1 Problem set-up

Suppose that $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ is a random sample from the distribution of X, where $\mathrm{X} = (X_1, \ldots, X_d)^t$ is a random vector. Let F_i be the distribution function of X_i for $i = 1, \ldots, d$ and C be the copula of X, which is the joint distribution function of $F_1(X_1)$, ..., $F_d(X_d)$. It is of interest to test

$$H_0 : C \in \mathcal{P} \text{ versus } H_1 :\notin \mathcal{P}, \tag{1}$$

where $\mathcal{P} = \{C_\theta : \theta \in \Theta\}$ is a given parametric family of copulas. For the testing problem in (1), various goodness-of-fit tests have been proposed. Chen and Huang (2007) have proposed a test based on the MISE of the difference between a nonparametric copula estimator and a semiparametric copula estimator. In Chen and Huang (2007), the nonparametric copula estimator is of the form $(d = 2)$

$$\hat{C}(u, v) = \tilde{C}(u, v) - b(u, v),$$

where $\tilde{C}(u, v)$ is a two-stage kernel estimator of $C(u, v)$ and $b(u, v)$ is an approximation of the bias of $\tilde{C}(u, v)$ obtained by direct calculation. However, such a calculation becomes complicated when the dimension d is large. To overcome this difficulty, in this study, a modified test statistic is considered, where the boundary bias correction term $b(u, v)$ is replaced by a statistic based on the semiparametric copula estimator. Such an approach have been used for boundary bias correction in density estimation in Fan (1994) and Fermanian (2005). The bias correction is valid if the underlying copula belongs to the given parametric family.

2 The test

The test statistic is an estimator of the quantity

$$\int_0^1 \cdots \int_0^1 E\left(\tilde{C}(u_1, \ldots, u_d) - b(\hat{\theta}, u_1, \ldots, u_d) - C_{\hat{\theta}}(u_1, \ldots, u_d)\right)^2 du_1 \cdots du_d,$$

where $\tilde{C}(u_1, \ldots, u_d)$, $b(\hat{\theta}, u_1, \ldots, u_d)$ and $C_{\hat{\theta}}(u_1, \ldots, u_d)$ are defined below.

- $\tilde{C}(u_1, \ldots, u_d)$: $\tilde{C}(u_1, \ldots, u_d)$ is a nonparametric copula estimator based on kernel estimation, which involves pre-determined kernel functions K_1 and K and bandwidthes b_0 and h_0. To describe this estimator, some notation will be introduced first. For $1 \le k \le d$, let

$$\hat{F}_k(x) = \frac{1}{n} \sum_{i=1}^{n} G_1\left(\frac{x - X_{k,i}}{b_0}\right),$$

  where $G_1(x) = \int_{-\infty}^{x} K_1(t)dt$. For $c \in [0, 1]$ and $h > 0$, let

$$K_{c,h}(x) = \frac{K(x)(a_2(c, h) - a_1(c, h)x)}{a_0(c, h)a_2(c, h) - a_1^2(c, h)} \text{ and } G_{c,h}(t) = \int_{-\infty}^{t} K_{c,h}(x)dx,$$

where

$$a_\ell(c,h) = \int_{-(c-1)/h}^{c/h} t^\ell K(t)dt \text{ for } \ell = 0,1,2,3.$$

For $1 \le k \le d$ and $1 \le i \le n$, let $X_{k,i}$ be the k-th component of X_i, then

$$\tilde{C}(u_1,\ldots,u_d) = \frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{d} G_{u_k,h_0}\left(\frac{u_k - \hat{F}_k(X_{k,i})}{h_0}\right).$$

- $C_{\hat{\theta}}(u_1,\ldots,u_d)$: $C_{\hat{\theta}}(u_1,\ldots,u_d)$ is a semiparametric estimator of $C(u_1,\ldots,u_d)$. Let c denote the copula density (existence assumed) corresponding to C. Then the joint probability density function of X_1, …, X_n is $\prod_{i=1}^n f(\mathrm{X}_i)$, where

$$f(x_1,\ldots,x_d) = c(F_1(x_1),\ldots,F_d(x_d))\prod_{k=1}^{d} f_k(x_k)$$

  and f_k is the marginal probability density function of X_k for $1 \le k \le d$. When C belongs to the parametric family $\mathcal{P} = \{C_\theta : \theta \in \Theta\}$, the"maximum likelihood estimator"of θ can be obtained if F_1, …, F_d are replaced by the empirical CDFs. Denote such an estimator by $\hat{\theta}$ and we have $C_{\hat{\theta}}(u_1,\ldots,u_d)$.

- $b(\hat{\theta},u_1,\ldots,u_d)$: $b(\hat{\theta},u_1,\ldots,u_d)$ is an estimator of the bias of $\tilde{C}(u_1,\ldots,u_d)$, which is approximately

$$\int\cdots\int C(u_1-s_1h,\ldots,u_d-s_dh)\prod_{k=1}^{d} K_{u_k,h}(s_k)ds_1\cdots ds_d - C(u_1,\ldots,u_d).$$

  Denote the above quantity by $b_1(u_1,\ldots,u_d,C)$. One can estimate $b_1(u_1,\ldots,u_d,C)$ by $b_1(u_1,\ldots,u_d,C_{\hat{\theta}})$, if $C \in \{C_\theta : \theta \in \Theta\}$.However, when d is large, it is time-consuming to compute the integral in $b_1(u_1,\ldots,u_d,C_{\hat{\theta}})$ and an estimator for the integral is needed. To obtain an estimator for the integral

$$I \stackrel{\text{def}}{=} \int\cdots\int C_{\hat{\theta}}(u_1 - s_1h,\ldots,u_d - s_dh)\prod_{k=1}^{d} K_{u_k,h}(s_k)ds_1\cdots ds_d,$$

  note that

$$\int\cdots\int C(u_1 - s_1h,\ldots,u_d - s_dh)\prod_{k=1}^{d} K_{u_k,h}(s_k)ds_1\cdots ds_d$$

  is the major term of $E(\tilde{C}(u_1,\ldots,u_d))$. Therefore, we can simulate IID random vectors S_1, …, S_m, where S_1 is a random sample from the distribution with cumulative distribution function $C_{\hat{\theta}}$. For $1 \le i \le$

m, let $\tilde{C}(u_1, \ldots, u_d, S_i)$ be the statistic $\tilde{C}(u_1, \ldots, u_d)$ with the sample $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ replaced by S_i. Then

$$m^{-1} \sum_{i=1}^{m} \tilde{C}(u_1, \ldots, u_d, S_i) \approx E\tilde{C}(u_1, \ldots, u_d, S_1) \approx I$$

and the bias estimator $b(\hat{\theta}, u_1, \ldots, u_d)$ is given by

$$\frac{1}{m} \sum_{i=1}^{m} \tilde{C}(u_1, \ldots, u_d, S_i) - C_{\hat{\theta}}(u_1, \ldots, u_d).$$

Here m is pre-determined.

To estimate

$$II \stackrel{\text{def}}{=} \int_0^1 \cdots \int_0^1 E\left(\tilde{C}(u_1, \ldots, u_d) - b(\hat{\theta}, u_1, \ldots, u_d) - C_{\hat{\theta}}(u_1, \ldots, u_d)\right)^2 du_1 \cdots du_d$$

to obtain a test statistic for the problem in (1), let $T = (\mathrm{X}_1, \ldots, \mathrm{X}_n)$, $u = (u_1, \ldots, u_d)$, and

$$g(T, u) = \left(\tilde{C}(u_1, \ldots, u_d) - b(\hat{\theta}, u_1, \ldots, u_d) - C_{\hat{\theta}}(u_1, \ldots, u_d)\right)^2.$$

Simulate IID random vectors (T_1, U_1), ..., (T_m, U_m) such that T_1, ..., T_m are bootstrap samples based on $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$, U_1 is a random sample of size d from the uniform distribution on $[0, 1]$, and T_1 and U_1 are independent. Then II can be estiamted by $W(\mathrm{X}_1, \ldots, \mathrm{X}_n) = m^{-1} \sum_{i=1}^{m} g(T_i, U_i)$, which is the test statistic considered in this study. The testing procedure is as follows. Simulate IID random vectors T_1^*, ..., T_m^* such that T_1^* is a random sample of size n from the distribution with copula $C_{\hat{\theta}}$ and marginals $\hat{F}_1$, ..., $\hat{F}_d$. Obtain the test statistic $W(T_i^*)$ for $1 \leq i \leq m$ and let c_α be the $1 - \alpha$ quantile of $W(T_1^*)$, ..., $W(T_m^*)$. Reject H_0 at level α if $W(\mathrm{X}_1, \ldots, \mathrm{X}_n) > c_\alpha$.

3 Summary

A goodness of test for copula modeling based on kernel estimation has been proposed and the boundary bias has been corrected. The power of the test needs be investigated through large scale simulation.

4 References

- Chen, S. and Huang, T. (2007), Nonparametric estimation of copula functions for dependence modeling, the Canadian Journal of Statistics, 35(2), 1-18.

- Fan, Y. (1994) Testing the goodness of fit of a parametric density function by kernel method, Econometric Theory, 10, 316V356.
- Fermanian, J. D. (2005) Goodness-of-fit tests for copulas, Journal of Multivariate Analysis, 95, 119-152