A Copula Goodness-of-fit Test Based on the Probability Integral Transform

Daniel Berg
daniel@danielberg.no

University of Oslo / Norwegian Computing Center

21st Nordic Conference on Mathematical Statistics
Rebild, Denmark, 14th June 2006

Outline

- 1. Introduction
- 2. Copula - Definitions and Theorems
- 3. Copula goodness-of-fit testing
  - 3.1. Probability integral transform
  - 3.2. Breymann, Dias and Embrecht's approach G
  - 3.3. New approach B
- 4. Power results from simulation
- 5. Application to daily return data
- 6. Summary

Introduction

- ▷ Copulae - a popular and flexible way of modelling dependence
- ▷ Copula choice may have huge impacts on e.g. capital allocation
- ▷ Is the data appropriately modelled by a given parametric copula?
- ▷ We propose a new copula goodness-of-fit approach.

2. Copula - Definitions and Theorems

Definition (Copula)

A d-dimensional copula is a multivariate distribution, $\mathcal{C}$, with standard uniform marginal distributions.

Theorem (Sklar)

Every multivariate distribution F, with margins, $F_1, F_2, \ldots, F_d$ can be written as

$$F(x_1, \ldots, x_d) = \mathcal{C}(F_1(x_1), \ldots, F_d(x_d)), \quad (2.1)$$

for some copula $\mathcal{C}$.

2. Copula - Definitions and Theorems

- Given a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ the copula of their joint distribution function may be extracted from equation (2.1):

$$\mathcal{C}(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)),$$

where the F_i^{-1}'s are the quantile functions of the margins.

- The copula is often represented by its density function $c(\mathbf{u})$:

$$\mathcal{C}(\mathbf{u}) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d) = \int_0^{u_1} \cdots \int_0^{u_d} c(\mathbf{u}) \mathrm{d}\mathbf{u},$$

2. Copula - Definitions and Theorems

- For the implicit copula of an absolutely continuous joint df F with strictly continuous marginal df's $F_1, \ldots, F_d$, the copula density is given by

$$c(\boldsymbol{u}) = \frac{f(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdots f_d(F_d^{-1}(u_1))}.$$

- Hence,

$$c(F_1(x_1), \ldots, F_d(x_d)) = \frac{f(x_1, \ldots, x_d)}{f_1(x_1) \cdots f_d(x_d)}.$$

- This means that a general d-dimensional density can be written as

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \cdot f_1(x_1) \cdots f_d(x_d)$$

for some copula density $c(\cdot)$.

2.1. Copula - Attractive features

- A copula describes how the marginals are tied together in the joint distribution
- The joint df is decomposed into the marginal dfs and a copula
- The marginal dfs and the copula can be modelled and estimated separately, independent of each other
- Given a copula, we can obtain many multivariate distributions by selecting different marginal dfs
- The copula is invariant under increasing and continuous transformations

3. Copula Goodness-of-fit Testing

- Determine whether a copula appropriately fits the data.
- Univariate distributions $\Rightarrow$ e.g. Anderson-Darling test or less quantitatively using QQ-plot.
- Multivariate domain $\Rightarrow$ fewer alternatives.
- Copula GOF is a special case of the more general problem of testing multivariate density models.
- Complicated due to the use of empirical margins. Hence, P-values are usually found by simulation.

3. Copula Goodness-of-fit Testing

- Several approaches proposed lately, e.g.
  - Breymann et al. (2003) - based on the probability integral transform (PIT)
  - Genest et al. (2006) - based on the empirical copula and Kendall's process
- Dimension reduction techniques reduce the multivariate problem to a univariate problem.

3.1. Probability Integral Transform

- Transforms a set of dependent variables into a new set of independent $U(0,1)$ variables, given the multivariate distribution.
- A universally applicable way of creating a set of iid $U(0,1)$ variables from any data set with known distribution.
- Given a test for multivariate, independent uniformity, this transformation can be used to test the fit of any assumed model.
- The concept was first introduced by Rosenblatt (1952) and can be interpreted as the inverse of simulation.

3.1. Probability Integral Transform

- The idea is to PIT the observed copula, assuming a $\mathcal{H}_0$ copula, and then test for independence. The null hypothesis may be a parametric copula family.
- An advantage with the PIT in this setting is that the null- and alternative hypotheses are the same, regardless of the distribution before the PIT.
- The PIT also enables weighting in a simple way since the data, under $\mathcal{H}_0$, is always iid $U(0, 1)$.

3.2. Breymann et al. (2003)'s approach: *G*

- Let $\boldsymbol{Z}$ be an iid $U(0,1)^d$ vector under $\mathcal{H}_0$. Now define

$$
\begin{aligned}
Y_G &= \sum_{i=1}^{d} \Phi^{-1}(z_i)^2, \\
W_G &= F_{\chi^2_d}(Y_G), \\
F_G(w) &= P(W_G \leq w), \qquad w \in [0,1].
\end{aligned}
$$

  Under $\mathcal{H}_0$ $F_G(w) = w$ and its density function $f_g(w) = 1$.
- Properties:
  - Coincides with the approaches proposed by Malevergne and Sornette (2003) when the latter is based on PIT. Also coincides with the second approach proposed by Chen et al. (2004).
  - Implicitly weights the tails of the copula through $\Phi^{-1}(\cdot)^2$
  - **NOT** consistent, some deviations may cancel out

3.3. New approach: B

- Extends G, solving the consistency issue by transforming the vector $\boldsymbol{Z}$. Decouples deviance measure from weighting functionality.
- Let $\boldsymbol{Z}$ be an iid $U(0,1)^d$ vector under $\mathcal{H}_0$. Define a new vector $\boldsymbol{Z}^*$ as

$$Z_i^* = \left(1 - \left(\frac{1 - \widetilde{z}_i}{1 - r_{i-1}}\right)^{d-(i-1)}\right),$$

for $i = 1, \ldots, d$, where $\widetilde{\boldsymbol{Z}} = (\widetilde{z}_1, \ldots, \widetilde{z}_d)$ is the sorted counterpart of $\boldsymbol{Z}$ and r_i is rank variable i from $\boldsymbol{Z}$.

3.3. New approach: B

- Next, let

$$Y_B = \sum_{i=1}^{d} \gamma(z_i; \boldsymbol{\alpha}) \cdot \Phi^{-1}(Z_i^*)^2,$$

where γ is a weight function used for weighting $\Phi^{-1}(z_i^*)^2$ depending on its corresponding value z_i, and α is the set of weight parameters.

- Further let $F_{Y_B}(\cdot)$ be the cdf of Y_B, i.e. the cdf of a linear combination of squared normal variables. Then

$$W_B = F_{Y_B}(Y_B),$$
$$F_B(w) = P\left(W_B \leq w\right), \qquad w \in [0,1].$$

Under $\mathcal{H}_0$ $F_B(w) = w$ and $f_b(w) = 1$.

3.3.1. Weighting functionality

The weight function may be of any form, for example:

- Power tail weighting: $\gamma(z_i;\alpha) = (z_i - 0.5)^\alpha$
- Left/Right power tail weighting:
  - Left power tail: $\gamma(z_i;\alpha) = 1 - z_i^{1/\alpha}$
  - Right power tail: $\gamma(z_i;\alpha) = 1 - (1 - z_i)^{1/\alpha}$
- Inverse Student's t tail weighting: $\gamma(z_i;\alpha) = t_\nu^{-1}(z_i)^2$

3.3.1. Weighting functionality

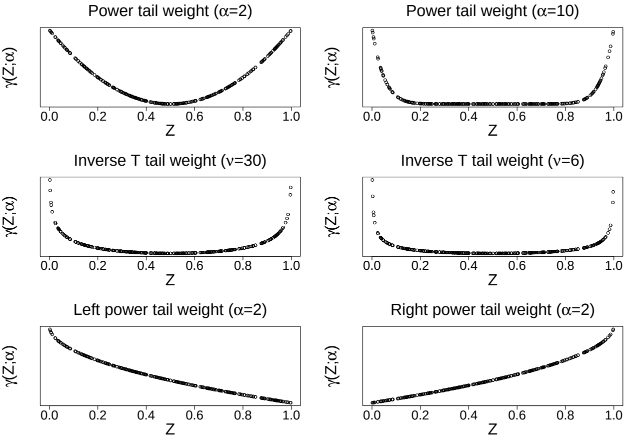

Figure: The effect of tail weighting.

3.3.2. Testing Procedure

Suppose we have n independent observations from a d-dimensional copula $\boldsymbol{X}$. The testing procedure would then be as follows:

1. PIT $\boldsymbol{X}$ under a $\mathcal{H}_0$ copula. This procedure usually involves estimating the parameters of the $\mathcal{H}_0$ copula, $\widehat{\boldsymbol{\theta}}$. The resulting copula, $\boldsymbol{Z}$, should be the independent copula if $\mathcal{H}_0$ is true.
2. Then, for each $j = 1, \ldots, n$, do:
   - ▷ From $\boldsymbol{Z}_j$, compute weights $\gamma(z_{ji}; \boldsymbol{\alpha})$,$i = 1, \ldots, d$.
   - ▷ Compute $\boldsymbol{Z}_j^*$. These variables are iid $U(0,1)^d$ under $\mathcal{H}_0$.
   - ▷ Compute the univariate variable Y_{Bj}.
   - ▷ Given F_{Y_B} (e.g. from simulations), compute W_{Bj}.
   - ▷ Given W_{Bj} compute $F_{Bj}(w)$, an iid $U(0,1)$ vector under $\mathcal{H}_0$.
3. Compute some univariate test $\widehat{\mathcal{T}}$ using $F_B(w)$ or $f_B(w)$.
4. Repeatedly (N times) perform step 1-3 using a simulated observed data set $\boldsymbol{X}^*$, simulated from the $\mathcal{H}_0$ distribution with parameter $\widehat{\boldsymbol{\theta}}$. The resulting N values of $\widehat{\mathcal{T}}^*$ form the distribution of $\mathcal{T}$.
5. Compute the p-value, $p = \frac{1+\sum_{k=1}^{N} I(\widehat{\mathcal{T}}^* \geq \widehat{\mathcal{T}})}{N+1}$.

4. Results

To assess the power of the test we performed so called 'Mixing' tests:

- $\mathcal{C}^{Mix} = (1-\beta)\cdot\mathcal{C}^{Ga} + \beta\cdot\mathcal{C}^{Alt}, \quad \beta\in[0,1], \quad \mathcal{C}^{Alt}\in[\mathcal{C}^{St},\, \mathcal{C}^{Cl}]$.
- $\mathcal{H}_0$: Gaussian copula
- PIT under $\mathcal{H}_0$ and compute *p*-value.
- Repeat 500 times to obtain rejection rates as a function of the mixing parameter β and the alternative copula.

4. Results

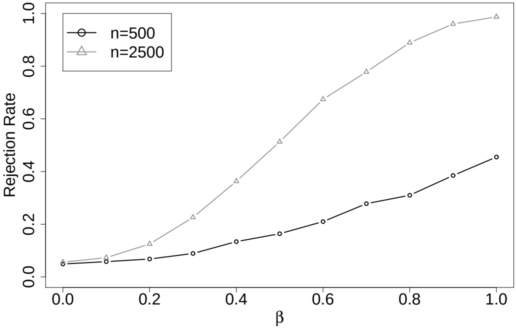

Figure: The effect of n - the number of observations. G/T mixing, power tail weighting, $d = 2,\ \alpha = 4,\ \rho = 0.5,\ \nu = 4,$ 5% significance level.

4. Results

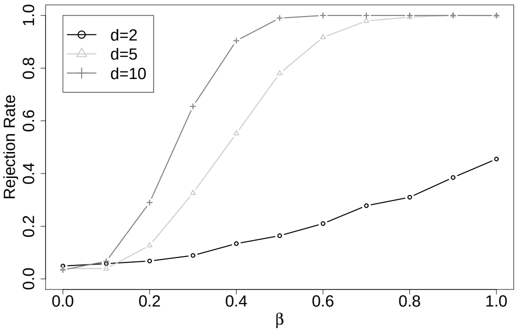

Figure: The effect of d - the dimension. G/T mixing, power tail weighting, $n = 500, \alpha = 4, \rho = 0.5, \nu = 4, 5\%$ significance level.

4. Results

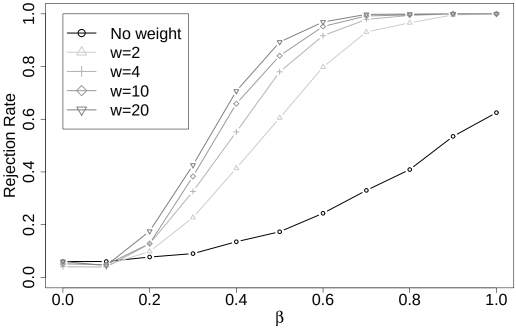

Figure: The effect of α - the power tail weighting parameter. Gaussian-Student-t mixing, power tail weighting, $d = 5, n = 500, \rho = 0.5, \nu = 4$, 5% significance level.

4. Results

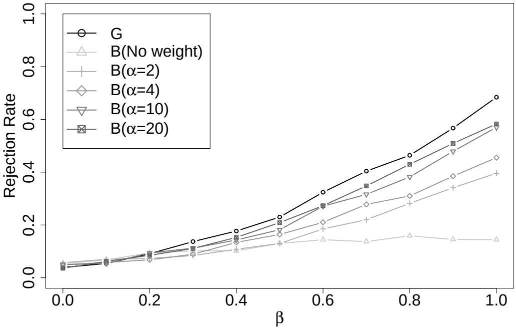

Figure: G test versus B test for $d = 2$ and $n = 500$. No weight and various power tail weights for the B test. Gaussian-Student's t mixing, $\rho = 0.5, \nu = 4$, 5% significance level

4. Results

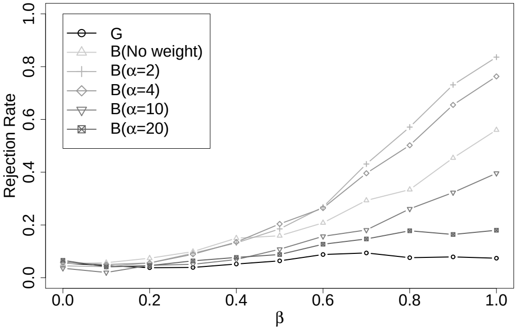

Figure: G test versus B test for $d = 5$ and $n = 500$. No weight and various power tail weights for the B test. Gaussian-Clayton mixing, $\rho = 0.5, \delta = 0.5$, 5% significance level

5. Application

- Portfolio of 50 large cap stocks. Daily log-returns from September 26th 2001 to September 16th 2005, i.e. $d = 50$ and $n = 1000$.
- Randomly select collections of 2 assets.
- PIT under Gaussian, Student-t and Clayton (one-parameter) $\mathcal{H}_0$ respectively.
- Compute P-value.
- Repeat 100 times $\Rightarrow$ rejection rates.
- Repeat for collections of 5 and 10 assets.

5. Application

Gaussian copula					
	No Weight / Power tail weight (parameter α)				
Dimension	No weight	$\alpha = 2$	$\alpha = 4$	$\alpha = 10$	$\alpha = 20$
2	0.076	0.132	0.176	0.466	0.512
5	0.700	0.930	0.930	0.920	0.910
10	0.740	1.000	1.000	1.000	1.000
Student-t copula					
	No Weight / Power tail weight (parameter w)				
Dimension	No weight	$\alpha = 2$	$\alpha = 4$	$\alpha = 10$	$\alpha = 20$
2	0.042	0.022	0.032	0.044	0.034
5	0.120	0.090	0.060	0.050	0.070
10	0.260	0.040	0.150	0.130	0.190
Clayton copula					
	No Weight / Power tail weight (parameter w)				
Dimension	No weight	$\alpha = 2$	$\alpha = 4$	$\alpha = 10$	$\alpha = 20$
2	0.622	0.354	0.792	0.434	0.396
5	0.980	0.990	0.980	0.970	0.950
10	1.000	1.000	1.000	1.000	1.000

Table: Rejection rates for the fit of the Gaussian, Student-t and Clayton copulae.

6. Summary

- New approach *B* merges the efficiency of one-dimensional tests with the consistency of multi-dimensional tests.
- The weighting functionality adds valuable flexibilities to the analyst.
- Mixing tests show that the approach has good power for tail heaviness and skewness. The weighting functionality also seem to be very powerful.
- Applied to daily log-returns of stock portfolios the Student-t copula outperforms the Gaussian and Clayton copulae, as expected and in accordance with the findings of other studies.

References

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

Chen, X., Y. Fan, and A. Patton (2004). Simple tests for models of dependence between multiple financial time series, with applications to U.S. equity returns and exchange rates. Financial Markets Group, London School of Economics, Discussion Paper 483. Revised July 2004.

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

Malevergne, Y. and D. Sornette (2003). Testing the gaussian copula hypothesis for financial assets dependence. *Quantitative Finance 3*, 231–250.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*, 470–472.