Copula goodness-of-fit testing: An overview and power comparison

DANIEL BERG

DnB NOR Asset Management

22nd Nordic Conference on Mathematical Statistics.

Vilnius, Lithuania – June 2008
Outline

- Introduction
- Copula goodness-of-fit testing
  - Introduction
  - Preliminaries
  - Proposed approaches
- Monte Carlo simulation results
- Conclusions and recommendations
Introduction

Motivation

Figure: Two simulated data sets - both with standard normal margins and correlation coefficient 0.7.
Figure: Nonzero precipitation values in two Norwegian cities and its copula.
Introduction

Definition & Theorem

Definition (Copula)

A \( d \)-dimensional copula is a multivariate distribution function \( C \) with standard uniform marginal distributions.

Theorem (Sklar, 1959)

Let \( H \) be a joint distribution function with margins \( F_1, \ldots, F_d \). Then there exists a copula \( C : [0, 1]^d \rightarrow [0, 1] \) such that

\[
H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).
\]
Introduction

Useful results

- A general \( d \)-dimensional density \( h \) can be expressed, for some copula density \( c \), as

\[
h(x_1, \ldots, x_d) = c\{F_1(x_1), \ldots, F_d(x_d)\}f_1(x_1) \cdots f_d(x_d).
\]

- Non-parametric estimate for \( F_i(x_i) \) commonly used to transform original margins into standard uniform:

\[
u_{ji} = \hat{F}_i(x_{ji}) = \frac{R_{ji}}{n + 1},
\]

where \( R_{ji} \) is the rank of \( x_{ji} \) amongst \( x_{1i}, \ldots, x_{ni} \).

- \( u_{ji} \) commonly referred to as pseudo-observations and models based on non-parametric margins and parametric copulas are referred to as semi-parametric copulas.
Copula GoF testing

Introduction

- $\mathcal{H}_0 : C \in \mathcal{C} = \{ C_\theta ; \theta \in \Theta \}$ vs. $\mathcal{H}_1 : C \notin \mathcal{C} = \{ C_\theta ; \theta \in \Theta \}$
- Univariate $\Rightarrow$ Anderson-Darling or QQ-plot,
  Multivariate $\Rightarrow$ fewer alternatives
- Pseudo-observations no longer independent. In addition, limiting distribution of many copula GoF test depends on null hypothesis copula and parameter value
- $p$-value estimation via parametric bootstrap procedures
- Focus in literature almost exclusively bivariate
- NOT model selection!
- Several techniques proposed: binning, multivariate smoothing, dimension reduction
Copula GoF testing

Preliminaries

Rosenblatt’s transform:

- Dependent variables $\Rightarrow$ independent $U[0, 1]$ variables, given multivariate distribution
- $\mathbf{v} = \mathcal{R}(\mathbf{z}) = (\mathcal{R}_1(z_1), \ldots, \mathcal{R}_d(z_d))$:
  - $v_1 = \mathcal{R}_1(z_1) = F_1(z_1) = z_1$,
  - $v_2 = \mathcal{R}_2(z_2) = F_{2|1}(z_2|z_1),$
  - $\vdots$
  - $v_d = \mathcal{R}_d(z_d) = F_{d|1\ldots d}(z_d|z_1, \ldots, z_d)$.

- Inverse of simulation (conditional inversion)
- GoF: $\mathbf{v} = \mathcal{R}(\mathbf{z}) \Rightarrow$ test $\mathbf{v}$ for independence
- $d!$ different permutation orders
Copula GoF testing
Proposed approaches: $A_1$ (1/9)

- $v = \mathcal{R}(z)$
- $W_{1j} = \sum_{i=1}^{d} \Gamma\{v_{ji}; \alpha\}, \quad j = 1, \ldots, n$
- Special case (a): $\sum_{i=1}^{d} \Phi^{-1}(v_{ji})^2$
- Special case (b): $\sum_{i=1}^{d} |v_{ji} - 0.5|$
- $S_1(t) = P\{F_1(W_1) \leq t\}, \quad t \in [0, 1]$
- CvM statistic:
  \[ \hat{T}_1 = n \int_{0}^{1} \left\{ \hat{S}_1(t) - S_1(t) \right\}^2 dS_1(t) \]

- References: Breymann et al. (2003); Malevergne and Sornette (2003); Berg and Bakken (2005)
Copula GoF testing

Proposed approaches: $A_2$ (2/9)

- **Empirical copula:**

\[
\widehat{C}(u) = \frac{1}{n+1} \sum_{j=1}^{n} I\{Z_{j1} \leq u_1, \ldots, Z_{jd} \leq u_d\}
\]

- **CvM statistic:**

\[
\widehat{T}_2 = n \int_{[0,1]^d} \left\{ \widehat{C}(z) - C_{\theta}(z) \right\}^2 d\widehat{C}(z)
\]

- **References:** Fermanian (2005); Genest and Rémillard (2008); Genest et al. (2008)
Copula GoF testing

Proposed approaches: $A_3$ (3/9)

▷ Approach $A_2$ on $v = R(z)$

▷ CvM statistic:

$$\hat{T}_3 = \frac{1}{n} \int_{[0,1]^d} \left\{ \hat{C}(v) - C_{\perp}(v) \right\}^2 d\hat{C}(v)$$

▷ References: Genest et al. (2008)
Copula GoF testing

Proposed approaches: $A_4$ (4/9)

- Cdf of empirical copula (Kendall’s dependence function):
  \[ S_4(t) = P\{C(z) \leq t\} \]

- CvM statistic:
  \[ \hat{T}_4 = n \int_0^1 \left\{ \hat{S}_4(t) - S_{4,\hat{\theta}}(t) \right\}^2 dS_{4,\hat{\theta}}(t) \]

- References: Genest and Rivest (1993); Wang and Wells (2000); Savu and Trede (2004); Genest et al. (2006)
Copula GoF testing

Proposed approaches: $\mathcal{A}_5$ (5/9)

- Spearman’s dependence function:
  \[
  S_5(t) = P\{C_\perp(z) \leq t\}
  \]

- CvM statistic:
  \[
  \hat{T}_5 = n \int_0^1 \left\{ \hat{S}_5(t) - S_{5,\hat{\theta}}(t) \right\}^2 dS_{5,\hat{\theta}}(t)
  \]

- References: Quessy et al. (2007)
Copula GoF testing

Proposed approaches: $A_6 (6/9)$

- Shih’s test for bivariate gamma frailty model (Clayton):
  \[
  \hat{T}_{Shih} = \sqrt{n} \left\{ \hat{\theta}_\tau - \hat{\theta}_W \right\}
  \]

- Extension to arbitrary dimension:
  \[
  \hat{T}_6 = \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} \left\{ \hat{\theta}_{\tau,ij} - \hat{\theta}_{W,ij} \right\}^2
  \]

- References: Shih (1998); Berg (2007)
Copula GoF testing

Proposed approaches: \( A_7 \) (7/9)

- Inner product of two vectors = 0 iff from the same family
  \[
  Q(z) = \langle z - z_{\hat{\theta}} | \kappa_d | z - z_{\hat{\theta}} \rangle
  \]

- \( \kappa \) a symmetric kernel, e.g. the gaussian kernel:
  \[
  \kappa_d(z, z_{\hat{\theta}}) = \exp \left\{ -\|z - z_{\hat{\theta}}\|^2 / (2dh^2) \right\}
  \]

- Statistic becomes:
  \[
  \hat{T}_7 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(z_i, z_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(z_i, z_{\hat{\theta},j}) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(z_{\hat{\theta},i}, z_{\hat{\theta},j})
  \]

- References: Panchenko (2005)
Copula GoF testing

Proposed approaches: \( A_8 \) (8/9)

- Approach \( A_7 \) on \( \mathbf{v} = R(z) \)
- Statistic:

\[
\hat{T}_8 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\mathbf{v}_i, \mathbf{v}_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\mathbf{v}_i, \mathbf{v}_{\hat{\theta},j}) \\
+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\mathbf{v}_{\hat{\theta},i}, \mathbf{v}_{\hat{\theta},j})
\]

- References: Berg (2007)
Each approach may detect deviations from $\mathcal{H}_0$ differently

Average approaches:

\[
\hat{T}_9^{(a)} = \frac{1}{9} \left\{ \hat{T}_1^{(a)} + \hat{T}_1^{(b)} + \sum_{k=2}^{8} \hat{T}_k \right\}
\]

\[
\hat{T}_9^{(b)} = \frac{1}{3} \left\{ \hat{T}_2 + \hat{T}_3 + \hat{T}_4 \right\}
\]

References: Berg (2007)
Monte Carlo simulations

Test procedure

1) \( x \sim n \) samples from the \( d \)-dimensional \( \mathcal{H}_1 \) copula with \( \theta(\tau) \).
2) \( z \sim \) pseudo-observations (normalized ranks)
3) \( \hat{\theta} \sim \) estimated parameter of the \( \mathcal{H}_0 \) copula
4) \( \hat{T}_i \sim \) test statistic \( i \) computed under the \( \mathcal{H}_0 \) copula using \( \hat{\theta} \).
5) Repeat steps 1-4 \( M \) times with \( \mathcal{H}_1 = \mathcal{H}_0 \) and \( \theta = \hat{\theta} \Rightarrow \hat{T}_{i,m}^0 \)
6) \( \hat{p} = \frac{1}{M} \sum_{m=1}^{M} 1(\hat{T}_{i,m}^0 > \hat{T}_i) \)
7) \( \hat{p} < 5\% \Rightarrow \) reject \( \mathcal{H}_0 \)
Monte Carlo simulations

Experimental setup

- $H_0$ copula (5 choices: Gaussian, Student, Clayton, Gumbel, Frank),
- $H_1$ copula (5 choices: Gaussian, Student ($\nu = 6$), Clayton, Gumbel, Frank),
- Kendall’s tau (2 choices: $\tau = \{0.2, 0.4\}$),
- Dimension (3 choices: $d = \{2, 4, 8\}$),
- Sample size (2 choices: $n = \{100, 500\}$)
- Student only considered as null in bivariate case.
- For each of these 240 cases, 10,000 repetitions $\Rightarrow$ size/power
Monte Carlo simulations

Testing the Gaussian copula

![Box plots comparing power differences for different copula models and sample sizes.](image)
Monte Carlo simulations

Testing the Student copula

![Box plot diagram showing power difference (%)]
Monte Carlo simulations

Testing the Clayton copula

![Box plots comparing power differences for different Clayton copula parameters and test procedures.](image)
Monte Carlo simulations

Testing the Gumbel copula
Monte Carlo simulations

Testing the Frank copula

![Box plot diagram showing power difference (%) for different copula models.]

1. Introduction
2. Copula GoF testing
3. MC simulations
4. Conclusions

Test procedure
1. Experimental setup
2. Power comparison

A_1 (a) A_1 (b) A_2 A_3 A_4 A_5 A_7 A_8 A_9 (a) A_9 (b)

0 20 40 60 80 100
Power difference (%)
Conclusions and recommendations

- Nominal levels all match prescribed size of 5%
- Power generally increases with dimension, sample size and dependence
- Clayton > Gumbel > Frank > Gaussian > Student (> : easier to test)
- No universally most powerful approach, but $A_2$, $A_4$ and $A_9^{(b)}$ perform very well in most cases
- $A_9^{(b)}$ is recommended in general, with special case exceptions:
  - For testing the Gaussian copula, if trying to detect heavy tails for $d > 2$ and large $n$ then $A_1$ very powerful
  - For testing the Clayton copula the generalized Shih’s test is most powerful
- Permutational variation of little concern for approaches based on Rosenblatt’s transform (see Berg (2007))


