Goodness-of-fit test for copulas

Valentyn Panchenko

CeNDEF, University of Amsterdam
Roetersstraat 11, 1018WB Amsterdam, The Netherlands

Abstract

Copulas are often used in finance to characterize the dependence between assets. However, a choice of the functional form for the copula is an open question in the literature. This paper develops a goodness-of-fit test for copulas based on positive definite bilinear forms. The suggested test avoids the use of plug-in estimators that is the common practice in the literature. The test statistics can be consistently computed on the basis of V-estimators even in the case of large dimensions. The test is applied to a dataset of US large cap stocks to assess the performance of the Gaussian copula for the portfolios of assets of various dimension. The Gaussian copula appears to be inadequate to characterize the dependence between assets.

Key words: copula, correlation, goodness-of-fit, multivariate time series, nonparametric statistics, V-statistics, bootstrap
PACS: 89.65.Gh

1 Introduction

The copula proved to be a handy instrument in the analysis of multivariate time series. It allows to capture the full dependence within multivariate time series without specifying a shape of the marginal distributions. This result is due to Sklar's theorem: any multivariate distribution can be decomposed into a copula and its marginals; if the marginal distributions are continuous the copula is unique. Moreover, the copula is invariant under strictly increasing transformations. For a thorough analysis of copulas see Nelsen [1].

Due to these favorable properties copulas proved to be useful in financial applications, e.g. risk management, portfolio aggregation, spillover effects (for

Email address: v.panchenko@uva.nl (Valentyn Panchenko).

review see Bouyé *et al.* [2]). In the Econophysics literature copulas were applied by Wise and Bhansali [3] and Malevergne and Sornette [4].

The major drawback in the copula approach is that there is no indication of what parametric form the copula is. Thus, to proceed with a traditional parametric analysis a specific functional form has to be assumed for the copula. Though many functional forms have been suggested [1], there are no general guidelines for optimal parametric copula selection.

Up to now there were a few studies trying to tackle the problem. Durrleman *et al.* [5] constructed the Deheuvels or empirical copula and compared it with various parametric copulas on the basis of bivariate data. The discrete L^2 norm was chosen as a criterion of fit. Malevergne and Sornette [4] investigated whether bivariate data dependences can be described by the Gaussian copula. Their tests are based on Kolmogorov and Anderson-Darling distances and their modifications. Patton *et al.* [6] considered multivariate data analysis to test the hypothesis of the Gaussian and Student copula. However, according to these authors their first test suffers the curse of dimensionality. The second test does not have this problem, but may be inconsistent.

This study develops an alternative goodness-of-fit test for bivariate and multivariate copulas. The test is based on a divergence measure first introduced by Diks *et al.* [7]. This measure, a kernel-based positive definite bilinear form, can be consistently estimated using V-statistics. It does not require the usage of plug-in estimators that is now a common practice in the field [8] and separates the problem of inference from consistent estimation of multivariate densities. The proposed test is nonparametric and may be applied to any functional form of the copula.

As an example of the empirical application, the fit of the Gaussian copula is evaluated on the US large cap stocks returns data. Both bivariate and multivariate portfolios of assets are considered in the analysis.

2 Estimation of copula parameters

Consider the case where a copula $C(\boldsymbol{u}),\ \boldsymbol{u} \in \mathbb{R}^N$ and marginals $F_n(x_n)$ are continuous. According to Sklar's theorem, the joint distribution $F(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^N$ can be represented as

$$F(\boldsymbol{x}) = C(F_1(x_1), F_2(x_2), \ldots, F_N(x_N)). \tag{1}$$

The corresponding density function is

$$f(\boldsymbol{x}) = c(F_1(x_1), F_2(x_2), \ldots, F_N(x_N)) \prod_{n=1}^{N} f_n(x_n), \tag{2}$$

where $f_n(x_n)$ is the density of the marginal $F_n(x_n)$ and $c(\boldsymbol{u})$ is the density of the copula $C(\boldsymbol{u})$

$$c(u_1, \ldots, u_N) = \frac{\partial C(u_1, \ldots, u_N)}{\partial u_1, \ldots, \partial u_N}. \tag{3}$$

The canonical maximum likelihood (CML) method [2] is used to estimate the vector of parameters α of the copula. First the data $\{x_1^t, x_2^t, \ldots, x_N^t\}_{t=1}^T$ are transformed into the corresponding empirical distributions $\hat{F}_n(x_n)$ through

$$\hat{F}_n(x_n) = \frac{1}{T} \sum_{t=1}^{T} I(x_n^t \leq x_n). \tag{4}$$

The vector of parameters α is estimated semi-parametrically maximizing log-likelihood for the copula density c, given the empirical marginals $\hat{F}_n(x_n)$

$$\hat{\alpha} = \arg\max \sum_{t=1}^{T} \ln c(\hat{F}_1(x_1^t), \ldots, \hat{F}_N(x_N^t); \alpha). \tag{5}$$

3 A notion of distance between probability distributions

Following Diks *et al.* [7] for an integrable functions f_1 and f_2 define the bilinear form

$$\langle f_1 | \kappa_N | f_2 \rangle = \int\int \kappa_N(\boldsymbol{s}_1, \boldsymbol{s}_2) f_1(\boldsymbol{s}_1) f_2(\boldsymbol{s}_2) d\boldsymbol{s}_1 d\boldsymbol{s}_2, \tag{6}$$

where $\boldsymbol{s}_1, \boldsymbol{s}_2 \in \mathbb{R}^N$ and $\kappa_N(\cdot, \cdot)$ is a positive definite symmetric kernel such as

$$\kappa_N(\boldsymbol{s}_1, \boldsymbol{s}_2) = e^{-\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|^2/(2Nd^2)}, \tag{7}$$

where $\|\cdot\|$ denotes Euclidean norm in $\mathbb{R}^N$ and $d > 0$ is a smoothing parameter, or a bandwidth. This kernel allows factorization

$$\kappa_N(\boldsymbol{s}_1, \boldsymbol{s}_2) = \prod_{i=1}^{N} \kappa_1(s_{1i}, s_{2i}). \tag{8}$$

The Gaussian kernel is chosen for convenience. In general, other positive definite kernel functions can be used.

According to a theorem [9] for integrable function g, $\langle g|\kappa_N|g\rangle \geq 0$ and $\langle g|\kappa_N|g\rangle = 0$ if and only if $g = 0$ almost everywhere. In fact, $\langle f_1|\kappa_N|f_2\rangle$ is an inner product of f_1 and f_2, which can be used as a measure of distance between f_1 and f_2. Note that the defined bilinear form (6) is an expectation of the kernel κ_N taken with respect to the independent random vectors $\boldsymbol{S}_1$ with the probability density function $f_1(\boldsymbol{s}_1)$ and $\boldsymbol{S}_2$ with the probability density function $f_2(\boldsymbol{s}_2)$.

$$E[\kappa_N(\boldsymbol{S}_1, \boldsymbol{S}_2)] = \int\int \kappa_N(\boldsymbol{s}_1, \boldsymbol{s}_2) f_1(\boldsymbol{s}_1) f_2(\boldsymbol{s}_2) d\boldsymbol{s}_1 d\boldsymbol{s}_2. \tag{9}$$

Define a squared distance Q between f_1 and f_2 as

$$Q = \langle f_1 - f_2|\kappa_N|f_1 - f_2\rangle . \tag{10}$$

It follows from the aforementioned theorem that Q becomes zero only when $f_1(\cdot)$ and $f_2(\cdot)$ are equal. Following the properties of the inner product, Q can be decomposed as follows

$$Q = Q_{11} - 2Q_{12} + Q_{22}, \tag{11}$$

where $Q_{ij} = \langle f_i|\kappa_N|f_j\rangle$. Each term of the above decomposition can be consistently estimated using V-statistics

$$\widehat{Q}_{ij} = \frac{1}{T^2} \sum_{t_1=1}^{T} \sum_{t_2=1}^{T} \kappa_N(\boldsymbol{S}_i^{t_1}, \boldsymbol{S}_j^{t_2}), \tag{12}$$

where $\boldsymbol{S}^t$ denotes a realization of the random vector $\boldsymbol{S}$ at a time t.

Relying on the theory of V-statistics it is also possible to develop asymptotics [10] for the functional of interest Q.

4 Testing procedure

Our goal is to test whether a specific functional copula can adequately describe the dependence between given series. Serial independence of the individual series is assumed throughout this section. First, the data series are transformed to the empirical marginal cumulative distributions according to Eq. (4). Jointly this transformed data can be viewed as a sample from the true copula. In the second stage, the parameters of the specific copula of interest are estimated as described in Section 2. Next, we sample from the parametric copula with the estimated parameters. Under the null hypothesis both series, the transformed empirical series $\boldsymbol{S_1}$ (parallel to the notation of Section 3) and the sampled series $\boldsymbol{S_2}$, originate from the same copula. The squared distance Q is used

to compare the series. Its estimate denoted by $\hat{Q}$ is computed element-wise according to Eq. (11) and Eq. (12).

Since the asymptotic theory for the statistics Q is still under development, the parametric bootstrap [11] is proposed to determine p-values of the test. We repeatedly (B-times) sample series $\boldsymbol{S}^*_{1j},\ 1 \leq j \leq B$ from the copula with estimated parameters using different seeds of the random number generator and compare them with the initially sampled series $\boldsymbol{S}_2$. The corresponding distances are denoted by Q^*_j. P-values are generated following the standard procedure of comparing $\widehat{Q}$, an estimated distance between the transformed empirical data $\boldsymbol{S}_1$ and the initially sampled series $\boldsymbol{S}_2$, with values of $\widehat{Q}^*_j$, an estimated distance between $\boldsymbol{S}^*_{1j}$ and $\boldsymbol{S}_2$, i.e.

$$\widehat{p} = \frac{\sum_{i=1}^{B} I(\widehat{Q} \geq \widehat{Q}^*_i) + 1}{B+1}, \tag{13}$$

where I is indicator function. The null hypothesis is rejected whenever $\widehat{p} \leq \alpha$, where α is a size of the test. The minimal number of replications is $B = \frac{1}{\alpha} - 1$ (for a 5%-test, the minimal value of B is 19). Note that $\widehat{Q}_{22}$, an estimated element of the decomposition (11) depends only on $\boldsymbol{S}_2$. Therefore, it takes the same values in $\widehat{Q}$ and $\widehat{Q}^*_j$ and may be abandoned without any effect on p-values.

An optimal value of the bandwidth parameter d is determined via simulations. Power of the test is used as a criterion for the selection. The highest power is achieved for the value of the bandwidth $d = 0.05$ (time series are standardized to the unit variance).

5 Testing the Gaussian copula hypothesis

In this section we apply the previously described procedure to test the Gaussian copula hypothesis. The N-variate Gaussian copula with the correlation matrix R is defined as

$$GaC^N_R(\boldsymbol{u}) = \Phi^N_R(\Phi^{-1}(u_1), ..., (\Phi^{-1}(u_N)), \tag{14}$$

where Φ^N_R denotes the joint distribution function of $N-$variate standard normal distribution with the linear correlation matrix R, and Φ denotes the distribution function of the univariate standard normal distribution. Due to relatively simple estimation procedure and a possibility to incorporate multivariate series, the Gaussian copula is now widely applied in finance. However, this copula is radially symmetric and has zero tail dependence [1]. Thus, it may be inadequate to characterize the dependence between the financial series.

The proposed goodness-of-fit test is used to determine whether the Gaussian copula can adequately describe the dependence between the series of (log)-returns of the US large cap stocks. The dataset consists of US traded securities with a volume over 10 million per day and market capitalization over one billion US dollars as of the 1st of April 2003, which comprises 33 stocks from different sectors. The selection of the large cap stocks ensures considerable interest from the side of investors. The length of the time series covering the period January 1997-March 2003 is 1607.

Collections of the assets of the dimension 2, 5 and 10 were randomly selected 100 times, the full set of 33 stocks was also included in the analysis. In most of the applications individual financial series are filtered with (G)ARCH process to remove long-term serial dependence in the variance [6]. Serial independence of the individual series is required by our test. However, it is possible that (G)ARCH filtering destroys some of the dependence structure between assets [4]. Therefore, we consider both the raw data and the GARCH(1,1) filtered series and compare the corresponding outcomes of the test. A presence of serial dependence in the raw data may weaken power of the test.

The CML parameter estimator (5) in case of the Gaussian copula reduces to

$$\hat{R} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t \mathbf{y}_t^{'}, \tag{15}$$

where $\mathbf{y}_t = (\Phi^{-1}(\hat{F}_1(x_1^t)), ..., \Phi^{-1}(\hat{F}_N(x_N^t))))^{'}$ and $\hat{F}_i(x_i^t)$ is computed according to Eq. (4). Sampling from the Gaussian copula is straightforward: a multivariate standard normal variable $\mathbf{z} \sim N(0, R)$ transformed to $(\Phi(z_1), ..., \Phi(z_N))$ yields a sample originating from the Gaussian copula with the correlation matrix R. Next, a testing procedure described in Section 4 is applied. Table 1 presents the rejection rates with the nominal size set to 0.05. The number of replications B was set to 19 and bandwidth $d = 0.05$.

Table 1
Rejection rates for the null hypothesis of the Gaussian copula for US large cap stock (log)-returns

procedure\dimension	2	5	10	33
raw returns	14/100	32/100	52/100	1/1
GARCH(1,1) filtered returns	10/100	19/100	45/100	1/1

The analysis reveals that the rejection rates are notably larger than the nominal size. The number of rejections increases with the dimension of the asset collections. GARCH(1,1) filtered data indicate less deviations from the Gaussian copula possibly because of the change in the dependence structure. Consequently, the analysis suggests that the Gaussian copula is inadequate

in characterization of the dependence between US large cap stocks, especially for multivariate collections of assets. Similar rejections rates were found by Malevergne and Sornette [4] for bivariate collections of stocks from the dataset similar to one applied in this paper. Patton *et al.* [6] report lower than the nominal rejection rates for bivariate collections and the rejection probability close to one for 5- and higher dimensional collections. The difference may be attributed to the difference in data selection and testing procedures.

6 Conclusions

In line with the current increasing use of nonlinear time series analysis the goodness-of-fit test for copulas is suggested. The test procedure remains consistent and applicable even in the case of higher dimensions. An asymptotic theory for the test is still under development and the bootstrap procedure is used instead. The application of the test to the US large cap stocks showed inadequacy of the Gaussian copula. As an alternative more flexible Student-t copula maybe used for the dependence modelling of multivariate collections of stocks. To reflect dynamical changes in the dependence structure more flexible copula forms are to be developed.

Acknowledgements

I would like to express my gratitude to Dr. Cees Diks for helpful discussions and enthusiastic supervision. The usual disclaimers apply.

References

[1] R. B. Nelsen, An Introduction to Copulas, Springer Verlag, New York, 1999.

[2] E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, T. Roncalli, Copulas for Finance: A Reading Guide and Some Applications, Manuscript, Financial Econometrics Research Center (2000).

[3] M. B. Wise, V. Bhansali, Implications of Correlated Default For Portfolio Allocation To Corporate Bonds, California Inst. of Tech. Working Paper No. CALT-68-2405 (2002).

[4] Y. Malevergne, D. Sornette, Testing the Gaussian copula hypothesis for financial assets dependences, Quantitative Finance 3 (2003) 231–250.

[5] V. Durrleman, A. Nikeghbali, T. Roncalli, Which copula is the right one?, Working paper, Groupe de Recherche Operationnelle, Credit Lyonnais (2000).

[6] A. Patton, X. Chen, Y. Fan, Simple Tests for Models of Dependence Between Multiple Financial Time Series, with Applications to U.S. Equity Returns and Exchange Rates, Discussion Paper 483, Financial Markets Group, London School of Economics (Feb. 2004).

[7] C. Diks, W. R. van Zwet, F. Takens, J. DeGoede, Detecting differences between delay vector distributions, Physical Review E 53 (1996) 2169–2176.

[8] C. W. Granger, E. Maasoumi, J. Racine, A dependence metric for possibly nonlinear processes, Journal of Time Series Analysis 25 (5) (2004) 649–669.

[9] C. Diks, H. Tong, A test for symmetry of multivariate probability distributions, Biometrika 86 (3) (1999) 605–614.

[10] M. Denker, G. Keller, On U-Statistics and v. Mises' Statistics for Weakly Dependent Processes, Z. Wahrscheinlichkeitstheorie verw. Gebiete 64 (1983) 505–522.

[11] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, New York: Chapman & Hall, 1993.