Goodness of Fit for Auto-Copulas: Testing the Adequacy of Time Series Models

RAKONCZAI, P.¹, MÁRKUS, L.¹, and ZEMPLÉNI, A.¹

Eötvös Loránd University, Department of Probability Theory and Statistics, Budapest, Hungary
(correspondence: paulo@math.elte.hu)

Abstract. Copula models proved to be powerful tools in describing the interdependence structure of multivariate data sets. The advantage of using copulas lays in the fact that - unlike Pearson-correlations - they represent nonlinear dependencies as well, and make it possible to study the interdependence of high (or low) values of the variables. So, it is very natural to extend the use of copulas to the interdependence structure of time series. To the analogy of the autocorrelation function the use of auto-copulas for the lagged series can reveal the specifics of the dependence structure in a much finer way. Therefore the fit of the corresponding auto-copulas can provide an important tool for evaluating time series models. For making comparisons among competing models the fit of copulas has to be measured. Our suggested goodness of fit test is based on the probability integral transformation of the joint distribution, which reduces the multivariate problem to one dimension. We apply the proposed methods for investigating the auto-dependence of river flow time series with particular focus on the synchronised appearance of high values and we also consider how these methods could be improved by choosing different weights for more efficient detecting.

Keywords: Copulas, Goodness of Fit Tests, Probability Integral Transformation, River Flow Time Series

1 Copula Theory and GOF Tests

It is proved by Sklar’s theorem (1959) that for any bivariate distribution function \( H(x, y) \), with \( F(x), G(y) \) univariate marginals there exits a copula \( C \) such that \( H(x, y) = C(F(x), G(y)) \).

1.1 Archimedean copulas

We recall briefly the basic elements of copula theory for the class of the bivariate Archimedean copulas. The main advantage of these family is the very convenient build up and in addition the capability for handling the dependence structure between extremes as well.

¹ Moreover, \( C \) is unique if the marginal distributions are continuous.
Let us consider a copula generator function: $\phi_\theta(u) : [0, 1] \to [0, \infty]$, which is continuous and strictly decreasing with $\phi_\theta(1) = 0$. Then the distribution function of the Archimedean copula is

$$C_{\phi_\theta}(u, v) = \phi_\theta^{-1}\left(\phi_\theta(u) + \phi_\theta(v)\right). \quad (1)$$

Since the focus of this paper is to introduce some appropriate methods for checking a given model’s adequacy and not to consider many different copula models, we review only the Gumbel family, which is eligible for our purpose. We should emphasize, however, that the presented methods can be adapted for more general copula models as well.

The generator function of the Gumbel copula has the form $\phi_\theta(u) = [-\ln(u)]^\theta$, where $\theta \in [1, +\infty)$. Hence the Gumbel bivariate copula distribution function is given by

$$C_{\text{Gumbel}}(u, v) = e^{-\left([-\ln(u)]^\theta + [-\ln(v)]^\theta\right)^\frac{1}{\theta}}. \quad (2)$$

Simulations for Monte Carlo applications can be performed by general methods, such as conditional sampling, that can be computed quite easily with the help of the derivatives of the function $\phi_\theta^{-1}(t)$, for further details see Cherubini et al. (2004). Aside of the simulation, another relevant question is the parameter estimation which can be achieved easily based on the measures of association, e.g., for Kendall’s $\tau$ it is known $\tau_{\text{Gumbel}}(\theta) = 4 \int_0^1 \frac{\phi_\theta(u)}{\phi'_\theta(u)} du + 1 = 1 - \frac{1}{\theta}$. By inversion of the above statistics we can get a consistent estimator for the parameter

$$\hat{\theta}_n = \tau_{\text{Gumbel}}^{-1}(\tau_n) = \frac{1}{1 - \tau_n}, \quad (3)$$

where $\tau_n = -1 + \frac{4}{n(n-1)} \sum_{i\neq j} 1(X_i \leq X_j, Y_i \leq Y_j)$ is the sample version.

1.2 Goodness of Fit Tests based on the copula distribution function

The GoF statistics, we discuss, are based on the probability integral transformation (PIT) of the copula distribution function. Let a random vector $(X, Y)$ possess a bivariate copula model $C_\theta$ with unknown marginal distribution functions $F, G$ and $(X_1, Y_1), ..., (X_n, Y_n), n \geq 2$ a random sample. The so called $K$-function $K(\theta, t)$ of a copula is the distribution function of the variable $H(X, Y)$, where $H$ is the joint distribution function of $(X, Y)$:

$$K(\theta, t) = P(H(X, Y) \leq t) = P(C_\theta(F(X), G(Y)) \leq t). \quad (4)$$

---

2 The open-source R language includes an entire copula package with all of the known classes of copulas. There can be found also options for estimating copula parameters and simulating from a given copula model. For description see http://cran.r-project.org/doc/packages/copula.pdf
In the case of Gumbel copula family (4) can be computed as
\[ K_{\text{Gumbel}}(\theta, t) = t - \frac{\phi(t)}{\phi'(t)} = t(1 - \frac{\ln(t)}{\theta}), \] (5)
where \( t \in (0, 1] \). For the general form for multivariate case we refer to Genest, et al. (2006).

![K−functions for Gumbel 2−copulas](image)

**Fig. 1.** \( K \)-functions for Gumbel copulas according to different Kendall’s \( \tau \)

Define the empirical version of the \( K \)-function as
\[ K_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1(E_{in} \leq t), t \in [0, 1], \] (6)
where \( E_{in} = \frac{1}{n} \sum_{k=1}^{n} 1(X_k \leq X_i, Y_k \leq Y_i) \). In order to decide the null-hypothesis whether our observations could arise from the Gumbel family, we first fit a copula model to the observations, then compute its \( K \)-function \( K_n(t) \), and compare it to the empirical counterpart \( K_n(t) \). Known tests for the bivariate case (see Genest at el. 2006) use continuous functionals of Kendall’s process \( \kappa_n(t) = \sqrt{n}(K(\theta_n, t) - K_n(t)) \) having favourable asymptotic properties. Based on the suggested Cramer-von Mises\(^3\) type statistics \( S_n = \int_0^1 (\kappa_n(t))^2 dt \) we propose the following weighted test statistics (second coloumn), which are expectingly more sensitive in detecting descrapancies occurring near the tail (see Rakonczai, 2007):

<table>
<thead>
<tr>
<th>Deviations</th>
<th>Weighted deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 = \sum_{t_i \in [0+\epsilon, 1-\epsilon]}</td>
<td>K(\theta_n, t_i) - K_n(t_i)</td>
</tr>
<tr>
<td>( S_2 = \sum_{t_i \in [0+\epsilon, 1-\epsilon]} (K(\theta_n, t_i) - K_n(t_i))^2 )</td>
<td>( S_4 = \sum_{t_i \in [0+\epsilon, 1-\epsilon]} \frac{(K(\theta_n, t_i) - K_n(t_i))^2}{K(\theta_n, t_i)^2} )</td>
</tr>
</tbody>
</table>

\(^3\) Other type of measures are also conceivable e.g. Kolmogorov-Smirnov type statistics \( T_n = \sup_{0\leq t \leq 1} |\kappa_n(t)| \), but these are proved to be generally less powerfull.
where \((t_i)_{i=1}^n\) is an appropriately fine division of the interval \((0, 1)\). The empirical distribution of these statistics can be obtained by Monte Carlo simulation from the fitted model, and one can use some high quantiles as a critical value for the hypothesis testing.

2 Interdependence structure of River Flow Series

An application of the aforementioned methods is aimed at the evaluation of models for Danube and Tisza Rivers in Hungary, fitted to daily river discharge data. Model construction is described in detail and is tested against extremal characteristics like the fit of quantiles, maxima, extremal index, time spent over thresholds (flood duration) and the aggregate excesses (flood volume) in (Elek and Markus (2008), Vasas et al. (2007)). Here we address the adequacy of the interdependence structure of the time series simulated from those fitted models, as compared to the observed discharge series of Tisza River at Tivadar gauge. Analogously to the autocorrelation function, we characterise the interdependence structure of time series through the auto-copulas, that is the copulas of the lagged series. The empirical auto-copulas for the observed data are shown in Figure 1.

First we briefly sketch the models given in the mentioned papers, where motivation and further analysis are also presented. The deseasonalised river flow series \(X_t - c_t\) (with seasonal component \(c_t\)) has skewed, leptokurtic and light-tailed marginal distribution. The autocorrelated squares and absolute values of the innovations of a fitted ARMA filter and its non-seasonal periods of high and low variances point to conditionally heteroscedastic (GARCH-type in wide sense) modelling. Elek and Markus (2008) suggests:

\[
X_t = c_t + \sum_{i=1}^p a_i(X_{t-i} - c_{t-i}) + \sum_{i=1}^q b_i \epsilon_{t-i},
\]

\[
\epsilon_t = \sigma(X_{t-1}) Z_t,
\]

\[
\sigma(x) = (\alpha_0 + \alpha_1(x - m)_+)^{1/2}.
\]

with positive constants \(a_i, b_i, \alpha_0, \alpha_1, m\), innovation \(\epsilon_t\) and noise \(Z_t\). The model differs from conventional ARMA-GARCH ones as the variance of innovations is conditioned on the lagged values of the generated process instead of the innovations themselves and is asymptotically proportional (instead of being a quadratic function) to its past values. Vasas et al. (2007) propose a regime switching approach, where the dynamics is governed by two regimes, along which both the autoregressive coefficients and the innovation distributions are altering, moreover, the hidden regime indicator process is allowed to be non-Markovian. Assume that the discharge process, now denoted by \(Y_t\) for distinction, is governed by the hidden regime process \(I_t\) the following
where $\epsilon_{1,t}$ is an i.i.d. sequence distributed as $\Gamma(\alpha, \lambda)$ (i.e. as a gamma distribution with shape parameter $\alpha$ and scale parameter $\lambda$) and $\epsilon_{2,t}$ is an i.i.d. Gaussian sequence with zero mean and variance $\sigma^2$. $\alpha$, $\lambda$ and $\sigma$ are positive real numbers and we assume that $0 < a < 1$.

The duration of the $I_t = 0$ regime is distributed as negative binomial with parameters $(b, p_0)$ and the duration of the $I_t = 1$ regime is geometrically distributed with parameter $p_1$, where $b > 0$ and $0 < p_i < 1$ ($i = 0, 1$). The tail behaviour and extremal clustering of this regime switching model was investigated in Elek and Zempléni (2008).

3 Results and Conclusions

Fitting the mentioned two models to the discharge series of Tivadar gauge we simulated (as described in the mentioned references) artificial discharges from both. This allowed for comparison of the auto-copulas of the observed river flow data with that of the simulated series at fixed lags by means of the GOF test proposed in section 2.

As a reference we also compare the Gumbel copulas fitted to the lagged observed series, with the corresponding empirical copula. The simulations from the fitted Gumbel model are illustrated in Figure 2. It is obvious that the symmetric copula model can not be perfect, because of the asymmetry
Fig. 3. Simulations from the estimated Gumbel models

arising from the dynamics of river flows. Even so, the simulated structure shows a visually quite appealing similarity with the observations (cf. Figure 1). Note that near the tails the dependence structure is particularly well captured. Performing the tests for checking the hypothesis whether the observations can arise from the Gumbel model we found that the fit is acceptable. None of the mentioned tests (based on weighted and non-weighted, absolute and squared deviation) rejects the hypothesis. Figure 3 displays a diagnostic plot of differences for the $K$-functions at 3 days lag together with the 95% confidence bounds. The bounds were determined by simulations from the theoretical model (100 repetitions).

Figure 4 and 5 shows the auto-copulas of the artificial discharges simulated from the two competing time series models. An asymmetry appears in the copulas, showing that Gumbel model is not quite adequate in this respect and the dependence decays faster with the lags than in the observed series. Basically both copula fits are weaker than that of the Gumbel, but we have to emphasise here that the Gumbel copula is not associated with a dynamical model, and so it can not be utilised for simulating flow series.

Figure 6 compares the auto-copula of the two dynamical models and the fitted Gumbel copula in terms of tail dependence. Although globally the Gumbel model outperforms by far the two dynamical ones, at the so important high values, at the quantile range of 0.7 - 0.9 the regime switching model is significantly better than the other two. This is very much in line with the fact that the theoretical extremal index of the regime switching model is less than one - unlike the heteroscedastic model - indicating the clustering of high values. In the range of 0.9 - 1 all three models perform similarly -
GoF for Auto-Copulas - Testing Adequacy of TS Models

Fig. 4. K-function differences at 3 days lag with 95% confidence bound

Fig. 5. Simulations from the Heteroscedastic Model

perhaps more simulations would be needed for a further reaching analysis in this range.

References


GENEST, C. and FAVRE, A.-C. (2006) "Everything you always wanted to know about copula modeling but were afraid to ask", Journal of Hydrologic Engineering.
RAKONCZAI, P. and ZEMPLÉNI, A. (2007) "Copulas and goodness of fit tests", in Recent Advances in Stochastic Modelling and Data Analysis, World Scientific Publishing