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

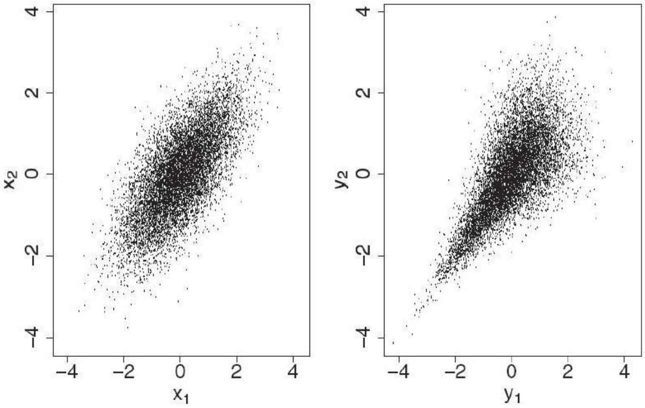


Figure: Two simulated data sets - both with standard normal margins and correlation coefficient 0.7.

Introduction

Motivation

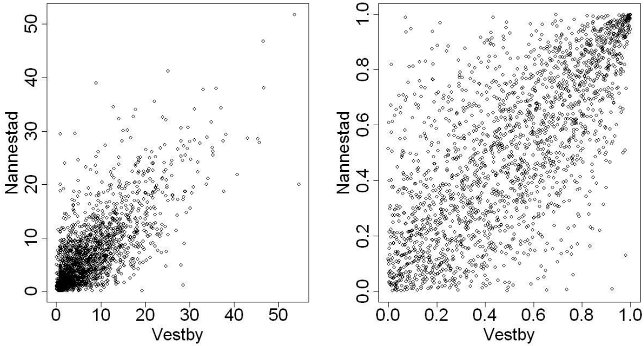


Figure: Nonzero precipitation values in two Norwegian cities and its copula.

Introduction

Definition & Theorem

Definition (Copula)

A d-dimensional copula is a multivariate distribution function $\mathcal{C}$ 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:

$$u_{ji} = \widehat{F}_i(x_{ji}) = \frac{R_{ji}}{n+1},$$

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
- $\boldsymbol{v} = \mathcal{R}(\boldsymbol{z}) = (\mathcal{R}_1(z_1), \ldots, \mathcal{R}_d(z_d))$:

$$\begin{aligned} v_1 &= \mathcal{R}_1(z_1) = F_1(z_1) = z_1, \\ v_2 &= \mathcal{R}_2(z_2) = F_{2|1}(z_2|z_1), \\ &\vdots \\ v_d &= \mathcal{R}_d(z_d) = F_{d|1\ldots d}(z_d|z_1, \ldots, z_d). \end{aligned}$$

- Inverse of simulation (conditional inversion)
- GoF: $\boldsymbol{v} = \mathcal{R}(\boldsymbol{z}) \Rightarrow$ test $\boldsymbol{v}$ for independence
- $d!$ different permutation orders

Copula GoF testing

Proposed approaches: $\mathcal{A}_1$ (1/9)

- $\boldsymbol{v} = \mathcal{R}(\boldsymbol{z})$
- $W_{1j} = \sum_{i=1}^{d} \Gamma\{v_{ji}; \alpha\}, \quad j = 1, \ldots, n$
- 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:

$$\widehat{T}_1 = n \int_0^1 \left\{ \widehat{S}_1(t) - S_1(t) \right\}^2 \mathrm{d}S_1(t)$$

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

Copula GoF testing

Proposed approaches: $\mathcal{A}_2$ (2/9)

- Empirical copula:

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

- CvM statistic:

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

- References: Fermanian (2005); Genest and Rémillard (2008); Genest et al. (2008)

Copula GoF testing

Proposed approaches: $\mathcal{A}_3$ (3/9)

- Approach $\mathcal{A}_2$ on $\boldsymbol{v} = \mathcal{R}(\boldsymbol{z})$
- CvM statistic:

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

- References: Genest et al. (2008)

Copula GoF testing

Proposed approaches: $\mathcal{A}_4$ (4/9)

- Cdf of empirical copula (Kendall's dependence function):

$$S_4(t) = P\{C(\mathbf{z}) \leq t\}$$

- CvM statistic:

$$\widehat{T}_4 = n \int_0^1 \left\{ \widehat{S}_4(t) - S_{4,\widehat{\theta}}(t) \right\}^2 \mathrm{d}S_{4,\widehat{\theta}}(t)$$

- 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:

$$\widehat{T}_5 = n \int_0^1 \left\{ \widehat{S}_5(t) - S_{5,\widehat{\theta}}(t) \right\}^2 \mathrm{d}S_{5,\widehat{\theta}}(t)$$

- References: Quessy et al. (2007)

Copula GoF testing

Proposed approaches: $\mathcal{A}_6$ (6/9)

- Shih's test for bivariate gamma frailty model (Clayton):

$$\widehat{T}_{Shih} = \sqrt{n}\left\{\widehat{\theta}_\tau - \widehat{\theta}_W\right\}$$

- Extension to arbitrary dimension:

$$\widehat{T}_6 = \sum_{i=1}^{d-1}\sum_{j=i+1}^{d}\left\{\widehat{\theta}_{\tau,ij} - \widehat{\theta}_{W,ij}\right\}^2$$

- References: Shih (1998); Berg (2007)

Copula GoF testing

Proposed approaches: $\mathcal{A}_7$ (7/9)

- Inner product of two vectors = 0 iff from the same family

$$Q(z) = \left\langle z - z_{\widehat{\theta}} \middle| \kappa_d \middle| z - z_{\widehat{\theta}} \right\rangle$$

- κ a symmetric kernel, e.g. the gaussian kernel:

$$\kappa_d(z, z_{\widehat{\theta}}) = \exp\left\{-\|z - z_{\widehat{\theta}}\|^2/(2dh^2)\right\}$$

- Statistic becomes:

$$\widehat{T}_7 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_d(z_i, z_j) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_d(z_i, z_{\widehat{\theta},j}) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa_d(z_{\widehat{\theta},i}, z_{\widehat{\theta},j})$$

- References: Panchenko (2005)

Copula GoF testing

Proposed approaches: $\mathcal{A}_8$ (8/9)

- Approach $\mathcal{A}_7$ on $\boldsymbol{v} = \mathcal{R}(\boldsymbol{z})$
- Statistic:

$$\widehat{T}_8 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\boldsymbol{v}_i, \boldsymbol{v}_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\boldsymbol{v}_i, \boldsymbol{v}_{\widehat{\theta},j}) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa_d(\boldsymbol{v}_{\widehat{\theta},i}, \boldsymbol{v}_{\widehat{\theta},j})$$

- References: Berg (2007)

Copula GoF testing

Proposed approaches: $\mathcal{A}_9$ (9/9)

- Each approach may detect deviations from $\mathcal{H}_0$ differently
- Average approaches:

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

$$\widehat{T}_9^{(b)} = \frac{1}{3}\left\{\widehat{T}_2 + \widehat{T}_3 + \widehat{T}_4\right\}$$

- References: Berg (2007)

Monte Carlso simulations

Test procedure

1) $\boldsymbol{x} \sim n$ samples from the d-dimensional $\mathcal{H}_1$ copula with $\theta(\tau)$.
2) $\boldsymbol{z} \sim$ pseudo-observations (normalized ranks)
3) $\hat{\theta} \sim$ estimated parameter of the $\mathcal{H}_0$ copula
4) $\widehat{T}_i \sim$ 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 $\theta = \hat{\theta} \Rightarrow \widehat{T}^0_{i,m}$
6) $\hat{p} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(\widehat{T}^0_{i,m} > \widehat{T}_i)$
7) $\hat{p} < 5\% \Rightarrow$ reject $\mathcal{H}_0$

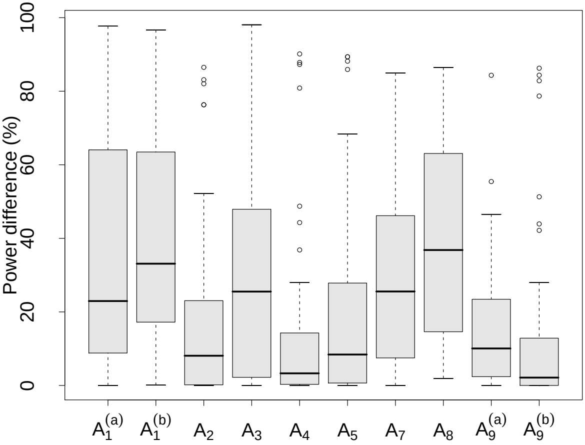
Monte Carlso simulations

Experimental setup

- $\mathcal{H}_0$ copula (5 choices: Gaussian, Student, Clayton, Gumbel, Frank),
- $\mathcal{H}_1$ 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

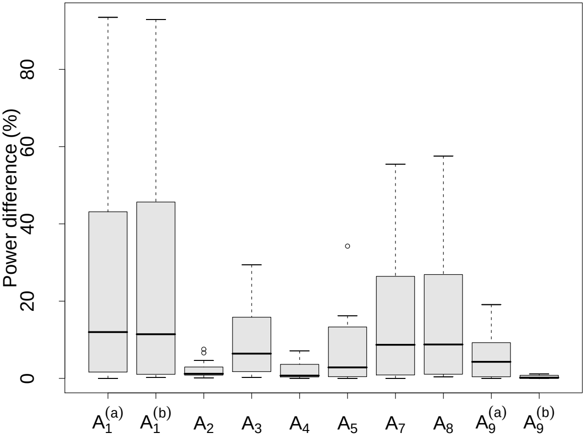
Monte Carlso simulations

Testing the Gaussian copula



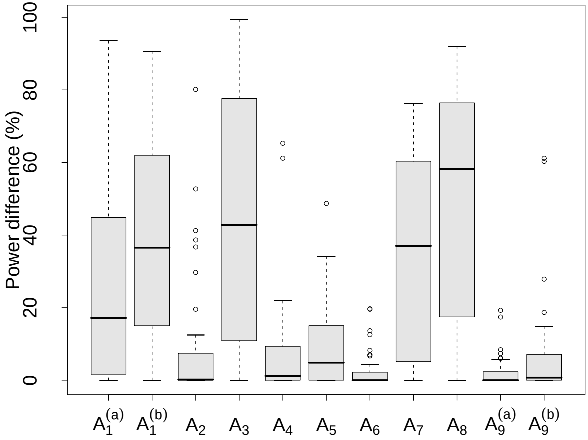
Monte Carlso simulations

Testing the Student copula



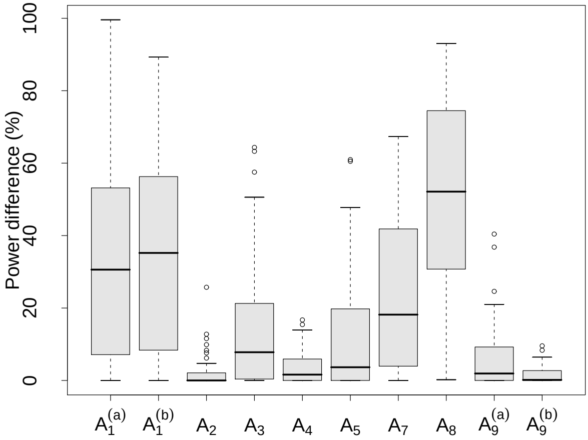
Monte Carlso simulations

Testing the Clayton copula



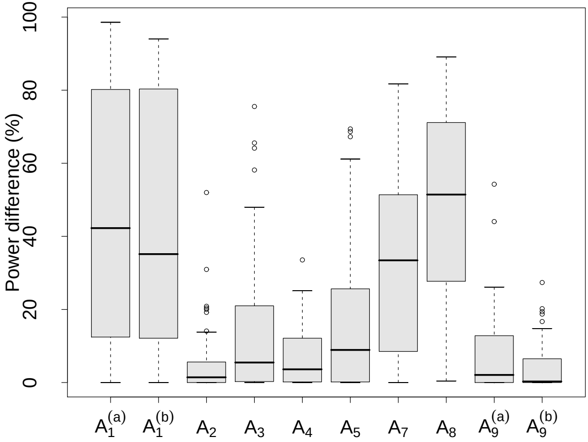
Monte Carlso simulations

Testing the Gumbel copula



Monte Carlso simulations

Testing the Frank copula



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 $\mathcal{A}_2$, $\mathcal{A}_4$ and $\mathcal{A}_9^{(b)}$ perform very well in most cases
- $\mathcal{A}_9^{(b)}$ 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 $\mathcal{A}_1$ 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))

Berg, D. (2007). Copula goodness-of-fit testing: an overview and power comparison. Technical report, University of Oslo. Statistical research report no. 5, ISSN 0806-3842.

Berg, D. and H. Bakken (2005). A goodness-of-fit test for copulae based on the probability integral transform. Technical report, University of Oslo. Statistical research report no. 10, ISSN 0806-3842.

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

Fermanian, J. (2005). Goodness of fit tests for copulas. *Journal of Multivariate Analysis 95*, 119–152.

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*, 337–366.

Genest, C. and B. Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Henri Poincaré 44*. In press.

Genest, C., B. Rémillard, and D. Beaudoin (2008). Omnibus goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics 42*. In press.

Genest, C. and L.-P. Rivest (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, 1034–1043.

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

Panchenko, V. (2005). Goodness-of-fit test for copulas. *Physica A 355(1)*, 176–182.

Quessy, J.-F., M. Mesfioui, and M.-H. Toupin (2007). A goodness-of-fit test based on Spearmans dependence function. Working paper, Université du Québec à Trois-Rivières.

Savu, C. and M. Trede (2004). Goodness-of-fit tests for parametric families of archimedean copulas. CAWM discussion paper, No. 6.

Shih, J. H. (1998). A goodness-of-fit test for association in a bivariate survival model. *Biometrika 85*, 189–200.

Wang, W. and M. T. Wells (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association 95*, 62–72.