A graphical goodness-of-fit test for dependence models in higher dimensions

Marius Hofert[1], Martin Mächler[2]

2012-06-05

Abstract

A graphical goodness-of-fit test for copulas in more than two dimensions is introduced. The test is based on pairs of variables and can thus be interpreted as a first-order approximation of the underlying dependence structure. The idea is to first transform pairs of data columns with Rosenblatt's transformation to bivariate standard uniform distributions under the null hypothesis. This hypothesis is then tested with a goodness-of-fit test for bivariate independence for each pair of variables and pairs that violate the null hypothesis can be identified. A global p-value can also be derived. Graphically, one can display a plot matrix of the Rosenblatt-transformed data with the p-values of the pairwise tests of independence encoded as background color. By allowing for transformations of each pair of Rosenblatt-transformed data columns, one can also depict the goodness of fit via Q-Q plots or P-P plots, for example. The presented graphical goodness-of-fit test is particularly suited to detect hierarchies since they are often derived from pairs of variables. Various examples are given and the methodology is applied to a financial data set. An implementation is provided by the R package `copula`.

Keywords
Copulas, graphical goodness-of-fit tests, test for bivariate independence, Rosenblatt transformation.

MSC2010
62H15, 62-09.

1 Introduction

A *copula* is a multivariate distribution function with standard uniform univariate margins. Copulas are an important tool to model dependencies between components of a random vector. In the realm of quantitative risk management, for example, these components are risk factor changes based on which losses are modeled; see McNeil et al. (2005). From a statistical point of view, it is thus important to have goodness-of-fit tests for copulas. Such tests have recently gained interest, see, for

[1]RiskLab, Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland, marius.hofert@math.ethz.ch. The author (Willis Research Fellow) thanks Willis Re for financial support while this work was being completed.
[2]Seminar für Statistik, ETH Zurich, 8092 Zurich, Switzerland, maechler@stat.math.ethz.ch.

example, Genest and Rivest (1993), Breymann et al. (2003), Fermanian (2005), Berg and Bakken (2006), Genest et al. (2006a), Berg and Bakken (2007), Dobrić and Schmid (2007), Berg (2009), Genest et al. (2009), and references therein. Graphical goodness-of-fit tests are rarely found in the literature, some references to work in this area are Fisher and Switzer (1985) and Genest and Boies (2003). The former authors investigate chi-plots, the latter so-called K-plots, that is, a Q-Q-like plot based on the Kendall distribution function usually denoted by K. For practical applications, the d-dimensional case (for $d > 2$) is mainly of interest. Although K-plots extend to this case, they always reduce the information of a d-dimensional sample to a one-dimensional distribution function making it difficult (if not impossible) to decide, for example, which pair of data columns do not follow the dependence given by the null hypothesis. Further, K-plots are only straightforward to apply when the Kendall distribution function under the null hypothesis takes on a simple form.

For goodness-of-fit tests of copulas in general, *Rosenblatt's transformation* is commonly applied; see Rosenblatt (1952). With the help of this transformation, (pseudo-)observations from the underlying copula model can be transformed to variates which follow, under the null hypothesis, a multivariate standard uniform distribution, hence are independent. The hypothesis of multivariate independence can be tested in different ways. In this work, we apply a goodness-of-fit test introduced by Genest and Rémillard (2004). Although both Rosenblatt's transformation and the goodness-of-fit test for multivariate independence apply to the general d-dimensional case ($d \geq 2$), it is typically numerically and computationally challenging (if not impossible) to compute them for realistic dependence models in large but also moderate dimensions such as $d = 10$ or more.

In this paper we suggest a graphical goodness-of-fit test for dependence models in higher dimensions based on Rosenblatt's transformation applied to all pairs of variables. In addition, we conduct tests for bivariate independence. This is inspired by the fact that multivariate data of dimension $d > 2$ is often depicted by scatter plot matrices, showing a matrix of scatter plots of pairs of data columns. This "bivariate" thinking is also predominant in the context of hierarchical copulas where two variables belonging to the same hierarchical level behave differently than two variables belonging to different levels. Besides the numerical and computational advantages, our suggested graphical goodness-of-fit test has the advantage of being able to identify pairs of variables that do not follow the tested dependence structure. This is important for practical applications in high dimensions since a high-dimensional model rarely fits given data perfectly and one is interested in finding the pairs of variables which statistically contradict the model assumption rather then computing a single p-value.

There are several possibilities to graphically display the result of our goodness-of-fit test. The basic framework consists of a matrix of pairwise plots where the background color of each panel is determined by the p-value of the corresponding test of independence. Each panel can then contain further information such as: a scatter plot of the Rosenblatt-transformed data; a Q-Q plot of the Rosenblatt-transformed data mapped to a χ^2 distribution; or none (in case the dimension is very large and one only wants to display the p-values of the tests for independence). By a combination of having a lot of information about the goodness-of-fit of the copula under the null hypothesis and having p-values as summary information, one can quickly obtain a graphical overview of the deviations from the null hypothesis for

given high-dimensional data.

The paper is organized as follows. Section 2 briefly provides some background information about goodness-of-fit tests for dependence models in general. In Section 3, we present the idea underlying our graphical goodness-of-fit test and present a procedure how it can be applied to high-dimensional data. Section 4 presents various examples including an application to historical data of the Swiss Market Index (SMI) constituents. Finally, Section 5 concludes.

Note that all examples, including the financial application, can be reproduced either by the supplementary material or the code in `demo(gof_graph)` from the R package `copula`.

2 Goodness-of-fit testing setup for copulas

Let $\boldsymbol{X} = (X_1, \dots, X_d)^\top$, $d \geq 2$, denote a random vector with distribution function H and continuous marginal distribution functions $F_1, \dots, F_d$. By Sklar's Theorem, see, for example, Sklar (1996), there exists a unique copula C which couples $F_1, \dots, F_d$ with H, that is,

$$H(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)),\ \boldsymbol{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^d.$$

In a copula model for $\boldsymbol{X}$, one usually would like to know if C is well-represented by a specific copula C_{H_0}. In other words, one wants to test the null hypothesis

$$H_0 : C = C_{H_0}$$

based on realizations of independent and identically distributed copies $\boldsymbol{X}_i$, $i \in \{1, \dots, n\}$, of $\boldsymbol{X}$. In practical applications, the copula C_{H_0} usually arises from a parametric family of copulas with estimated parameter vector. For testing H_0, the marginal distributions are treated as unknown nuisance parameters and are replaced by their scaled empirical counterparts, the *pseudo-observations* $\hat{\boldsymbol{U}}_i = (\hat{U}_{i1}, \dots, \hat{U}_{id})^\top$, $i \in \{1, \dots, n\}$, with

$$\hat{U}_{ij} = \frac{n}{n+1} \hat{F}_{nj}(X_{ij}) = \frac{R_{ij}}{n+1},\ i \in \{1, \dots, n\},\ j \in \{1, \dots, d\}, \tag{1}$$

where X_{ij} denotes the jth component of $\boldsymbol{X}_i$, $\hat{F}_{nj}(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}\{X_{kj} \leq x\}$ denotes the empirical distribution function of the jth data column (the data matrix consisting of the entries X_{ij}, $i \in \{1, \dots, n\}$, $j \in \{1, \dots, d\}$), and R_{ij} denotes the rank of X_{ij} among X_{kj}, $k \in \{1, \dots, n\}$. The rank-based pseudo-observations are interpreted as observations of C (besides the known issues of this interpretation, that is, for any $j \in \{1, \dots, d\}$, the pseudo-observations are neither independent nor perfectly following a univariate standard uniform distribution; see, for example, Genest et al. (2009)) and are therefore used to test H_0. In the following, we consider a test based on Rosenblatt's transform. For other goodness-of-fit tests, see Genest et al. (2009).

3 A graphical goodness-of-fit test in higher dimensions

One way of testing H_0 is based on Rosenblatt's transformation; for others, see Genest et al. (2009). This transformation, introduced by Rosenblatt (1952), can be used to obtain standard uniform random vectors $\boldsymbol{U}_i'$, $i \in \{1, \dots, n\}$, from given random vectors $\boldsymbol{U}_i$, $i \in \{1, \dots, n\}$, from C. One can thus test H_0 by testing the Rosenblatt-transformed random vectors $\boldsymbol{U'}$, $i \in \{1, \dots, n\}$, for multivariate standard uniformity. As mentioned before (including possible limitations of this approach), in practical applications where the margins are unknown, the $\boldsymbol{U}_i$'s are replaced by the pseudo-observations $\hat{\boldsymbol{U}}_i$, $i \in \{1, \dots, n\}$, and H_0 is tested based on the Rosenblatt-transformed pseudo-observations $\hat{\boldsymbol{U}}_i'$, $i \in \{1, \dots, n\}$.

Consider a representative d-dimensional random vector $\boldsymbol{U}$ distributed according to the copula C. To obtain a standard uniform random vector $\boldsymbol{U'}$ on $[0,1]^d$, *Rosenblatt's transformation* $R : \boldsymbol{U} \to \boldsymbol{U'}$ is given by

$$\begin{aligned} U_1' &= U_1, \\ U_2' &= C_2(U_2 \,|\, U_1), \\ &\vdots \\ U_d' &= C_d(U_d \,|\, U_1, \dots, U_{d-1}), \end{aligned}$$

where for $j \in \{2, \dots, d\}$, $C_j(u_j \,|\, u_1, \dots, u_{j-1})$ denotes the conditional distribution function of U_j given $U_1 = u_1, \dots, U_{j-1} = u_{j-1}$. This bijection is quite general in that it applies to any copula. The drawback is that the conditional copula functions are often difficult to obtain. The typical approach (assuming C admits continuous partial derivatives with respect to the first $d-1$ arguments) is to compute $C_j(u_j \,|\, u_1, \dots, u_{j-1})$, $j \in \{2, \dots, d\}$, via (see Schmitz (2003, p. 20))

$$C_j(u_j \,|\, u_1, \dots, u_{j-1}) = \frac{\mathrm{D}_{j-1,\dots,1}\, C^{(1,\dots,j)}(u_1, \dots, u_j)}{\mathrm{D}_{j-1,\dots,1}\, C^{(1,\dots,j-1)}(u_1, \dots, u_{j-1})}, \; j \in \{2, \dots, d\}, \tag{2}$$

where $C^{(1,\dots,k)}$ denotes the k-dimensional marginal copula of C corresponding to the first k arguments and $\mathrm{D}_{j-1,\dots,1}$ denotes the mixed partial derivative of order $j-1$ with respect to the first $j-1$ arguments. The problem when applying (2) in large dimensions is that it is often difficult to compute the high-order derivatives, the price one has to pay for such a general transformation. Furthermore, numerically evaluating the derivatives can be time-consuming and prone to errors so this is often not a feasible approach either.

3.1 The basic idea

For a pair $(U_j, U_k)^\top$, Rosenblatt's transformation based on the H_0 copula C_{H_0} is given by (for simplicity, we drop the index $j = 2$ of the conditional copula function in this case)

$$\begin{aligned} U_j' &= U_j, \\ U_k' &= C_{H_0}(U_k | U_j) = \mathrm{D}_1\, C_{H_0}(U_j, U_k), \end{aligned}$$

where the last equality holds almost surely without any further assumption on the copula of $(U_j, U_k)^\top$; see Schmitz (2003, p. 17). This *pairwise Rosenblatt transform* is typically easy to compute for the copula at hand, see Section 4 for some examples. It can be stored in an $n \times d \times d$ array, where n denotes the sample size and d the dimension. The entry $(\cdot, k, j)$ of this array contains $C_{H_0}(U_k|U_j)$ for $k \neq j$ and U_k for $k = j$.

Similar to a first-order approximation, our suggested graphical goodness-of-fit test utilizes a *plot matrix*, that is, a matrix of (pairwise) plots, in order to plot the pairwise Rosenblatt transformed data contained in this array. One can then visually check whether the (j,k)th pair of data columns

$$(U'_{ij}, U'_{ik})^\top = (U_{ij}, C_{H_0}(U_{ik} \,|\, U_{ij}))^\top,\ i \in \{1, \ldots, n\}, \tag{3}$$

can be viewed as realizations of the bivariate independence copula, hence whether the (j,k)th margin of C is the one of the H_0 copula C_{H_0}. Let us remark that such a scatter plot matrix is not necessarily symmetric (that is, the (j,k)th plot is not necessarily equal to the (k,j)th plot in the matrix) since the Rosenblatt-transformed data in (3) may be different from

$$(U'_{ik}, U'_{ij})^\top = (U_{ik}, C_{H_0}(U_{ij} \,|\, U_{ik}))^\top,\ i \in \{1, \ldots, n\}.$$

Instead of just plotting the pairs (3) in the (j,k)th entry of the plot matrix, one can also provide other useful graphical illustrations. For this, consider the data (3) corresponding to the pair of variables with indices k and j. Now let us introduce transformations

$$g_1 : [0,1]^n \to [0,1]^n \quad \text{and} \quad g_2 : [0,1]^{n \times 2} \to [0,1]^n.$$

The idea is to display in the (j,k)th entry of the plot matrix the points (viewed as an $n \times 2$ matrix)

$$\Big(g_1(U'_{1j}, \ldots, U'_{nj}), g_2((U'_{1j}, U'_{1k}), \ldots, (U'_{nj}, U'_{nk}))\Big) \tag{4}$$

for suitable choices of g_1 and g_2. In particular, the following choices of g_1, g_2 turn out to be useful.

Scatter plots
By choosing

$$\begin{aligned} g_1(u'_1, \ldots, u'_n) &= (u'_1, \ldots, u'_n)^\top, \\ g_2((u'_1, v'_1), \ldots, (u'_n, v'_n)) &= (v'_1, \ldots, v'_n)^\top, \end{aligned}$$

it follows from (4) that this transformation will just lead to a scatter plot of the data given in (3).

Q-Q plots
Q-Q plots are well-known as powerful tools for graphical goodness-of-fit assessment. They are typically simpler to interpret visually than scatter plots, especially if the

number of observations is either small or very large. Choosing

$$\begin{aligned} g_1(u_1', \dots, u_n') &= (F_{\chi_2^2}^{-1}(p_1), \dots, F_{\chi_2^2}^{-1}(p_n))^\top, \\ g_2((u_1', v_1'), \dots, (u_n', v_n')) &= \Big(\Big((\Phi^{-1}(u_i')^2 + \Phi^{-1}(v_i')^2)_{i=1}^n \Big)_{(1)}, \dots, \\ &\qquad \Big((\Phi^{-1}(u_i')^2 + \Phi^{-1}(v_i')^2)_{i=1}^n \Big)_{(n)} \Big)^\top \end{aligned}$$

will lead to a matrix of pairwise Q-Q plots based on the chi-square distribution with two degrees of freedom and corresponding distribution function $F_{\chi_2^2}$. Here, $p_1, \dots, p_n$ are suitable points in $(0,1)$, $p_i \approx i/n$ but symmetric inside $(0,1)$, as determined by the function `ppoints` in R. Furthermore, Φ^{-1} denotes the quantile function of the standard normal distribution and $w_{(1)} \le w_{(2)} \le \cdots \le w_{(n)}$ are the order statistics of $(w_1, w_2, \dots, w_n)^\top$.

Similarly, by using

$$\begin{aligned} g_1(u_1', \dots, u_n') &= (F_{\Gamma_2}^{-1}(p_1), \dots, F_{\Gamma_2}^{-1}(p_n))^\top, \\ g_2((u_1', v_1'), \dots, (u_n', v_n')) &= \Big(\Big((-\log(u_i') - \log(v_i'))_{i=1}^n \Big)_{(1)}, \dots, \\ &\qquad \Big((-\log(u_i') - \log(v_i'))_{i=1}^n \Big)_{(n)} \Big)^\top, \end{aligned}$$

where $F_{\Gamma_2}^{-1}$ denotes the quantile function of a $\Gamma(2,1)$ distribution (that is, a Gamma distribution with shape parameter 2 and scale parameter 1), one can build a matrix of Q-Q plots based on this distribution.

P-P plots

Another choice is

$$\begin{aligned} g_1(u_1', \dots, u_n') &= (p_1, \dots, p_n)^\top, \\ g_2((u_1', v_1'), \dots, (u_n', v_n')) &= \Big(F_{\chi_2^2}\Big(\Big((\Phi^{-1}(u_i')^2 + \Phi^{-1}(v_i')^2)_{i=1}^n \Big)_{(1)} \Big), \dots, \\ &\qquad F_{\chi_2^2}\Big(\Big((\Phi^{-1}(u_i')^2 + \Phi^{-1}(v_i')^2)_{i=1}^n \Big)_{(n)} \Big) \Big)^\top, \end{aligned}$$

which leads to a P-P plot based on the transformation via a chi-square distribution with two degrees of freedom.

Similarly, using

$$\begin{aligned} g_1(u_1', \dots, u_n') &= (p_1, \dots, p_n)^\top, \\ g_2((u_1', v_1'), \dots, (u_n', v_n')) &= \Big(F_{\Gamma_2}\Big(\Big((-\log(u_i') - \log(v_i'))_{i=1}^n \Big)_{(1)} \Big), \dots, \\ &\qquad F_{\Gamma_2}\Big(\Big((-\log(u_i') - \log(v_i'))_{i=1}^n \Big)_{(n)} \Big) \Big)^\top \end{aligned}$$

leads to a P-P plot based on the transformation via a $\Gamma(2,1)$ distribution.

3.2 A test for bivariate independence

Additionally to the information plotted, we further add a summary information to the (j,k)th panel plot background given by the p-value of a test of bivariate

independence based on the data given in (3). The $d \times d$ matrix of p-values (with empty diagonal) is transformed to colors which are then used as background colors for the panels. The colors can be carefully chosen to mainly reflect p-values below the chosen significance level; see Section 4 for example illustrations.

The test of bivariate independence we utilize is suggested by Genest and Rémillard (2004); see also Genest et al. (2006b) for a theoretical treatment of its properties and Kojadinovic and Yan (2010) for an implementation in the R package `copula`. In our setup, the test statistic for the (j, k)th pair is given by

$$I_n^{(j,k)} = \int_{[0,1]} \int_{[0,1]} n(C_n^{(j,k)}(u_1, u_2) - u_1 u_2)^2 \, du_1 \, du_2,$$

where $C_n^{(j,k)}$ denotes the empirical copula based on the Rosenblatt transformed pseudo-observations $(U'_{ij}, U'_{ik})^\top$, $i \in \{1, \ldots, n\}$, that is, the empirical distribution function based on these pseudo-observations. First, the distribution of $I_n^{(\cdot,\cdot)}$ is simulated under independence (this has to be done only once, not for all pairs, of course). Then, for each pair of data columns $(U'_{ij}, U'_{ik})^\top$, $i \in \{1, \ldots, n\}$, an approximate p-value for the test of bivariate independence based on the simulated (empirical) distribution of $I_n^{(\cdot,\cdot)}$ is computed.

The panels in the plot matrix become smaller and smaller the larger the dimension d is. For high-dimensional data, instead of displaying any points from transformations g_1 and g_2, one can just graphically display a small square containing the corresponding background color. In the limit (for d large), this leads to an image plot, where each pixel is colorized according to the corresponding p-value of the test of bivariate independence of the Rosenblatt transformed pseudo-observations.

3.3 A global p-value

Although we suggest a *graphical* goodness-of-fit test to identify pairs of variables that do not follow the dependence structure specified by the null hypothesis, the question arises how a global p-value could be obtained from the $d(d-1)$ pairwise tests above. The p-values themselves need to be adjusted for multiple testing and we propose to take the minimum of the adjusted pairwise p-values as global p-value. There are several methods available to correct for multiple comparisons, and we use those available by R function `p.adjust()`, which includes proposals of Bonferroni, Holm, Hochberg, Benjamini, etc, see the function documentation. Together with the graphical goodness-of-fit test described above, one obtains a good overview about the dependence structure between (pairs of) the variables; see the following section for examples.

3.4 A word concerning the technical tools in the background

The main idea behind the plot function to create our graphical goodness of fit test is based on the scatter plot matrix as implemented in R function `pairs()`, specifically its default method `pairs.default`. We create a matrix of plots, however, due to the placement of the title, sub-title, and the color key (implemented similar to the color key in `filled.contour`), we use the function `layout` to create the skeleton of plots. Another difference to `pairs()` is that the (j, k)th entry of the plot matrix is

typically different from the (k, j)th after interchanging the axes. The reason for this is that flipping the two columns

$$(U'_{ik}, U'_{ij})^\top = (U_{ik}, C_{H_0}(U_{ij} \,|\, U_{ik}))^\top,\ i \in \{1, \dots, n\}, \tag{5}$$

is not necessarily equal to (3). In the plot matrix we create, we thus opt for the columns to display the same variable. Therefore, all x values of the plots in a fixed column are equal, whereas the y values differ. We indicate this by the simplified notation $\cdot \,|\, \boldsymbol{u}_k$ on the diagonal, meaning that the corresponding kth column is based on $\boldsymbol{u}'_k$ as conditioning variable, and all plots in the kth column are against $g_1(\boldsymbol{u}'_k)$.

Additionally, all p values are drawn (as "ticks") on the left side of the color key. For solutions to the more challenging technical details we solved, we refer the interested reader to the R package `copula` which contains all the necessary code and documentation, currently via `demo(gof_graph)` (search for `JCGS`).

4 Examples

In this section we consider several examples and an application of our suggested graphical goodness-of-fit test. In practical applications, the Gaussian and t copulas (as members of the class of elliptical copulas) and the Clayton, Frank, and Gumbel copulas (as members of the class of Archimedean copulas) frequently appear in various different contexts. Recently, hierarchical copulas have become of interest. For elliptical copulas, hierarchies can be modeled through a block correlation matrix so that pairs belonging to the same hierarchical level share the same entry. For Archimedean copulas, their corresponding nested versions (obtained by plugging in Archimedean copulas into each other) are able to model hierarchies. In the following, we pick out some examples that demonstrate how our graphical goodness-of-fit test can be applied. We start with the technical tools needed to compute the pairwise Rosenblatt transforms.

4.1 Conditional copula functions

One of the most well-known copulas is the (bivariate) *Gaussian copula*

$$\begin{aligned} C(u_1, u_2) &= \Phi_P(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \\ &= \int_{-\infty}^{\Phi^{-1}(u_2)} \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right) dx_1\, dx_2, \end{aligned}$$

where Φ_P is the distribution function of the bivariate normal distribution with mean vector zero and 2×2 correlation matrix P with off-diagonal entry $\rho \in [0, 1)$, and Φ^{-1} is the quantile function of the univariate standard normal distribution. For this copula, we have

$$\mathrm{D}_1\, C(u_1, u_2) = \Phi\left(\frac{\Phi^{-1}(u_2) - \rho\, \Phi^{-1}(u_1)}{\sqrt{1-\rho^2}}\right).$$

Another important example is the (bivariate) t_ν *copula* (t copula with ν degrees of freedom), given by

$$\begin{aligned} C(u_1, u_2) &= \mathrm{t}_{\nu,P}(\mathrm{t}_\nu^{-1}(u_1), \mathrm{t}_\nu^{-1}(u_2)) \\ &= \int_{-\infty}^{\mathrm{t}_\nu^{-1}(u_2)} \int_{-\infty}^{\mathrm{t}_\nu^{-1}(u_1)} \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\pi\nu\sqrt{1-\rho^2}} \left(1 + \frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{\nu(1-\rho^2)}\right)^{-\frac{\nu+2}{2}} dx_1\, dx_2, \end{aligned}$$

where $\mathrm{t}_{\nu,P}$ is the distribution function of the bivariate t distribution with ν degrees of freedom, mean vector zero, and dispersion matrix P (a 2×2 correlation matrix P with off-diagonal entry $\rho \in [0, 1)$) and t_ν^{-1} is the quantile function of the univariate t distribution with ν degrees of freedom. For this copula, we have

$$\mathrm{D}_1 C(u_1, u_2) = \mathrm{t}_{\nu+1}\left(\frac{\mathrm{t}_\nu^{-1}(u_2) - \mu}{\sigma}\right), \quad \text{where } \mu = \rho\, \mathrm{t}_\nu^{-1}(u_1), \quad \sigma^2 = (1-\rho^2)\frac{\nu + \mathrm{t}_\nu^{-1}(u_1)^2}{\nu+1}.$$

For Archimedean copulas with differentiable generator ψ (see, for example, McNeil and Nešlehová (2009)), we have

$$\mathrm{D}_1 C(u_1, u_2) = \frac{\psi'(\psi^{-1}(u_1) + \psi^{-1}(u_2))}{\psi'(\psi^{-1}(u_1))},$$

which is usually straightforward to evaluate for a given parametric Archimedean family.

Since our suggested graphical goodness-of-fit test is based on pairs of copulas only, it is especially simple to apply to hierarchical elliptical or Archimedean copulas. The reason for this is that all pairs of (hierarchical) elliptical copulas are bivariate elliptical copulas and all pairs of nested Archimedean copulas are Archimedean copulas again. We can therefore apply the above results in order to compute our graphical goodness-of-fit test.

4.2 Five-dimensional Gumbel and nested Gumbel copula

For the first example, we simulate $n = 1000$ realizations of a Gumbel copula in $d = 5$ dimensions with parameter $\theta = 2$ corresponding to a Kendall's tau of $\tau = 0.5$. We then build the pseudo-observations and compute the pairwise Rosenblatt transformed data under

$$H_0 : C \text{ is a Gumbel copula with (the true) parameter } \theta = 2.$$

Afterwards, we apply the pairwise test of independence to compute a matrix of p-values. This matrix is converted to colors (with orange to white colors for p-values above the chosen significance level 0.05 and blue to red colors for p-values below it). Figure 1 (left-hand side) displays the result with the pairwise Rosenblatt transformed pseudo-observations as scatter plots and the background colors as computed from the pairwise tests of independence. The global p-values for the different methods as provided by `p.adjust` are shown in the sub-title.

Next, we sample ($n = 1000$ and $d = 5$ as before) a nested Archimedean copula of the form

$$C(\boldsymbol{u}) = C_0(C_1(u_1, u_2), C_2(u_3, u_4, u_5)), \tag{6}$$

where C_k is a Gumbel copula with parameter θ_k, $k \in \{0, 1, 2\}$. As parameters we choose $\theta_0 = 5/4$, $\theta_1 = 5/3$, and $\theta_2 = 5/2$, corresponding to Kendall's tau of $\tau_0 = 0.2$, $\tau_1 = 0.4$, and $\tau_2 = 0.6$, respectively. We then compute the pseudo-observations and the pairwise Rosenblatt transformed data, but this time under

$$H_0 : C \text{ is a nested Gumbel copula of the form as in (6)}$$
$$\text{with parameters } \theta_0 = 5/4 \text{ and } \theta_1 = \theta_2 = 5/3.$$

This means, we choose the true parameters for θ_0 and θ_1, but a different than the true value for θ_2. The result is clearly visible both from the scatter plots and the p-values (background colors) on the right-hand side of Figure 1.

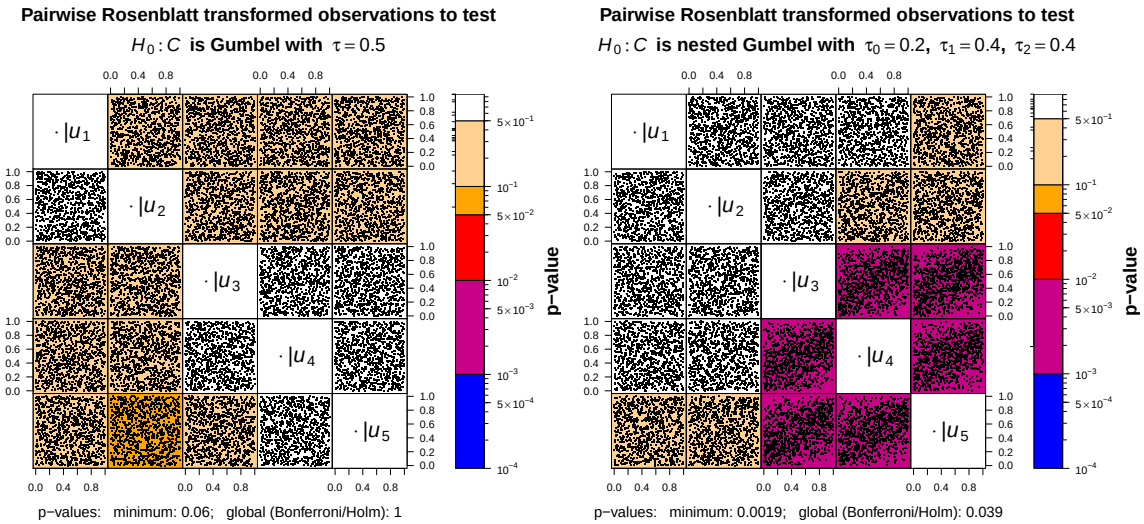

Figure 1: Graphical goodness-of-fit test based on pairwise Rosenblatt transformed pseudo-observations for 1000 simulated samples of five-dimensional Gumbel (left-hand side) and nested Gumbel (right-hand side) copulas.

4.3 A five-dimensional estimated t_4 copula

We now consider a five-dimensional t copula with four degrees of freedom. Similarly as before, we simulate $n = 1000$ realizations of this copula for the dispersion matrix P (a correlation matrix)

$$P = \begin{pmatrix} 1 & \rho_1 & \rho_0 & \rho_0 & \rho_0 \\ \rho_1 & 1 & \rho_0 & \rho_0 & \rho_0 \\ \rho_0 & \rho_0 & 1 & \rho_2 & \rho_2 \\ \rho_0 & \rho_0 & \rho_2 & 1 & \rho_2 \\ \rho_0 & \rho_0 & \rho_2 & \rho_2 & 1 \end{pmatrix},$$

where $\rho_0 = 0.3090$, $\rho_1 = 0.5878$, and $\rho_2 = 0.8090$, corresponding to Kendall's tau of $\tau_0 = 0.2$, $\tau_1 = 0.4$, and $\tau_2 = 0.6$, respectively (note that $\tau = \frac{2}{\pi} \arcsin \rho$). We then build the pseudo-observations as before. This time, the pairwise Rosenblatt transformed data is computed under

$H_0 : C$ is a t_4 copula with dispersion matrix P being estimated
with the pairwise Kendall's tau estimator; see Demarta and McNeil (2005).

This means that we first estimate the entries of P by pairwise inverting Kendall's

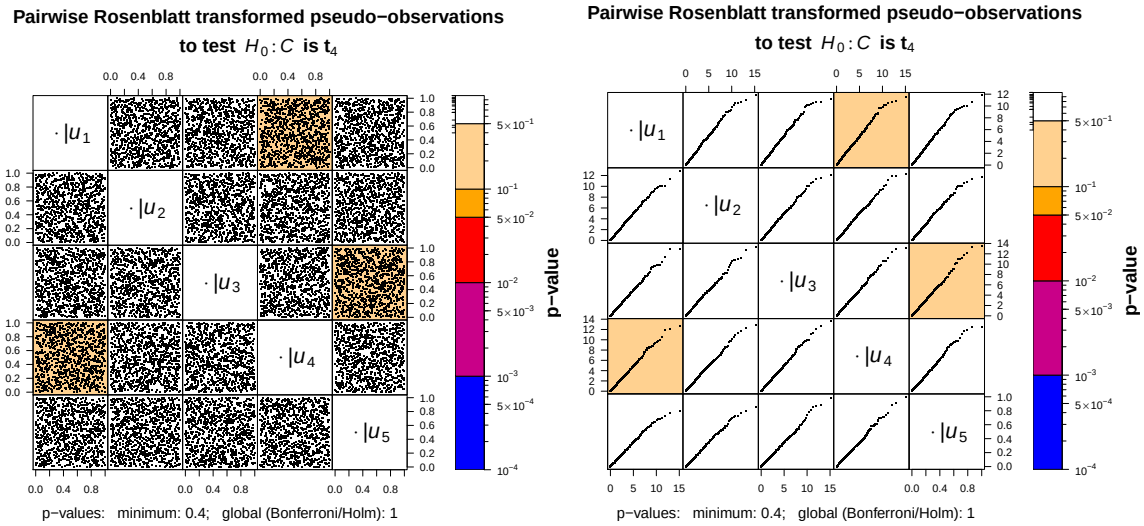

Figure 2: Graphical goodness-of-fit test for an estimated t_4 copula with the identity transformation leading to pairwise scatter plots (left) and the transformation to pairwise Q-Q plots (right).

tau (so the (j,k)th entry ρ_{jk} of P is estimated by $\hat{\rho}_{n,jk} = \sin(\frac{\pi}{2}\hat{\tau}_{n,jk})$, where $\hat{\tau}_{n,jk}$ denotes the sample version of Kendall's tau for the data column (j,k)) then adjust this estimated matrix $\hat{P} = (\hat{\rho}_{n,jk})_{j,k}$ to be positive definite (this is accomplished via function `nearPD` from R package `Matrix`), and finally apply the pairwise Rosenblatt transformation for a t_4 copula with this adjusted, estimated dispersion matrix. As before, the pairwise test of independence is computed to determine the background colors. Figure 2 displays the result.

4.4 A financial data example

In this example, we consider all $d = 20$ constituents of the Swiss Market Index (SMI), with Yahoo finance (http://finance.yahoo.com) ticker symbols ABBN.VX, ATLN.VX, ADEN.VX, CSGN.VX, GIVN.VX, HOLN.VX, BAER.VX, NESN.VX, NOVN.VX, CFR.VX, ROG.VX, SGSN.VX, UHR.VX, SREN.VX, SCMN.VX, SYNN.VX, SYST.VX, RIGN.VX, UBSN.VX, ZURN.VX.

We build log-returns of close-day prices of all constituents from 2011-09-09 to 2012-03-28 ($n = 141$ trading days); in this period, all constituents in the index remained the same. Figure 3 shows the corresponding pseudo-observations of the log-returns which clearly indicate dependencies between the SMI constituents.

We estimate t copulas to the pseudo-observations of the log-returns with the approach as described above. After estimating the dispersion matrix P, we determine the degree of freedom parameter ν by its maximum likelihood estimator; see Demarta and McNeil (2005). Since the likelihood is increasing in ν, we go for

H_0 : C is a Gaussian copula with correlation matrix P being estimated with the pairwise Kendall's tau estimator.

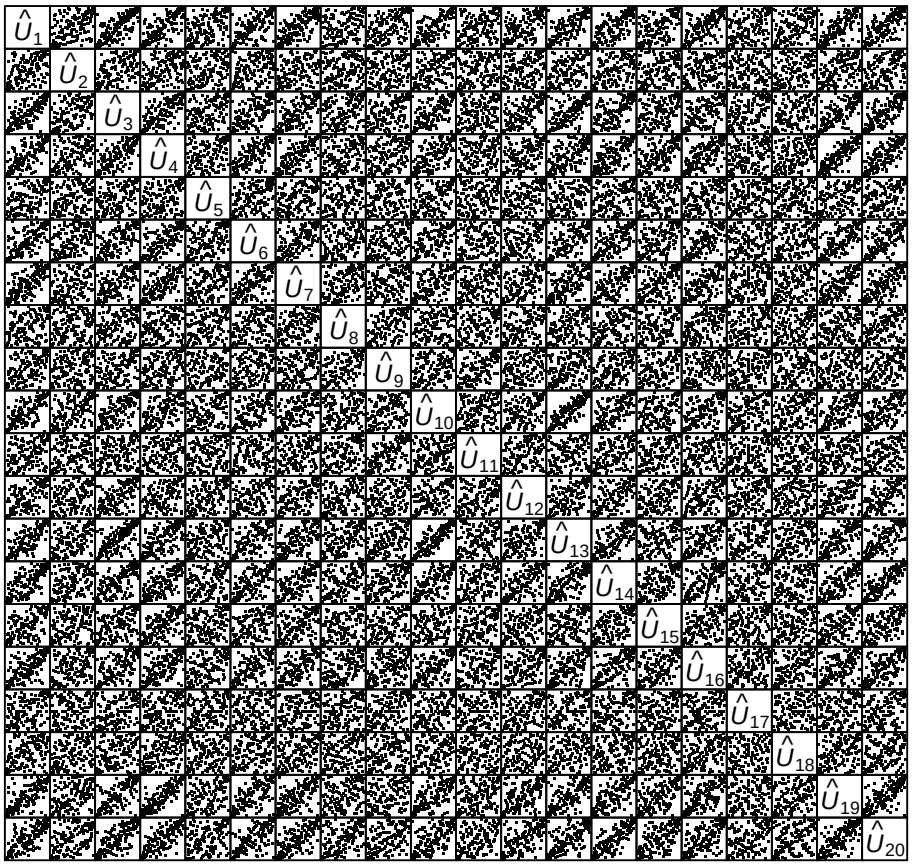

Figure 3: Scatter plot matrix of the pseudo-observations of the log-returns for the twenty constituents of the Swiss Market Index (SMI) from 2011-09-09 to 2012-03-28.

As before, we use `nearPD()` to make the matrix of pairwise estimated correlation coefficients positive definite. Computing our graphical goodness-of-fit test leads to Figure 4. No pair of variables was rejected according to the 0.05 significance level.

Finally, let us remark that the p-value of the test of multivariate normality of the log-returns themselves as provided by the R function `mshapiro.test` from the package `mvnormtest` led to $p \leq 2.2 \cdot 10^{-16}$. Although the dependence structure of the multivariate normal distribution (that is, the Gaussian copula) seems to be an adequate model for the dependence between the SMI constituents over the given time horizon, the joint distribution (hence the dependence structure and the marginal distributions) is clearly not multivariate normal.

5 Conclusion

We presented a graphical goodness-of-fit test for dependence models in higher dimensions. This test can help to discover pairs of observations with deviations from the null hypothesis. For this, we introduced a plot matrix of scatter, Q-Q, or P-P plots (more types of plots can easily be constructed) of pairwise Rosenblatt-transformed

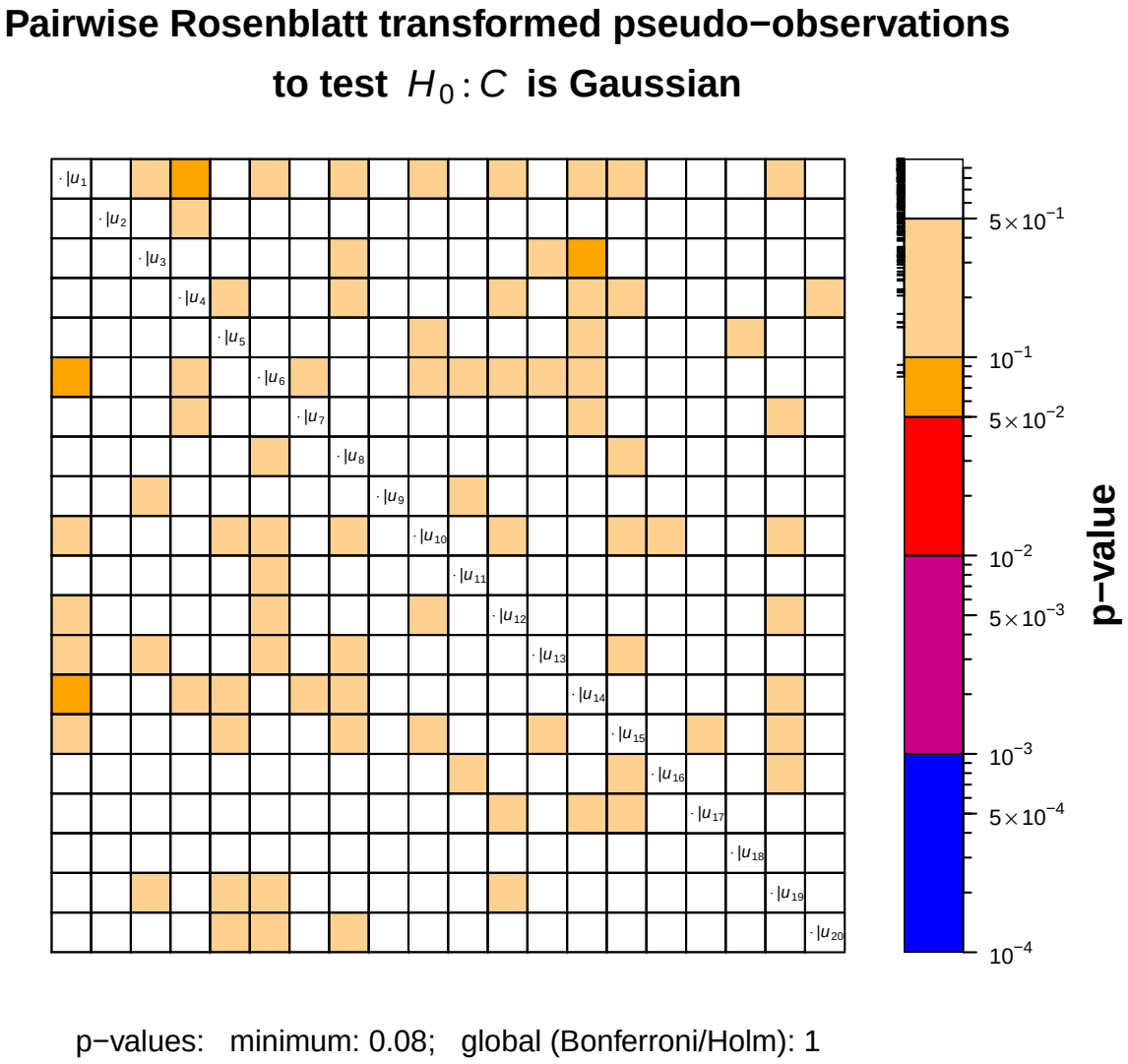

Figure 4: Graphical goodness-of-fit test based on pairwise Rosenblatt transformed pseudo-observations of the log-returns for the twenty constituents of the SMI.

(pseudo-)observations. The plot matrix is further enhanced by visualizing deviations from the null hypothesis by colors derived from the (pairwise) p-values of a non-parametric test of independence. This allows one to easily detect pairs showing large deviations from the null hypothesis. A global p-value can also be obtained. The suggested graphical test is particularly useful when the data is high-dimensional since a graphical assessment may be more meaningful in this case than just a single p-value. An implementation is provided via by the R package `copula` and detailed examples including an application to financial data were given.

Acknowledgements

We would like to thank Niels Hagenbuch (Karolinska Institutet) for detailed feedback on this work and Yongsheng Wang for help with the data example.

Supplementary materials

R package `copula`: Version 0.99-2 (source; May 2012) of the R package `copula` which provides the base functionality required for reproducing the presented graphical goodness-of-fit tests (copula_0.99-2.tar.gz).

Reproducing R script: R script reproducing the computations and figures of this paper (ggraph.R; obtained by `R CMD Stangle ggraph.Rnw`).

Auxiliary R code: Three auxiliary R scripts `source()`d by ggraph.R (wrapper.R, ggraph-tools.R, ggraph-graphics.R).

financial data: R data file containing the SMI data; use `load()` (SMI_yh.rda).

References

Berg, D. (2009), Copula goodness-of-fit testing: an overview and power comparison, *The European Journal of Finance*, http://www.informaworld.com/10.1080/13518470802697428 (2009-03-25).

Berg, D. and Bakken, H. (2006), Copula Goodness-of-fit Tests: A Comparative Study, http://www.danielberg.no/publications/CopulaGOF.pdf (2008-11-01).

Berg, D. and Bakken, H. (2007), A copula goodness-of-fit approach based on the conditional probability integral transformation, http://www.danielberg.no/publications/Btest.pdf (2008-11-01).

Breymann, W., Dias, A., and Embrechts, P. (2003), Dependence structures for multivariate high-frequency data in finance, *Quantitative Finance*, 3, 1–14.

Demarta, S. and McNeil, A. J. (2005), The t Copula and Related Copulas, *International Statistical Review*, 73.1, 111–129.

Dobrić, J. and Schmid, F. (2007), A goodness of fit test for copulas based on Rosenblatt's transformation, *Computational Statistics & Data Analysis*, 51, 4633–4642.

Fermanian, J.-D. (2005), Goodness of fit tests for copulas, *Journal of Multivariate Analysis*, 95.1, 119–152.

Fisher, N. I. and Switzer, P. (1985), Chi-plots for assessing dependence, *Biometrika*, 72.2, 253–265.

Genest, C. and Boies, J.-C. (2003), Detecting dependence with Kendall plots, *The American Statistician*, 57, 275–284.

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

Genest, C., Quessy, J. F., and Rémillard, B. (2006b), Local efficiency of a Cramér–von Mises test of independence, *Journal of Multivariate Analysis*, 97, 274–294.

Genest, C. and Rémillard, B. (2004), Tests of Independence and Randomness Based on the Empirical Copula Process, *TEST*, 13.2, 335–369.

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

Genest, C. and Rivest, L.-P. (1993), Statistical Inference Procedures for Bivariate Archimedean Copulas, *Journal of the American Statistical Association*, 88.423, 1034–1043.

Kojadinovic, I. and Yan, J. (2010), Modeling Multivariate Distributions with Continuous Margins Using the copula R Package, *Journal of Statistical Software*, 34.9, 1–20.

McNeil, A. J., Frey, R., and Embrechts, P. (2005), *Quantitative Risk Management: Concepts, Techniques, Tools*, Princeton University Press.

McNeil, A. J. and Nešlehová, J. (2009), Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions, *The Annals of Statistics*, 37.5b, 3059–3097.

Rosenblatt, M. (1952), Remarks on a Multivariate Transformation, *The Annals of Mathematical Statistics*, 23.3, 470–472.

Schmitz, V. (2003), Copulas and Stochastic Processes, PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen.

Sklar, A. (1996), Random variables, distribution functions, and copulas – a personal look backward and forward, *Distributions with Fixed Marginals and Related Topics*, 28, 1–14.