A SIMPLE GENERALISATION OF THE BLOCK BOOTSTRAP AND ITS APPLICATION FOR COPULA TESTS

László Varga[1] and András Zempléni[2]

Department of Probability Theory and Statistics, Eötvös Loránd University

[1]vargal4@cs.elte.hu, [2]zempleni@ludens.elte.hu

For personal consultation, the authors will be available on Friday.

Abstract

In the last decades the bootstrap methodology has become more and more widespread in different areas of statistical applications. In an earlier paper (Rakonczai et al. [2014]), we have emphasized the effective sample size for autocorrelated data. The simulations were based on the block bootstrap methodology. However, the discreteness of the usual block size did not allow for exact calculations. We propose a generalisation of the block bootstrap methodology, which overcomes this problem, by allowing any real number $b > 1$ as block size. We combine this approach with the VAR modelling for testing the homogeneity of copulas (for the test procedure see Rémillard and Scaillet [2009]) and apply it to a temperature data set of the Carpathian Basin.

Vector autoregression (VAR) processes

The time series $\{\mathbf{X}_t\}_{t\in\mathbb{Z}} = \{(X_{1,t}, X_{2,t})^T\}_{t\in\mathbb{Z}}$ is called a zero-mean two-dimensional VAR(p) process if

$$\mathbf{X}_t = A_1\mathbf{X}_{t-1} + A_2\mathbf{X}_{t-2} + \ldots + A_p\mathbf{X}_{t-p} + \varepsilon_t,$$

where $A_1, \ldots, A_p$ are 2×2 parameter matrices and the $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ independent innovation process is a two-dimensional white noise with $E(\varepsilon_t) = \mathbf{0} = (0,0)^T$ and $\mathrm{Cov}(\varepsilon_t) = C$ symmetric positive definite covariance matrix. The VAR(p) process is stationary if the roots of the $P(x) = \det(I_2 - A_1x - \ldots - A_px^p)$ characteristic polynomial lie outside the unit circle.

Any VAR(p) process can be rewritten as a VAR(1) process in the following way: $\mathbf{Y}_t = A\mathbf{Y}_{t-1} + \mathbf{e}_t$, where

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{X}_t \\ \mathbf{X}_{t-1} \\ \vdots \\ \mathbf{X}_{t-p+1} \end{pmatrix}, \quad \mathbf{e}_t = \begin{pmatrix} \varepsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \quad A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & I_2 & \mathbf{0} \end{pmatrix}.$$

This representation is more convenient in calculating the autocovariances.

Let us assume that $\mathbf{X}_t$ is stationary, we denote with $\Gamma_X(h) = E(\mathbf{X}_{1+h}\mathbf{X}_1^T)$ the autocovariance function of the process $\mathbf{X}_t$. $\Gamma_X(h)$ is a 2×2 matrix valued function, the symbols $\gamma_{i,j}(h)$ stand for its elements. We denote with $\Gamma_Y(h) = E(\mathbf{Y}_{1+h}\mathbf{Y}_1^T)$ the $2p \times 2p$ matrix valued autocovariance function of the process $\mathbf{Y}_t$. The covariance matrix of $\mathbf{Y}_t$ is $\Gamma_Y(0)$ which can be determined by solving the matrix equation $\Gamma_Y(0) - A\Gamma_Y(0)A^T = \mathrm{Cov}(\mathbf{e_t})$. It is easy to see that for $1 \leq h \in \mathbb{Z}$, the autocovariances can be calculated by $\Gamma_Y(h) = A^h\Gamma_Y(0)$. The powers of the matrix A can be computed using the spectral decomposition. Lastly, we need the autocovariance matrix of the original process, and by the construction, it is the upper left 2×2 submatrix of $\Gamma_Y(h)$.

In the applications we will use the covariance matrix of the sample mean. The following asymptotic result will be crucial: if $\sum_{h=-\infty}^{\infty} |\gamma_{i,i}(h)| < \infty$ for $i = 1, 2$, then

$$n \cdot \mathrm{tr}\left(\mathrm{Cov}(\overline{\mathbf{X}}_n)\right) \longrightarrow \sum_{i=1}^{2}\sum_{h=-\infty}^{\infty} \gamma_{i,i}(h) \quad \text{as } n \to \infty,$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

Homogeneity of copulas (1)

Let $\mathbf{X} = (X_1, \ldots, X_d)^T$ be a random vector with joint distribution function $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,\ldots,X_d}(x_1, \ldots, x_d)$ and marginal distribution functions $F_1(x_1) := F_{X_1}(x_1), \ldots, F_d(x_d) := F_{X_d}(x_d)$.

Sklar's theorem: there exists a copula $\mathbf{C}$, a distribution over the d-dimensional unit cube, with uniform margins, such that

$$F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) = \mathbf{C}\left(F_1(x_1), \ldots, F_d(x_d)\right).$$

Moreover the copula $\mathbf{C}$ is unique if the marginal distribution functions are continuous.

We focus on testing the homogeneity of copulas, motivated by the question whether the climate change has also an effect on the dependence between pairs of observations.

Let us suppose we have two independent samples of $\mathbb{R}^d$-valued vectors: $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$. We intend to test the hypothesis that the dependence structure of the two copulas has arisen from the same copula $\mathbf{C}_0$.

Homogeneity of copulas (2)

Rémillard and Scaillet [2009] have developed a method for this problem. Their approach is based on the empirical copula, defined for the first sample as

$$\mathbf{C}_{1,n}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{U}_i \leq \mathbf{u}).$$

where $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{U}_i$ denotes the d-dimensional vector of the rank based pseudo-observations: $\mathbf{U}_i = \mathbf{U}_{i,n} = (U_{i1,n}, \ldots, U_{id,n})$, where n refers to the size of the first sample and $U_{ij,n} = \frac{n}{n+1}F_j(X_{ij})$. Similarly, based on the pseudo-observations $\mathbf{V}_i$ of the second sample, we can define the empirical copula $\mathbf{C}_{2,m}(\mathbf{u})$.

The proposed test statistic is a Cramér-von Mises type statistic based on the empirical copula process, which can be written in the following form:

$$S_{n,m} = \left(\frac{1}{n} + \frac{1}{m}\right)^{-1} \cdot \left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\prod_{s=1}^{d}(1 - U_{is,n} \vee U_{js,n}) + \right.$$
$$+ \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\prod_{s=1}^{d}(1 - V_{is,m} \vee V_{js,m} -$$
$$\left. - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\prod_{s=1}^{d}(1 - (U_{is,n} \vee V_{js,m})\right].$$

As the limit distribution of the above statistic is not distribution-free, a simulation algorithm is needed to get critical values.

Generalized block bootstrap

The bootstrap is a usually computer-intensive resampling method for estimating the distribution of a statistic of interest. In the presence of serial dependence, one of the most commonly used methods is the block bootstrap.

The *circular block bootstrap* (CBB) sample can be defined as follows. 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 division "modulo n". For some m, let $i_1, \ldots, i_m$ be a uniform sample from the set $\{1, 2, \ldots, n\}$. Then, for a given block size b, we construct $n' = m \cdot b \quad (n' \approx n)$ resampled data:

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

We generalised the CBB in the following way. Let k be a random integer between 1 and the sample size n, and let us wrap the sample around the circle. The generalized bootstrap blocks are either of length $\lfloor b \rfloor$ or $\lceil b \rceil$ $(1 < b \in \mathbb{R})$:

$X_k, X_{k+1}, \ldots, X_{k+\lfloor b \rfloor}$ with probability $1 - b + \lfloor b \rfloor$

$X_k, X_{k+1}, \ldots, X_{k+\lceil b \rceil}$ with probability $b - \lfloor b \rfloor$

At last, we put the blocks together. This procedure ensures that for integer-valued b the new definition coincides with the traditional one, so this is indeed a generalisation.

Let us denote with $L_1, L_2, \ldots$ the block sizes – they are random variables independent from each other with common conditional distribution $P_*(L_1 = \lceil b \rceil) = 1 - P_*(L_1 = \lfloor b \rfloor) = b - \lfloor b \rfloor$. We can also write $L_i = \lfloor b \rfloor + J_i$, where $J_i|\mathcal{X}_n$ follows a Bernoulli distribution with parameter $p = b - \lfloor b \rfloor$. Let N be the random variable, which gives the number of blocks with block size $\lfloor b \rfloor$. If we have N, we can calculate the number of blocks with block size $\lceil b \rceil$, we denote it with $g(N)$; and the remainder block size with $r(N)$. The conditional covariance matrix of the bootstrap mean can be calculated explicitly:

$$\mathrm{Cov}_*(\overline{X}^*_b) = \frac{\lfloor b \rfloor^2}{n^2}\left[\mathrm{Cov}_*(\overline{X}^*_{\lfloor b \rfloor,i}) \cdot E_*N + \mathrm{Cov}_*N \cdot \overline{X}_n(\overline{X}_n)^T\right] +$$
$$+ \frac{\lceil b \rceil^2}{n^2}\left[\mathrm{Cov}_*(\overline{X}^*_{\lceil b \rceil,i}) \cdot E_*(g(N)) + \mathrm{Cov}_*(g(N)) \cdot \overline{X}_n(\overline{X}_n)^T\right] +$$
$$+ \frac{1}{n^2}\left[\sum_{i=0}^{\lfloor b \rfloor} i^2 P_*(r(N) = i) \cdot \mathrm{Cov}_*(\overline{X}^*_{i,1}) + \mathrm{Cov}_*(r(N)) \cdot \overline{X}_n(\overline{X}_n)^T\right].$$

Block length plays an important role, and it is not trivial to determine its optimal value. Our idea was that we tried to find the best block size by fitting a VAR model and then checking the variance of $\overline{X}$ with the help of the block bootstrap. The block size was determined as the b, for which the estimated trace of the covariance matrix of the mean was equal to the one using the fitted VAR model:

$$\mathrm{tr}\left(\mathrm{Cov}(\overline{X}_{VAR})\right) = \mathrm{tr}\left(\mathrm{Cov}_*(\overline{X}^*_b)\right).$$

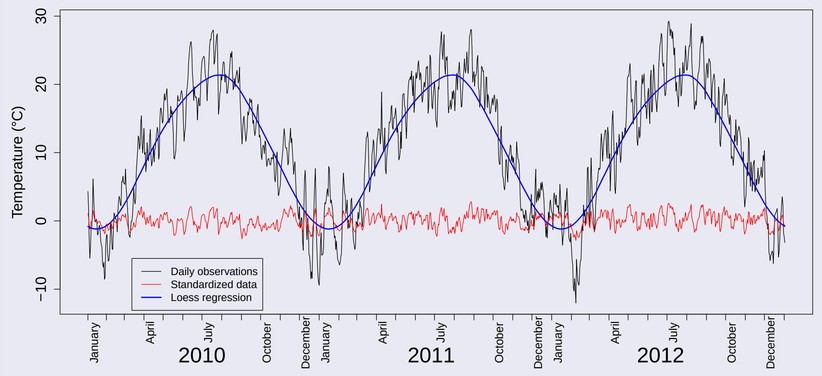
Applications

The used observations are the 63 years of daily temperature data of the European Climate Assessment (E-OBS).

We have worked with the part of the 0.5-grade grid, which lies in the Carpathian Basin. The map depicts the used five grid points (blue marks).

As we intend to use models, suitable for stationary data, first the stationarity had to be ensured. We have first subtracted the smoothed daily averages from the observations. The smoothing was made by loess regression. It turned out that the second-order stationarity is still far from being true (in winter the variances were substantially larger than in summer), so we have divided the observations with the smoothed estimated standard deviation for the given day. In order to reduce the effect of the outliers of our results, we have finally computed the ten-days moving averages.

The original and the standardized data for Budapest:



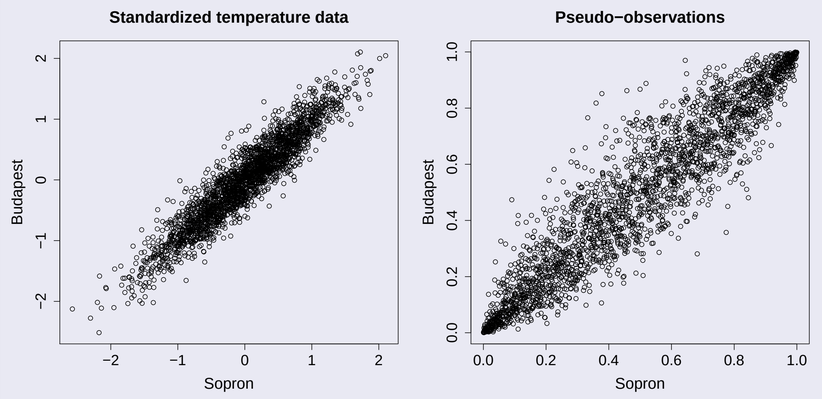
In the next step, we examined the fixed grid point Budapest paired with the other five grid points of the database. We chose the lags of the vector autoregressions to model our data pairs using the Akaike information criterion.

Our main goal was to detect if there is a significant change in the dependence structure of the data. We separated the pairs of points into two equal parts. We wanted to test the null hypothesis if the copula of the first half of the sample is equal to the copula of the second half of the sample.

The next table summarizes the main results:

Pairs of grid points	Chosen VAR-lag	$n \cdot \mathrm{tr}\left(\mathrm{Cov}(\overline{X}_{VAR})\right)$	Optimal block size	p-values of the test
Bp. & Sopron	3	1.9	20.4	0.064
Bp. & Apatovac	1	1.9	11.2	**0.028**
Bp. & Zaránd	9	2.0	6.5	**0.034**
Bp. & Nyíregyh.	4	1.8	9.3	0.116
Bp. & Püspökh.	4	1.9	32.4	0.848

Bivariate data and the corresponding pseudo-observations:



References

- **L. Varga, and A. Zempléni.** Generalised block bootstrap and its use in meteorology. Advances in Statistical Climatology, Meteorology and Oceanography, 2016. *Submitted.*
- **P. Rakonczai, L. Varga, and A. Zempléni.** Copula fitting to autocorrelated data with applications to wind speed modelling. Annales Universitatis Scientarium de Rolando Eotvos Nominatae, Sectio Computatorica, 43: 3–20, 2014.
- **B. Rémillard and O. Scaillet.** Testing for equality between two copulas. Journal of Multivariate Analysis, 100(3): 377–386, 2009.