A goodness-of-fit test for copulas

Prokhorov, Artem

Concordia University

2008

A Goodness-of-fit Test for Copulas

Artem Prokhorov*

August 13, 2008

Abstract

A new goodness-of-fit test of copulas is proposed. It is based on restrictions on certain elements of the information matrix and so relates to the White (1982) specification test. The test avoids the need to correctly specify and consistently estimate a parametric model for the marginal distributions. It does not involve kernel weighting and bandwidth selection or parametric bootstrap and is relatively simple compared to other available tests.

JEL Classification: C13

Keywords: Copula, MLE, Information Matrix, Goodness-of-fit

*Department of Economics, Concordia University, Montreal, PQ H3G1M8 Canada; email: artem.prokhorov@concordia.ca

1 Introduction

Copulas are useful because they allow to model dependence between random variables separately from their marginal distributions. Consider two continuous random variables X_1 and X_2 with cdf's F_1 and F_2 and pdf's f_1 and f_2, respectively. Suppose the joint cdf of (X_1, X_2) is H and the joint pdf is h. A copula is a function $C(u, v)$ such that $H = C(F_1, F_2)$ or, in densities, $h = c(F_1, F_2)f_1f_2$. The marginal densities f_1 and f_2 are now "extracted" from the joint density and the copula density c captures the entire dependence between X_1 and X_2. Sklar (1959) showed that given H, F_1, F_2 there exists a unique C. So, given F_1 and F_2, the choice is which copula C to use.

If the chosen copula is correct, $C(F_1, F_2)$ is the correct joint distribution of (X_1, X_2). Then one may base an estimation of the dependence parameters (parameters of the copula function) on the correctly specified joint likelihood without worrying about modeling the marginal distributions (they can be estimated nonparametericallly). Such likelihood-based estimators are consistent. They have been used extensively in applications in finance (e.g., Patton, 2006; Breymann et al., 2003), in risk management (e.g., Embrechts et al., 2003, 2002) and in health and labor economics (Smith, 2003; Cameron et al., 2004).

However, if the copula function is incorrect, the joint distribution is misspecified. This generally means that estimators based on the joint likelihood will be inconsistent. In particular, the copula dependence parameter will be inconsistent whether the marginal distributions are estimated parametrically or nonparametrically. Moreover, copula misspecification may affect consistency of marginal parameter estimates. Suppose interest is in efficient estimation of the marginal distribution parameters using copula-based likelihood. Under copula misspecification, such estimators are generally inconsistent (see Prokhorov and Schmidt, 2008).

It is therefore important to have a simple and reliable test of copula correctness.

There exist several copula goodness-of-fit tests. Panchenko (2005) proposes a test based on a V-statistic. His test has an unknown asymptotic distribution and depends on the choice of bandwidth. Nikoloulopoulos and Karlis (2008) propose a test based on the Mahalanobis squared distance between the original and the simulated likelihoods. Their test uses parametric bootstrap. Fermanian (2005) proposes two tests based on a kernel estimation of the copula function. Dobric and Schmid (2007) propose a test based on Rosenblatt's transform. Their test procedure is not directly applicable if the marginal distributions are unknown. Prokhorov and Schmidt (2008) propose a conditional moment test, which tests if the copula-based score function has zero mean. Their test does not distinguish between the correct copula and any

other copula that has a zero mean score function.

The test proposed in this paper is based on the information matrix equality which involves the copula-based Hessian and outer-product of the score. The statistic has a standard distribution and accounts for the use of empirical marginal distributions in place of the true ones. The test is proposed in Section 3. Section 2 discusses the connection between copulas and the information matrix equality. As an illustration, Section 4 tests goodness-of-fit of the Gaussian copula in a model with two stock indices.

2 Copulas and Information Matrix Equivalence

Consider an N-dimensional copula $C(u_1, \ldots, u_n)$ and N univariate marginals $F_n(x_n)$, $n = 1, \ldots, N$. Then, by Sklar's theorem, the joint distribution of $(X_1, \ldots, X_N)$ is given by

$$H(x_1, \ldots, x_N) = C(F_1(x_1), \ldots, F_N(x_N)). \tag{1}$$

Assume F_n are continuous, so $C(u_1, \ldots, u_n)$ is unique.

The joint density of $(X_1, \ldots, X_N)$ is

$$\begin{aligned} h(x_1, \ldots, x_N) &= \left. \frac{\partial^N C(u_1, \ldots, u_N)}{\partial u_1 \ldots \partial u_N} \right|_{u_n = F_n(x_n), n=1,\ldots,N} \prod_{n=1}^{N} f_n(x_n) \\ &= c(F_1(x_1), \ldots, F_N(x_N)) \prod_{n=1}^{N} f_n(x_n), \end{aligned} \tag{2}$$

where $c(u_1, \ldots, u_N)$ is the copula density.

Copula functions usually include parameters. For example, the N-variate Gaussian copula includes $N(N-1)/2$ parameters. This copula has the form

$$\Phi_N(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_N)), \tag{3}$$

where Φ_N is the joint distribution function of N standard normal covariates with a given correlation matrix and Φ^{-1} is the inverse of the standard normal cdf. For the Gaussian copula, the copula parameters are simply the distinct elements of the correlation matrix used to construct the multivariate normal distribution Φ_N. (See Nelsen, 2006; Joe, 1997, for other examples).

Let subscript θ denote the dependence parameter vector of a copula function and let p denote its dimension. It is well known that if there exists a value θ_o such that $H(x_1, \ldots, x_N) = C_{\theta_o}(F_1(x_1), \ldots, F_N(x))$ then we have a correctly specified likelihood model and, under regularity conditions, the MLE is consistent for θ_o. Moreover, in this case White (1982)'s information matrix equivalence theorem holds: the Fisher information matrix can be equivalently calculated as minus the expected Hessian or as the expected outer product of the score function.

More notation is needed. Assume that the likelihood is (three times) continuously differentiable and the relevant expectations exist. Let H_θ denote the expected Hessian matrix of $\ln c_\theta$ and let C_θ denote the expected outer product of the corresponding score function. Then,

$$\begin{aligned} H_\theta &= E\nabla_\theta^2 \ln c_\theta(F_1(x_1), \ldots, F_N(x_N)) \\ C_\theta &= E\nabla_\theta \ln c_\theta(F_1(x_1), \ldots, F_N(x_N))\nabla_\theta' \ln c_\theta(F_1(x_1), \ldots, F_N(x_N)), \end{aligned}$$

where "∇" denotes partial derivative with respect to θ.

The White (1982) information matrix equivalence theorem essentially says that, under correct specification of the copula,

$$-H_{\theta_o} = C_{\theta_o}. \tag{4}$$

The copula misspecification test we propose uses this equality.

3 Testing Procedure

In practice we do not observe θ_o. Moreover the matrices H_θ and C_θ contain the marginals F_n which are often unknown. We can, however, easily estimate these quantities. In particular, it is common to use the empirical distribution function $\hat{F}_n$ in place of F_n, a consistent estimate $\hat{\theta}$ in place of θ_o, the sample averages $\hat{H}$ and $\hat{C}$ in place of the expectations H and C.

Given T observations on $(x_1, \ldots, x_N)$, the empirical distribution function is given by

$$\hat{F}_n(s) = T^{-1} \sum_{t=1}^{T} I\{x_{nt} \leq s\}, \tag{5}$$

where $I\{\cdot\}$ is the indicator function and s takes values in the observed set of x_n. Then, $\hat{\theta}$ – a consistent estimator of θ_o sometimes called the Canonical Maximum Likelihood estimator – is the solution to

$$\max \sum_{t=1}^{T} \ln c_\theta(\hat{F}_1(x_{1t}), \ldots, \hat{F}_N(x_{Nt})).$$

To introduce the sample versions of H and C, we define new notation. Let

$$\begin{aligned} H_t(\theta) &= \nabla^2_\theta \ln c_\theta(\hat{F}_1(x_{1t}), \ldots, \hat{F}_N(x_{Nt})), \\ C_t(\theta) &= \nabla_\theta \ln c_\theta(\hat{F}_1(x_{1t}), \ldots, \hat{F}_N(x_{Nt})) \nabla'_\theta \ln c_\theta(\hat{F}_1(x_{1t}), \ldots, \hat{F}_N(x_{Nt})). \end{aligned}$$

Then, we can write the sample equivalents of H_θ and C_θ as

$$\begin{aligned} \hat{H}_\theta &= T^{-1} \sum_{t=1}^{T} H_t(\theta), \\ \hat{C}_\theta &= T^{-1} \sum_{t=1}^{T} C_t(\theta). \end{aligned}$$

We can now base the test on the distinct elements of the testing matrix $\hat{H}_{\hat{\theta}} + \hat{C}_{\hat{\theta}}$. Given that the dimension of θ is p, there are $p(p+1)/2$ such elements. Under correctness of copula they are all zero. This is in essence the likelihood misspecification test of White (1982). However, he deals with the full but possibly incorrect parametric log-density. So the elements of his testing matrix (he calls them "indicators") do not contain empirical estimates of the marginal distributions. It turns out that this difference precludes a direct application of his test statistic in our setting.

White (1982) points out that it is sometimes appropriate to drop some of the indicators because they are identically zero or represent a linear combination of the others. When $p = 1$ – the case of a bivariate one-parameter copula – this problem does not arise. Whether it arises in higher dimensional models is a copula-specific question that is not addressed in this paper. Assume that no indicators need be dropped.

Following White (1982) define

$$d_t(\theta) = vech(H_t(\theta) + C_t(\theta))$$

so that the indicators of interest are

$$\hat{D}(\theta) = T^{-1} \sum_{t=1}^{T} d_t(\theta).$$

Let $D_\theta = E d_t(\theta)$.

What differs our setting from White (1982) is that nonparametric estimates of the marginals are used to construct the joint density. It is well known that the empirical distribution converges to the true distribution at the rate $\sqrt{T}$ so the CMLE estimate $\hat{\theta}$ that uses empirical distributions $\hat{F}_n$ is still $\sqrt{T}$-consistent.

However, the asymptotic variance matrix of $\sqrt{T}\hat{\theta}$ will be affected by the nonparametric estimation of the marginals. Therefore, the asymptotic variance of $\sqrt{T}\hat{D}_{\hat{\theta}}$ will also be affected. The proper adjustments of the variance matrix for the general two-step semiparametric estimation are given in Newey (1994); Chen and Fan (2004). Specifically, we will use the variance formula derived by Chen and Fan (2004) for the case when empirical marginal distributions are used in the parametric estimation of copulas.

Proposition 1 *Under the correct copula specification, the information matrix test statistic*

$$\mathscr{I} = T\hat{D}'_{\hat{\theta}} V_{\theta_o}^{-1} \hat{D}_{\hat{\theta}}, \tag{6}$$

where V_{θ_o} is defined in the appendix, is distributed asymptotically as $\chi^2_{p(p+1)/2}$.

In practice, a consistent estimate of V_{θ_o} will be used.

4 An application of the test

To demonstrate how the test procedure of the previous section can be applied in practice we test whether the Gaussian copula is appropriate for modeling dependence between an American and an European stock index.

We use the FTSE100 and DJIA close from June 26, 2000 to June 28, 2008. We have 1972 pairs of returns after eliminating holidays. Table 1 contains descriptive statistics of the returns.

An AR-GARCH filter that we apply to the return data accounts for most of the observed autocorrelation in the level and squared returns. The preferred models contain Normal innovations – allowing for Student-t innovations resulted in a relatively high estimate of the degrees of freedom (over 9) and did not improve the fit substantially. Table 2 reports the results of AR-GARCH modeling.

Table 3 contains the results of the testing procedure. The estimated parameter θ, which is just the correlation coefficient in the case of the Gaussian copula in (3), is high, positive and significant. There are two test statistics. One is called *unadjusted*. It is incorrect because it ignores the fact that the cdf's were estimated semiparametrically in the first step. The other is called *adjusted*. This is the statistic that uses the correct variance formula. Note that adjusting for estimation of cdf's makes the statistic larger. At 5%, we reject the hypothesis that the Gaussian copula is appropriate to model dependence between the two time series.

Table 1: Summary statistics of returns series

	FTSE	DJIA
$mean$	0.0001	-.0001
$st.d.$	0.107	0.103
m_3	0.104	0.020
m_4	6.101	6.590
$Q(20)$	52.97	33.29

5 Concluding remarks

Unlike many available alternatives, the test proposed in this paper is simple and easy to implement. Essentially it is a special case of White's information equivalence test with the complication of a first-step empirical density estimation. However, as such, it also inherits a number of drawbacks.

Horowitz (1994), for example, points out to large deviations of the finite-sample size of the White test from its nominal size based on asymptotic critical values and suggests using bootstrapped critical values instead.

Another complication is the need to evaluate the third derivative of the log-copula density function. Lancaster (1984) shows how one can construct the test statistic without using the third order derivatives.

Clearly all these considerations apply to our test statistic.

6 Appendix

Sketch of proof of Proposition 1. Provided that the derivatives and expectation exist, let

$$\nabla D_\theta = E\nabla_\theta d_t(\theta)$$

and

$$\nabla \hat{D}_\theta = T^{-1}\sum_{t=1}^{T}\nabla_\theta d_t(\theta).$$

Start by MVT for $\sqrt{T}\hat{D}_{\hat{\theta}}$

$$\sqrt{T}\hat{D}_{\hat{\theta}} = \sqrt{T}\hat{D}_{\theta_o} + \nabla\hat{D}_{\bar{\theta}}\sqrt{T}(\hat{\theta}-\theta_o), \tag{7}$$

Table 2: AR-GARCH estimates and standard errors

	FTSE	DJIA
μ	-0.0004(0.0002)	-0.0004(0.0002)
AR(1)	-0.0703(0.0230)	-
ω	0.0000(0.0000)	0.0000(0.0000)
α	0.1154(0.0176)	0.0738(0.0170)
β	0.8743(0.0199)	0.9191(0.0199)
ll	6393.6	6433.21
m_3	-0.0138	-0.098
m_4	3.343	3.736
$Q(20)$	23.69	26.71
$Q^2(20)$	15.44	31.05

Table 3: Testing the Gaussian copula

$\hat{\theta}$	0.4785(0.0188)
$\mathscr{I}$ unadjusted	2.751
$\mathscr{I}$ adjusted	3.528
$p-value$ for $\mathscr{I}_a$	0.0603

for $\bar{\theta}$ between θ_o and $\hat{\theta}$. Then,

$$\sqrt{T}\hat{D}_{\hat{\theta}} = \sqrt{T}\hat{D}_{\theta_o} + \nabla D_{\theta_o}\sqrt{T}(\hat{\theta} - \theta_o) + o_p(1). \tag{8}$$

Now, Chen and Fan (2004) show that

$$\sqrt{T}(\hat{\theta} - \theta_o) \to N(0, B^{-1}\Sigma B^{-1}), \tag{9}$$

where

$$\begin{aligned} B &= -H_{\theta_o} \\ \Sigma &= \lim_{T\to\infty} Var(\sqrt{T}A_T^*) \\ A_T^* &= \frac{1}{T}\sum_{t=1}^{T}(\nabla \ln c(U_t, V_t, \theta_o) + W_1(U_t) + W_2(V_t)) \end{aligned} \tag{10}$$

Here terms $W_1(U_t)$ and $W_2(V_t)$ are the adjustments needed to account for the empirical distributions used in place of the true distributions. These terms are calculated as follows:

$$W_1(U_t) = \int_0^1\int_0^1 [I\{U_t \le u\} - u]\nabla^2_{\theta,u} \ln c(u, v; \theta_o)\, c(u, v; \theta_o) dvdu$$
$$W_1(V_t) = \int_0^1\int_0^1 [I\{V_t \le v\} - v]\nabla^2_{\theta,v} \ln c(u, v; \theta_o)\, c(u, v; \theta_o) dvdu$$

So,

$$\sqrt{T}(\hat{\theta} - \theta_o) = B^{-1}\sqrt{T}A_T^* + o_p(1). \tag{11}$$

Then,

$$\sqrt{T}\hat{D}_{\hat{\theta}} = \sqrt{T}\hat{D}_{\theta_o} + \nabla D_{\theta_o}B^{-1}\sqrt{T}A_T^* + o_p(1), \tag{12}$$

and we have the asymptotic distribution of $\sqrt{T}\hat{D}_{\hat{\theta}}$:

$$\sqrt{T}\hat{D}_{\hat{\theta}} \to N(0, V(\theta_o)), \tag{13}$$

where

$$\begin{aligned} V_{\theta_o} &= E[d_t(\theta_o) + \nabla D_{\theta_o}B^{-1}(\nabla \ln c(U_t, V_t, \theta_o) + W_1(U_t) + W_2(V_t))] \\ &\quad \times [d_t(\theta_o) + \nabla D_{\theta_o}B^{-1}(\nabla \ln c(U_t, V_t, \theta_o) + W_1(U_t) + W_2(V_t))]'. \end{aligned}$$

References

Breymann, W., A. Dias, and P. Embrechts (2003): "Dependence structures for multivariate high-frequency data in finance," *Quantitative Finance*, 3, 1–14, http://www.iop.org/EJ/abstract/1469-7688/3/1/301/.

Cameron, A. C., T. Li, P. K. Trivedi, and D. M. Zimmer (2004): "Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts," *Econometrics Journal*, 7, 566–84.

Chen, X. and Y. Fan (2004): "Estimation of Copula-Based Semiparametric Time Series Models." .

Dobric, J. and F. Schmid (2007): "A goodness of fit test for copulas based on Rosenblatt's transformation," *Computational Statistics and Data Analysis*, 51, 4633–4642.

Embrechts, P., A. Höing, and A. Juri (2003): "Using copulae to bound the Value-at-Risk for functions of dependent risks," *Finance and Stochastics*, 7, 145–167.

Embrechts, P., A. McNeil, and D. Straumann (2002): "Correlation and dependence in risk management: properties and pitfalls," in *Risk Management: Value at Risk and Beyond*, ed. by M. Dempster, Cambridge: Cambridge University Press, 176–223.

Fermanian, J.-D. (2005): "Goodness-of-fit tests for copulas," *J. Multivar. Anal.*, 95, 119–152.

Horowitz, J. L. (1994): "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Joe, H. (1997): *Multivariate models and dependence concepts*, vol. 73 of *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Lancaster, T. (1984): "The Covariance Matrix of the Information Matrix Test," *Econometrica*, 52, 1051–1053.

Nelsen, R. B. (2006): *An Introduction to Copulas*, vol. 139 of *Springer Series in Statistics*, Springer, 2 ed.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Nikoloulopoulos, A. K. and D. Karlis (2008): "Copula model evaluation based on parametric bootstrap," *Computational Statistics and Data Analysis*, 52, 3342–3353.

Panchenko, V. (2005): "Goodness-of-fit test for copulas," *Physica A*, 355, 176 (7 pages).

Patton, A. (2006): "Modelling Asymmetric Exchange Rate Dependence," *International Economic Review*, 47, 527–556.

Prokhorov, A. and P. Schmidt (2008): "Robustness, redundancy and validity of copulas in likelihood models," .

Sklar, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Smith, M. D. (2003): "Modelling sample selection using Archimedian copulas," *Econometrics Journal*, 6, 99–123.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–26.