Annales Univ. Sci. Budapest., Sect. Comp. 43 (2014) 3–19

COPULA FITTING TO AUTOCORRELATED DATA, WITH APPLICATIONS TO WIND SPEED MODELLING

Pál Rakonczai (Budapest, Hungary)

László Varga (Budapest, Hungary)

András Zempléni (Budapest, Hungary)

Dedicated to András Benczúr on the occasion of his 70th birthday

Communicated by László Lakatos

(Received June 1, 2014; accepted July 1, 2014)

Abstract. Copulas became a popular tool in multivariate modelling, with several fitting methods readily available. In an earlier paper [19] we focused on the goodness of fit for copulas. These tests are based on independent samples. To assume complete independence for time series data is usually too optimistic. Now, as in real applications time dependence is a common feature, we turn to the investigation of the effect of this phenomenon to the proposed test-statistics, especially to the Kendall's process approach of [5] and [6]. The block bootstrap methodology is used for defining the effective sample size for time dependent bivariate observations. The critical values are then computed by simulation from independent samples with the adjusted size, determined by these bootstrap methods. The methods are illustrated by 2-dimensional modelling of the weekly maxima of 50-years observations of wind data for German sites. We also propose methods for assessing the reliability of the prediction regions, introduced in [18].

Key words and phrases: block bootstrap, copula, effective sample size, goodness of fit, Kendall's process, wind speed maxima

2010 Mathematics Subject Classification: 62P12

Pál Rakonczai's research was supported by an OTKA mobility grant (OTKA registration number: MB08A 84576 PKR registration number: HUMAN-MB08-1-2011-0007).

https://doi.org/10.71352/ac.43.003

1. Introduction

In the last decade the question of multivariate modelling of high-dimensional data has become also tractable, mainly due to the vast number of recorded data and the powerful computing equipment readily available. However, most of the methodology has been developed for the case of independent (multivariate) observations. In this paper, we focus on the effect of the serial dependence, naturally arising in many time series data. In an earlier paper, [19] we investigated the possibilities for checking the validity of the copula models. Copulas are simple yet powerful tools, which ensure the separation of marginal modelling and the dependence structure. They have been re-invented in the 1990s and their use has been expanded rapidly since then. One natural area of their applications is in the analysis of environmental data, where they are often used to model the dependence structure of extreme events at different sites. Here we apply the methodology to weekly maxima of wind gusts at 2 different German sites. The data spans about 50 years, from 1957 to 2007.

In Section 2, we first briefly review the needed elements of copula theory and present the notations. In Section 3, we summarize the most recent approaches for measuring the goodness of fit (GoF) for copula models, including the modifications suggested in [19]. Our proposed weighted GoF test is based on the Kendall's transform of the joint distribution (see [5] and [6]), which reduces the multivariate problem to one dimension. Section 4 is devoted to the bootstrap resampling method, including the block bootstrap approach, which is suitable for the case of serial dependence. In Section 5 we model the bivariate dependence structure of the wind data set. We show the effect of the serial dependence on the model selection. The prediction regions are useful tools in visualisation of the estimated model (see [18]). We use the block bootstrap methodology of Section 4 to investigate the uncertainty in the model estimation. The conclusion summarizes our findings and gives ideas for future research.

2. Copula concepts

Consider a random vector $\mathbf{X} = (X_1, ..., X_d)$ with joint distribution function $\mathbf{H}$ and margins $F_1(x_1), ..., F_d(x_d)$. Due to Sklar's theorem to any continuous d-variate distribution function $\mathbf{H}$, with univariate margins F_i there exists a copula $\mathbf{C}$, a distribution over the d-dimensional unit cube with uniform margins, such that

$$\mathbf{H}(x_1, ..., x_d) = \mathbf{C}(F_1(x_1), ..., F_d(x_d)). \tag{2.1}$$

Moreover, the copula $\mathbf{C}$ is unique if the marginal distributions are continuous. This construction allows capturing the dependence structure without specifying the marginal

distributions. In the recent literature various families of copulas have been introduced, for an overview and examples see e.g. the textbooks of [2], [13] and [14].

In this paper, we concentrate on some of the most widely used copulas from the Archimedean family. In our cases, these provided the best results when fitting to the data, but of course the proposed methodology can be carried out for other copulas (e.g. elliptical ones) as well.

2.1. Archimedean copulas

A broad class of copulas is called the Archimedean copula family. It is frequently used due to its very convenient structure. Let us consider a so-called copula generator function: $\phi(u) : [0,1] \to [0,\infty]$, which is continuous and strictly decreasing with $\phi(1) = 0$. Then a d-variate Archimedean copula function is

$$\mathbf{C}_\phi(\mathbf{u}) = \phi^{-1}\left(\sum_{i=1}^{d} \phi(u_i)\right).$$

In the course of the next sections we will present the Clayton and Gumbel copula family, but we emphasize that the presented methods can be adapted to any Archimedean models exactly in the same way. The Gumbel copula has the generator $\phi_\theta(u) = [-\log(u)]^\theta$, where $\theta \in [1, +\infty)$. Thus, the Gumbel d-copula function is given by

$$\mathbf{C}_{\text{Gumbel}}(\mathbf{u}) = \exp\left(-\left(\sum_{i=1}^{d}(-\log u_i)^\theta\right)^{\frac{1}{\theta}}\right).$$

We should notice that the Gumbel copula belongs to another important family, too. A copula $\mathbf{C}$ is called *extreme value copula* if $\mathbf{C}(u_1^t, ..., u_d^t) = \mathbf{C}^t(u_1, ..., u_d)$ for all $t > 0$. This family consists of copulas of multivariate extreme value distributions. The Gumbel copula is the only Archimedean copula, which is also included in the extreme value copula family and actually it coincides with the often used logistic dependence structure. The generator function of the Clayton copula (also known as Cook and Johnson's family) is given by $\phi_\theta(u) = u^{-\theta} - 1$, where $\theta > 0$. Thus, the Clayton d-copula function is the following

$$\mathbf{C}_{\text{Clayton}}(\mathbf{u}) = \left(\sum_{i=1}^{d} u_i^{-\theta} - d + 1\right)^{-\frac{1}{\theta}}.$$

For parameter estimation variants of the method of moments or pseudo-maximum likelihood estimation are the most widely accepted method in the above cases. For more details and for simulation methods see the Chapters 5-6 in [2]. For simulation

and parameter estimation in practice the `copula` package of the open source **R** language may be used. An illustration of the described copula models is shown by Figure 1. Here we can see the scatter plots of 2 dimensional simulations with the same sample size $n = 2000$ and given strength of association (Kendall's $\tau = 0.47$).

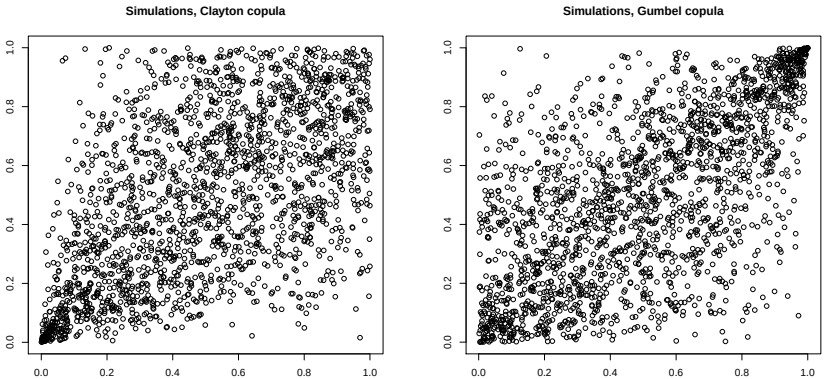

Figure 1. *Left panel*: simulations from the Clayton family ($\theta = 1.25$). *Right panel*: simulations from the Gumbel family ($\theta = 1.9$). The Kendall's τ is 0.47 in both cases.

2.2. Statistical inference from copula models

After fitting a copula model there is an important question how our choice influences the joint distribution. To tackle this, one has to transform back the results to the original scale. As in most cases the most important questions are the quantile estimators, a suitable parametric model may be preferable (especially in case of high quantiles). In our case we assumed generalized extreme value (GEV) distributed margins having cdf as

$$F(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}, \tag{2.2}$$

where $1+\xi\frac{x-\mu}{\sigma} > 0$. $\mu \in \mathbb{R}$ is called the location parameter, $\sigma > 0$ the scale parameter and $\xi \in \mathbb{R}$ the shape parameter. (For instance Gumbel copula and GEV margins together are equivalent to the well-known bivariate logistic extreme value distribution.) Instead of using the usual quantile curves obtained from the bivariate distribution function, we propose constructing compact quantile-like regions by integrating the bivariate density, as they are easily understandable graphical representations of the

model. The bivariate density, based on the formula (2.1) can be written as

$$h(x_1, x_2) = c(F(x_1), F(x_2)) f_1(x_1) f_2(x_2), \tag{2.3}$$

where $f_1(x) = F_1'(x)$ and $f_2(x) = F_2'(x)$ are the marginal densities and $c(u, v) = \frac{\partial^2}{\partial u \partial v} C(u, v)$. Let

$$\hat{\mathcal{R}}(u) \quad = \quad \{(x, y) : \hat{h}(x, y) \geq u\},$$

where $\hat{h}$ is an estimator of the bivariate density h of (2.3). Then, following the notation in [8], the **prediction region** for a given probability γ is defined to be $R(\hat{u}_\gamma)$, where $u = \hat{u}_\gamma$ is the solution to the equation

$$\int\limits_{\hat{\mathcal{R}}(u)} \hat{h}(x, y) dx dy = \gamma.$$

Prediction regions provide a rather intuitive method for visualizing the effect of the copula choice on the original distribution.

3. Goodness-of-Fit tests

After estimating the model parameters one must be able to check the fit of the results. Formally we intend to test the hypothesis

$$\mathcal{H}_0 : \mathbf{C} \in \mathcal{C_0} = \{\mathbf{C}_\theta, \theta \in \mathbf{\Theta}\}, \tag{3.1}$$

that the dependence structure of the copula arises from a specific parametric family $\mathcal{C}_0$ of copulas. The most obvious way for testing GoF is to consider multidimensional χ^2 approaches, but in this case we need to discretize the data, losing valuable information. In order to avoid its use, dimension reducing methods can be utilized. As usual in this context, we consider the F_j marginal distributions as nuisance parameters and base all of the tests on ranks. Basically in a preliminary step we perform the probability integral transformation (PIT) for the observations mapping them into the d-dimensional unit cube as

$$\overbrace{\mathbf{X}_i = (X_{i1}, ..., X_{id})}^{\text{Observations}} \sim \mathbf{H} \longrightarrow_{\text{PIT}} \overbrace{\mathbf{U}_i = (U_{i1}, ..., U_{id})}^{\text{Pseudo-observations}} \sim \mathbf{C}, \text{ for } i = 1, ..., n.$$

The PIT is defined by $U_{ij} = \frac{n\hat{F}_j(X_{ij})}{n+1} = \frac{R_{ij}}{n+1}$, where $\hat{F}_j$ denotes the empirical distribution function of the jth margin, R_{ij} is the rank of X_{ij} among $X_{1j}, ..., X_{nj}$ and

$\frac{n}{n+1}$ is just a scaling factor avoiding possible problems at the boundary of $[0,1]^d$. Therefore the *pseudo-observations* $\mathbf{U}_1, ..., \mathbf{U}_n$ can be interpreted as a sample from the underlying copula $\mathbf{C}$. In the review paper [7] it is strongly emphasized that the pseudo-observations are of course not really mutually independent and their components are only approximately independent, so any construction of GoF tests should take this fact into account, otherwise the testing procedures fail to hold their nominal level.

One of our main aims in this paper is to give methods for investigating the effect of (even slight or stronger) autocorrelation in the marginal components on the GoF-tests. We have chosen Cramér-von Mises type tests as they are generally proven to be one of the most powerful GoF methods formulated (not denoting the dependence on the parameters) as

$$T = n \int_{-\infty}^{\infty} (\hat{F}(x) - F(x))^2 \Phi(x) dF(x),$$

where $\hat{F}$ is the empirical cdf, F is the cdf which is to be fitted and $\Phi(x)$ is a weight function. In the simplest case, when $\Phi(x) = 1$ we get the Cramér-von Mises statistics, or if the focus is on the tails we may set the weight function as $\Phi(x) = \frac{1}{F(x)(1-F(x))}$, leading to the Anderson-Darling statistics

$$T_{AD} = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dF(x). \tag{3.2}$$

In many cases when only one of the tails is important (usually maximum for environmental or insurance loss data), the $\Phi(x) = \frac{1}{1-F(x)}$ weight function is suggested with $\Phi(x) = \frac{1}{F(x)}$ if the minima are in the focus of attention. The advantage of using these weights in comparison to standard Anderson-Darling in (3.2) is that the sensitivity is concentrated to discrepancies at the relevant tail of the distribution: see Zempléni's test in [11]. The computation of these statistics is straightforward. We come back to the critical value estimation in Section 4.

3.1. Kendall's transform

An important, widely used class of multivariate GoF statistics can be based on the Kendall's transform, which is the distribution function of the probability integral transformation of the joint distribution

$$\mathcal{K}(\theta, t) = P(C_\theta(F_1(X_1), ..., F_d(X_d)) \leq t) = P(C_\theta(U_1, ..., U_d) \leq t). \tag{3.3}$$

In the case of the Archimedean copula family, (3.3) can be computed as

$$\mathcal{K}(\theta, t) = t + \sum_{i=1}^{d-1} \frac{(-1)^i}{i!} \left[\phi_\theta(t)^i\right] f_i(\theta, t),$$

where $f_i(\theta, t) = \frac{d^i}{dx^i} \phi_\theta^{-1}(x)|_{x=\phi_\theta(t)}$. Note that actually $f_{i+1}(\theta, t) = f_1(\theta, t) \frac{\partial}{\partial t} f_i(\theta, t)$, $i \in \{1, ..., d-1\}$. The $\mathcal{K}$ function defined this way is invariant on the marginal distributions, hence it depends only on the copula of $\mathbf{X}$.

The empirical version of $\mathcal{K}$ can be computed by the rank based pseudo-observations as

$$\mathcal{K}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(E_{in} \leq t), \qquad t \in [0, 1],$$

where

$$E_{in} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}(U_{j1} \leq U_{i1}, ..., U_{jd} \leq U_{id}).$$

For illustrations see Figure 2.

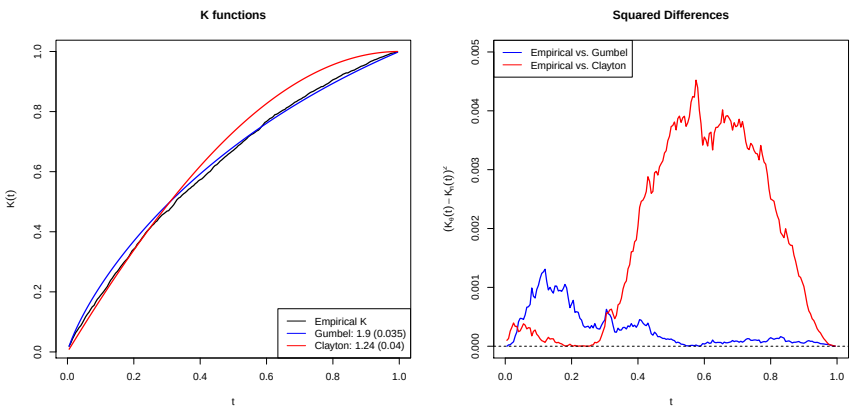

Figure 2. Comparison between parametric $\mathcal{K}_\theta$ of fitted copula families and the empirical version $\mathcal{K}_n$ for the wind data.

Known tests for checking the match of the theoretical and empirical version of the Kendall's transform $\mathcal{K}$ use continuous functionals of Kendall's process

$$\kappa_n(t) = \sqrt{n}(\mathcal{K}(\theta_n, t) - \mathcal{K}_n(t))$$

having favorable asymptotic properties. There are two different kind of approaches

investigated in [4], Cramér-von Mises type

$$S_n = \int_0^1 (\kappa_n(t))^2 dt$$

and Kolmogorov-Smirnov type

$$T_n = \sup_{0 \le t \le 1} |\kappa_n(t)|$$

statistics. As the second approach is generally less powerful in detecting discrepancies near the tails, we based our inference on the test statistics summarized by Table 1, where $(t_i)_{i=1}^m$ is an appropriately fine division of the interval $(0, 1)$.

Table 1. Numerically approximated test statistics.

Focused Regions	**Test Statistics**
Global	$K_1 = \frac{1}{m} \sum_{t_i \in [\varepsilon, 1-\varepsilon]} (\mathcal{K}(\theta_n, t_i) - \mathcal{K}_n(t_i))^2$
Upper Tail	$K_2 = \frac{1}{m} \sum_{t_i \in [\varepsilon, 1-\varepsilon]} \frac{(\mathcal{K}(\theta_n, t_i) - \mathcal{K}_n(t_i))^2}{1-\mathcal{K}(\theta_n, t_i)}$
Lower Tail	$K_3 = \frac{1}{m} \sum_{t_i \in [\varepsilon, 1-\varepsilon]} \frac{(\mathcal{K}(\theta_n, t_i) - \mathcal{K}_n(t_i))^2}{\mathcal{K}(\theta_n, t_i)}$
Lower and Upper Tail	$K_4 = \frac{1}{m} \sum_{t_i \in [\varepsilon, 1-\varepsilon]} \frac{(\mathcal{K}(\theta_n, t_i) - \mathcal{K}_n(t_i))^2}{\mathcal{K}(\theta_n, t_i)(1-\mathcal{K}(\theta_n, t_i))}$

As the limit distribution of the above statistics is not distribution-free, a simulation algorithm is needed to get critical values. The algorithm can be performed as follows:

1. Simulate a sample from the copula model C_θ under the null-hypothesis.

2. Re-estimate $\hat{\theta}$ from the simulation.

3. Calculate the test statistics.

Finally one should repeat the above steps as many times as needed to an accurate estimation of the p values, which can be compared to the computed test statistics. As this procedure may be time consuming (especially in higher dimensions), it is worth mentioning that in the paper [10] a quicker procedure was proposed, which is based on a new bootstrap approach.

4. Bootstrap methods

The bootstrap is a computer-intensive, usually non-parametric or semi-parametric statistical method to estimate the distribution of a statistics of interest. The basic bootstrap concept was introduced by Efron in 1979 (see [3]) and has become a popular tool in solving many statistical problems.

4.1. The bootstrap principle

Let X_1,X_2,... be i.i.d. random variables with unknown common distribution F. Suppose that we have a random sample $\mathbf{X}_n = \{X_1, \ldots, X_n\}$ and let $T_n = t_n(\mathbf{X}_n; F)$ be a statistics of interest. Let G_n denote the sampling distribution of T_n. Our main purpose is to approximate the unknown distribution of T_n or its function of interest, for example the standard error.

The (mostly referred as i.i.d.) basic bootstrap method is the following. For given $\mathbf{X}_n$, we draw a simple random sample $\mathbf{X}_m^* = \{X_1^*, \ldots, X_m^*\}$ of size m (usually $m \approx n$) with replacement from $\mathbf{X}_n$. Therefore, the common distribution of the X_i^* 's is given by the empirical distribution $\hat{F}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, where δ_z is the probability measure having unit mass at z. In the next step, we define the bootstrap alterego of T_n: $T_{m,n}^* = t_m(\mathbf{X}_m^*; \hat{F}_n)$. By repeating this procedure, we can approximate the unknown distribution G_n by its bootstrap counterpart G_n^*.

4.2. Serial dependence

Up to now, we have considered independent, identically distributed observations for copula estimation or in the simulation of test statistics. However, in realistic cases there is a serial correlation between the neighbouring observations. This phenomena is widely investigated in time series analysis, and has got substantial attention in the field of extreme value modelling. As a new utilization of copulas, Rakonczai et al. [17] have recently published a paper on the so-called autocopulas, as a powerful tool in investigating the serial dependence in univariate time series. The approaches above are different from our case, where we are interested in the effect of serial dependence on the GoF tests and on copula modelling in general.

If we have dependent data, one of the most commonly used methods is the so-called block bootstrap, see [12] for details. In our work, we use the circular block bootstrap (CBB) which can be defined as follows. First, we wrap the data $X_1, \ldots, X_n$ around a circle, i.e., define the series $Y_t = X_{t_{\mathrm{mod(n)}}}$ ($t \in \mathbb{N}$), where $\mathrm{mod(n)}$ denotes "modulo n". For some m, let $i_1, \ldots, i_m$ be a uniform sample from the set $\{1, 2, ..., n\}$.

After that, for a specific block size (or block length) b, we construct $n' = m \cdot b$ ($n' \approx n$) pseudo-data:

$$Y^*_{(k-1)b+j} = Y_{i_k+j-1}, \qquad \text{where } j = 1, \ldots, b \quad \text{and } k = 1, \ldots, m.$$

At last, we can calculate the function of interest, for example the bootstrap sample mean: $\overline{Y}^*_{n'} = \frac{Y^*_1 + \ldots + Y^*_{n'}}{n'}$.

Block length plays an important role in the process, and it is not trivial to determine its optimal value. For instance, [16] suggests an "automatic" block length selection algorithm (its correction was published in [15]) - but the practical applications of this method is far from obvious due to the needed parameter selection.

As practically sample size is the single parameter we can modify in the critical value simulation algorithm, we suggest to use the notion of effective sample size, denoted by n_e. It originates from the survey sampling literature (see [9]) and is used widely in different subjects, [20] being one of the recent applications, in the area of genetics. As an illustration of the method for determining the effective sample size, let $X_1, ..., X_n$ be univariate stationary observations with expectation μ and standard deviation σ and let our statistics of interest be the sample mean $\overline{X}$. In case of i.i.d. observations we get $\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$. However, if the data are serially dependent, then the variance of the sample mean is greater than $\frac{\sigma^2}{n}$. In its original definition, the effective sample size practically means the size of an **independent** sample from the distribution of $X_1, \ldots, X_n$, for which the variance of the sample mean coincides with its observed variance

$$\text{Var}(\overline{X}) = \frac{\sigma^2}{n_e}.$$

From an empirical sample, σ has to be estimated as well as the variance of the sample mean. For estimating the latter we use the circular block bootstrap method.

If the observations are multivariate, the definition of the effective sample size changes a bit. Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be multivariate stationary observations with expectation μ and covariance matrix Σ. Then n_e is the $n_e \leq n$ sample size for which

$$\text{tr}\left(\Sigma(\overline{\mathbf{X}})\right) = \frac{\text{tr}(\Sigma)}{n_e}, \tag{4.1}$$

where tr(.) denotes the trace operator. In our case we have another statistics in mind, the Kendall's transform from above. The distribution of (3.3) can be approximated by the circular block bootstrap as well. We use the bootstrap as defined above for this investigation. In the next Section we present an example as an illustration to the above methodology.

5. Application to wind speed maxima

We applied the described modelling procedures to a 2-dimensional wind speed dataset. The observations are daily maxima measured for the recent 50 years at Hamburg and Fehmarn, two locations in North Germany (from 1958 till 2007). First, the periodicity was removed by standard local regression: the minimum of the daily averages was 5.6 m/s and the maximum 6.8 m/s for both locations. Next, the trend was removed by a linear regression (which was weak, but significant – pointing downwards – in both of the cases). The effect of this decrease was around 0.8 m/s during the 50 years period for both locations. Finally, we calculated weekly maxima of the residuals in order to achieve a better fit of the marginal GEV model. The autocorrelation of the resulting approximately stationary sequence $\mathbf{X}_t$ is shown in Figure 3.

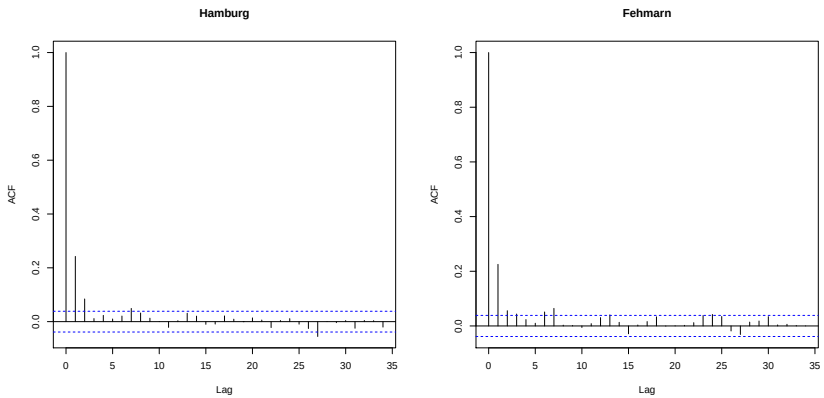

Figure 3. Autocorrelation functions of weekly maxima of wind speed after trend removal in Hamburg and Fehmarn.

The weekly maxima of the original data are shown on the left panel of Figure 4. The points in the upper right corner represent weeks when there was extremely high wind measured at both of the places. These kind of events could cause rather dangerous consequences (e.g. from insurance point of view) so we were more focused on whether the fitted model was appropriate in these upper regions. As it was mentioned in Section 3, for fitting the different models the pseudo-observations have been used. Therefore, in the first step of the analysis we transformed the data into the unit square with the help of the empirical univariate margins. The dependence structure among the pseudo-observations is shown on the right panel of Figure 4.

Next we fitted a 2 dimensional VAR(1) model to $\mathbf{X}_t$

$$\mathbf{X}_t = A\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t,$$

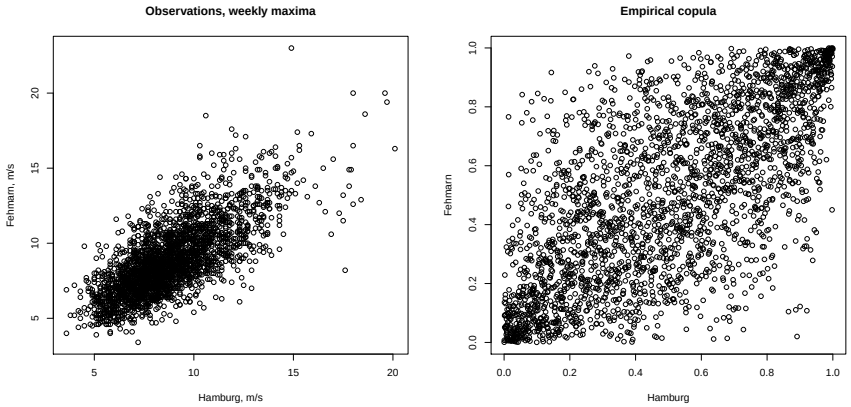

Figure 4. *Left panel*: Weekly maxima of wind speed data (m/s). *Right panel*: pseudo-observations after marginal transformations by the empirical distribution functions.

where A and $C = \Sigma(\varepsilon_t)$ are 2×2 parameter matrices. For details about the model, see [1] for example. The fit was reasonable, so this model was used to determine the optimal block length for the block bootstrap as follows. From the estimated parameters of the VAR(1) model we can calculate the trace of the covariance matrix of the sample mean, which will be denoted by $\mathrm{tr}(\Sigma_{\mathrm{VAR}}(\overline{\mathbf{X}}))$. After that we took for each block size in the range of 1 to 30, 1000 circular block bootstrap samples and estimated the trace of the covariance matrix of the sample mean, denoted by $\mathrm{tr}(\Sigma^{*i}(\overline{\mathbf{X}}))$ $(i = 1, \ldots, 30)$. The values are given in Table 2. The estimated trace derived from the VAR model is 0.6086, so we can see from Table 2 that the closest value is for block size 8. The final results are represented in Table 3. The Σ in the fourth column is the covariance matrix of i.i.d. observations, simply estimated with the sample covariance matrix. The original sample size for weekly observations was 2580 therefore the effective sample size becomes 1571.

Table 2. The estimated trace of the covariance matrix of the bootstrapped sample mean for different block sizes.

Block size (i)	$\mathrm{tr}(\Sigma^{*i}(\overline{\mathbf{X}}))$	Block size (i)	$\mathrm{tr}(\Sigma^{*i}(\overline{\mathbf{X}}))$
3	0.5084	10	0.6268
4	0.5361	11	0.6355
5	0.5612	12	0.6481
6	0.5790	13	0.6593
7	0.5941	14	0.6647
8	0.6101	15	0.6742
9	0.6159	16	0.6727

Table 3. Optimal block length and sample size reduction for the pair Hamburg and Fehmarn.

Block length	$\mathrm{tr}(\Sigma^{*8}(\overline{\mathbf{X}}))$	$\mathrm{tr}(\Sigma_{\mathrm{VAR}}(\overline{\mathbf{X}}))$	$\frac{\mathrm{tr}(\Sigma)}{n}$	Reduction factor	Sample size	Effective sample size
8	0.6101	0.6086	0.3715	1.6422	2580	1571

We used the two copula models of Section 2 as candidates for our data. The parameters of the fitted copulas were $\theta = 1.904$ for the Gumbel, and $\theta = 1.247$ for the Clayton copula. See Figure 1 for simulations from these. (The models were fitted by the copula package of $\mathbf{R}$, using the widely proposed pseudo-ML estimators.) As the next step we investigated the GoF by the K-tests, as described in Section 3. The fitted $\mathcal{K}(\theta_n, t)$ functions are displayed in Figure 2 together with the empirical $\mathcal{K}_n(t)$ of the observations.

After performing the tests we found that our approach played an important role as the critical values have increased substantially (Table 4 and Table 5). While the Clayton copula was rejected for all investigated levels, the Gumbel model was accepted for $\alpha = 0.002$ when the sample size correction was taken into account (2nd and 4th columns in Table 5). This relatively poor fit is not unusual in data analysis for sample sizes over 1000, so we carried on the modelling with the relative best option, the Gumbel copula.

Table 4. Bootstrapped statistics with observed and critical values for Clayton copula. The original critical values were calculated with sample size 2580, the adjusted critical values with the effective sample size 1571.

		Global	Global (Adj.)	Upper Tail	Upper Tail (Adj.)
Observed statistics		0.00174	0.00174	0.02534	0.02534
Critical values	95%	0.00007	0.00012	0.00039	0.00065
	99%	0.00011	0.00017	0.00058	0.00093
	99.8%	0.00025	0.00034	0.00116	0.00166

We calculated the prediction regions to the nominal levels of 50%, 75%, 95% and 99%, as described in Section 2. These regions are depicted in Figure 5. We were also interested in the reliability of these estimations. The possible fluctuation of the regions were estimated based on bivariate block bootstrap samples. The block size was 8, because this was found as the optimal size. We used the following algorithm:

1. Resampling from the wind database by block bootstrap with given $b = 8$ block size.

Table 5. Bootstrapped statistics with observed and critical values for Gumbel copula. The original critical values were calculated with sample size 2580, the adjusted critical values with the effective sample size 1571.

		Global	Global (Adj.)	Upper Tail	Upper Tail (Adj.)
Observed statistics		0.00028	0.00028	0.00084	0.00084
Critical values	95%	0.00006	0.00010	0.00021	0.00036
	99%	0.00009	0.00014	0.00030	0.00051
	99.8%	0.00016	0.00032	0.00055	0.00089

2. Estimation of its marginal GEV parameters.

3. Estimation of the parameter of the Gumbel copula (by pseudo-ML) for the bootstrap sample.

4. Computation of the prediction regions based on the fitted model's density.

5. After completing steps 1-4 200 times, the upper and lower confidence bounds (inner and outer black curves on Figure 5) were computed as the 5% and 95% level curves of the bootstrapped regions.

The number of replicas shown on Figure 5 was chosen as low as 20 just to allow the comparison of the individual regions. The computation was quick enough to allow for several hundreds of runs in a reasonable time.

6. Conclusions

We can summarize our findings as follows.

The serial dependence has a substantial impact on the critical values of the GoF tests, and by the block bootstrap methodology we were able to estimate its effect. We do hope that the effective sample size, as an easily interpretable notion will find its place in the publications of the field. The theoretical background – in the flavour of similar consistency results for the bootstrap – is a conjecture at the moment, we hope to be able to prove and publish it soon. There is a straightforward generalization of the methods to the multivariate cases, although the curse of dimensionality will definitely hinder its applications in really high dimensions.

Another important use of the block bootstrap is in the estimation of the variability of the prediction regions. We plan to implement it as part of the `mgpd` package, within the framework of the bivariate threshold models, which is maintained by the

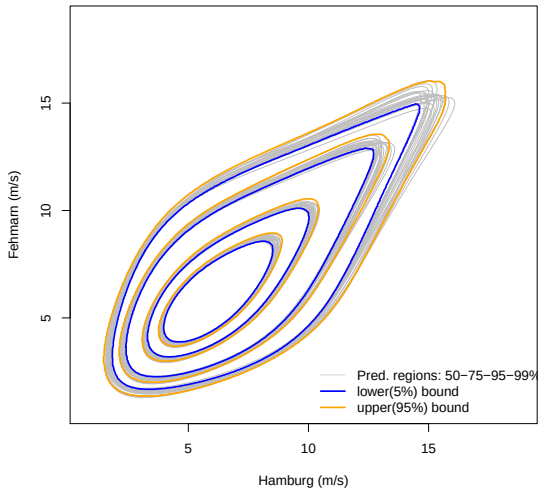

Figure 5. Prediction regions estimation by bootstrap.

first author of this paper. And we do hope that other relevant packages will soon provide similar options as they are indeed useful visualizations of the variability of the estimated regions.

References

[1] **Brockwell, P.J. and R.A. Davis,** *Time Series: Theory and Methods*, Springer, 2009.

[2] **Cherubini, U., E. Luciano and W. Vecchiato,** *Copula Methods in Finance*, John Wiley & Sons, 2004.

[3] **Efron, B.,** Bootstrap methods: Another look at the jackknife, *The Annals of Statistics,* (1979), 1–26.

[4] **Genest, C. and A.-C. Favre,** Everything you always wanted to know about copula modeling but were afraid to ask, *Journal of Hydrologic Engineering*, **12** (2007), 347–368.

[5] **Genest, C., J.-F. Quessy, and B. Rémillard,** Goodnes-of-fit procedures for copula models based on the probability integral transformation, *Scandinavian J. of Statistics*, **33** (2006), 337–366.

[6] **Genest, C. and B. Rémillard,** Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models, *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, **44** (2008), 1096–1127.

[7] **Genest, C., B. Rémillard and D. Beaudoin,** Goodness-of-fit tests for copulae: A review and a power study, *Insurance: Mathematics and Economics*, **44** (2009), 199–213.

[8] **Hall, P. and N. Tajvidi,** Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data, *Statistical Science*, (2000), 153–167.

[9] **Kish, L.,** *Survey Sampling*, John Wiley & Sons, New York, 1965.

[10] **Kojadinovic, I. and J. Yan,** A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems, *Statistics and Computing*, **21** (2011), 17–30.

[11] **Kotz, S. and S. Nadarajah,** *Extreme Value Distributions*, World Scientific, 2000.

[12] **Lahiri, S.N.,** *Resampling Methods for Dependent Data*, Springer, 2003.

[13] **McNeil, A.J., R. Frey and P. Embrechts,** *Quantitative Risk Management*, Princeton University Press, 2005.

[14] **Nelsen, R.B.,** *An Introduction to Copulas*, Springer, 2007.

[15] **Patton, A., D.N. Politis, and H. White,** Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White, *Econometric Reviews*, **28** (2009), 372–375.

[16] **Politis, D.N. and H. White,** Automatic block-length selection for the dependent bootstrap, *Econometric Reviews*, **23** (2004), 53–70.

[17] **Rakonczai, P., L. Márkus, and A. Zempléni,** Autocopulas: Investigating the interdependence structure of stationary time series, *Methodology and Computing in Applied Probability*, **14** (2012), 149–167.

[18] **Rakonczai, P. and N. Tajvidi,** On prediction of bivariate extremes, *Int. J. of Intelligent Technologies and Applied Statistics*, **3** (2010), 115–139.

[19] **Rakonczai, P. and A. Zempléni,** Copulae and goodness of fit tests, *Recent Advances in Stochastic Modeling and Data Analysis*, ed. C.H.Skiadas, World Scientific, 2007, 198–206.

[20] **Yang, Y.B., E.L. Remmers, C. Ogunwole, D. Kastner, P.K. Gregersen, and W.F. Li,** Effective sample size: Quick estimation of the effect of related samples in genetic case–control association analyses, *Computational Biology and Chemistry*, **35** (2011), 40–49.

Pál Rakonczai
Department of Probability Theory and Statistics,
Eötvös Loránd University
Budapest, Hungary
paulo@math.elte.hu

László Varga
Department of Probability Theory and Statistics,
Eötvös Loránd University
Budapest, Hungary
vargal4@math.elte.hu

András Zempléni
Department of Probability Theory and Statistics,
Eötvös Loránd University
Budapest, Hungary
zempleni@math.elte.hu