*63.44
t	AIC	8.38	5.80	1.04	8.68	5.02	7.40	∗**63.68**

TABLE A.2: ^{p}AIC v.s. AIC $N = 100$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*95.74	1.30	0	0	0	1.28	1.68
clayton	AIC	∗**98.42**	0.30	0	0	0	0.52	0.76
frank	^{p}AIC	2.34	*70.76	0.98	2.80	2.92	14.02	6.18
frank	AIC	0.92	∗**83.14**	0.32	1.76	1.28	9.86	2.72
galambos	^{p}AIC	0.08	3.18	11.62	30.82	*38.60	7.86	7.84
galambos	AIC	0	1.68	**14.36**	28.20	*44.32	5.90	5.54
gumbel	^{p}AIC	0	3.06	11.18	*40.64	27.74	6.28	11.10
gumbel	AIC	0	2.26	12.70	∗**42.64**	29.66	5.08	7.66
huslerReiss	^{p}AIC	0.06	2.08	10.24	17.66	*55.60	10.32	4.04
huslerReiss	AIC	0	1.10	10.66	13.16	∗**66.08**	6.90	2.10
normal	^{p}AIC	4.02	8.54	2.40	2.94	13.78	*55.60	12.72
normal	AIC	1.22	8.40	1.04	1.80	7.48	∗**72.94**	7.12
t	^{p}AIC	4.38	3.46	1.96	9.92	3.66	9.66	*66.96
t	AIC	1.88	3.14	1.46	5.98	3.02	13.84	∗**70.68**

TABLE A.3: ^{p}AIC v.s. AIC $N = 100$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*98.46	0.84	0	0	0	0.22	0.48
clayton	AIC	***99.96**	0	0	0	0	0.02	0.02
frank	^{p}AIC	0.24	*89.14	0.34	1.32	0.60	5.94	2.42
frank	AIC	0.04	***97.72**	0.06	0.24	0.12	1.40	0.42
galambos	^{p}AIC	0	1.66	15.18	*37.26	28.28	7.76	9.86
galambos	AIC	0	0.20	**20.04**	34.42	*39.48	2.46	3.40
gumbel	^{p}AIC	0	1.68	14.30	*44.12	22.76	6.30	10.84
gumbel	AIC	0	0.28	18.44	***45.12**	29.62	2.00	4.54
huslerReiss	^{p}AIC	0	1.40	13.46	16.96	*52.40	11.16	4.62
huslerReiss	AIC	0	0.08	11.80	8.30	***76.24**	2.74	0.84
normal	^{p}AIC	0.64	4.34	1.94	2.28	8.78	*64.38	17.64
normal	AIC	0.04	0.70	0.46	0.64	2.64	***87.76**	7.76
t	^{p}AIC	1.16	2.28	1.64	8.44	2.62	13.54	*70.32
t	AIC	0.10	0.50	0.92	3.46	1.46	16.34	***77.22**

TABLE A.4: ^{p}AIC v.s. AIC $N = 250$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*95.30	1.20	0	0.02	0	2.12	1.36
clayton	AIC	***95.94**	0.96	0	0.02	0	1.94	1.14
frank	^{p}AIC	3.60	*70.90	0.88	3.00	1.94	17.42	2.26
frank	AIC	2.44	***75.82**	0.68	2.12	1.38	16.24	1.32
galambos	^{p}AIC	0.02	2.34	15.40	30.70	*41.18	7.34	3.02
galambos	AIC	0	2.14	**16.72**	27.42	*43.76	7.56	2.40
gumbel	^{p}AIC	0.12	3.20	13.64	*49.40	21.80	5.46	6.38
gumbel	AIC	0.06	2.92	14.32	***49.78**	22.22	5.96	4.74
huslerReiss	^{p}AIC	0.08	1.70	13.88	19.50	*54.96	7.86	2.02
huslerReiss	AIC	0.02	1.50	12.86	16.16	***60.14**	7.66	1.66
normal	^{p}AIC	5.32	14.52	2.64	4.28	10.74	*58.56	3.94
normal	AIC	4.08	15.34	1.64	3.66	6.92	***66.42**	1.94
t	^{p}AIC	2.76	1.36	0.96	5.92	0.86	2.38	*85.76
t	AIC	1.68	1.30	1.02	4.88	0.74	2.78	***87.60**

TABLE A.5: ^{p}AIC v.s. AIC $N = 250$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	***99.84**	0.04	0	0	0	0.10	0.02
clayton	AIC	*100	0	0	0	0	0	0
frank	^{p}AIC	0.06	*93.22	0.12	0.46	0.06	5.36	0.72
frank	AIC	0	***96.60**	0.02	0.26	0.06	2.74	0.32
galambos	^{p}AIC	0	0.40	26.82	*35.10	31.90	3.40	2.38
galambos	AIC	0	0.10	**29.18**	32.46	*35.70	1.24	1.32
gumbel	^{p}AIC	0	0.56	21.56	*54.90	16.26	2.66	4.06
gumbel	AIC	0	0.12	23.92	***55.30**	17.82	0.82	2.02
huslerReiss	^{p}AIC	0	0.14	18.18	9.34	*66.94	4.90	0.50
huslerReiss	AIC	0	0.04	16.52	6.84	***74.68**	1.66	0.26
normal	^{p}AIC	0.16	2.66	1.44	1.06	5.30	*83.48	5.90
normal	AIC	0	1.88	0.64	0.42	1.68	***92.74**	2.64
t	^{p}AIC	0.16	0.54	0.84	4.56	0.30	3.68	*89.92
t	AIC	0.04	0.30	0.52	2.00	0.22	4.16	***92.76**

TABLE A.6: ^{p}AIC v.s. AIC $N = 250$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	AIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**99.54**	0	0.02	0	0.26	0.18
frank	AIC	0	*100	0	0	0	0	0
galambos	^{p}AIC	0	0.08	29.96	*45.28	19.98	1.68	3.02
galambos	AIC	0	0	**34.76**	*39.32	25.28	0.20	0.44
gumbel	^{p}AIC	0	0.12	26.82	∗**55.48**	12.32	1.46	3.80
gumbel	AIC	0	0	32.62	*51.56	15.18	0.10	0.54
huslerReiss	^{p}AIC	0	0.08	20.14	6.40	*69.02	3.60	0.76
huslerReiss	AIC	0	0	14.16	2.36	∗**83.32**	0.10	0.06
normal	^{p}AIC	0	0.48	0.64	0.50	1.92	*88.50	7.96
normal	AIC	0	0.02	0.12	0.04	0.24	∗**97.26**	2.32
t	^{p}AIC	0	0.10	0.56	2.46	0.10	4.22	*92.56
t	AIC	0	0	0.10	0.52	0.04	3.92	∗**95.42**

TABLE A.7: ^{p}AIC v.s. AIC $N = 500$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	*99.32	0.16	0	0	0	0.40	0.12
clayton	AIC	∗**99.62**	0.06	0	0	0	0.18	0.14
frank	^{p}AIC	0.20	*86.86	0.14	1.10	0.24	11.08	0.38
frank	AIC	0.16	∗**88.86**	0.16	0.94	0.26	9.30	0.32
galambos	^{p}AIC	0	0.46	26.72	30.78	*38.00	3.40	0.64
galambos	AIC	0	0.46	**28.46**	28.68	*39.38	2.60	0.42
gumbel	^{p}AIC	0	0.80	19.22	*62.68	13.14	2.38	1.78
gumbel	AIC	0	0.64	19.64	∗**62.82**	13.54	1.92	1.44
huslerReiss	^{p}AIC	0	0.24	19.82	12.38	*63.56	3.80	0.20
huslerReiss	AIC	0	0.10	19.44	10.30	∗**67.18**	2.92	0.06
normal	^{p}AIC	1.34	8.62	1.54	1.52	4.10	*81.96	0.92
normal	AIC	0.82	8.76	1.00	1.10	2.50	∗**85.30**	0.52
t	^{p}AIC	0.26	0.22	0.22	2.38	0.04	0.72	*96.16
t	AIC	0.26	0.16	0.20	1.50	0.02	0.72	∗**97.14**

TABLE A.8: ^{p}AIC v.s. AIC $N = 500$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	AIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	*99.06	0	0	0	0.86	0.08
frank	AIC	0	∗**99.60**	0	0	0	0.36	0.04
galambos	^{p}AIC	0	0	*41.76	32.02	24.88	0.92	0.42
galambos	AIC	0	0	∗**43.00**	30.00	26.72	0.12	0.16
gumbel	^{p}AIC	0	0	27.84	*63.42	7.30	0.52	0.92
gumbel	AIC	0	0	28.20	∗**63.92**	7.64	0.06	0.18
huslerReiss	^{p}AIC	0	0.02	20.42	3.20	*74.98	1.34	0.04
huslerReiss	AIC	0	0	16.86	2.14	∗**80.90**	0.10	0
normal	^{p}AIC	0.02	0.48	0.28	0.10	0.70	*97.18	1.24
normal	AIC	0	0.24	0.08	0.04	0.08	∗**98.98**	0.58
t	^{p}AIC	0	0	0.10	1.24	0.02	0.74	*97.90
t	AIC	0	0	0.02	0.18	0	0.70	∗**99.10**

TABLE A.9: ^{p}AIC v.s. AIC $N = 500$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	AIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	AIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	*44.86	43.20	11.38	0.24	0.32
galambos	AIC	0	0	∗**48.58**	36.78	14.64	0	0
gumbel	^{p}AIC	0	0	32.68	∗**61.72**	5.04	0.04	0.52
gumbel	AIC	0	0	36.90	*57.20	5.88	0	0.02
huslerReiss	^{p}AIC	0	0	17.20	1.50	*80.84	0.44	0.02
huslerReiss	AIC	0	0	9.26	0.46	∗**90.28**	0	0
normal	^{p}AIC	0	0.02	0.06	0.02	0.16	*98.22	1.52
normal	AIC	0	0	0.02	0	0	∗**99.62**	0.36
t	^{p}AIC	0	0	0.06	0.44	0	0.74	*98.76
t	AIC	0	0	0	0.02	0	0.54	∗**99.44**

TABLE A.10: ^{p}AIC v.s. AIC $N = 1000$, $\tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**99.94**	0	0	0	0	0.06	0
clayton	AIC	*100	0	0	0	0	0	0
frank	^{p}AIC	0	*95.02	0.04	0.04	0.02	4.88	0
frank	AIC	0	∗**95.66**	0	0.02	0	4.32	0
galambos	^{p}AIC	0	0.02	*41.02	24.46	33.76	0.74	0
galambos	AIC	0	0	∗**42.20**	23.04	34.34	0.40	0.02
gumbel	^{p}AIC	0	0.10	20.44	∗**73.42**	5.26	0.48	0.30
gumbel	AIC	0	0.06	20.80	*73.24	5.40	0.24	0.26
huslerReiss	^{p}AIC	0	0.02	23.54	5.28	*70.32	0.84	0
huslerReiss	AIC	0	0.02	22.24	4.32	∗**73.10**	0.32	0
normal	^{p}AIC	0.06	3.32	0.28	0.34	0.66	*95.32	0.02
normal	AIC	0.02	3.34	0.30	0.10	0.26	∗**95.98**	0
t	^{p}AIC	0.02	0	0	0.24	0	0	*99.74
t	AIC	0	0	0	0.12	0	0.04	∗**99.84**

TABLE A.11: ^{p}AIC v.s. AIC $N = 1000$, $\tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	AIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	AIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	*56.42	28.04	15.42	0.08	0.04
galambos	AIC	0	0	∗**57.04**	26.20	16.74	0	0.02
gumbel	^{p}AIC	0	0	25.46	∗**72.82**	1.60	0.02	0.10
gumbel	AIC	0	0	26.28	*71.88	1.84	0	0
huslerReiss	^{p}AIC	0	0	14.42	0.36	*85.18	0.04	0
huslerReiss	AIC	0	0	11.84	0.20	∗**87.96**	0	0
normal	^{p}AIC	0	0.02	0.04	0	0.06	*99.82	0.06
normal	AIC	0	0	0	0	0.02	∗**99.98**	0
t	^{p}AIC	0	0	0	0.12	0	0.02	*99.86
t	AIC	0	0	0	0	0	0.02	∗**99.98**

TABLE A.12: ^{p}AIC v.s. AIC $N = 1000$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	^{p}AIC	∗**100**	0	0	0	0	0	0
clayton	AIC	∗**100**	0	0	0	0	0	0
frank	^{p}AIC	0	∗**100**	0	0	0	0	0
frank	AIC	0	∗**100**	0	0	0	0	0
galambos	^{p}AIC	0	0	*56.54	39.36	4.10	0	0
galambos	AIC	0	0	∗**60.56**	34.30	5.14	0	0
gumbel	^{p}AIC	0	0	31.90	∗**67.24**	0.86	0	0
gumbel	AIC	0	0	35.62	*63.14	1.24	0	0
huslerReiss	^{p}AIC	0	0	8.52	0.04	*91.44	0	0
huslerReiss	AIC	0	0	4.08	0.02	∗**95.90**	0	0
normal	^{p}AIC	0	0	0	0	0	∗**99.88**	0.12
normal	AIC	0	0	0	0	0	*100	0
t	^{p}AIC	0	0	0	0	0	0.04	*99.96
t	AIC	0	0	0	0	0	0.02	∗**99.98**

TABLE A.13: ^{p}AIC v.s. AIC — hit rate for selection — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**82.24**	**98.42**	∗**99.96**
100	^{p}AIC	81.28	95.74	*98.46
250	AIC	**95.94**	∗**100**	∗**100**
250	^{p}AIC	95.30	99.84	∗**100**
500	AIC	**99.62**	∗**100**	∗**100**
500	^{p}AIC	99.32	∗**100**	∗**100**
1000	AIC	∗**100**	∗**100**	∗**100**
1000	^{p}AIC	99.94	∗**100**	∗**100**

TABLE A.14: ^{p}AIC v.s. AIC — confidence in conclusion — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**71.72**	**96.07**	∗**99.82**
100	^{p}AIC	65.05	89.79	*97.97
250	AIC	**92.05**	**99.96**	∗**100**
250	^{p}AIC	88.89	99.62	∗**100**
500	AIC	**98.77**	∗**100**	∗**100**
500	^{p}AIC	98.21	99.98	∗**100**
1000	AIC	**99.98**	∗**100**	∗**100**
1000	^{p}AIC	99.92	∗**100**	∗**100**

TABLE A.15: ^{p}AIC v.s. AIC — hit rate for selection — copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**56.62**	**83.14**	∗**97.72**
100	^{p}AIC	46.66	70.76	*89.14
250	AIC	**75.82**	**96.60**	∗**100**
250	^{p}AIC	70.90	93.22	*99.54
500	AIC	**88.86**	**99.60**	∗**100**
500	^{p}AIC	86.86	99.06	∗**100**
1000	AIC	**95.66**	∗**100**	∗**100**
1000	^{p}AIC	95.02	∗**100**	∗**100**

TABLE A.16: ^{p}AIC v.s. AIC — confidence in conclusion — copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**51.14**	**83.12**	∗**98.23**
100	^{p}AIC	49.79	76.59	*87.96
250	AIC	**75.83**	**97.53**	∗**99.98**
250	^{p}AIC	74.45	95.55	*99.14
500	AIC	**89.72**	**99.75**	∗**100**
500	^{p}AIC	89.21	99.49	*99.98
1000	AIC	**96.54**	∗**100**	∗**100**
1000	^{p}AIC	96.48	99.98	∗**100**

TABLE A.17: ^{p}AIC v.s. AIC
— hit rate for selection —
copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	7.70	**14.36**	∗**20.04**
100	^{p}AIC	**8.00**	11.62	*15.18
250	AIC	**16.72**	**29.18**	∗**34.76**
250	^{p}AIC	15.40	26.82	*29.96
500	AIC	**28.46**	**43.00**	∗**48.58**
500	^{p}AIC	26.72	41.76	*44.86
1000	AIC	**42.20**	**57.04**	∗**60.56**
1000	^{p}AIC	41.02	56.42	*56.54

TABLE A.18: ^{p}AIC v.s. AIC
— confidence in conclusion —
copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**30.24**	**35.42**	∗**38.74**
100	^{p}AIC	30.12	30.27	*32.39
250	AIC	**35.39**	**41.21**	∗**42.51**
250	^{p}AIC	32.48	*38.89	38.35
500	AIC	**41.30**	**48.77**	∗**51.26**
500	^{p}AIC	39.49	46.19	*47.29
1000	AIC	**49.33**	**59.94**	∗**60.40**
1000	^{p}AIC	48.07	*58.56	58.31

TABLE A.19: ^{p}AIC v.s. AIC
— hit rate for selection —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	32.84	**42.64**	∗**45.12**
100	^{p}AIC	**34.70**	40.64	*44.12
250	AIC	**49.78**	∗**55.30**	51.56
250	^{p}AIC	49.40	54.90	∗**55.48**
500	AIC	**62.82**	∗**63.92**	57.20
500	^{p}AIC	62.68	*63.42	**61.72**
1000	AIC	*73.24	71.88	63.14
1000	^{p}AIC	∗**73.42**	**72.82**	**67.24**

TABLE A.20: ^{p}AIC v.s. AIC
— confidence in conclusion —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**35.27**	**45.58**	∗**48.94**
100	^{p}AIC	32.55	38.78	*39.97
250	AIC	**47.84**	∗**56.84**	**54.96**
250	^{p}AIC	43.78	*52.07	50.37
500	AIC	**59.63**	∗**66.38**	**60.55**
500	^{p}AIC	56.54	*63.43	57.74
1000	AIC	**72.62**	∗**73.13**	**64.78**
1000	^{p}AIC	70.74	*71.85	63.05

TABLE A.21: ^{p}AIC v.s. AIC
— hit rate for selection —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**51.70**	**66.08**	∗**76.24**
100	^{p}AIC	47.08	*55.60	52.40
250	AIC	**60.14**	**74.68**	∗**83.32**
250	^{p}AIC	54.96	66.94	*69.02
500	AIC	**67.18**	**80.90**	∗**90.28**
500	^{p}AIC	63.56	74.98	*80.84
1000	AIC	**73.10**	**87.96**	∗**95.90**
1000	^{p}AIC	70.32	85.18	*91.44

TABLE A.22: ^{p}AIC v.s. AIC
— confidence in conclusion —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**35.31**	**43.51**	∗**50.97**
100	^{p}AIC	32.53	39.07	*45.39
250	AIC	**44.49**	**57.37**	∗**67.16**
250	^{p}AIC	41.80	55.43	*66.78
500	AIC	**54.67**	**70.14**	*81.48
500	^{p}AIC	53.37	69.50	∗**82.98**
1000	AIC	**64.63**	82.54	*93.76
1000	^{p}AIC	63.91	**83.29**	∗**94.85**

TABLE A.23: ^{p}AIC v.s. AIC
— hit rate for selection —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**41.48**	**72.94**	∗**87.76**
100	^{p}AIC	30.80	55.60	*64.38
250	AIC	**66.42**	**92.74**	∗**97.26**
250	^{p}AIC	58.56	83.48	*88.50
500	AIC	**85.30**	**98.98**	∗**99.62**
500	^{p}AIC	81.96	97.18	*98.22
1000	AIC	**95.98**	**99.98**	∗**100**
1000	^{p}AIC	95.32	99.82	*99.88

TABLE A.24: ^{p}AIC v.s. AIC
— confidence in conclusion —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**39.58**	**63.40**	∗**77.85**
100	^{p}AIC	36.99	52.94	*58.90
250	AIC	**61.18**	**89.72**	∗**95.74**
250	^{p}AIC	57.89	80.59	*88.74
500	AIC	**82.86**	**98.66**	∗**99.46**
500	^{p}AIC	79.00	95.68	*98.53
1000	AIC	**94.74**	**99.98**	∗**99.98**
1000	^{p}AIC	93.15	99.83	*99.95

TABLE A.25: ^{p}AIC v.s. AIC
— hit rate for selection —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**63.68**	**70.68**	∗**77.22**
100	^{p}AIC	63.44	66.96	*70.32
250	AIC	**87.60**	**92.76**	∗**95.42**
250	^{p}AIC	85.76	89.92	*92.56
500	AIC	**97.14**	**99.10**	∗**99.44**
500	^{p}AIC	96.16	97.90	*98.76
1000	AIC	**99.84**	∗**99.98**	∗**99.98**
1000	^{p}AIC	99.74	99.86	*99.96

TABLE A.26: ^{p}AIC v.s. AIC
— confidence in conclusion —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**60.69**	**73.18**	∗**81.97**
100	^{p}AIC	52.76	*60.58	60.52
250	AIC	**86.90**	**93.39**	∗**96.59**
250	^{p}AIC	81.87	*86.87	85.48
500	AIC	**97.10**	**99.04**	∗**99.61**
500	^{p}AIC	95.96	97.31	*97.64
1000	AIC	**99.72**	**99.98**	∗**100**
1000	^{p}AIC	99.68	99.80	*99.88

Appendix B

AIC **vs.** TIC

This appendix is included in order to give some support to a couple of comments in section 4.2, with regard to the use of AIC and TIC in the parametric case.

Lack of generality: Since we instead of general observations $\boldsymbol{\mathcal{X}}_n$ from $\mathbb{R}^2$ will use the "idealized" observations ${}^{u}\boldsymbol{\mathcal{X}}_n$ that we got from the procedure described in section 4.1, the line of argument employed in the analysis below will alas not be general in nature.

In a practial setting, the only way we could obtain ${}^{u}\boldsymbol{\mathcal{X}}_n$ from some observations $\boldsymbol{\mathcal{X}}_n$ would be if the idealized condition of complete knowledge of the marginal distributions were satisfied. This implies that we cheat when we use ${}^{u}\boldsymbol{\mathcal{X}}_n$ as the basis for our fully-parametrically model-selection strategy, since the assumption that the marginal distributions are uniform tremendously simplifies our estimation procedure.

In order to do this analysis in a general setting, we should have simulated general observations $\boldsymbol{\mathcal{X}}_n$ by specifying both the copula model C and the marginal distributions F_1 and F_2 – and then we would need to create an assorted collection of bivariate models (by varying the choices of the copula models and the marginals), $F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$ and fit all these to $\boldsymbol{\mathcal{X}}_n$. Thereafter AIC and TIC should be used to rank our models, and finally our conclusions should be based upon these results.

This special case: Since an analysis along the general lines mentioned above would incur a large computational load, we will use the less general approach were we only consider idealized samples ${}^{u}\boldsymbol{\mathcal{X}}_n$ with uniform margins. It is not a general approach, but we do at least have proper independent observations at our disposal – and the noise from the transformation ${}^{u}\boldsymbol{\mathcal{X}}_n \to {}^{p}\boldsymbol{\mathcal{X}}_n$ is absent. Our analysis can thus be conducted using a maximum-likelihood approach, and therefore the parametric model selection methods AIC and TIC can be used to rank the models.

The point of interest is to check the sanity of the "folklore" that tells us that it is safe to apply AIC instead of the computationally more expensive TIC. Remember from the discussion in section 3.3.2 that the simple form of the bias-correcting term of the AIC is under the assumption that we are using it on a correctly specified model. This implies that the use of AIC as a model selection tool on the copula models discussed in this appendix, at most could be true for one of the models.

If we do not want to make an assumption with regard to whether or not we have found the correct model, the bias-correcting term should be estimated by the more complicated expression used in the TIC-formula.

Our investigation in this appendix are identical to the procedure used in section 4.2 to compare ^{p}AIC vs. xv-CIC on the dependent pseudo-observations ${}^{p}\boldsymbol{\mathcal{X}}_n$, the only difference being that we now are considering AIC and TIC on the independent observations ${}^{u}\boldsymbol{\mathcal{X}}_n$.

The most interesting tables are tables B.1 and B.2 which corresponds to tables 4.29 and 4.30, so we present them first. The other tables are included at the end, and the diligent reader can inspect them in order to see that we do have the same kind of closeness here as the one observed in the tables related to ^{p}AIC vs. xv-CIC. In particular, the extra computational effort invested in the computation of the TIC have not resulted in a payoff that can justify its cost.

TABLE B.1: Difference in hit rates for $N = 500$: AIC − TIC

$N = 500$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	0.00	0.00	0.00
frank	-0.28	-0.02	0.06
galambos	0.20	0.26	1.50
gumbel	-2.70	-2.86	-3.26
huslerReiss	2.78	2.54	2.10
normal	0.82	0.14	0.16
t	-0.30	-0.10	-0.04

TABLE B.2: Difference in confidence in conclusion for $N = 500$: AIC − TIC

$N = 500$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
clayton	0.04	0.00	0.00
frank	0.38	0.02	0.00
galambos	-0.01	0.12	0.53
gumbel	1.41	1.28	0.86
huslerReiss	-0.81	-1.53	-1.69
normal	-0.43	-0.08	-0.06
t	0.53	0.12	0.16

Conclusion: We see from tables B.1 and B.2 that the use of AIC as selection model, in spite of it obviously not being formally valid to use for all the copula models, performs just as good as the TIC. In order to reduce the total computational load, and keep things as simple as possible, it is thus clear that the preferred selection method to pick from our toolbox will be the AIC.

Note: The author would like to apologize for the lack of references to those that already have considered the question of which *model selection method* to select in the parametric case. With regard to their results being ingrained into the folklore, it was deemed easier

to just do the required computations in the present situation, than to go on a quest for references to earlier work upon this.

TABLE B.3: AIC v.s. TIC $N = 100,\ \tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**82.24**	5.02	0	0.14	0.34	6.88	5.38
clayton	TIC	*81.44	5.06	0.04	0.12	0.38	6.68	6.28
frank	AIC	8.52	∗**56.62**	0.82	5.28	6.38	16.88	5.50
frank	TIC	8.22	*55.46	0.98	5.32	6.38	16.82	6.82
galambos	AIC	1.14	7.44	**7.70**	24.32	*41.12	10.92	7.36
galambos	TIC	1.12	7.10	7.54	26.80	*38.22	10.28	8.94
gumbel	AIC	1.20	8.54	7.16	*32.84	28.78	9.54	11.94
gumbel	TIC	1.38	8.32	6.36	∗**34.78**	26.16	8.96	14.04
huslerReiss	AIC	0.96	6.50	7.12	16.72	∗**51.70**	11.68	5.32
huslerReiss	TIC	1.00	6.26	7.08	19.86	*48.46	11.12	6.22
normal	AIC	12.22	20.78	1.62	5.12	13.04	∗**41.48**	5.74
normal	TIC	12.00	20.12	1.80	5.24	12.84	*40.64	7.36
t	AIC	8.38	5.80	1.04	8.68	5.02	7.40	*63.68
t	TIC	7.78	5.04	0.96	8.26	4.50	6.16	∗**67.30**

TABLE B.4: AIC v.s. TIC $N = 100,\ \tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**98.42**	0.30	0	0	0	0.52	0.76
clayton	TIC	*98.20	0.36	0	0	0	0.52	0.92
frank	AIC	0.92	*83.14	0.32	1.76	1.28	9.86	2.72
frank	TIC	0.92	∗**83.30**	0.20	1.92	1.30	9.14	3.22
galambos	AIC	0	1.68	**14.36**	28.20	*44.32	5.90	5.54
galambos	TIC	0	1.74	13.58	33.62	*39.34	5.18	6.54
gumbel	AIC	0	2.26	12.70	*42.64	29.66	5.08	7.66
gumbel	TIC	0	2.22	11.58	∗**46.70**	25.80	4.58	9.12
huslerReiss	AIC	0	1.10	10.66	13.16	∗**66.08**	6.90	2.10
huslerReiss	TIC	0	1.12	11.08	16.92	*61.94	6.38	2.56
normal	AIC	1.22	8.40	1.04	1.80	7.48	∗**72.94**	7.12
normal	TIC	1.18	8.94	1.24	2.16	7.10	*69.80	9.58
t	AIC	1.88	3.14	1.46	5.98	3.02	13.84	*70.68
t	TIC	1.78	3.06	1.04	6.16	2.80	11.18	∗**73.98**

TABLE B.5: AIC v.s. TIC $N = 100,\ \tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**99.96**	0	0	0	0	0.02	0.02
clayton	TIC	∗**99.96**	0	0	0	0	0.02	0.02
frank	AIC	0.04	*97.72	0.06	0.24	0.12	1.40	0.42
frank	TIC	0.04	∗**97.92**	0.04	0.26	0.12	1.12	0.50
galambos	AIC	0	0.20	**20.04**	34.42	*39.48	2.46	3.40
galambos	TIC	0	0.20	18.98	*41.20	33.58	2.30	3.74
gumbel	AIC	0	0.28	18.44	*45.12	29.62	2.00	4.54
gumbel	TIC	0	0.28	17.06	∗**50.90**	24.92	1.84	5.00
huslerReiss	AIC	0	0.08	11.80	8.30	∗**76.24**	2.74	0.84
huslerReiss	TIC	0	0.10	13.76	11.68	*70.88	2.66	0.92
normal	AIC	0.04	0.70	0.46	0.64	2.64	∗**87.76**	7.76
normal	TIC	0.04	0.76	0.48	0.82	2.48	*85.06	10.36
t	AIC	0.10	0.50	0.92	3.46	1.46	16.34	*77.22
t	TIC	0.10	0.48	0.80	3.58	1.18	13.40	∗**80.46**

TABLE B.6: AIC v.s. TIC $N = 250,\ \tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**95.94**	0.96	0	0.02	0	1.94	1.14
clayton	TIC	*95.74	0.96	0	0.02	0	1.92	1.36
frank	AIC	2.44	*75.82	0.68	2.12	1.38	16.24	1.32
frank	TIC	2.44	∗**75.86**	0.64	2.24	1.30	15.84	1.68
galambos	AIC	0	2.14	16.72	27.42	*43.76	7.56	2.40
galambos	TIC	0	2.12	**16.92**	30.60	*40.24	7.34	2.78
gumbel	AIC	0.06	2.92	14.32	*49.78	22.22	5.96	4.74
gumbel	TIC	0.08	2.86	12.94	∗**52.78**	20.08	5.66	5.60
huslerReiss	AIC	0.02	1.50	12.86	16.16	∗**60.14**	7.66	1.66
huslerReiss	TIC	0.02	1.50	14.32	18.34	*56.54	7.34	1.94
normal	AIC	4.08	15.34	1.64	3.66	6.92	∗**66.42**	1.94
normal	TIC	4.16	15.84	1.60	4.04	6.76	*65.08	2.52
t	AIC	1.68	1.30	1.02	4.88	0.74	2.78	*87.60
t	TIC	1.46	1.32	0.84	4.52	0.62	2.28	∗**88.96**

TABLE B.7: AIC v.s. TIC $N = 250,\ \tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**100**	0	0	0	0	0	0
clayton	TIC	∗**100**	0	0	0	0	0	0
frank	AIC	0	*96.60	0.02	0.26	0.06	2.74	0.32
frank	TIC	0	∗**96.72**	0.02	0.26	0.06	2.54	0.40
galambos	AIC	0	0.10	**29.18**	32.46	*35.70	1.24	1.32
galambos	TIC	0	0.12	28.82	*36.42	31.88	1.18	1.58
gumbel	AIC	0	0.12	23.92	*55.30	17.82	0.82	2.02
gumbel	TIC	0	0.12	22.06	∗**59.58**	15.34	0.74	2.16
huslerReiss	AIC	0	0.04	16.52	6.84	∗**74.68**	1.66	0.26
huslerReiss	TIC	0	0.04	18.58	8.44	*70.98	1.66	0.30
normal	AIC	0	1.88	0.64	0.42	1.68	∗**92.74**	2.64
normal	TIC	0	2.16	0.64	0.52	1.58	*91.74	3.36
t	AIC	0.04	0.30	0.52	2.00	0.22	4.16	*92.76
t	TIC	0.04	0.28	0.40	1.90	0.20	3.28	∗**93.90**

TABLE B.8: AIC v.s. TIC $N = 250,\ \tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	***100**	0	0	0	0	0	0
clayton	TIC	***100**	0	0	0	0	0	0
frank	AIC	0	*100	0	0	0	0	0
frank	TIC	0	***99.98**	0	0	0.02	0	0
galambos	AIC	0	0	**34.76**	*39.32	25.28	0.20	0.44
galambos	TIC	0	0	33.94	*43.60	21.80	0.20	0.46
gumbel	AIC	0	0	32.62	*51.56	15.18	0.10	0.54
gumbel	TIC	0	0	30.16	***56.42**	12.76	0.10	0.56
huslerReiss	AIC	0	0	14.16	2.36	***83.32**	0.10	0.06
huslerReiss	TIC	0	0	16.68	3.36	*79.78	0.08	0.10
normal	AIC	0	0.02	0.12	0.04	0.24	***97.26**	2.32
normal	TIC	0	0.02	0.14	0.06	0.22	*96.54	3.02
t	AIC	0	0	0.10	0.52	0.04	3.92	*95.42
t	TIC	0	0	0.06	0.50	0.06	3.06	***96.32**

TABLE B.9: AIC v.s. TIC $N = 500,\ \tau = 0.25$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	***99.62**	0.06	0	0	0	0.18	0.14
clayton	TIC	***99.62**	0.06	0	0	0	0.18	0.14
frank	AIC	0.16	*88.86	0.16	0.94	0.26	9.30	0.32
frank	TIC	0.18	***89.14**	0.12	1.00	0.26	8.96	0.34
galambos	AIC	0	0.46	**28.46**	28.68	*39.38	2.60	0.42
galambos	TIC	0	0.46	28.26	31.50	*36.74	2.54	0.50
gumbel	AIC	0	0.64	19.64	*62.82	13.54	1.92	1.44
gumbel	TIC	0	0.62	18.04	***65.52**	12.30	1.88	1.64
huslerReiss	AIC	0	0.10	19.44	10.30	***67.18**	2.92	0.06
huslerReiss	TIC	0	0.12	20.70	11.90	*64.40	2.80	0.08
normal	AIC	0.82	8.76	1.00	1.10	2.50	***85.30**	0.52
normal	TIC	0.84	9.22	1.16	1.18	2.36	*84.48	0.76
t	AIC	0.26	0.16	0.20	1.50	0.02	0.72	*97.14
t	TIC	0.26	0.16	0.12	1.42	0.02	0.58	***97.44**

TABLE B.10: AIC v.s. TIC $N = 500,\ \tau = 0.5$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	***100**	0	0	0	0	0	0
clayton	TIC	***100**	0	0	0	0	0	0
frank	AIC	0	*99.60	0	0	0	0.36	0.04
frank	TIC	0	***99.62**	0	0	0	0.34	0.04
galambos	AIC	0	0	***43.00**	30.00	26.72	0.12	0.16
galambos	TIC	0	0	*42.74	32.82	24.16	0.12	0.16
gumbel	AIC	0	0	28.20	*63.92	7.64	0.06	0.18
gumbel	TIC	0	0	26.22	***66.78**	6.74	0.06	0.20
huslerReiss	AIC	0	0	16.86	2.14	***80.90**	0.10	0
huslerReiss	TIC	0	0	18.76	2.78	*78.36	0.10	0
normal	AIC	0	0.24	0.08	0.04	0.08	***98.98**	0.58
normal	TIC	0	0.26	0.10	0.04	0.08	*98.84	0.68
t	AIC	0	0	0.02	0.18	0	0.70	*99.10
t	TIC	0	0	0.02	0.14	0	0.64	***99.20**

TABLE B.11: AIC v.s. TIC $N = 500$, $\tau = 0.75$ — based on $R = 5000$ replicates.

d.cop	IC	clayton	frank	galambos	gumbel	huslerReiss	normal	t
clayton	AIC	∗**100**	0	0	0	0	0	0
clayton	TIC	∗**100**	0	0	0	0	0	0
frank	AIC	0	*100	0	0	0	0	0
frank	TIC	0	∗**99.94**	0	0	0.06	0	0
galambos	AIC	0	0	∗**48.58**	36.78	14.64	0	0
galambos	TIC	0	0	*47.08	40.20	12.72	0	0
gumbel	AIC	0	0	36.90	*57.20	5.88	0	0.02
gumbel	TIC	0	0	34.48	∗**60.46**	5.04	0	0.02
huslerReiss	AIC	0	0	9.26	0.46	∗**90.28**	0	0
huslerReiss	TIC	0	0	11.22	0.60	*88.18	0	0
normal	AIC	0	0	0.02	0	0	∗**99.62**	0.36
normal	TIC	0	0	0.02	0	0	*99.46	0.52
t	AIC	0	0	0	0.02	0	0.54	*99.44
t	TIC	0	0	0	0.02	0.02	0.48	∗**99.48**

TABLE B.12: AIC v.s. TIC — hit rate for selection — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**82.24**	**98.42**	∗**99.96**
100	TIC	81.44	98.20	∗**99.96**
250	AIC	**95.94**	∗**100**	∗**100**
250	TIC	95.74	∗**100**	∗**100**
500	AIC	**99.62**	∗**100**	∗**100**
500	TIC	**99.62**	∗**100**	∗**100**

TABLE B.13: AIC v.s. TIC — confidence in conclusion — copula = `clayton`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	71.72	96.07	∗**99.82**
100	TIC	**72.10**	**96.19**	∗**99.82**
250	AIC	92.05	**99.96**	∗**100**
250	TIC	**92.14**	**99.96**	∗**100**
500	AIC	**98.77**	∗**100**	∗**100**
500	TIC	98.73	∗**100**	∗**100**

TABLE B.14: AIC v.s. TIC — hit rate for selection — copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**56.62**	83.14	*97.72
100	TIC	55.46	**83.30**	∗**97.92**
250	AIC	75.82	96.60	∗**100**
250	TIC	**75.86**	**96.72**	*99.98
500	AIC	88.86	99.60	∗**100**
500	TIC	**89.14**	**99.62**	*99.94

TABLE B.15: AIC v.s. TIC — confidence in conclusion — copula = `frank`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	51.14	**83.12**	∗**98.23**
100	TIC	**51.65**	82.68	*98.17
250	AIC	**75.83**	**97.53**	∗**99.98**
250	TIC	75.51	97.26	*99.98
500	AIC	**89.72**	**99.75**	∗**100**
500	TIC	89.33	99.73	∗**100**

TABLE B.16: AIC v.s. TIC — hit rate for selection — copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**7.70**	**14.36**	∗**20.04**
100	TIC	7.54	13.58	*18.98
250	AIC	16.72	**29.18**	∗**34.76**
250	TIC	**16.92**	28.82	*33.94
500	AIC	**28.46**	**43.00**	∗**48.58**
500	TIC	28.26	42.74	*47.08

TABLE B.17: AIC v.s. TIC — confidence in conclusion — copula = `galambos`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	30.24	**35.42**	∗**38.74**
100	TIC	**30.45**	35.07	*37.12
250	AIC	35.39	**41.21**	∗**42.51**
250	TIC	**35.80**	40.86	*41.91
500	AIC	41.30	**48.77**	∗**51.26**
500	TIC	**41.31**	48.65	*50.73

TABLE B.18: AIC v.s. TIC
— hit rate for selection —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	32.84	42.64	*45.12
100	TIC	**34.78**	**46.70**	∗**50.90**
250	AIC	49.78	*55.30	51.56
250	TIC	**52.78**	∗**59.58**	**56.42**
500	AIC	62.82	*63.92	57.20
500	TIC	**65.52**	∗**66.78**	**60.46**

TABLE B.19: AIC v.s. TIC
— confidence in conclusion —
copula = `gumbel`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**35.27**	**45.58**	∗**48.94**
100	TIC	34.64	43.44	*46.93
250	AIC	**47.84**	∗**56.84**	**54.96**
250	TIC	46.89	*55.61	54.28
500	AIC	**59.63**	∗**66.38**	**60.55**
500	TIC	58.22	*65.11	59.69

TABLE B.20: AIC v.s. TIC
— hit rate for selection —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**51.70**	**66.08**	∗**76.24**
100	TIC	48.46	61.94	*70.88
250	AIC	**60.14**	**74.68**	∗**83.32**
250	TIC	56.54	70.98	*79.78
500	AIC	**67.18**	**80.90**	∗**90.28**
500	TIC	64.40	78.36	*88.18

TABLE B.21: AIC v.s. TIC
— confidence in conclusion —
copula = `huslerReiss`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	35.31	43.51	*50.97
100	TIC	**35.38**	**44.79**	∗**53.22**
250	AIC	44.49	57.37	*67.16
250	TIC	**45.03**	**59.13**	∗**69.59**
500	AIC	54.67	70.14	*81.48
500	TIC	**55.47**	**71.66**	∗**83.17**

TABLE B.22: AIC v.s. TIC
— hit rate for selection —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**41.48**	**72.94**	∗**87.76**
100	TIC	40.64	69.80	*85.06
250	AIC	**66.42**	**92.74**	∗**97.26**
250	TIC	65.08	91.74	*96.54
500	AIC	**85.30**	**98.98**	∗**99.62**
500	TIC	84.48	98.84	*99.46

TABLE B.23: AIC v.s. TIC
— confidence in conclusion —
copula = `normal`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	39.58	63.40	*77.85
100	TIC	**40.37**	**65.36**	∗**79.94**
250	AIC	61.18	89.72	*95.74
250	TIC	**61.71**	**90.70**	∗**96.55**
500	AIC	82.86	98.66	*99.46
500	TIC	**83.29**	**98.74**	∗**99.51**

TABLE B.24: AIC v.s. TIC
— hit rate for selection —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	63.68	70.68	*77.22
100	TIC	**67.30**	**73.98**	∗**80.46**
250	AIC	87.60	92.76	*95.42
250	TIC	**88.96**	**93.90**	∗**96.32**
500	AIC	97.14	99.10	*99.44
500	TIC	**97.44**	**99.20**	∗**99.48**

TABLE B.25: AIC v.s. TIC
— confidence in conclusion —
copula = `t`

N	IC	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$
100	AIC	**60.69**	**73.18**	∗**81.97**
100	TIC	57.54	69.84	*79.66
250	AIC	**86.90**	**93.39**	∗**96.59**
250	TIC	84.85	92.33	*95.87
500	AIC	**97.10**	**99.04**	∗**99.61**
500	TIC	96.57	98.92	*99.46

Appendix C

Some comments on the code

This appendix gives some comments on the code that was used to test the small-sample performance of the cross-validation copula information criterion, xv-CIC, for the combinations of copula models, values of Kendall's τ and sample sizes given in section 4.1.

The explicit code that was used will not be included in this appendix, but instead there will be some remarks with regard to the challenges encountered when executing such a simulation task - with emphasize on the `R`-program, R Development Core Team [29], and some some packages that the author would like to recommend.

Simulations and fitting of models: The main package that everything rests upon is the `copula`-package. This package contains functions like `rCopula` that creates simulated data from a given *copula and parameter* combination, the function `pobs` that create the corresponding pseudo-observations and the function `fitCopula`, which computes ml- or mpl-estimates of the parameters and the corresponding maximum of the (pseudo)log-likelihood.

Computation of bias-correcting terms: The three bias-correcting terms of the xv-CIC, i.e. $\widehat{p}_n$, $\widehat{q}_n$ and $\widehat{r}_n$ from eq. (3.97), can be computed numerically by the help of the function `genD` from the package `numDeriv`. However, with regard to the time used on these computations, it is highly recommendable to use another approach for the Archimedean copula models. Remember from page 14 that the Archimedean copulas are given by generators $\psi(t) : [0, \infty] \to [0, 1]$, and the point of interest is that the `R`-function `D` symbolically can compute the required derivatives of the corresponding pdfs.

Note that we do not need to bother with the specification of the required generators, since this information can be found in the copula-objects we are considering. When the desired derivatives have been created and stored e.g. as a `call`, it is only necessary to specify the parameter-value $\boldsymbol{\theta}$ and the coordinates of the point $\boldsymbol{u}$ – and then use the

function `eval` to get our value. A sanity check of the code can easily be obtained by comparing with the result of the `genD` function.

Bookkeeping: When we want to apply the same program on many combinations, like in section 4.1 where we have the combination of 7 data-generating copula-models and 7 proposed copula models, 4 sample-sizes and 3 parameter settings, altogether a total of 588 combinations, we really need some kind of automatic bookkeeping.

With regard to this, the `R`-package `plyr` is a package that this author would like to recommend. The approach used in the code for this thesis primarily used the function `aaply` to produce the multi-dimensional arrays containing all the desired information. The main points of interest to mention with regard to this is that we by using e.g. `aaply` can avoid a lot of the inefficient `for` loops in `R`.

In most of the applications in this thesis, the `aaply`-function worked upon a matrix whose rows were used as labels for the extraction of data from one or two previously computed multi-dimensional arrays – and these values where then used in functions that gave the entries in a new multi-dimensional array.

A nice feature of the `aaply`-function is that the dimension-names on the resulting array will be inherited from the array it works upon, and in addition, if the function that is used delivers a result in the form of a named vector/matrix/array, then these names too will be inherited to the final array.

The argument array for `aaply` was made by the help of `expand.grid`, which creates a matrix containing all the possible combinations of the content of the vectors or lists it is given as its arguments. A note of warning with regard to the use of `expand.grid` is that one of the standard defaults of this functions can mess things up if we want to extract information from *two* different arrays. To avoid any potential errors due to the chance that the two arrays under consideration does not have its content labeled in the exact same way, it is paramount that we in the argument of `expand.grid` use the setting `stringsAsFactors = FALSE`.

Wrapping of errors: Even though the code is without errors, it can sometimes happen that the result of a computation still turns out to be a `NA` or `NaN`, cf. the discussion around table 4.31. If this happens within a loop like the one performed by the `aaply`-function, then the loop will terminate.

To prevent this undesirable consequence, it is necessary to wrap our functions within some protecting layer, that will "hide" such errors from the loop, and instead return some default value that later on can be used to count the amount of such problems. The

two wrapper-functions `try` and `failwith` have been used to prevent the code in this thesis from terminating due to such errors.

Limitations on memory size, partitioning of problem: A detail that should be considered from the start, is how much memory that will be required in order to store the information that `R` is supposed to work upon. If the amount of required memory is too large, the problem must be partitioned into smaller chunks. For the case of this thesis, the solution was to divide the simulations and estimations into four separate cases according to the sample-sizes, and each of these cases then had 25 chunks of 200 replicates each. This approach ensured that the memory could handle the separate tasks, but the cost was a need for a lot of file-handling in the course of the code.

Extracting relevant data: The two functions `melt` and `cast` from the `reshape` package was used when it was time to extract information from the arrays produced by the `aaply`-function. In addition, the function `abind` from the package with the same name, was found to be very useful when the results from all the different chunks were to be pasted together. These three functions were used to create the content of all the tables in this thesis in a simple and uniform way.

Presenting the data The tables in this thesis were converted from arrays in `R` to tables in LaTeX by the help of the `Sweave`-program introduced in Leisch [30], and the `R`-function `xtable`. For those unfamiliar with `Sweave`, the important thing to know is that this program take cares of the boring task of writing the LaTeX-code for the tables, and in a similar fashion it simplifies the task of including graphical elements in the document.

Although `Sweave` tremendously simplifies the writing of a document in LaTeX, it does have some quirks that can take some time to get used to. It might thus be advisable to learn the more "mature" version `knitr`, introduced in Xie [31], instead of `Sweave`, but that advise is subject to the disclaimer that the author still need to learn `knitr` first.

Bibliography

[1] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis* -. Pearson Prentice Hall, New Jersey, 6th ed. edition, 2007. ISBN 978-0-131-87715-3.

[2] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

[3] Roger B. Nelsen. *An Introduction to Copulas* -. Springer, Berlin, Heidelberg, 2nd edition, 2006. ISBN 978-0-387-28659-4.

[4] Harry Joe. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997. ISBN 0-412-07331-5.

[5] Kjersti Aas and Daniel Berg. Models for construction of multivariate dependence - a comparison study. *European Journal of Finance*, 15(7-8):639–659, 2009. URL http://EconPapers.repec.org/RePEc:taf:eurjfi:v:15:y:2009:i:7-8:p:639-659.

[6] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44: 182–198, 2009.

[7] Eike Christian Brechmann. Truncated and simplified regular vines and their applications. Master's thesis, Technische Universität München, 2010. URL http://mediatum.ub.tum.de/doc/1079285/1079285.pdf.

[8] Harry Joe. *Distributions with Fixed Marginals and Related Topics*, chapter Families of m-variate distributions with given margins and $m(m-1)/2$ dependence parameters. IMS, Hayward, CA, 1996.

[9] Ingrid Hobæk Haff. *Pair-copula constructions - an inferential perspective*. PhD thesis, University of Oslo, Faculty of Mathematics and Natural Sciences, September 2012.

[10] Ingrid Hobæk Haff, Kjersti Aas, and Arnoldo Frigessi. On the simplified pair-copula construction - simply useful or too simplistic? *J. Multivariate Analysis*, 101 (5):1296–1310, 2010.

[11] Tim Bedford and Roger M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268, August 2001. ISSN 1012-2443. doi: 10.1023/A:1016725902970. URL http://dx.doi.org/10.1023/A:1016725902970.

[12] Tim Bedford and Roger M. Cooke. Vines - a new graphical model for dependent random variables. *Ann. Statist*, 30:1031–1068, 2002.

[13] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on automatic control*, 19:716–723, 1974.

[14] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):pp. 179–195, 1975. ISSN 00390526. URL http://www.jstor.org/stable/2987782.

[15] Steffen Grønneberg and Nils Lid Hjort. The copula information criterion. Technical Report 7, Department of Matehmatics, University of Oslo, 2008.

[16] Steffen Grønneberg. *Some applications of stochastic process techniques to statistics.* PhD thesis, University of Oslo, Faculty of Mathematics and Natural Sciences, November 2011. URL http://urn.nb.no/URN:NBN:no-30622.

[17] Jan Marius Hofert. *Sampling Nested Archimedean Copulas with Applications to CDO Pricing.* PhD thesis, Ulm Univeristy, Institute of Number Theory and Probability Theory, January 2010. URL http://vts.uni-ulm.de/docs/2010/7242/vts_7242_10223.pdf.

[18] Harry Joe and James J. Xu. The estimation method of inference functions for margins for multivariate models. Technical report, Department of Statistics, University of British Columbia, 1996.

[19] Jun Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007. URL http://www.jstatsoft.org/v21/i04/.

[20] Gerda Claeskens and Nils Lid Hjort. *Model Selection And Model Averaging* -. Cambridge University Press, Cambridge, illustrated edition edition, 2008. ISBN 978-0-521-85225-8.

[21] Dorota Kurowicka and Harry Joe. *Dependence Modeling - Vine Copula Handbook.* World Scientific, Singapore, 2010. ISBN 978-9-814-29988-6.

[22] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

[23] K. Takeuchi. Distribution of informational statistics and a criterion of model ïňĄting. *Suri-Kagaku (Mathematical Sciences)*, 153:12–18, 1976. In Japanese.

[24] Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. Technical report, Department of Mathematic, University of Oslo, 1993. URL http://www.stat.yale.edu/~pollard/Papers/convex.pdf.

[25] Frederik Hendrik Ruymgaart. Asymptotic normality of nonparametric tests for independence. *Annals of Statistics*, 2(5):892–910, 1974. URL http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176342812.

[26] C. Genest, K. Ghoudi, and L. P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82 (3):543–552, September 1995. URL http://biomet.oxfordjournals.org/cgi/content/abstract/82/3/543.

[27] Alexander J. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN BULLETIN*, 27(1):117–137, 1999. URL http://www.casact.org/library/astin/vol27no1/117.pdf.

[28] Geir Drage Berentsen, Bård Støve, Dag Tjøstheim, and Tommy Nordbø. Recognizing and visualizing copulas: an approach using local gaussian approximation. University of Bergen, Department of Mathematics, 2013. URL http://folk.uib.no/gbe062/local-gaussian-correlation/Recognizing-and-Visualizing-copulas-an%20approach-using-local-gaussian-correlation.pdf.

[29] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[30] Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. URL http://www.stat.uni-muenchen.de/~leisch/Sweave. ISBN 3-7908-1517-9.

[31] Yihui Xie. *knitr: A general-purpose package for dynamic report generation in R*, 2013. URL http://yihui.name/knitr/. R package version 1.2.