Journal of Statistical Software
May 2010, Volume 34, Issue 9. http://www.jstatsoft.org/

Modeling Multivariate Distributions with Continuous Margins Using the copula R Package

Ivan Kojadinovic
The University of Auckland

Jun Yan
University of Connecticut

Abstract

The copula-based modeling of multivariate distributions with continuous margins is presented as a succession of rank-based tests: a multivariate test of randomness followed by a test of mutual independence and a series of goodness-of-fit tests. All the tests under consideration are based on the empirical copula, which is a nonparametric rank-based estimator of the true unknown copula. The principles of the tests are recalled and their implementation in the **copula** R package is briefly described. Their use in the construction of a copula model from data is thoroughly illustrated on real insurance and financial data.

Keywords: goodness of fit, multivariate independence, pseudo-observations, rank-based tests, serial independence.

1. Introduction

Copulas are being increasingly used to model multivariate distributions with continuous margins in fields such as hydrology (Salvadori, Michele, Kottegoda, and Rosso 2007), actuarial sciences (Frees and Valdez 1998) or finance (Cherubini, Vecchiato, and Luciano 2004; McNeil, Frey, and Embrechts 2005). The quite recent enthusiasm for the use of this modeling approach (see e.g., Genest, Gendron, and Bourdeau-Brien 2009a, for an analysis of this phenomenom in finance) finds its origin in an elegant representation theorem due to Sklar (1959). Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector with continuous marginal cumulative distribution functions (c.d.f.s) $F_1, \ldots, F_d$. Sklar (1959) showed that the c.d.f. H of $\mathbf{X}$ can be represented as

$$H(\mathbf{x}) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad \mathbf{x} \in \mathbb{R}^d, \tag{1}$$

in terms of a unique function $C : [0,1]^d \to [0,1]$ called a *copula*, which is merely a d-dimensional c.d.f. with standard uniform margins.

The aim in many applications is to estimate the unknown c.d.f. H from available data $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Sklar's representation then suggests breaking the construction of a model for H into two parts: the estimation of the marginal c.d.f.s $F_1, \ldots, F_d$, and the estimation of the copula C. As a model for C, one could for instance consider the Gumbel-Hougaard family of copulas, parameterized by a real $\theta \geq 1$, and defined by

$$C_\theta^{\mathrm{Gu}}(\mathbf{u}) = \exp\left(- \left[\sum_{i=1}^d \{-\log(u_i)\}^\theta \right]^{1/\theta} \right), \quad \mathbf{u} \in [0,1]^d,$$

or the normal copula family, parameterized by a correlation matrix Σ, and defined by

$$C_\Sigma^{\mathrm{N}}(\mathbf{u}) = \Phi_\Sigma\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\}, \quad \mathbf{u} \in [0,1]^d,$$

where Φ_Σ and Φ are the c.d.f.s of the multivariate standard normal with correlation Σ and the univariate standard normal, respectively. A comprehensive list of copula families can be found in Joe (1997) and Nelsen (2006).

An excellent review of the concepts and statistical issues involved in the previously mentioned model-building is given in Genest and Favre (2007). In particular, in the latter paper as in an increasing proportion of the literature, it is argued that the estimation of C should be solely based on the vectors of ranks $\mathbf{R}_1, \ldots, \mathbf{R}_n$, where $\mathbf{R}_i = (R_{i1}, \ldots, R_{id})$ and R_{ij} is the rank of X_{ij} among $X_{1j}, \ldots, X_{nj}$. The use of ranks makes the estimation of C margin-free, which implies that a misspecification of one of the marginals $F_1, \ldots, F_d$ will have no consequences on the copula estimate (see e.g., Fermanian and Scaillet 2005; Kim, Silvapulle, and Silvapulle 2007, for empirical arguments in favor of the use of ranks). The aim of this article is to present how all the steps involved in such a rank-based estimation of C can be practically carried out in R (R Development Core Team 2009) using the **copula** package (Yan and Kojadinovic 2010), which is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=copula`.

As we continue, we will assume that the data at hand consist of n copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of the random vector $\mathbf{X}$ whose c.d.f. H admits representation (1). The first practical step in the construction of a model for C is to test whether $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent, i.e., if they can be regarded as a random sample from H. This is of particular importance in fields such as finance where the data are typically time series. If the i.i.d. hypothesis is rejected, one may attempt to fit a time series model to each margin and work on the residuals. When dealing with financial log-returns, GARCH models are a frequent choice for attempting to remove serial dependence in the component time series, as discussed in Grégoire, Genest, and Gendron (2008) and Giacomini, Härdle, and Spokoiny (2009). If the i.i.d. hypothesis is not rejected, a sensible second step is to test against the presence of dependence among the components of $\mathbf{X}$. In the context under consideration, this amounts to testing

$$H_0 : C = \Pi \qquad \text{against} \qquad H_1 : C \neq \Pi,$$

where $\Pi(\mathbf{u}) = \prod_{i=1}^d u_i$, $\mathbf{u} \in [0,1]^d$, is the *independence copula*. If independence is rejected, a typical next step is to fit an appropriate parametric copula family to the available data. In practice, this amounts to performing goodness-of-fit tests of the form

$$H_0 : C \in \mathcal{C} \qquad \text{against} \qquad H_1 : C \notin \mathcal{C},$$

for several parametric families $\mathcal{C} = \{C_\theta\}$. The final step involves choosing one of the candidate families that were not rejected, if any, and possibly providing standard errors for the parameter estimates.

All the steps mentioned above can be carried out by means of the **copula** R package. The basic functionalities of the package were described in Yan (2007). Since then, significant development has been added to the package which enables the user, among other things, to perform tests of independence, serial independence and goodness of fit.

An important building block of the tests under consideration is the *empirical copula* of the data (Deheuvels 1979, 1981b) which is a consistent estimator of the unknown copula C. Let $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_n$ be *pseudo-observations* from C defined by $\hat{\mathbf{U}}_i = \mathbf{R}_i/(n+1)$, $i \in \{1, \ldots, n\}$. The components of the pseudo-observations can equivalently be rewritten as $\hat{U}_{ij} = n\hat{F}_j(X_{ij})/(n+1)$, where $\hat{F}_j$ is the empirical c.d.f. computed from $X_{1j}, \ldots, X_{nj}$, and where the scaling factor $n/(n+1)$ is introduced to avoid problems at the boundary of $[0,1]^d$. The empirical copula is then classically defined as the empirical c.d.f. computed from the pseudo-observations, i.e.,

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{\mathbf{U}}_i \leq \mathbf{u}), \qquad \mathbf{u} \in [0,1]^d. \tag{2}$$

The rest of the article is organized as follows. The principles of the tests of multivariate independence based on the empirical copula studied by Deheuvels (1981a) and Genest and Rémillard (2004) are presented in Section 2. The third section briefly describes the serial analogues of the latter tests that can be used to test randomness. Section 4 is then devoted to an overview of a powerful "blanket" goodness-of-fit test for copulas. The computation of an approximate p-value for the test statistic can be performed using either the parametric bootstrap (Genest and Rémillard 2008) or a fast multiplier approach (Kojadinovic, Yan, and Holmes 2010). The usage of the various tests in the copula-based modeling process is demonstrated in Section 5 through the analysis of real insurance and financial data. Concluding remarks are given in Section 6.

2. Tests of multivariate independence

Inspired by the work of Blum, Kiefer, and Rosenblatt (1961), Dugué (1975), Deheuvels (1981b), and more recently Ghoudi, Kulperger, and Rémillard (2001), Genest and Rémillard (2004) suggested to base a test of the mutual independence of the components of $\mathbf{X}$ on the statistic

$$I_n = \int_{[0,1]^d} n \left\{ C_n(\mathbf{u}) - \prod_{i=1}^{d} u_i \right\}^2 \mathrm{d}\mathbf{u}.$$

An interesting aspect of the test under consideration comes from the fact that, under the mutual independence of the components $X_1, \ldots, X_d$ of $\mathbf{X}$, the empirical process $\sqrt{n}\{C_n - \Pi\}$ can be decomposed, using the *Möbius transform* (Rota 1964), into $2^d - d - 1$ sub-processes $\sqrt{n}\mathcal{M}_A(C_n)$, $A \subseteq \{1, \ldots, d\}$, $|A| > 1$, that converge jointly to tight centered mutually independent Gaussian processes. One fundamental property of this decomposition, whose form is precisely given for instance in Genest and Rémillard (2004), is that mutual independence among $X_1, \ldots, X_d$ is equivalent to having $\mathcal{M}_A(C)(\mathbf{u}) = 0$, for all $\mathbf{u} \in [0,1]^d$ and all $A \subseteq \{1, \ldots, d\}$ such that $|A| > 1$. Instead of the single test statistic I_n, this suggests considering $2^d - d - 1$ test statistics of the form

$$M_{A,n} = \int_{[0,1]^d} n\{\mathcal{M}_A(C_n)(\mathbf{u})\}^2 \mathrm{d}\mathbf{u},$$

where $A \subseteq \{1,\ldots,d\}$, $|A| > 1$, that are asymptotically mutually independent under the null hypothesis of independence. Each statistic $M_{A,n}$ can be seen as focusing on the dependence among the components of $\mathbf{X}$ whose indices are in A. The above decomposition has been recently extended by Kojadinovic and Holmes (2009) to the situation where one wants to test the mutual independence of several continuous random vectors.

As an alternative to the statistic I_n, Genest and Rémillard (2004) (see also Genest, Quessy, and Rémillard 2007) studied several ways to combine the $2^d - d - 1$ statistics $M_{A,n}$ into one global statistic for testing independence. Two combination rules implemented in the **copula** package are those of Fisher (1932) and Tippett (1931). The test based on the former tends to give the best results and was found to frequently outperform the test based on I_n in the Monte Carlo experiments carried out by Genest and Rémillard (2004) and Kojadinovic and Holmes (2009).

To visualize the results of the independence test when based on all the $2^d - d - 1$ statistics $M_{A,n}$, a graphical representation, called a *dependogram*, can be used. For each subset $A \subseteq \{1,\ldots,d\}$, $|A| > 1$, a vertical bar is drawn whose height is proportional to the value of $M_{A,n}$. The approximate critical values of $M_{A,n}$ are represented on the bars by black bullets. Subsets for which the bar exceeds the critical value can be considered as being composed of dependent variables. Examples of such a representation are given in Figures 2 and 3.

The tests described above are implemented in the functions `indepTestSim` and `indepTest` of the **copula** package. The function `indepTestSim` does not take data as input. It returns an object of class `indepTestDist` which contains a large number of approximate independent realizations of the test statistics under mutual independence. The function `indepTest` then takes in the data and the previously returned object of class `indepTestDist`, and returns the statistics and their approximate p-values. The dependogram can be plotted via the function `dependogram`. The function implementing the extension of Kojadinovic and Holmes (2009) allowing to test independence among continuous random vectors is called `multIndepTest`, and its usage is similar to that of the function `indepTest`.

3. Tests of randomness

The test of multivariate independence of Deheuvels (1981b) can be extended to test randomness as suggested by Genest and Rémillard (2004) (see also Ghoudi *et al.* 2001). Given a stationary and ergodic univariate sequence of continuous random variables $X_1, X_2, \ldots$ and an integer $p > 1$, first form p-dimensional vectors of observations $\mathbf{Y}_i = (X_i, \ldots, X_{i+p-1})$, $i \in \{1,\ldots,n\}$, where p is the *embedding dimension*. The value of p, which determines the maximum lag, has to be chosen by the user. Departure from serial independence can then be measured using the statistic

$$I_n^s = \int_{[0,1]^p} n \left\{ C_n^s(\mathbf{u}) - \prod_{k=1}^{p} u_k \right\}^2 \mathrm{d}\mathbf{u},$$

where C_n^s is the serial analogue of the empirical copula and is computed from $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. As for the test of multivariate independence described in the previous section, the Möbius transform can be used to derive a collection of statistics $M^s_{A,n}$, $A \subseteq \{1,\ldots,p\}$, $A \ni 1$, $|A| > 1$, that converge jointly to mutually independent random variables under serial independence. Each of the statistics $M^s_{A,n}$ can be seen as focusing on the departure from serial independence arising from the lags $\{i+1 : i \in A\}$. As in the previous section, these statistics can be combined into one global statistic using the combination rules of Fisher (1932) or Tippett (1931) to give a potentially more powerful test than that based on I^s_n.

The functions implementing these tests are `serialIndepTestSim` and `serialIndepTest`. Their usage is similar to those of their non-serial counterparts.

The previous approach has been extended by Kojadinovic and Yan (2010b) to the situation where one wants to test against serial dependence in continuous multivariate time series. The corresponding function is called `multSerialIndepTest`.

4. Goodness-of-fit tests

The goodness-of-fit tests implemented in the **copula** package are all based on the empirical process

$$\mathbb{C}_n(\mathbf{u}) = \sqrt{n}\{C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u})\}, \qquad \mathbf{u} \in [0,1]^d, \tag{3}$$

where C_n is the empirical copula defined in (2) and C_{θ_n} is an estimator of C under the hypothesis that $H_0 : C \in \{C_\theta\}$ holds. The estimator θ_n of θ appearing in (3) is again solely based on ranks. It is either one of two method-of-moment estimators based respectively on the inversion of Kendall's tau and Spearman's rho, or the maximum pseudo-likelihood estimator of Genest, Ghoudi, and Rivest (1995).

In the large scale Monte Carlo experiments carried out by Berg (2009) and Genest, Rémillard, and Beaudoin (2009b), the statistic

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\mathbf{u})^2 \mathrm{d}C_n(\mathbf{u}) = \sum_{i=1}^{n} \{C_n(\hat{\mathbf{U}}_i) - C_{\theta_n}(\hat{\mathbf{U}}_i)\}^2 \tag{4}$$

gave the best results overall.

An approximate p-value for S_n can be obtained by means of a parametric bootstrap-based procedure which is recalled in the next subsection, and whose validity was recently shown by Genest and Rémillard (2008). The main inconvenience of this approach is its very high computational cost, as each iteration requires both random number generation from the fitted copula and estimation of the copula parameters. As the sample size increases, the application of the parametric bootstrap-based goodness-of-fit test becomes prohibitive. In order to circumvent this very high computational cost, a fast large-sample testing procedure based on multiplier central limit theorems was proposed in Kojadinovic *et al.* (2010) (see also Kojadinovic and Yan 2010a). Its principles are recalled in Section 4.2.

4.1. Parametric bootstrap-based goodness-of-fit test

An approximate p-value for the test based on the statistic defined in (4) can be obtained by means of the following procedure (see Genest and Rémillard 2008, for more details):

1. Compute C_n from the pseudo-observations $\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_n$ and estimate θ from $\hat{\mathbf{U}}_1,\ldots,\hat{\mathbf{U}}_n$ by means of a rank-based estimator θ_n.

2. Compute the test statistic S_n defined in (4).
3. For some large integer N, repeat the following steps for every $k \in \{1, \dots, N\}$:
   (a) Generate a random sample $\mathbf{X}_1^{(k)}, \dots, \mathbf{X}_n^{(k)}$ from copula C_{θ_n} and compute the associated pseudo-observations $\hat{\mathbf{U}}_1^{(k)}, \dots, \hat{\mathbf{U}}_n^{(k)}$.
   (b) Let
   $$C_n^{(k)}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{\mathbf{U}}_i^{(k)} \le \mathbf{u}), \qquad \mathbf{u} \in [0,1]^d,$$
   and compute an estimate $\theta_n^{(k)}$ of θ from $\hat{\mathbf{U}}_1^{(k)}, \dots, \hat{\mathbf{U}}_n^{(k)}$ using the same rank-based estimator as in Step 1.
   (c) Compute an approximate independent realization of S_n under H_0 by
   $$S_n^{(k)} = \sum_{i=1}^{n} \{C_n^{(k)}(\hat{\mathbf{U}}_i^{(k)}) - C_{\theta_n^{(k)}}(\hat{\mathbf{U}}_i^{(k)})\}^2.$$
4. An approximate p-value for the test is given by $N^{-1}\sum_{k=1}^{N} \mathbf{1}(S_n^{(k)} \ge S_n)$.

As one can see, this procedure is very computationally intensive as each iteration in Step 3 involves random number generation from the hypothesized copula and estimation of the copula parameters. This is particularly true if θ_n is the maximum pseudo-likelihood estimator. As n reaches 300, the extensive Monte Carlo experiments carried out for $d = 2$, 3 and 4 in Kojadinovic and Yan (2010a) indicate that one can alternatively safely use the fast multiplier approach described in the next subsection.

Note that, in the package, approximate p-values are computed using the slightly modified expression
$$\frac{1}{N+1}\left\{\sum_{k=1}^{N} \mathbf{1}(S_n^{(k)} \ge S_n) + \frac{1}{2}\right\}$$
to ensure that they are in the open interval $(0,1)$ so that transformations by inverse c.d.f.s of continuous distributions are always well-defined. This convention is adopted for all tests for which approximate p-values are computed using the empirical c.d.f. Consequently, approximate p-values are numbers in the set
$$\left\{\frac{1}{N+1}\left(0+\frac{1}{2}\right), \frac{1}{N+1}\left(1+\frac{1}{2}\right), \dots, \frac{1}{N+1}\left(N+\frac{1}{2}\right)\right\}.$$

4.2. Goodness-of-fit test based on multiplier central limit theorems

Inspired by the seminal work of Scaillet (2005) and Rémillard and Scaillet (2009), a valid and much faster alternative to the previously presented parametric bootstrap-based procedure was recently proposed in Kojadinovic *et al.* (2010). The computational efficiency of the procedure follows from the fact that, under suitable regularity conditions, the goodness-of-fit process $\mathbb{C}_n$ can be written as
$$\mathbb{C}_n(\mathbf{u}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{J}_i(\mathbf{u}) + R_n(\mathbf{u}), \tag{5}$$

where $\mathbb{J}_1, \ldots, \mathbb{J}_n$ are independent and identically distributed processes whose form depends on the estimator θ_n and on the hypothesized copula family $\{C_\theta\}$, and $\sup |R_n(\mathbf{u})|$ tends to 0 in probability. Let $\hat{\mathbb{J}}_{i,n}$ be the version of $\mathbb{J}_i$ in which all the unknown quantities are replaced by their estimates. The multiplier approach modifies Step 3 of the parametric bootstrap-based procedure as follows:

3. For some large integer N, repeat the following steps for every $k \in \{1, \ldots, N\}$:
   (a) Generate n i.i.d. random variates $Z_1^{(k)}, \ldots, Z_n^{(k)}$ with expectation 0 and variance 1.
   (b) Form an approximate independent realization of $\mathbb{C}_n$ under H_0 by
   $$\mathbb{C}_n^{(k)}(\mathbf{u}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i^{(k)} \hat{\mathbb{J}}_{i,n}(\mathbf{u}).$$
   (c) Compute an approximate independent realization of S_n under H_0 by
   $$S_n^{(k)} = \int_{[0,1]^d} \{\mathbb{C}_n^{(k)}(\mathbf{u})\}^2 \mathrm{d}C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \{\mathbb{C}_n^{(k)}(\hat{\mathbf{U}}_i)\}^2.$$

This procedure is fast because the terms $\hat{\mathbb{J}}_{i,n}$ need to be computed only once. The derivation and the computation of these terms, however, is not trivial, as they involve partial derivatives of the c.d.f. and the p.d.f. of the hypothesized copula with respect to the u_i and the parameters (see Kojadinovic *et al.* 2010; Kojadinovic and Yan 2010a, for more details).

4.3. Function usage

Both methods are implemented in the **copula** package with a common interface function called `gofCopula`. The important arguments of `gofCopula` are as follows.

- `copula`: an object representing the hypothesized copula whose attribute `parameters` is ignored.
- `x`: the observed data matrix, each row of which is a multivariate observation.
- `simulation`: the simulation method; can be either `"pb"` (parametric bootstrap) or `"mult"` (multiplier). The default simulation method is `"pb"`.
- `N`: the number of bootstrap/multiplier iterations.
- `method`: estimation method for the copula parameter(s); can be either `"mpl"` (maximum pseudo-likelihood), `"itau"` (inversion of Kendall's tau) or `"irho"` (inversion of Spearman's rho).

4.4. Computational aspects

For the simulation method `"mult"` to work for a given copula family, it is necessary that appropriate functions to compute the terms $\hat{\mathbb{J}}_{i,n}$ are implemented. This non-trivial step is the price to pay for the computational efficiency of the multiplier approach. These functions have been implemented for six copula families: the Clayton, Gumbel-Hougaard, Frank, Plackett, normal and t. In dimension two, all three methods for estimating the copula parameter (`"mpl"`, `"itau"`, and `"irho"`) can be used. In dimension three or higher, only maximum pseudo-likelihood estimation (`"mpl"`) is available. This approach however cannot be used for dimensions higher than 10 for the Clayton and Gumbel-Hougaard families, and 6 for the Frank family, mainly because the expressions of the p.d.f.s of these copulas (and their partial derivatives) have not been obtained and stored for higher dimensions. For the normal and t copulas, there is no limit on the dimension code-wise.

The speed of the function `gofCopula` with arguments `simulation = "mult"` and `method = "mpl"` depends on the hypothesized copula family. For instance, testing for the Clayton dependence structure is generally much faster than testing for the normal copula. As could be expected, the computing time increases as the dimension increases, but again in a copula-dependent way. As reported in Kojadinovic and Yan (2010a, Section 6), for $n = 500$ and $d = 4$, to test whether the normal copula (resp. the Clayton copula) may be considered as an appropriate model for the data, the multiplier procedure based on the maximum pseudo-likelihood estimator takes about 1.73 seconds (resp. 0.20 seconds) on one 2.2 GHz processor with $N = 1000$. If the dimension d is increased to five, all other parameters being unchanged, the procedure takes approximately 0.22 seconds if the Clayton copula is hypothesized but 4.28 seconds if the normal is hypothesized. If d is increased to six, the approximate execution times become 0.26 and 9.34 seconds, respectively. It is however important to keep in mind that these measures are based on our mixed R and C implementation which is not optimal in terms of speed, especially when the hypothesized family is the normal or the t.

Let us conclude this section with a few words about the function `gofCopula` with arguments `simulation = "pb"` (parametric bootstrap) and `method = "mpl"`. Unlike the method-of-moment estimators which usually require more programming work, maximum pseudo-likelihood estimation only requires the density of the hypothesized copula to be implemented. For user-defined copula families, a combination of estimation method `"mpl"` and simulation method `"pb"` yields a ready-to-use goodness-of-fit test as long as the density and random number generation functions are implemented.

5. Illustration

This section presents two applications of the tests described in the previous sections to data sets originally considered by Frees and Valdez (1998) and McNeil *et al.* (2005). Both data sets are available in the package.

5.1. Insurance loss

The insurance data of Frees and Valdez (1998) are frequently used for illustration purposes in copula modeling (see e.g. Klugman and Parsa 1999; Genest, Quessy, and Rémillard 2006; Ben Ghorbal, Genest, and Nešlehová 2009). The two variables of interests are `loss`, an indemnity payment, and `alae`, the corresponding allocated loss adjustment expense. They were observed for 1500 claims of an insurance company. Following Genest *et al.* (2006), we restrict ourselves to the 1466 uncensored claims.

```
R> library("copula")
```

```
R> data("loss")
R> myLoss <- subset(loss, censored == 0, select = c("loss", "alae"))
R> nrow(myLoss)

[1] 1466
```

It can also be verified that the variables `loss` and `alae` take only 541 and 1401 unique values, respectively, which means that there is a non-negligible number of ties in the data. The presence of ties here can be attributed to rounding and precision issues. As we shall see later in this subsection, ignoring the ties, for instance, by using midranks to compute pseudo-observations, can affect the conclusions of the analysis qualitatively. Indeed, all the tests described in the previous sections were derived under the assumption of continuous margins, which implies that ties occur with probability zero. To deal with ties in a slightly more satisfactory way, we propose to construct pseudo-observations by randomly breaking the ties. The resulting pseudo-observations and those obtained using average ranks for ties are plotted in Figure 1.

```
R> set.seed(123)
R> pseudoLoss <- sapply(myLoss, rank, ties.method = "random") /
+    (nrow(myLoss) + 1)
R> pseudoLoss.ave <- sapply(myLoss, rank)/(nrow(myLoss) + 1)
R> par(mfrow = c(1, 2), mgp = c(1.5, 0.5, 0), mar = c(3.5, 2.5,
+    0, 0))
R> plot(pseudoLoss, sub = "(a) random rank for ties")
R> plot(pseudoLoss.ave, sub = "(b) average rank for ties")
```

The "vertical lines" appearing in Figure 1 (b) confirm that the variable `loss` is the most affected by ties.

By repeating the analysis based on `ties.method = "random"` a large number of times, we will verify at the end of this subsection that the randomization does not change the results qualitatively.

As there is no temporal element in the data (which are sorted according to the variable `loss`), there is no reason to test for serial dependence. Our first step in the construction of a copula model is thus to test the independence between `loss` and `alae`, which can be done using the test described in Section 2. With 1466 observations, generating 1000 approximate independent realizations of the test statistics under independence takes a while. The execution times reported hereafter and in the rest of the paper are in seconds and were obtained on one 2.33 GHz processor.

```
R> system.time(empsamp <- indepTestSim(nrow(pseudoLoss), p = 2,
+    N = 1000, print.every = 0))

   user  system elapsed
280.597   1.179 281.862

R> indepTest(pseudoLoss, empsamp)
```

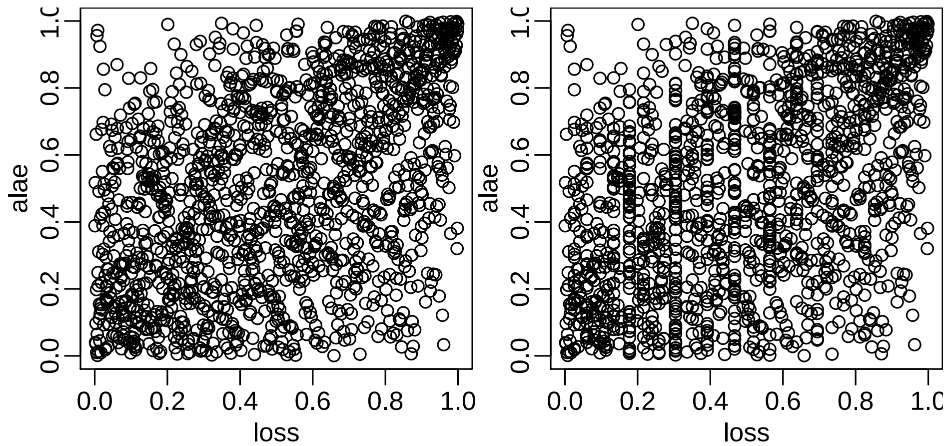

Figure 1: Pseudo-observations constructed from the `loss` and `alae` insurance data. (a) The ties are broken at random. (b) Average ranks are used for ties.

```
Global Cramer-von Mises statistic: 2.784981 with p-value 0.0004995005
Combined p-values from the Mobius decomposition:
   0.0004995005  from Fisher's rule,
   0.0004995005  from Tippett's rule.
```

Since $d = 2$, the only statistic obtained from the Möbius decomposition, $M_{\{1,2\},n}$, coincides with the global statistic I_n. As there is very strong evidence against the null hypothesis of independence, it makes sense to consider different copula families to model the dependence between `loss` and `alae`.

Our next step therefore is to perform several goodness-of-fit tests. As candidate families, we consider the Gumbel-Hougaard, Clayton, Frank, normal, Plackett and t with 4 degrees of freedom. Notice that, for technical reasons discussed for instance in Demarta and McNeil (2005) or in Kojadinovic and Yan (2010a), the number of degrees of freedom of the t copula has to be fixed and will therefore not be considered as a parameter to be estimated. Also, if not specified, the number of degrees of freedom is set to four by default in the package. For the six families under consideration, all three estimation methods, `"mpl"` (maximum pseudo-likelihood), `"itau"` (inversion of Kendall's tau), and `"irho"` (inversion of Kendall's tau), can be used. Let us choose `"itau"` and compare, for the Gumbel-Hougaard family, the results and execution times of the two possible simulation methods `"pb"` (parametric bootstrap) and `"mult"` (multiplier).

```
R> system.time(lossGof.gumbel.pb <- gofCopula(gumbelCopula(1), pseudoLoss,
+    method = "itau", simulation = "pb", N = 1000, print.every = 0))
```

```
   user  system elapsed
250.040   0.006 250.084
```

```
R> lossGof.gumbel.pb
```

```
Parameter estimate(s): 1.440397
Cramer-von Mises statistic: 0.02100463 with p-value 0.2312687

R> system.time(lossGof.gumbel.mult <- gofCopula(gumbelCopula(1),
+    pseudoLoss, method = "itau", simulation = "mult", N = 1000))

   user  system elapsed
 30.048   0.045  30.099

R> lossGof.gumbel.mult

Parameter estimate(s): 1.440397
Cramer-von Mises statistic: 0.02100463 with p-value 0.2402597
```

Note that the first argument of `gofCopula`, `gumbelCopula(1)`, indicates which copula family is being tested. The value of its attribute `parameters` is ignored.

As can be observed, the multiplier approach and the parametric bootstrap-based procedure give very similar p-values but, as implemented in the **copula** package, the former is about 8 times faster. As shall be seen in the next subsection, it becomes orders of magnitude faster when the estimation is based on the maximization of the pseudo-likelihood. Notice that, had this not been an illustration, we would have set the number of bootstrap/multiplier iteration N to 10 000 to obtain more accurate p-values.

For the five remaining families, goodness-of-fit results are obtained by successively replacing the object `gumbelCopula(1)` with `claytonCopula(1)`, `frankCopula(1)`, `normalCopula(0)`, `plackettCopula(1)`, and `tCopula(0, df = 4, df.fixed = TRUE)`. For more details, see the function `myAnalysis` below. For all five families, the obtained approximate p-value is `0.0004995005`. Therefore, among all the families that we have tested, the Gumbel-Hougaard is the only one that is not rejected at the 5% significance level. This is in accordance with the results obtained for instance in Chen and Fan (2005) using pseudo-likelihood ratio tests, or in Genest *et al.* (2006) using a goodness-of-fit procedure based on Kendall's process.

Let us finally compute the standard error of the parameter estimate for the Gumbel-Hougaard copula:

```
R> fitCopula(gumbelCopula(1), pseudoLoss, method = "itau")

The estimation is based on the inversion of Kendall's tau
and a sample of size 1466.
      Estimate Std. Error  z value Pr(>|z|)
param 1.440397 0.03326981 43.29443        0
```

As the analysis presented thus far was based on the use of randomization to break ties, we repeat the previous steps a large number of times to see how different randomizations affect the results. This is done with the hope that many different configurations will be obtained for the parts of the data affected by ties. Given a `matrix` of observations, the function `myAnalysis` hereafter returns the p-values of the tests of independence and goodness of fit, as well as the parameter estimate of the Gumbel-Hougaard copula and the corresponding standard error.

```
R> myAnalysis <- function(myLoss) {
+    pseudoLoss <- sapply(myLoss, rank, ties.method = "random") /
+      (nrow(myLoss) + 1)
+    indTest <- indepTest(pseudoLoss, empsamp)$global.statistic.pvalue
+    gof.g <- gofCopula(gumbelCopula(1), pseudoLoss, method = "itau",
+      simulation = "mult")$pvalue
+    gof.c <- gofCopula(claytonCopula(1), pseudoLoss, method = "itau",
+      simulation = "mult")$pvalue
+    gof.f <- gofCopula(frankCopula(1), pseudoLoss, method = "itau",
+      simulation = "mult")$pvalue
+    gof.n <- gofCopula(normalCopula(0), pseudoLoss, method = "itau",
+      simulation = "mult")$pvalue
+    gof.p <- gofCopula(plackettCopula(1), pseudoLoss, method = "itau",
+      simulation = "mult")$pvalue
+    gof.t <- gofCopula(tCopula(0, df = 4, df.fixed = TRUE), pseudoLoss,
+      method = "itau", simulation = "mult")$pvalue
+    fit.g <- fitCopula(gumbelCopula(1), pseudoLoss, method = "itau")
+    c(indep = indTest, gof.g = gof.g, gof.c = gof.c, gof.f = gof.f,
+      gof.n = gof.n, gof.t = gof.t, gof.p = gof.p, est = fit.g@estimate,
+      se = sqrt(fit.g@var.est))
+  }
R> myReps <- t(replicate(100, myAnalysis(myLoss)))
R> round(apply(myReps, 2, summary), 3)

        indep gof.g gof.c gof.f gof.n gof.t gof.p   est    se
Min.        0 0.189     0     0     0     0     0 1.440 0.033
1st Qu.     0 0.222     0     0     0     0     0 1.442 0.033
Median      0 0.232     0     0     0     0     0 1.442 0.033
Mean        0 0.235     0     0     0     0     0 1.442 0.033
3rd Qu.     0 0.248     0     0     0     0     0 1.443 0.033
Max.        0 0.290     0     0     0     0     0 1.444 0.033
```

As can be seen for the above summary of 100 replications, the randomization step does not affect the conclusions qualitatively.

It is important to notice that ignoring the ties in the computation of the pseudo-observations, for instance by using the default method `"average"` in the function `rank`, leads to the rejection of all the families including the Gumbel-Hougaard.

```
R> gofCopula(gumbelCopula(1), pseudoLoss.ave, method = "itau",
+    simulation = "mult", N = 1000)

Parameter estimate(s): 1.446450
Cramer-von Mises statistic: 0.0882599 with p-value 0.0004995005
```

As the tests considered in this paper were derived under the assumption of continuous margins, we believe that it is important not to ignore ties when their number is large. The proposed randomization-based approach, although it may not be completely satisfactory, constitutes a first way forward. Further research in that direction is necessary.

5.2. Stock returns

McNeil *et al.* (2005, Chapter 5) analyzed five years of daily log-return data (1996–2000) for the Intel, Microsoft and General Electric stocks. As for most stock return data, the number of ties is very small.

```
R> data("rdj")
R> nrow(rdj)

[1] 1262

R> apply(rdj[, 2:4], 2, function(x) length(unique(x)))

INTC MSFT   GE
1244 1244 1227
```

For the sake of simplicity, we shall therefore ignore the ties as we continue, although we could easily proceed as in the previous subsection at the expense of more computation.

```
R> pseudoSR <- apply(rdj[, 2:4], 2, rank)/(nrow(rdj) + 1)
```

As we are dealing with financial time series data, we might expect that the assumption of serial independence does not hold. The first step is thus to perform a test of randomness. The standard tools for this are the test of Ljung and Box (1978) and its multivariate extension studied by Hosking (1980) (see also Johansen 1995). Tests based on the sample autocorrelation function are however known not to maintain their nominal level for heavy-tailed distributions (see e.g. Kojadinovic and Yan 2010b, for empirical evidence of this phenomenon). The sample excess kurtoses for the three univariate return series are 5.11, 5.49, and 1.18, respectively. A safer alternative hence consists of using a multivariate rank-based test such as the one described in Section 3. Following the common practice in finance, we apply the test on the squared returns, which will be transformed into ranks within the function.

```
R> set.seed(123)
R> system.time(srMultSerialIndepTest <- multSerialIndepTest(rdj[,
+    2:4]^2, lag.max = 4, print.every = 0))

    user   system  elapsed
2145.572    0.080 2145.979

R> srMultSerialIndepTest

Global Cramer-von Mises statistic: 0.000878193 with p-value 0.0004995005
Combined p-values from the Mobius decomposition:
   0.0004995005  from Fisher's rule,
   0.0004995005  from Tippett's rule.
```

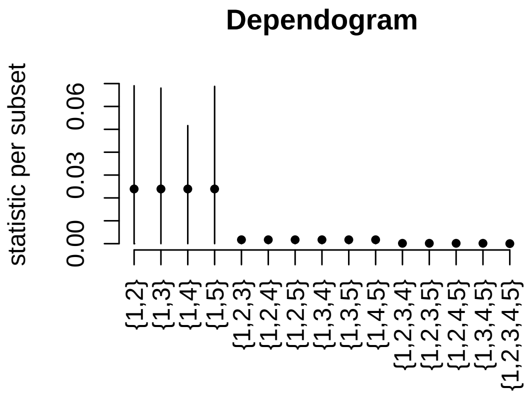

Figure 2: Dependogram summarizing the results of the test of serial independence for the stock return data with `lag.max=4`.

```
R> dependogram(srMultSerialIndepTest)
```

The dependogram represented in Figure 2 flags most of the "subset of lags" for serial dependence (more details can be obtained by calling `dependogram(srMultSerialIndepTest, print = TRUE)`). All three p-values computed within the function `multSerialIndepTest` indicate very strong evidence against the null hypothesis of serial independence, which, in turn, provides very strong evidence against the geometric Brownian motion model commonly adopted in finance for stock prices. At this point, one possibility would consist of fitting a GARCH model to each of the marginal daily log-returns series and of working on the residuals as suggested in Grégoire *et al.* (2008). As our objective is to attempt to recover the results obtained by McNeil *et al.* (2005, Chapter 5), we will nonetheless ignore the serial dependence as they did and continue with the analysis. While this is not completely satisfactory, it might be justified by the stylized fact of financial time series that serial dependence in daily log-returns, although present, is generally very weak. The extent to which serial dependence might affect the copula modeling has not been studied in the literature.

Proceeding as if the trivariate observations at hand were i.i.d., the next step is to test the mutual independence of the three series using, for instance, the test described in Section 2.

```
R> system.time(empsamp <- indepTestSim(nrow(pseudoSR), p = 3, N = 1000,
+    print.every = 0))

   user  system elapsed
496.365   0.024 496.463

R> srMultIndepTest <- indepTest(pseudoSR, empsamp)
R> srMultIndepTest

Global Cramer-von Mises statistic: 5.081161 with p-value 0.0004995005
Combined p-values from the Mobius decomposition:
   0.0004995005  from Fisher's rule,
   0.0004995005  from Tippett's rule.
```

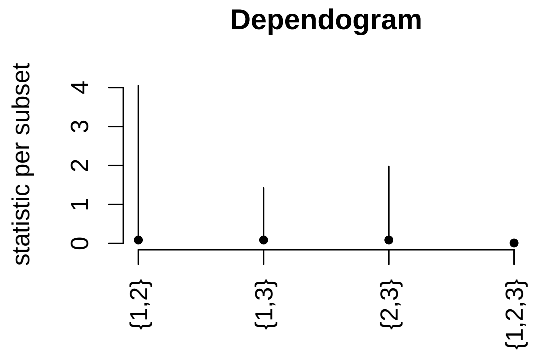

Figure 3: Dependogram summarizing the results of the test of mutual independence for the trivariate stock return data.

```
R> dependogram(srMultIndepTest)
```

As we could have expected, there is very strong evidence against the null hypothesis of mutual independence. Unsurprisingly, the dependogram, which is given in Figure 3, indicates that independence is rejected for all subsets of variables.

The last but one step then consists of performing several goodness-of-fit tests. As candidate families, we consider three one-parameter families, *viz.* the Clayton, the Frank and the Gumbel-Hougaard, and five three-parameter families, *viz.* the normal and the t with 5, 10, 15 and 20 degrees of freedom. Let us first compare for the t family with 5 degrees of freedom, the results and execution times of the two possible simulation methods "pb" (parametric bootstrap) and "mult" (multiplier). We use the default value for N which is 1000.

```
R> system.time(srGof.t.pb <- gofCopula(tCopula(c(0, 0, 0), dim = 3,
+    dispstr = "un", df = 5, df.fixed = TRUE), pseudoSR, method = "mpl",
+    print.every = 0))

     user    system   elapsed
11944.082     2.459 11948.410

R> srGof.t.pb

Parameter estimate(s): 0.5807332 0.3531494 0.4153485
Cramer-von Mises statistic: 0.03189032 with p-value 0.1713287

R> system.time(srGof.t.mult <- gofCopula(tCopula(c(0, 0, 0), dim = 3,
+    dispstr = "un", df = 5, df.fixed = TRUE), pseudoSR, method = "mpl",
+    simulation = "mult"))

   user  system elapsed
 37.382   0.039  37.426
```

```
R> srGof.t.mult

Parameter estimate(s): 0.5807332 0.3531494 0.4153485
Cramer-von Mises statistic: 0.03192471 with p-value 0.1893107
```

As already observed in the previous subsection, the multiplier approach and the parametric bootstrap-based procedure give very similar p-values but the former is this time about 320 times faster. The small difference between the values of the test statistic is due to the fact that the computation of S_n defined in (4) involves the use of the function `pmvt` from the **mvtnorm** package (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2009). The latter uses randomized quasi-Monte Carlo methods for computing the c.d.f. of the multivariate t.

For the other families, we only use the multiplier simulation method. The approximate p-values are 0.000 for the Clayton, Frank and Gumbel-Hougaard, 0.070 for the normal, and 0.568, 0.552, and 0.450 for the t with 10, 15 and 20 degrees of freedom, respectively. As can be observed, the normal and the t families with 5, 10, 15 and 20 degrees of freedom are not rejected at the 5% significance level. The fact that the test does not seem to be able to clearly distinguish between these families might not be very surprising. Indeed, the results of the Monte Carlo experiments reported in Berg (2009), Genest *et al.* (2009b) and Kojadinovic and Yan (2010a) indicate that available goodness-of-fit tests for copulas generally have difficulty distinguishing between elliptical copulas with main differences in the tails.

We arbitrarily opt for the family with the highest p-value and end this illustration with the computation of standard errors for the parameter estimates.

```
R> fitCopula(tCopula(c(0, 0, 0), dim = 3, dispstr = "un", df = 10,
+    df.fixed = TRUE), pseudoSR, method = "mpl")

The estimation is based on the maximum pseudo-likelihood
and a sample of size 1262.
       Estimate Std. Error  z value Pr(>|z|)
rho.1 0.5923770 0.01830657 32.35870        0
rho.2 0.3625107 0.02410258 15.04033        0
rho.3 0.4264413 0.02312991 18.43679        0
The maximized loglikelihood is  415.2775
The convergence code is  0
```

The selected t copula is very similar to that selected in McNeil *et al.* (2005, Chapter 5). The only noticeable difference comes from the degrees of freedom (6.5 in their analysis versus 10 here). Note that, for the t family with 6.5 degrees of freedom, the multiplier goodness-of-fit test gives a p-value of about 40%.

6. Conclusion

We have illustrated how functions from the **copula** package can be used to successively test randomness, independence and goodness-of-fit in the construction of a copula model from data. The package provides not only a tool for pedagogical purposes but can also be used as a platform for real data analysis and as a test bed for new methodological developments.

Let us finally mention that its functionalities are currently being augmented by functions for manipulating extreme-value copulas and performing various kinds of tests related to these dependence structures.

References

Ben Ghorbal M, Genest C, Nešlehová J (2009). "On the Test of Ghoudi, Khoudraji, and Rivest for Extreme-Value Dependence." *The Canadian Journal of Statistics*, **37**(4), 534–552.

Berg D (2009). "Copula Goodness-of-Git Testing: An Overview and Power Comparison." *The European Journal of Finance*, **15**, 675–701.

Blum J, Kiefer J, Rosenblatt M (1961). "Distribution Free Tests of Independence Based on the Sample Distribution Function." *Annals of Mathematical Statistics*, **32**, 485–498.

Chen X, Fan Y (2005). "Pseudo-Likelihood Ratio Tests for Semiparametric Multivariate Mopula Model Selection." *Canadian Journal of Statistics*, **33**, 389–414.

Cherubini G, Vecchiato W, Luciano E (2004). *Copula Models in Finance*. John Wiley & Sons, New-York.

Deheuvels P (1979). "La fonction de dépendance empirique et ses propriétés: un test non paramétrique d'indépendance." *Bulletin de la Classe des Sciences, V. Série, Académie Royale de Belgique*, **65**, 274–292.

Deheuvels P (1981a). "An Asymptotic Decomposition for Multivariate Distribution-Free Tests of Independence." *Journal of Multivariate Analysis*, **11**, 102–113.

Deheuvels P (1981b). "A Non Parametric Test for Independence." *Publications de l'Institut de Statistique de l'Université de Paris*, **26**, 29–50.

Demarta S, McNeil A (2005). "The *t* Copula and Related Copulas." *International Statistical Review*, **73**(1), 111–129.

Dugué D (1975). "Sur les tests d'indépendance 'indépendants de la loi'." *Comptes rendus de l'Académie des Sciences de Paris, Série A*, **281**, 1103–1104.

Fermanian JD, Scaillet O (2005). "Some Statistical Pitfalls in Copula Modelling for Financial Applications." In E Klein (ed.), *Capital Formation, Governance and Banking*, pp. 59–74. Nova Science.

Fisher R (1932). *Statistical Methods for Research Workers*. Olivier and Boyd, London.

Frees E, Valdez E (1998). "Understanding Relationships Using Copulas." *North American Actuarial Journal*, **2**, 1–25.

Genest C, Favre AC (2007). "Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask." *Journal of Hydrological Engineering*, **12**, 347–368.

Genest C, Gendron M, Bourdeau-Brien M (2009a). "The Advent of Copulas in Finance." *European Journal of Finance*, **15**, 609–618.

Genest C, Ghoudi K, Rivest LP (1995). "A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions." *Biometrika*, **82**, 543–552.

Genest C, Quessy JF, Rémillard B (2006). "Goodness-of-Fit Procedures for Copulas Models Based on the Probability Integral Transformation." *Scandinavian Journal of Statistics*, **33**, 337–366.

Genest C, Quessy JF, Rémillard B (2007). "Asymptotic Local Efficiency of Cramér-von Mises Tests for Multivariate Independence." *The Annals of Statistics*, **35**, 166–191.

Genest C, Rémillard B (2004). "Tests of Independence and Randomness Based on the Empirical Copula Process." *Test*, **13**(2), 335–369.

Genest C, Rémillard B (2008). "Validity of the Parametric Bootstrap for Goodness-of-Fit Testing in Semiparametric Models." *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques*, **44**, 1096–1127.

Genest C, Rémillard B, Beaudoin D (2009b). "Goodness-of-Fit Tests for Copulas: A Review and a Power Study." *Insurance: Mathematics and Economics*, **44**, 199–213.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2009). ***mvtnorm**: Multivariate Normal and t Distribution.* R package version 0.9-7, URL http://CRAN.R-project.org/package=mvtnorm.

Ghoudi K, Kulperger R, Rémillard B (2001). "A Nonparametric Test of Serial Independence for Times Series and Residuals." *Journal of Multivariate Analysis*, **79**, 191–218.

Giacomini E, Härdle W, Spokoiny V (2009). "Inhomogeneous Dependence Modeling with Time Varying Copulae." *Journal of Business & Economic Statistics*, **27**(2), 224–234.

Grégoire V, Genest C, Gendron M (2008). "Using Copulas to Model Price Dependence in Energy Markets." *Energy Risk*, **5**(5), 58–64.

Hosking J (1980). "The Multivariate Portmanteau Statistic." *Journal of the American Statistical Association*, **75**(371), 602–608.

Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

Johansen S (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press, New York.

Kim G, Silvapulle M, Silvapulle P (2007). "Comparison of Semiparametric and Parametric Methods for Estimating Copulas." *Computational Statistics and Data Analysis*, **51**(6), 2836–2850.

Klugman S, Parsa R (1999). "Fitting Bivariate Loss Distributions with Copulas." *Insurance: Mathematics and Economics*, **24**, 139–148.

Kojadinovic I, Holmes M (2009). "Tests of Independence among Continuous Random Vectors Based on Cramér-von Mises Functionals of the Empirical Copula Process." *Journal of Multivariate Analysis*, **100**(6), 1137–1154.

Kojadinovic I, Yan J (2010a). "A Goodness-of-Fit Test for Multivariate Multiparameter Copulas Based on Multiplier Central Limit Theorems." *Statistics and Computing*. Forthcoming.

Kojadinovic I, Yan J (2010b). "Tests of Serial Independence for Continuous Multivariate Time Series Based on a Möbius Decomposition of the Independence Empirical Copula Process." *Annals of the Institute of Statistical Mathematics*. Forthcoming.

Kojadinovic I, Yan J, Holmes M (2010). "Fast Large-Sample Goodness-of-Fit for Copulas." *Statistica Sinica*. Forthcoming.

Ljung G, Box G (1978). "On a Measure of Lack of Fit in Time Series Models." *Biometrika*, **65**, 297–303.

McNeil A, Frey R, Embrechts P (2005). *Quantitative Risk Management*. Princeton University Press, New Jersey.

Nelsen R (2006). *An Introduction to Copulas*. 2nd edition. Springer–Verlag, New-York.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rémillard B, Scaillet O (2009). "Testing for Equality Between Two Copulas." *Journal of Multivariate Analysis*, **100**(3), 377–386.

Rota GC (1964). "On the Foundations of Combinatorial Theory. I. Theory of Möbius Functions." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **2**, 340–368.

Salvadori G, Michele CD, Kottegoda N, Rosso R (2007). *Extremes in Nature: An Approach Using Copulas*. Water Science and Technology Library, Vol. 56. Springer–Verlag.

Scaillet O (2005). "A Kolmogorov-Smirnov Type Test for Positive Quadrant Dependence." *Canadian Journal of Statistics*, **33**, 415–427.

Sklar A (1959). "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de Statistique de l'Université de Paris*, **8**, 229–231.

Tippett L (1931). *The Method of Statistics*. Williams and Norgate, London.

Yan J (2007). "Enjoy the Joy of Copulas: With a Package **copula**." *Journal of Statistical Software*, **21**(4), 1–21. URL http://www.jstatsoft.org/v21/i04/.

Yan J, Kojadinovic I (2010). ***copula**: Multivariate Dependence with Copulas*. R package version 0.9-5, URL http://CRAN.R-project.org/package=copula.

Affiliation:
Ivan Kojadinovic
Department of Statistics
The University of Auckland
Private Bag 92019
Auckland 1142, New Zealand
Email: ivan@stat.auckland.ac.nz
URL: http://www.stat.auckland.ac.nz/~ivan/

Jun Yan
Department of Statistics
University of Connecticut
215 Glenbrook Rd. Unit 4120
Storrs, CT 06269, United States of America
Email: jun.yan@uconn.edu
URL: http://www.stat.uconn.edu/~jyan/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 34, Issue 9 *Submitted:* 2009-07-07
May 2010 *Accepted:* 2010-04-23