Adv Data Anal Classif (2010) 4:287–299
DOI 10.1007/s11634-010-0072-5

REGULAR ARTICLE

Detecting atypical observations in financial data: the forward search for elliptical copulas

Tiziano Bellini

Received: 30 November 2009 / Revised: 21 May 2010 / Accepted: 23 June 2010 /
Published online: 2 September 2010

Abstract In the last few years, copulas have been widely applied in many field of studies. Concentrating our attention on financial applications, we pursue the goal to detect multivariate atypical observations by extending to elliptical copulas the forward search originally introduced in linear and nonlinear regression by Atkinson and Riani (Robust diagnostic regression analysis. Springer, New York, 2000). Considering that, in the forward search, observations are ranked according to their closeness to the fitted data, we need to define a measure through which to initialize, progress and monitor the search. We achieve this goal building up the forward search for elliptical copulas relying on the squared Mahalanobis distance. Stressing the need to find theoretical boundaries for the inference on outliers, we introduce a procedure for computing envelopes as in Riani and Atkinson (Adv Data Anal Classif 1:123–141, 2007). Once defined our framework, we apply the forward search to a simulated environment where contaminations are exogenously introduced then, we carry out the analysis on n equity log-return real time series.

Keywords Copulas · Forward search · Squared Mahalanobis distance

Mathematics Subject Classification (2000) 62F35 · 62-07

1 Introduction

Copulas have become a popular multivariate tool in numerous areas where the multivariate dependence is of great interest (Nelsen 1999). Focusing on financial applications, as described among all by Embrechts et al. (2001), copulas are used in asset allocation, default risk modelling, derivative pricing and risk management.

T. Bellini (✉)
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: tiziano.bellini@unipr.it

When copulas are fitted to data, outlier detection becomes an important issue. In order to detect atypical observations, deletion methods usually do not lead to the identification of the uncontaminated observations. Then, attempting to overcome this drawback, the forward search was proposed, originally in linear and nonlinear regression by Atkinson and Riani (2000), as powerful general method for detecting multiple masked outliers and for determining their effect on inferences about models fitted to data. In the forward search the evolution of residuals, parameter estimates and inferences are monitored as the subset size increases. Results are presented as forward plots which show the evolution of the quantities of interest as a function of the sample size (Riani et al. 2009).

For the first time in the literature, we extend the forward search to copulas concentrating on the elliptical ones. Even considering that, particularly in finance, data marginal distributions may vary on a wide range of distributions, we focus on elliptical copulas because they are candidate for representing the dependence structure of data from a large number of areas.

The key element of the forward search is to specify a measure through which to rank observations. Starting from the research of Malevergne and Sornette (2003), we exploit the squared Mahalanobis distance to apply the forward search to Gaussian and Student T copulas. We begin our data analysis on a simulated uncontaminated dataset, then we introduce contaminations and, stressing the need to find theoretical boundaries for the inference on outliers, we check the effectiveness of our procedure exploiting envelopes obtained through Monte Carlo simulations. The final step of this research is to apply our framework to time series which are not obtained through a simulation algorithm. We focus on n equity log-return real time series of Italian companies belonging to various economic sectors.

The paper is organised as follows. In Sect. 2 we introduce copulas, we describe how to estimate parameters and compute the squared Mahalanobis distance for the elliptical ones. In Sect. 3 we show the elliptical copula forward search. In Sect. 4 we apply the forward search to a simulated uncontaminated and contaminated dataset. In Sect. 5 we extend our analysis to n equity log-return real time series. Section 6 contains concluding remarks and directions for future research.

2 Copulas, parameter estimation and squared Mahalanobis distance

A copula is a mathematical function that combines marginal probabilities into a joint distribution. An n-dimensional copula is a function $C : [0, 1]^n \to [0, 1]$. The basic idea behind copulas is to separate dependence and marginal behavior of the elements constituting the random vector. It follows that the choice of the copula does not constrain the choice of the marginal distributions. Sklar (1959) showed that any multivariate distribution function can be written in the form of a copula function. An important corollary of this theorem is that, letting H be an n-dimensional distribution function with continuous margins $F_1, \ldots, F_n$, considering the copula C, for any $\mathbf{u}$ in $[0, 1]^n$, the following equation holds

$$C(u_1, \ldots, u_n) = H(F^{-1}(u_1), \ldots, F^{-1}(u_n)). \tag{1}$$

In our work we focus on the Gaussian and the Student T copulas because they are candidate for representing the dependence structure of data from several areas. In the Gaussian environment instead of the generic distribution function H we consider the multivariate distribution Φ_R with linear correlation matrix R and instead of $F^{-1}(u_i)$ we denote $\Phi^{-1}(u_i)$ the inverse of the standard univariate Gaussian distribution function. In the case of Student T copula we consider, respectively, the multivariate distribution function $\xi_{R,\vartheta}$ with correlation matrix R and ϑ degrees of freedom, while $\xi_{\vartheta}^{-1}(u_i)$ represents the inverse of the standard univariate Student T distribution with ϑ degrees of freedom.

Considering a time series $\mathbf{x}_t = (x_{1,t}, \ldots, x_{n,t})'$ where n stands for the number of assets and $t = 1, \ldots, T$ is the discrete time, in the literature several methods to estimate copula parameters have been proposed. In our research we concentrate on the canonical maximum likelihood method (CML). We exploit this method because it is not based on a priori assumption on the distributional form of the marginals, it is easy to implement even for high dimensional copulas and it is not computationally intensive as other procedures (Durrleman et al. 2000).

We can summarize the CML estimation as follows:

- Transformation of the initial dataset $\mathbf{x}_t = (x_{1,t}, \ldots, x_{n,t})'$ into uniform variates, using the empirical marginal distribution. It means that for $t = 1, \ldots, T$ we compute $\hat{\mathbf{u}}_t = (\hat{u}_{1,t}, \ldots, \hat{u}_{n,t})'$.
- Estimation of the vector of the copula parameters. In our analysis the parameters are R for the Gaussian copula and R, ϑ for the Student T copula.

The CML method relies on the concept of the marginal transformation of empirical observations x_i into pseudo observations $\hat{u}_i$. The pseudo observations are computed as follows

$$\hat{u}_{i,t} = \frac{1}{T} \Sigma_{j=1}^{T} \mathbf{1}_{x_{i,j} \leq x_{i,t}}, \tag{2}$$

where $\mathbf{1}_{x_{i,j}} \leq x_{i,t}$ is the indicator function which assumes value 1 if $x_{i,j} \leq x_{i,t}$ and 0 otherwise.

In order to estimate the correlation matrix R we exploit the following equation

$$\tau(X_i, X_j) = \frac{2}{\pi} \arcsin\left(R_{i,j}\right), \tag{3}$$

where $\tau(X_i, X_j)$ and $R_{i,j}$ indicate respectively the Kendall τ and Pearson linear correlation coefficient for the random variables (X_i, X_j) (for a proof, see Lindskog et al. 2001). The estimator of R, $\sin[\frac{\pi}{2}\hat{\tau}(X_i, X_j)]$, inherits the robustness of the Kendall's τ estimator and is an efficient estimator of R for both elliptical distributions and non elliptical distributions with elliptical copulas (Embrechts et al. 2001). As pointed out by Lindskog (2000) there is no guarantee that the transformation of the empirical Kendall's τ matrix is positive definite. In this case, it can be adjusted using the eigenvalue method of Rousseeuw and Molenberghs (1993).

For the Student T copula, in order to obtain the degrees of freedom ϑ, Mashal and Zeevi (2002) proposed the following algorithm:

– Starting from the random sample $\mathbf{x}$, transform the initial dataset into the set of uniform variate $\hat{\mathbf{u}}$ using the empirical marginal transformations described above.
– Estimate the correlation matrix.
– Find the CML estimate of degrees of freedom by maximizing the log-likelihood copula density function as follows

$$\hat{\vartheta}_{\text{CML}} = \text{argmax}_{\vartheta} \sum_{t=1}^{T} ln \left[c_{\text{Student}}(\hat{u}_{1,t}, \ldots, \hat{u}_{n,t} | \hat{R}, \vartheta) \right]. \tag{4}$$

Looking for an effective measure of unit closeness, starting from the research of Malevergne and Sornette (2003) and concentrating on financial applications, we exploit the squared Mahalanobis distance. In order to compute this distance, considering $\hat{u}_{i,t}$ of Eq. (2), we need to calculate

$$\tilde{u}_{i,t} = F^{-1}\left(\hat{u}_{i,t}\right), \tag{5}$$

where, for the Gaussian copula $F^{-1}(\cdot)$ is the inverse of the standard univariate normal cdf $\Phi^{-1}(\cdot)$ and for the Student T copula we consider the standard univariate cdf $\xi_{\vartheta}^{-1}(\cdot)$. It is useful to remark that, in order to avoid $F^{-1}\left(\hat{u}_{i,t}\right)$ to reach its extreme limits, considering that $\hat{u}_{i,t}$ is the rank of $x_{i,t}$ divided by T, we set $\hat{u}_{i,t}$ to $\frac{\text{rank}-0.5}{T}$ which never reaches the boundaries 0 and 1.

Starting from $\hat{R}$ and $\tilde{u}_{i,t}$, we obtain the squared Mahalanobis distance as follows

$$\hat{z}_t = \tilde{\mathbf{u}}_t^{'}(\hat{R})^{-1}\tilde{\mathbf{u}}_t. \tag{6}$$

It is interesting to notice that the transformation of Eq. (2) could limit the effect of univariate atypical observations on the distance of Eq. (6). However, emphasizing that copulas are exploited for multivariate purposes, we stress that we concentrate on multivariate outlier detection where different marginals show extreme values, than our analysis is substantially not affected by the bounding effect caused by the pseudo observation transformation.

In the next section we describe how to carry out the search on elliptical copulas.

3 The forward search framework for elliptical copulas

The forward search is made up of the following three main steps: initialize, progress and monitor. The first task is to find the appropriate starting subset of observations. Considering that we are studying time series, the initial subset can be chosen among q blocks of contiguous observations of a predefined dimension b. To find the initial subset, we perform the search over all possible blocks and we choose the one that is considered more compact according to a certain distance. The second step is the way we progress in the search. At each step we rank units according to a specified distance continuing until all units are included in the subset. The third task is to monitor some suitable quantities along the search. In what follows we describe how these steps are performed.

1. Division of the dataset into q blocks. We split our time series in blocks and, in order to retain some dependence structure, we include the first observation as first unit of each initial subset. As pointed out in Riani (2004), the choice of the number of blocks does not dramatically affect the procedure, then we define the number of units of each block according to the pragmatic rule $b \approx \sqrt{T}$.
2. Transformation of each unit into uniform variate. We compute the empirical distribution of Eq. (2).
3. Correlation matrix estimate. Exploiting Eq. (3), we estimate the correlation matrix considering only units belonging to the subset.
4. Cdf transformation. According to the copula that we choose, starting from the empirical distribution above specified, we obtain, for all units of the dataset, the transformation described in Eq. (5).
5. Squared Mahalanobis distance computation. We compute, for all units of the dataset, the squared Mahalanobis distance of Eq. (6) considering the correlation matrix above estimated on the subset.
6. Initial subset S_b. For each of the q blocks, we compute the median squared Mahalanobis distance. We choose as initial b-dimensional subset, S_b, the one with the lowest median. This is a generalization of the least median of squares criterion in regression (Rousseeuw 1984).
7. Subset S_{b+1}. We add to S_b the unit with the lowest squared Mahalanobis distance obtained considering the correlation matrix estimated on S_b.
8. Progressing the search, subset $S_{m>b+1}$. We perform steps from 2 to 5 and the subset S_m is made up by the m units with the lowest squared Mahalanobis distances. In order to compute these distances the correlation matrix is estimated considering the $m-1$ units belonging to S_{m-1}.

Pursuing the goal to detect atypical observations, at each step m of the search, we focus on both the maximum distance within the subset, $d_{\max}(m)$, and the minimum distance of units not belonging to the subset $d_{\min}(m)$. Then, in order to check for the presence of outliers, we exploit envelopes as in Riani and Atkinson (2007).

In the next section we apply our forward search framework to a simulated dataset explaining how to compute $d_{\max}(m)$, $d_{\min}(m)$ and their corresponding envelopes.

4 The forward search applied to a simulated dataset

In order to check the effectiveness of our procedure we concentrate on a simulated dataset obtained exploiting the algorithm described by Embrechts et al. (2001) for both the Gaussian and the Student T copulas. The analysis is carried out through routines implemented in Matlab 2008.

We concentrate, first of all, on Gaussian copula with $n = 6$, $T = 81$ and low correlation: $R_{i,j} = 0.01 \; \forall \, i \neq j = 1, 2, \ldots, 6$. We focus on this scenario of very low correlation because realizations are potentially more unstable than in the case of high correlation.

On the top panel of Fig. 1 we show the squared Mahalanobis distances at each step of the search. According to our expectations we notice that the majority of units are very closed to each others, while only a few units have a greater distance. It is now interesting to check whether these units can be considered outliers.

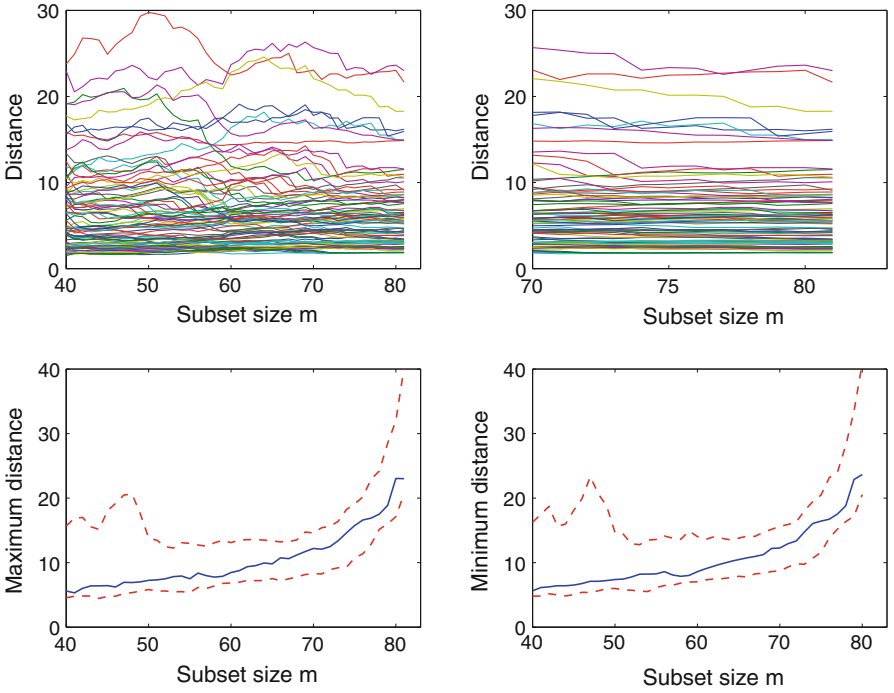

Fig. 1 Forward distance plots and envelopes: uncontaminated setting. In the *top panel* forward distance plots show the overall evolution and a focus on the last few steps of squared Mahalanobis distances for each unit along the search. In the *left bottom plot* the maximum distance in the subset, $d_{\max}(m)$, is compared to envelopes while on the *right* the minimum distance outside the subset, $d_{\min}(m)$, is compared to envelopes. According to these plots no outliers are detected in the uncontaminated environment

When we carry out the forward search, we need to find theoretical boundaries for the inference on outliers. For this reason we introduce envelopes which provide an objective basis for outlier detection.

The idea underlying envelopes is to compare boundaries to quantities obtained from the analysis on the examined dataset. In our setting we concentrate on both the maximum distance in the subset $d_{\max}(m)$ and the minimum distance outside the subset $d_{\min}(m)$. In order to obtain envelopes, we exploit the same algorithm as above (Embrechts et al. 2001). We simulate time series from copulas with parameters estimated on the whole set and, for each simulated time series, we compute $d_{\max}(m)$ and $d_{\min}(m)$. Then, for each subset size m we have a distribution of $d_{\max}(m)$ and $d_{\min}(m)$. Lower and upper envelopes are the collection of point-wise values, corresponding to lower and upper percentiles of these distributions at each subset size. In our analysis we focus on envelopes at a confidence level 99%.

On the bottom panels of Fig. 1 we notice that both the curves of the maximum distance in the subset and the minimum distance outside the subset are within envelopes. According to our expectations, no outliers are detected.

As we anticipated, we contaminate our initial dataset in order to verify whether our procedure is really capable to detect outliers. We exogenously fix at extreme value observations from 20 to 24, and we carry out the same analysis as above.

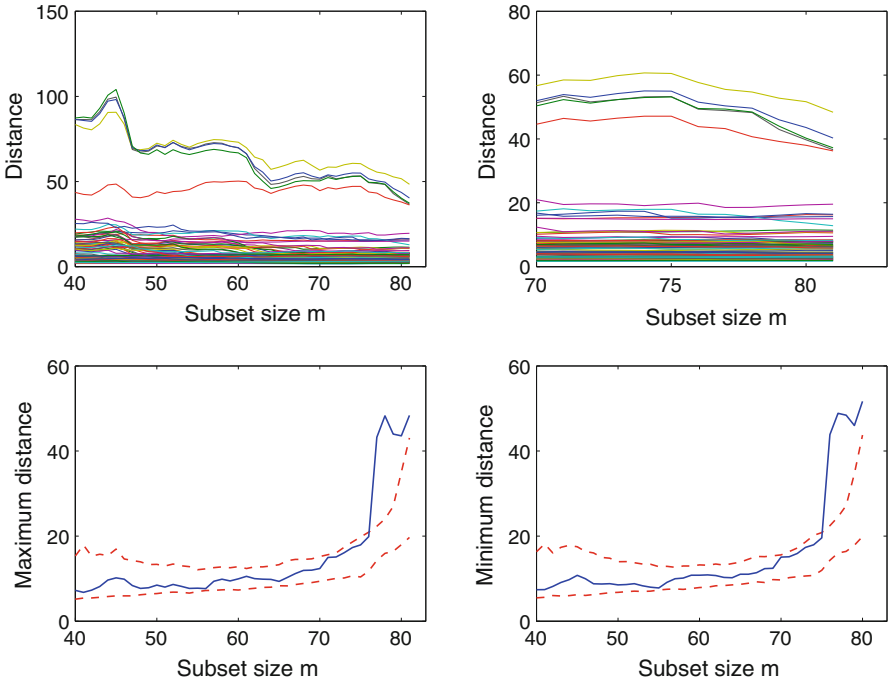

Fig. 2 Forward distance plots and envelopes: contaminated setting. In the *top left panel* the overall forward distance plot is shown, while the *top right plot* focuses on the last steps of the search showing that 5 contaminated units have higher distances than others. In the *bottom panel* the maximum distance in the subset, $d_{\max}(m)$, and the minimum distance outside the subset, $d_{\min}(m)$, are compared to their envelopes. Contaminated units cross envelopes highlighting atypical observations

In Fig. 2 we notice that, on the top panel, the 5 contaminated units show higher distances than others. Then, we examine envelopes on the bottom panel of the same figure. We notice that in the last few steps of the search both the curves of maximum distance in the subset and minimum distance out of the subset cross envelopes.

In order to check for the presence of outliers, as proposed by Riani et al. (2009), we superimpose envelopes. From Fig. 3 we can state that all contaminated observations are detected as outliers. The top panel of Fig. 3 shows that at step 76 the curve of maximum distance in the subset is within the envelopes. At the same time, when we examine the minimum distance outside the subset, the first picture on the right panel shows the first crossing. It means that at step 76 the minimum distance of units not belonging to the subset becomes higher than what is expected at a confidence level 99%. In fact at step 77, when the first outlier belongs to the subset, the second picture on the left panel shows, according to our expectations, that the first outlier enters in the subset. Other pictures confirm what we expected: all contaminated units are detected.

We carried out the same analysis as above in the high correlation setting and considering Student T copula. We obtained the same result as above. In addition, we introduced contaminations at the beginning, at the end and in other positions of the time series obtaining the same findings as above.

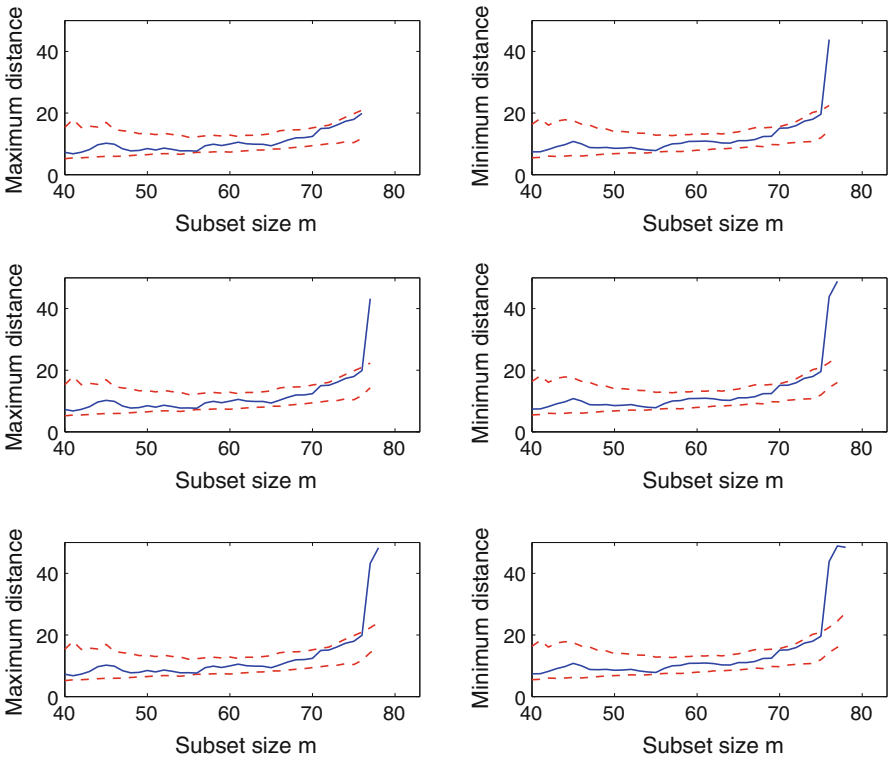

Fig. 3 Envelopes superimposed: contaminated setting. In the *top panel* envelopes are superimposed at subset size 76. In the *left plot* (maximum distance in the subset) no outliers are detected while in the *right plot* there is the first crossing of envelopes. It shows that at the next step of the search, the first outlier will enter in the subset. As anticipated, increasing the subset size by one unit, in the central panel, at step 77, the *left plot* (maximum distance in the subset) shows the first outlier, while in the *right plot* two outliers are detected. In the *bottom panel*, at step 78, the next outlier is detected

In order to verify the effectiveness of our framework in a setting which is not obtained exploiting a simulation algorithm, in the next section we apply the forward search to real market financial time series.

5 The forward search applied to real financial data

We perform our real data analysis considering, from January 2009 to January 2010, the daily equity log-returns of the following Italian companies: Bialetti, Cattolica, Gabetti, Mediobanca, Snai, Unipol.[1]

As we can see from Fig. 4, these log-returns are close to zero showing some upward, downward picks and volatility clusters.

Focusing, first of all, on the Gaussian copula, we start our analysis examining squared Mahalanobis distances. From the top panel of Fig. 5, we notice that two units,

[1] Data are available on the website http://economia.unipr.it/docenti/BELLINIT.

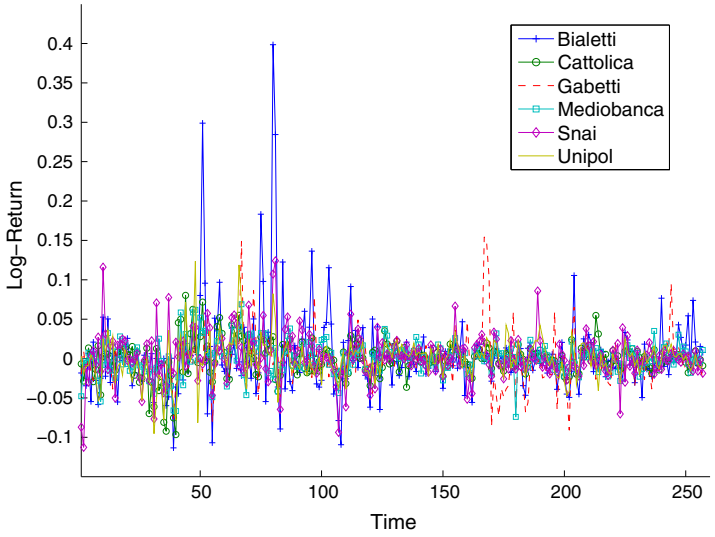

Fig. 4 Equity log-returns from January 2009 to January 2010

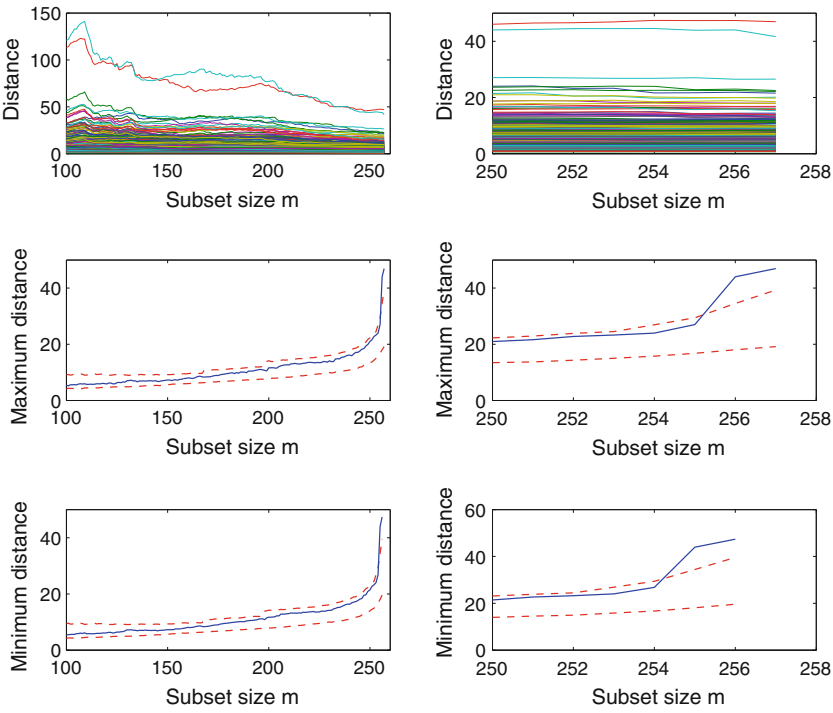

Fig. 5 Gaussian copula forward distance plots and envelopes: real market dataset. In the *top left plot* the overall forward distance plot is shown. In the *top right plot* only the last steps are highlighted showing that units 81 and 80 have higher distances than the others. In the *central* and *bottom panels* $d_{\max}(m)$ and $d_{\min}(m)$ are compared to their envelopes. It is shown that units 81 and 80 (respectively) cross envelopes on the penultimate and last step of the search

81 and 80 respectively, show higher distances than the others along the whole search. It is now interesting to check whether they can be considered outliers. For this reason, we compare $d_{\max}(m)$ and $d_{\min}(m)$ with their envelopes. From the central and the bottom panels of Fig. 5, where unit 81 enters in the subset at the penultimate step and unit 80 is the last unit of the search, it is evident that units 81 and 80 are outliers and this statement is further supported through superimposition.

It is now interesting to carry out the same analysis as above applying the Student T copula. Looking at the forward distance plot on the top panel of Fig. 6 we notice that, as in the Gaussian copula, units 81 and 80 show the highest squared Mahalanobis distances. When we consider both the central and the bottom panels where $d_{\max}(m)$ and $d_{\min}(m)$ are compared to their envelopes we can state that units 81 and 80 are outliers.

Once examined squared Mahalanobis distances, we focus our attention on parameter estimates. Figure 7 shows the evolution of the $n(n-1)/2$ correlation parameters along the search in the case of Gaussian copula. In particular on the left panel we consider the overall evolution while, on the right, we focus on the last steps of the search. It

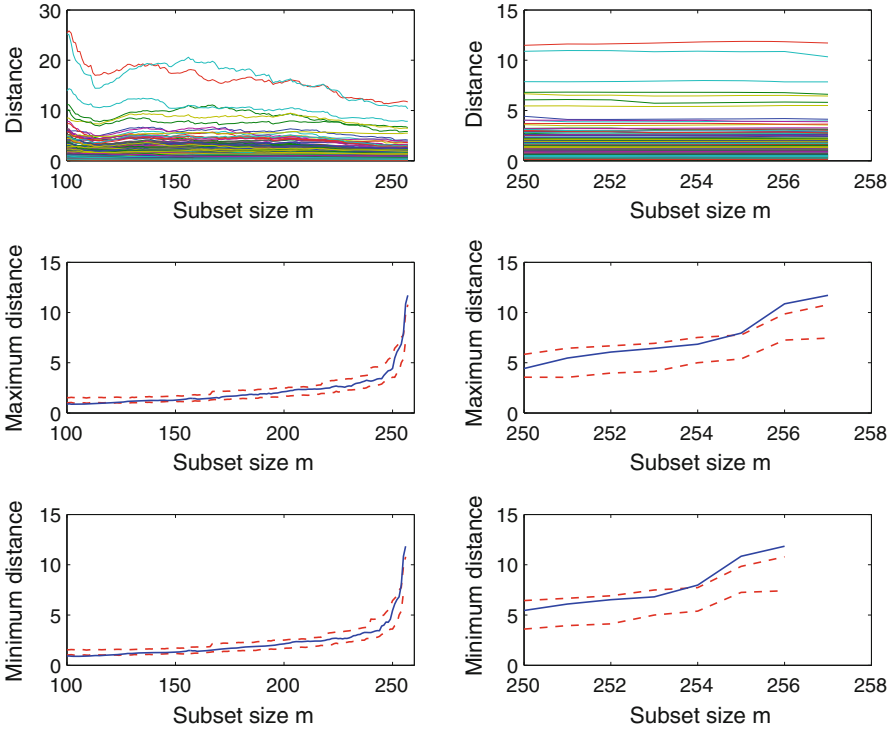

Fig. 6 Student T copula forward distance plots and envelopes: real market dataset. In the *top left plot* the overall forward distance plot is shown, while on the *right plot* the last few steps highlight units 81 and 80 highest distances. In the *central* and *bottom panels* $d_{\max}(m)$ and $d_{\min}(m)$ are compared to their envelopes. As in the Gaussian copula it is shown that units 81 and 80 (respectively) cross envelopes on the penultimate and last step of the search

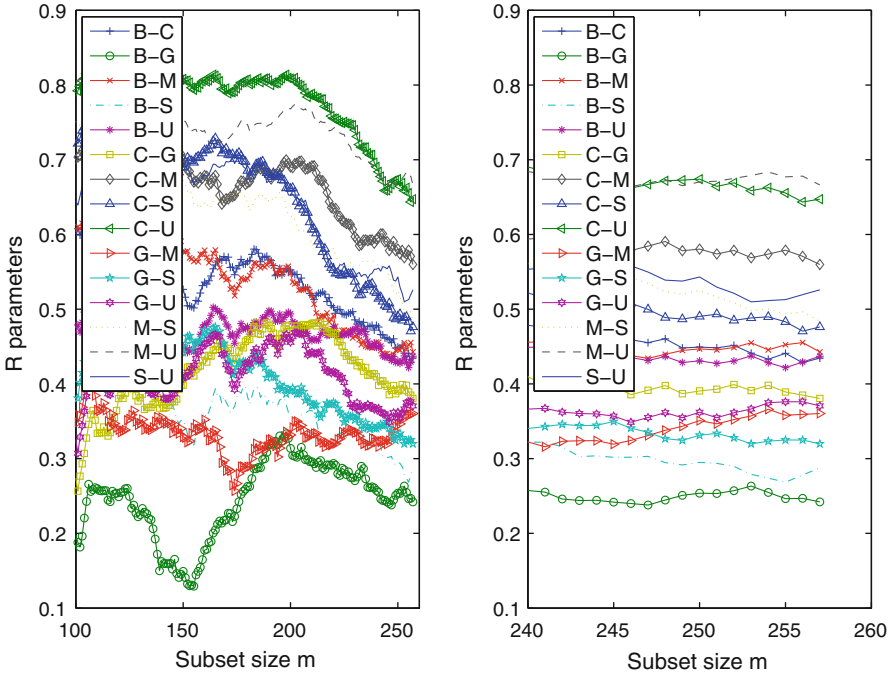

Fig. 7 Gaussian copula correlation parameters: real market dataset. In the *left panel* the overall evolution of $n(n-1)/2$ correlation parameters is shown while on the *right panel* the focus is on the last few steps of the search. Company names are shortened through their initial characters

is interesting to notice that higher correlation values are associated to companies which belong to the same economic sector. In particular Cattolica, Mediobanca and Unipol which are insurance and banking companies are more correlated than other companies. The lowest correlation is between Bialetti and Gabetti which belong, respectively, to the industry and the real estate sectors.

When we carry out the analysis on Student T copula, we obtain substantially the same evolution described in the Gaussian copula setting. Then, focusing on the last few steps of the search, as it is shown in Fig. 8, correlation parameters are very similar in both analysis. Furthermore, the procedure enables to examine the effect of atypical units. In our analysis we notice that when units 81 and 80 (respectively) are included in the subset, correlation parameters are subject to a small change.

It is useful to stress that ϑ, the parameter which represents the degrees of freedom for the Student T copula, is estimated step by step according to Eq. (4) and, apart from the very beginning of the search, it is substantially stable around the value 13.

It is finally interesting to notice that the most important findings of our procedure are concentrated on the last steps of the search. We carried out the analysis even without considering the block procedure described in previous Sections and, apart from the very beginning of the search, we obtained the same results as above.

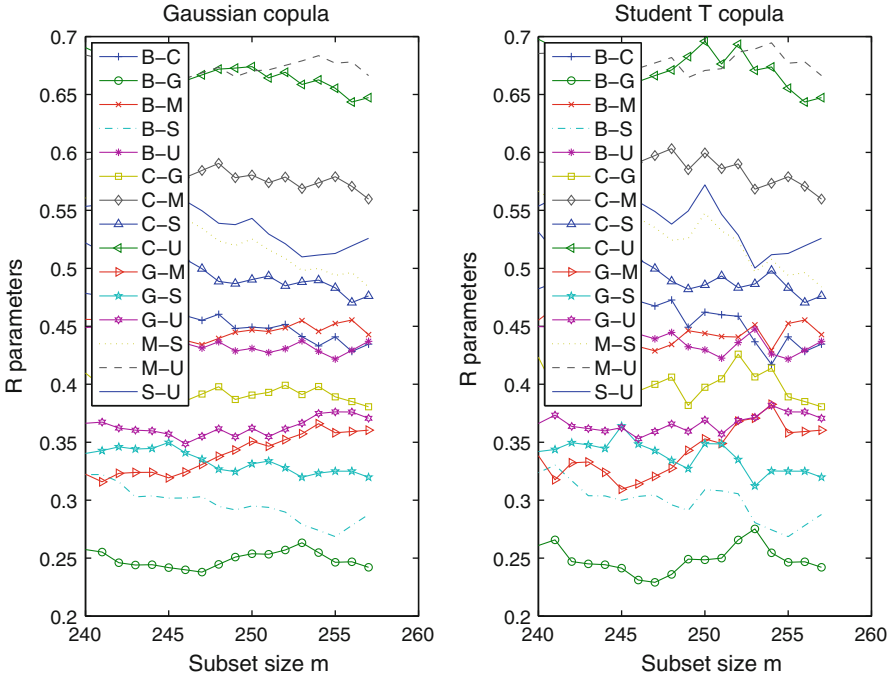

Fig. 8 Gaussian and Student T copula correlation parameters: real market dataset. Gaussian correlation parameters on the last few steps of the search (on the *left panel*) are compared to the Student T parameters (*right panel*). The evolution of these parameters is very similar along the search and, in particular, on the last few steps

6 Concluding remarks

We are pioneer in extending to elliptical copulas the forward search originally proposed by Atkinson and Riani (2000) in linear and nonlinear regression. The key issue of the research is the definition of an effective measure of unit closeness. We set up the analysis originating from the definition of the squared Mahalanobis distance, then we formalize how to carry out the search. Considering the need to find theoretical boundaries for the inference on outliers, we propose a Monte Carlo simulation to obtain envelopes.

We accomplish our data analysis starting from a simulated environment where units are generated from the algorithm described by Embrechts et al. (2001). We begin from an uncontaminated dataset, then we introduce contaminations to verify whether our approach is effective in detecting atypical observations and we show that our procedure achieves this goal regardless of their time chart.

The final issue of the research is to apply our framework to time series which are not obtained through a simulation algorithm. We carry out the search on a real market time series dataset with upward, downward picks and volatility clusters showing the effectiveness of the procedure in outlier detection and monitoring their effects on parameter estimates.

This work is the first step in the forward search applied to copulas. Further studies need to be devoted to extend this research to non elliptical copulas and to obtain general techniques to build up envelopes without using Monte Carlo simulations.

Acknowledgments I am grateful to Associate Editors for very constructive suggestions and to two anonymous referees for valuable comments on earlier drafts.

References

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer-Verlag, New York
Durrleman V, Nikeghbali A, Roncalli T (2000) Which copula is the right one? Groupe de Recherche Operationnelle, Credit Lyonnais, France
Embrechts P, Lindskog F, McNeil A (2001) Modelling dependence with copulas and applications to risk management. Department of Mathematics, ETH Zurich
Lindskog F (2000) Modelling dependence with copulas. RiskLab Report, ETH Zurich
Lindskog F, McNeil A, Schmock U (2001) A note on Kendall's tau for elliptical distributions. ETH preprint
Malevergne Y, Sornette D (2003) Testing the Gaussian copula hypothesis for financial assets dependence. Quant Finance 3:231–250
Mashal R, Zeevi A (2002) Beyond correlation: extreme co-movements between financial assets. Working paper, Columbia Graduate School of Business
Nelsen RB (1999) An introduction to copulas. Springer, New York
Riani M (2004) Extensions of the forward search to time series. Stud Nonlinear Dyn Econom 8(2):1–23
Riani M, Atkinson AC (2007) Fast calibrations of the forward search for testing multiple outliers in regression. Adv Data Anal Classif 1:123–141
Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J R Stat Soc Ser B 71:201–221
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Rousseeuw PJ, Molenberghs G (1993) Transformation of non positive semidefinite correlation matrices. Commun Stat Theory Methods 22:965–984
Sklar A (1959) Fonctions de repartition a n dimensions et leur marges. Publications de l'Institut de Statistique de l'Universite' de Paris 8:229–231