R-package gofCopula

Ostap Okhrin, Simon Trimborn

Qian M. Zhou, Shulin Zhang

Technische Universität Dresden

Humboldt–Universität zu Berlin

Simon Fraser University

Southwestern University of Finance and Economics

https://tu-dresden.de/bu/verkehr/ivw/osv

http://lvb.wiwi.hu-berlin.de

http://stat.sfu.ca/

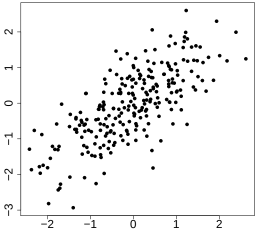
http://csr.swufe.edu.cn

Applications

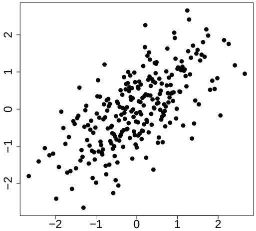
1. medicine (Vandenhende (2003))
2. hydrology (Genest and Favre (2006))
3. biometrics (Wang and Wells (2000, JASA), Chen and Fan (2006, CanJoS))
4. economics
   - portfolio selection (Patton (2004, JoFE), Xu (2004, PhD thesis), Hennessy and Lapan (2002, MathFin))
   - time series (Chen and Fan (2006a, 2006b, JoE), Fermanian and Scaillet (2003, JoR), Lee and Long (2005, JoE))
   - risk management (Junker and May (2002, EJ), Breyman et. al. (2003, QF))

How to be sure, that we use a proper copula?

Different tests ⇒ Different outcomes



(a) Gaussian copula



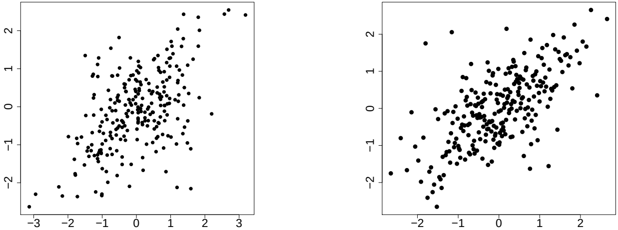
(b) year 2004

Figure 1: Sample from Gauss copula with $\mathrm{N}(0,1)$ margins, $\theta = 0.71$, $N = 250$ and residuals transformed to standard normal for Citygroup/BoA for 2004.

Visually - Gaussian copula

Test 1: Gumbel, Test 2: Gauss, Test 3: Gauss

Different tests ⇒ Different outcomes



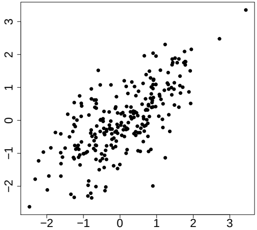
(a) t-copula (b) year 2006

Figure 2: Sample from t-copula with $\mathrm{N}(0,1)$ margins, $\theta = 0.6$, $N = 250$ and residuals transformed to standard normal for Citygroup/BoA for 2006.

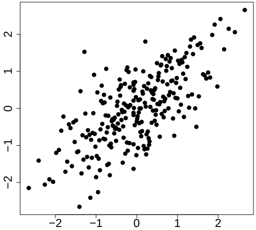
Visually - t-copula

Test 1: t-copula, Test 2: Gauss, Test 3: t-copula

Different tests ⇒ Different outcomes



(a) Gumbel copula



(b) year 2009

Figure 3: Sample from Gumbel copula with N(0,1) margins, $\theta = 2$, $N = 250$ and residuals transformed to standard normal for Citygroup/BoA for 2009.

Visually - Gumbel copula

Test 1: Gumbel, Test 2: Gumbel, Test 3: Gauss

Goodness-of-Fit Tests

$$\mathcal{H}_0 : C_0 \in \mathcal{C} \quad vs. \quad \mathcal{H}_1 : C_0 \notin \mathcal{C}$$

where $\mathcal{C} = \{C(\cdot;\theta) : \theta \in \Theta\}$.

- $X_1 = (X_{11}, \ldots, X_{1d})^\top, \ldots, X_n = (X_{n1}, \ldots, X_{nd})^\top$ random sample of size n drawn from multivariate distribution $H(x) = H(x_1, x_2, \ldots, x_d)$
- Continuous marginal cdf $F(x) = \{F_1(x_1), \ldots, F_d(x_d)\}$

$$H(x_1, x_2, \ldots, x_d) = C_0\{F(x)\} = C_0\{F_1(x_1), \ldots, F_d(x_d)\}.$$

Goodness-of-Fit tests for copulae

- Several tests available
- Different hypothesis
- Overall optimal test missing (for blanket tests)
- Computationally demanding

Outline

R-Paket gofCopula

- Goodness of Fit tests for copulae
  - Most tests in one package
  - Several margin structures
  - Most tests at least 3 dim
  - Automatized parallelization
- Hybrid test
  - p-value:

$$p_n^{hybrid} = \min\left(q \cdot \min\left(p_n^{(1)}, \ldots, p_n^{(q)}\right), 1\right)$$

with q tests and $p_n^{(i)}$ p-value of test i

Covered single tests

- gofRosenblattSnB ▸ Details
- gofRosenblattSnC ▸ Details

  ▸ Genest et al (2009)
- gofRosenblattChisq ▸ Details
- gofRosenblattGamma ▸ Details

  ▸ Breymann et al (2003)
- gofPIOSRn ▸ Details
- gofPIOSTn ▸ Details

  ▸ Zhang et al (2015)

▸ General Definitions

- gofKendallCvM ▸ Details
- gofKendallKS ▸ Details

  ▸ Genest et al (2006)
- gofSn ▸ Details

  ▸ Genest (2008)
- gofKernel ▸ Details

  ▸ Scaillet (2007)
- gofWhite ▸ Details

  ▸ Huang and Prokhorov (2011)

Single tests - computation

- Provided dataset

```
data(IndexReturns)
x = IndexReturns[c(1:100),c(1:2)]
```

- Cramer-von Mises test for normality

```
gofRosenblattSnB("normal", x, M = 1000)
```

- Computation time

```
[1] "The computation will take approximately 0 d
  , 0 h, 1 min and 36 sec."
```

Single tests - result

```
Parametric bootstrap goodness-of-fit test with
  RosenblattSnB test and normal copula
Tests results:
                  p.value test statistic     rho.1
RosenblattSnB      0.626      0.3300644 0.6265811
```

- Automatic margins estimation

```
Warning message:
In .gofRosenblattSnB(copula = copula, x = x, M =
    M, param = param,  :
The observations are not in [0,1]. The margins
  will be estimated by the ranks of the
  observations.
```

Hybrid test - computation

- Hybrid test for normality with 4 tests

```r
gofHybrid("normal", x, testset = c("gofRosenblattSnB", "gofRosenblattSnC", "gofSn", "gofPIOSRn"), M = 1000)
```

- Computation time

```
[1] "The computation will take approximately 0 d, 0 h, 4 min and 0 sec."
```

Hybrid test - result

```
Parametric bootstrap goodness-of-fit test with hybrid test and normal
  copula

Tests results:
                  p.value   test statistic     rho.1
RosenblattSnB       0.651       0.33006438 0.6265811
RosenblattSnC       0.590       0.49268271 0.6265811
Sn                  0.998       0.02592004 0.6265811
PIOSRn              0.006       0.73877935 0.6265811
hybrid(1, 2)        1.000               NA 0.6265811
hybrid(1, 3)        1.000               NA 0.6265811
hybrid(2, 3)        1.000               NA 0.6265811
hybrid(1, 2, 3)     1.000               NA 0.6265811
hybrid(1, 4)        0.012               NA 0.6265811
hybrid(2, 4)        0.012               NA 0.6265811
hybrid(1, 2, 4)     0.018               NA 0.6265811
hybrid(3, 4)        0.012               NA 0.6265811
hybrid(1, 3, 4)     0.018               NA 0.6265811
hybrid(2, 3, 4)     0.018               NA 0.6265811
hybrid(1, 2, 3, 4)  0.024               NA 0.6265811
```

Parallelization

- Hybrid test for normality with 4 tests and 4 parallel processes

```
gofHybrid("normal", x, testset = c("gofRosenblattSnB", "gofRosenblattSnC", "gofSn", "gofPIOSRn"), M = 1000, processes = 4)
```

- Computation time

```
[1] "The computation will take approximately 0 d, 0 h, 2 min and 34 sec."
```

Tests for copulae

- Available tests for normal copula?

```
gofWhich("normal", d = 2)
[1]"gofHybrid" "gofRosenblattSnB" "gofRosenblattSnC"
[4] "gofRosenblattChisq" "gofRosenblattGamma"
[6] "gofSn" "gofKendallCvM" "gofKendallKS"
[9] "gofKernel" "gofWhite"
[11] "gofPIOSRn" "gofPIOSTn"
```

- Use all for hybrid test

```
gofHybrid("normal", x, testset = gofWhich("normal",
  d = 2)[-1], M = 1000)
[1] "The computation will take approximately 0 d, 2
  h, 20 min and 15 sec."
```

Flexible testing structure

- Adjust margins

```
margins = "norm"
gofHybrid("normal", x, testset = c("gofRosenblattSnB", "gofRosenblattSnC"), M = 10, margins = margins)
```

- Fix parameter

```
parameter = 0.2
gofHybrid("normal", x, testset = c("gofRosenblattSnB", "gofRosenblattSnC"), M = 10, param.est = F, param = parameter)
```

Copulae for tests

- Available copulae for test?

```
gofWhichCopula("gofRosenblattSnB")
[1] "normal" "t" "clayton" "frank" "gumbel"

gofWhichCopula("gofWhite")
[1] "normal" "clayton" "frank" "gumbel"
```

Copulae for tests

- Use all tests with all copulae

```
copulae = gofWhichCopula("gofRosenblattSnB")
for (i in copulae){
print(gofHybrid(i, x, testset = gofWhich(i, d = 2)
  [-1], M = 10))
}
```

OR

```
gof(x, priority = "tests", M = 10)
```

gof

- Options of gof
  - priority $\in$ {"tests", "copula"}
    - "tests": all tests for their shared copulae
    - "copula": all tests which support {"normal", "t", "gumbel", "clayton", "frank"}
  - copula: which copulae to use
  - tests: which tests to use
- priority just in effect if copula = tests = NULL

Interface to copula package

- Connection to copula package
- Usage of both packages may be desirable

```
copulaobject = normalCopula(param = 0.2, dim = 2)
gofco(copulaobject, x, testset = c("gofRosenblattSnB", "gofRosenblattSnC"), M = 1000)

Tests results:
                  p.value     test statistic
RosenblattSnB  0.9475524          0.05416868
RosenblattSnC  0.8366633          0.07630096
hybrid(1, 2)   1.0000000                 NaN
```

Computation time

	normal	t	clayton	gumbel	frank
gofKendallCvM	21.81	31.21	40.88	58.90	39.02
gofKendallKS	20.80	30.60	31.18	54.60	27.38
gofKernel	17325.72	22556.60	17407.40	32082.61	14245.62
gofPIOSRn	21.73	33.65	19.24	43.90	13.90
gofPIOSTn	32.25	3170.06	2177.93	2550.00	1847.61
gofRosenblattChisq	19.70	29.14	18.94	45.35	13.20
gofRosenblattGamma	19.26	28.50	18.98	45.10	13.53
gofRosenblattSnB	99.57	107.94	101.28	123.73	95.27
gofRosenblattSnC	32.98	41.53	33.15	59.95	27.27
gofSn	48.64	59.16	19.57	30.09	13.77
gofWhite	26.41		26.67	56.73	21.18

Table 1: The table shows the estimated computation time in seconds with the function gofCheckTime for every available test and copula for the mean over 50 iterations for simulated 2 dimensional normally distributed data with 100 observations.

Real data - cryptocurrencies

- Market index CRIX
- Bitcoin
- Data period: 2014-08-02 - 2016-07-24
- Variance correction with GARCH(1,1) model

Result

	normal	t	gumbel	clayton	frank
RosenblattSnC	0.30	0.51	0.55	0.35	0.61
KendallCvM	0.78	0.14	0.82	0.82	0.75
KendallKS	0.70	0.12	0.99	0.89	1.00
Sn	0.91	0.28	0.94	0.31	0.64
hybrid(1, 2, 3, 4)	1.00	0.37	1.00	1.00	1.00

Table 2: The table shows the p-values for 4 test and the hybrid test for 5 copula for CRIX and Bitcoin data.

R-package gofCopula

Ostap Okhrin, Simon Trimborn

Qian M. Zhou, Shulin Zhang

Technische Universität Dresden

Humboldt–Universität zu Berlin

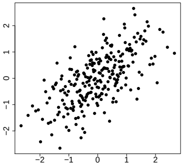
Simon Fraser University

Southwestern University of Finance and Economics

https://tu-dresden.de/bu/verkehr/ivw/osv

http://lvb.wiwi.hu-berlin.de

http://stat.sfu.ca/



Definitions

- Pseudo observations U_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$
- Empirical copula: $C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(U_{i1} \leq u_1, \ldots, U_{id} \leq u_d)$
- Test statistic from bootstrapping round b: T_b
- Unknown parameter vector: θ
- Estimate of θ: θ_n

Back

gofRosenblattSnB

- Hypothesis: $C \in \mathcal{C}_0$
- Rosenblatt probability integral transform:

  $E_1 = \mathcal{R}(U_1), \ldots, E_n = \mathcal{R}(U_n)$
- Mapping for vector $\mathbf{u}$ with $e_1 = u_1$:

$$e_i = \frac{\partial^{i-1} C(u_1, \ldots, u_i, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C(u_1, \ldots, u_{i-1}, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}}$$

- Test statistic T with $C_\perp(\mathbf{u}) = u_1 \cdot \cdots \cdot u_d$ and

  $D_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(E_i \leq \mathbf{u})$

$$T = n \int_{[0,1]^d} \{D_n(\mathbf{u}) - C_\perp(\mathbf{u})\}^2 d(\mathbf{u})$$

- p-value: $\frac{1}{M} \sum_{b=1}^M \mathbf{I}(T_b \geq T)$

gofRosenblattSnC

- Hypothesis: $C \in \mathcal{C}_0$
- Rosenblatt probability integral transform:

  $E_1 = \mathcal{R}(U_1), \ldots, E_n = \mathcal{R}(U_n)$
- Mapping for vector $\mathbf{u}$ with $e_1 = u_1$:

$$e_i = \frac{\partial^{i-1} C(u_1, \ldots, u_i, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C(u_1, \ldots, u_{i-1}, 1, \ldots, 1)}{\partial u_1 \cdots \partial u_{i-1}}$$

- Test statistic T with $C_\perp(\mathbf{u}) = u_1 \cdot \cdots \cdot u_d$ and

  $D_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(E_i \leq \mathbf{u})$

$$T = n \int_{[0,1]^d} \{D_n(\mathbf{u}) - C_\perp(\mathbf{u})\}^2 dD_n(\mathbf{u})$$

- p-value: $\frac{1}{M} \sum_{b=1}^{M} \mathbf{I}(T_b \geq T)$

Back

gofKendallCvM

- Rescaled pseudo observations: $V_1 = C_n(U_1), \dots, V_n = C_n(U_n)$
- Empirical distribution function K: $K_n(v) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(V_i \leq v)$
- $H_0^{''} : K \in \mathcal{K}_0,\ H_0 \subset H_0^{''}$
- Test statistic:

$$T = n \int_0^1 (K_n(v) - K_{\theta_n})^2 dK_{\theta_n}(v)$$

- p-value: $\frac{1}{M}\sum_{b=1}^{M} \mathbf{I}(T_b \geq T)$
- Note: for bivariate Archimedean copulae H_0 and $H_0^{''}$ are equivalent

Back

gofKendallKS

- Rescaled pseudo observations: $V_1 = C_n(U_1), \dots, V_n = C_n(U_n)$
- Empirical distribution function K: $K_n(v) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(V_i \leq v)$
- $H_0^{''} : K \in \mathcal{K}_0,\ H_0 \subset H_0^{''}$
- Test statistic:

$$T = \sqrt{n} \sup_{[0,1]} |(K_n(v) - K_{\theta_n})| dK_{\theta_n}(v)$$

- p-value: $\frac{1}{M}\sum_{b=1}^{M} \mathbf{I}(T_b \geq T)$
- Note: for bivariate Archimedean copulae H_0 and $H_0^{''}$ are equivalent

Back

gofPIOSRn

- Hypothesis: $H_0 : C_0 \in \mathcal{C}$
- Define

$$S(\theta) = -E_0[\frac{\partial^2}{\partial\theta\partial\theta^\top} l\{U_1;\theta\}]$$
$$V(\theta) = -E_0[\frac{\partial}{\partial\theta} l\{U_1;\theta\} l^\top\{U_1;\theta\}]$$

with l the likelihood.

- Test statistic:

$$T = tr\{S(\theta^*)^{-1} - V(\theta^*)\}$$

- p-value: $p = \frac{1}{M}\sum_{b=1}^{M} \mathbf{I}(|T_b| \geq T)$

Back

gofPIOSTn

- Hypothesis: $H_0 : C_0 \in \mathcal{C}$
- With l the likelihood, define

$$\theta_n = \arg\min_{\theta} \sum_{i=1}^{n} l(U_i; \theta)$$

$$\theta_n^{-b} = \arg\min_{\theta} \sum_{b' \neq b}^{M} \sum_{i=1}^{m} l(U_i^{b'}; \theta), b = 1, \ldots, M$$

- Test statistic:

$$T = \sum_{b=1}^{M} \sum_{i=1}^{m} [l\{U_i^b; \theta_n\} - l\{U_i^b; \theta_n^{-b}\}]$$

- p-value: $p = \frac{1}{M} \sum_{b=1}^{M} \mathbf{I}(|T_b| \geq T)$

Back

gofRosenblattChisq

- Hypothesis: $C \in \mathcal{C}_0$
- Rosenblatt transformation, see gofRosenblattSnB
- Use Anderson-Darling test statistic: $G(x_i) = \chi^2_d(x_i)$ with $x_i = \sum_{j=1}^{d} (\Phi^{-1}(e_{ij}))^2$
- χ^2_d Chi-Squared distribution with d degrees of freedom and $\Phi \sim N(0,1)$
- Test statistic:

$$T = -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\log(G(x_i)) + \log(1 - G(x_{n+1-i}))]$$

- p-value: $p = \frac{1}{M} \sum_{b=1}^{M} \mathbf{I}(|T_b| \geq T)$

gofRosenblattGamma

- Hypothesis: $C \in \mathcal{C}_0$
- Rosenblatt transformation, see gofRosenblattSnB
- Use Anderson-Darling test statistic: $G(x_i) = \Gamma_d\,(x_i)$ with $x_i = \sum_{j=1}^{d}(-\log e_{ij})$
- $\Gamma_d()$ Gamma distribution with shape parameter d and 1
- Test statistic:

$$T = -n - \sum_{i=1}^{n} \frac{2i-1}{n}[\log(G(x_i)) + \log(1 - G(x_{n+1-i}))]$$

- p-value: $p = \frac{1}{M}\sum_{b=1}^{M} \mathrm{I}(|T_b| \geq T)$

Back

gofSn

- Hypothesis: $C \in \mathcal{C}_0$
- Test statistic:

$$T = n \int_{[0,1]^d} \{C_n(\mathbf{u}) - C_{\theta_n}(\mathbf{u})\}^2 dC_n(\mathbf{u})$$

- p-value: $p = \frac{1}{M} \sum_{b=1}^{M} \mathbf{I}(|T_b| \geq T)$

Back

gofKernel

- Hypothesis: $C \in \mathcal{C}_0$
- Copula density kernel estimator:

$$c_n(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} K_H[\mathbf{u} - (U_{i1}, \ldots, U_{id})^\top]$$

- $K_H(y) = K(H^{-1}y)/\det(H)$ with K bivariate quadratic kernel
- $H = 2.6073 n^{-1/6}\hat{\Sigma}^{1/2}$ with $\hat{\Sigma}$ sample covariance matrix
- Test statistic:

$$T = \int_{[0,1]^d} \{c_n(\mathbf{u}) - K_H * c(\mathbf{u}, \theta_n)\}\omega(\mathbf{u})d\mathbf{u}$$

  with $*$ convolution operator and $\omega(\mathbf{u})$ a weight function.
- p-value: $p = \frac{1}{M}\sum_{b=1}^{M} \mathbf{I}(|T_b| \geq T)$

Back

gofWhite I

- Hypothesis: $H_0 : \mathbf{H}(\theta) + \mathbf{S}(\theta) = 0$ with $\mathbf{H}(\theta)$ the Hessian matrix and $\mathbf{S}(\theta)$ the outer product of the score function
- With

$$\bar{d}(\theta_n) = \frac{1}{n}\sum_{i=1}^{n} vech(\mathbf{H}_n(\theta_n|\mathbf{u}) + \mathbf{S}_n(\theta_n|\mathbf{u}))$$

$$d(\theta_n) = vech(\mathbf{H}_n(\theta_n|\mathbf{u}) + \mathbf{S}_n(\theta_n|\mathbf{u}))$$

$$V_{\theta_n} = \frac{1}{n}\sum_{i=1}^{n}(d(\theta_n) - D_{\theta_n}\mathbf{H}_n(\theta_n)^{-1}\delta l(\theta_n))(d(\theta_n) - D_{\theta_n}\mathbf{H}_n(\theta_n)^{-1}\delta l(\theta_n))^\top$$

Back

gofWhite II

- Further define

$$D_{\theta_n} = \frac{1}{n} \sum_{i=1}^{n} [\delta_{\theta_k} d_l(\theta_n)]_{l=1,\dots,\frac{p(p+1)}{2},k=1,\dots,p}$$

with $l(\theta_n)$ log likelihood function and p length of θ

- Test statistic:

$$T_n = n(\bar{d}(\theta_n))^\top V_{\theta_n}^{-1} \bar{d}(\theta_n)$$

- p-value: $T > (1-\alpha)(\chi^2_{p(p+1)/2})^{-1}$

Back

Bibliography

Shulin Zhang, Ostap Okhrin, Qian M. Zhou and Peter Song (2015)
Goodness-of-Fit test for specification of semiparametric copula dependence models
SFB 649 Discussion Paper

Christian Genest, Bruno Rémillard and David Beaudoin (2007)
Goodness-of-Fit tests for copulas: A review and a power study
Insurance: Mathematics and Economics

Bibliography

Jun Yan (2007)
Enjoy the Joy of Copulas: With a Package copula
Journal of Statistical Software