TIME SERIES CLUSTERING ON LOWER TAIL DEPENDENCE

Giovanni De Luca [1] and Paola Zuccolotto [2]

[1] University of Napoli *Parthenope*
(e-mail: giovanni.deluca@uniparthenope.it)

[2] University of Brescia
(e-mail: paola.zuccolotto@sis-statistica.org)

ABSTRACT: In this paper we analyse a case study based on the procedure introduced by De Luca and Zuccolotto (2011), whose aim is to cluster time series of financial returns in groups being homogeneous in the sense that their joint bivariate distributions exhibit high association in the lower tail. The dissimilarity measure used for such clustering is based on tail dependence coefficients estimated by means of copula functions. We carry out the clustering using an algorithm requiring a preliminary transformation of the dissimilarity index into a distance metric by means of a geometric representation of the time series, obtained with Multidimensional Scaling. The results of the clustering could be used for a portfolio selection purpose, when the goal is to protect investments from the effects of a financial crisis.

KEYWORDS: Time series clustering, tail dependence, copula function.

1 Introduction

Several approaches to time series clustering are present in the literature. After the first studies, where dissimilarities between time series were merely derived by the comparison between observations or some simple statistics computed on the data, more complex solutions have been proposed (see for example Piccolo, 1990; Corduas and Piccolo, 2008; Otranto, 2008; Galeano and Peña, 2000; Caiado et al., 2006; Alonso et al., 2006; Vilar et al., 2010; Kakizawa et al., 1998; Taniguchi and Kakizawa, 2000; Weng and Shen, 2008; Pattarin et al., 2004). In this paper we show a case study based on the use of the procedure proposed by De Luca and Zuccolotto (2011) to cluster time series of returns of financial assets according to their association in the lower tail. Then, we show how this approach can be employed for portfolio selection, especially from a financial crisis perspective. The paper is organized as follows: in Section 2 the clustering procedure is briefly recalled, while the main results of the case study are summarized in Section 3.

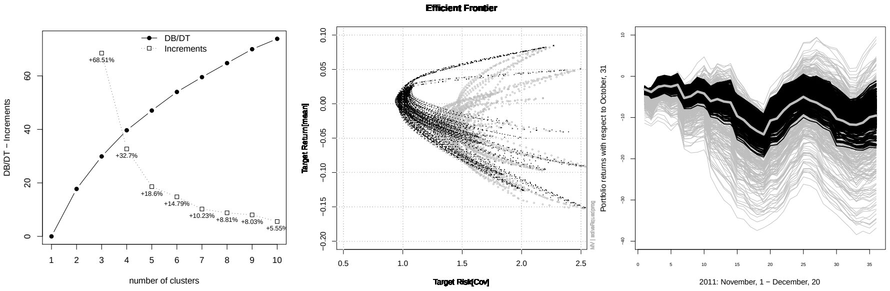
2 Tail dependence based clustering procedure

The interest of researchers in modelling the occurring of extreme events has several empirical motivations, especially in contexts where it can be directly associated to risk measurement, such as, for example, financial markets. Recently a great deal of attention has been devoted also to the study of association between extreme values of two or more variables. From a methodological point of view, the problem of quantifying this association has been addressed in different ways. One of the proposed approaches consists in analyzing the probability that one variable assumes an extreme value, given that an extreme value has occurred to the other variables (see Cherubini et al. 2004). This probability is known as lower or upper tail dependence and we will restrict its analysis to the bivariate case. Let Y_1 and Y_2 be two random variables and let $U_1 = F_1(Y_1)$ and $U_2 = F_2(Y_2)$ be their distribution functions. The lower and upper tail dependence coefficients are defined respectively as $\lambda_L = \lim_{v\to 0^+} P[U_1 \leq v | U_2 \leq v]$ and $\lambda_U = \lim_{v\to 1^-} P[U_1 > v | U_2 > v]$. In practice, the tail dependence coefficients have to be estimated from observed data. In general, a distributional assumption is needed, but this is usually hard, especially with financial data, which this paper is concerned with. A very effective alternative way is the use of copula functions, thanks to which tail dependence estimation is both simple and flexible. A two-dimensional copula function for two random variables Y_1 and Y_2 is defined as a function $C : [0,1]^2 \to [0,1]$ such that $F(y_1, y_2; \theta) = C(F_1(y_1; \vartheta_1), F_2(y_2; \vartheta_2); \tau)$, for all y_1, y_2, where $F(y_1, y_2; \theta)$ is the joint distribution function of Y_1 and Y_2 (see Nelsen 2006) and $\theta = (\vartheta_1, \vartheta_2, \tau)$. It is straightforward to show that the tail dependence coefficients can be expressed in terms of the copula function. In particular, the lower tail dependence coefficient, which will be hereafter the focus of the paper, is given by $\lambda_L = \lim_{v\to 0^+} C(v,v)/v$.

In the analysis of the relationship between financial returns, the lower tail dependence coefficient gives an idea of the risk of investing on assets for which extremely negative returns could occur simultaneously. So, the lower tail dependence is strictly linked to the diversification of investments, especially in financial crisis periods. For this reason, De Luca and Zuccolotto (2011) proposed to cluster time series of financial returns according to a dissimilarity measure defined as $\delta(\{y_{it}\}, \{y_{jt}\}) = -\log(\hat{\lambda}_L)$, where $\{y_{it}\}_{t=1,\dots,T}$ and $\{y_{jt}\}_{t=1,\dots,T}$ denote the time series of returns of two assets i and j, and $\hat{\lambda}_L$ is their estimated tail dependence coefficient. In this way we obtain clusters of assets characterized by high tail dependence in the lower tail. From a portfolio selection perspective, it should then be avoided portfolios containing assets belonging to the same cluster. Given p assets, the clustering procedure proposed by De Luca and Zuccolotto (2011) is composed by two steps: firstly, starting from the dissimilarity matrix $\Delta = (\delta_{ij})_{i,j=1,\dots,p}$, an *optimal* representation of the p time series $\{y_{1t}\}, \dots, \{y_{pt}\}$ as p points $\mathbf{y}_1, \dots, \mathbf{y}_p$ in R^q is found by means of Multidimensional Scaling (MDS); secondly, the k-means clustering algorithm is performed using the obtained geometric representation of the p time series. The above mentioned term *optimal* means that the Euclidean distance matrix $D = (d_{ij})_{i,j=1,\dots,p}$, with $d_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|$, of the points to be defined in the first step has to fit as closely as possible the dissimilarity matrix Δ. The extent to which the interpoint distances d_{ij} "match" the dissimilarities δ_{ij} is measured by an index called *stress*, which should be as low as possible. MDS works for a given value of the dimension q, which has to be given in input. So, it is proposed to start with the dimension $q = 2$ and then to repeat the analysis by increasing q until the minimum stress of the corresponding optimal configuration is lower than a given threshold $\bar{s}$.

3 Case study

In this case study we analyse the time series of the prices of the 24 stocks which have been included in FTSE MIB index during the whole period from January 3, 2006 to October 31, 2011. After transforming prices into log-returns, we preliminary removed autocorrelation and heteroskedasticity from the data by means of univariate Student-*t* AR-GARCH models. For each couple of stocks we estimated a bivariate Joe-Clayton copula function, $C(u_1,u_2) = 1-\{1-[(1-(1-u_1)^{\kappa})^{-\theta}+(1-(1-u_2)^{\kappa})^{-\theta}-1]^{-1/\theta}\}^{1/\kappa}$ using the estimated distribution functions of the standardized residuals. After estimating the 276 lower tail dependence coefficients, which in the case of the Joe-Clayton copula are given by $\hat{\lambda}_L = 2^{-1/\hat{\theta}}$, we carried out MDS using the dissimilarity matrix $\Delta = (\delta_{ij})_{i,j=1,\ldots,24}$. We set $\bar{s} = 0.005$. The minimum dimension allowing a final configuration with minimum stress lower than $\bar{s}$ resulted $q = 14$. In the second step, we performed a *k*-means clustering algorithm using the 14-dimensional MDS point configuration in the Euclidean space, $\mathbf{y}_1, \ldots \mathbf{y}_{24}$. The graph in the left of the Figure shows the pattern of the ratio of deviance between clusters over total deviance, as a function of the number of clusters *k* and its increments when considering the solution with *k* clusters, with respect to $k-1$ clusters. Looking at these increments, we decide for the solution with $k = 4$ (Cluster 1: ATLANTIA, ENEL, ENI, SAIPEM; Cluster 2: AUTOGRILL, FIAT, FINMECC, LOTTOMAT, LUXOTTICA, PIRELLI, STM, TELECOM; Cluster 3: MPS, FONDIARIA, GENERALI, INTESA, MEDIOBANCA, MEDIOLANUM, MEDIASET, BPM, UBI, UNICREDIT; Cluster 4: SNAM, TERNA).



We used the obtained clustering to construct a portfolio composed by as many stocks as the number *k* of clusters. The stocks are selected by imposing the restriction that each stock belongs to a different cluster; with the $k = 4$ above mentioned clusters, 640 different selections can be made according to this criterion. This strategy should protect the investments from parallel extreme losses during crisis periods, because the clustering solution is characterized by a moderate lower tail dependence between clusters. Using the popular Markowitz portfolio selection procedure, we plotted the efficient frontiers of all the possible 640 selections (black), compared to those of 1000 portfolios (gray) built with 4 randomly selected stocks (graph in the middle of the Figure). After selecting the minimum variance portfolio of each frontier, the returns of the 640 portfolios (black) are compared to the others (gray) in the period from November 1, 2011 to December 20, 2011, and to the returns of the naive minimum variance portfolio (bold gray) built using all the stocks.

References

ALONSO, A.M., BERRENDERO J.R. HERNÁNDEZ A., & JUSTEL, A. 2006. Time series clustering based on forecast densities. *Computational Statistics and Data Analysis*, **51**, 762–776.

CAIADO, J., CRATO N., & PEÑA, D. 2006. A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, **50**, 2668–2684.

CHERUBINI, U., LUCIANO E., & VECCHIATO, W. 2004. *Copula methods in finance*. New York: Wiley.

CORDUAS, M., & PICCOLO, D. 2008. Time series clustering and classification by the autoregressive metrics. *Computational Statistics and Data Analysis*, **52**, 1860–1872.

DE LUCA, G., & ZUCCOLOTTO, P. 2011. A tail dependence-based dissimilarity measure for financial time series clustering. *Advances in Classification and Data Analysis*, **5**, 323–340.

GALEANO, P., & PEÑA, D. 2006. Multivariate analysis in vector time series. *Resenhas*, **4**, 383–404.

JOE, H. 1997. *Multivariate models and dependence concepts*. New York: Chapman & Hall/CRC.

KAKIZAWA, Y., SHUMWAY R.H., & TANIGUCHI, M. 1998. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, **93**, 328–340.

NELSEN, R. 2006. *An introduction to copulas*. New York: Springer.

OTRANTO, E. 2008. Clustering heteroskedastic time series by model-based procedures. *Computational Statistics and Data Analysis*, **52**, 4685–4698.

PATTARIN, F., PATERLINI S., & MINERVA, T. 2004. Clustering financial time series: an application to mutual funds style analysis. *Computational Statistics and Data Analysis*, **47**, 353–372.

PICCOLO, D. 1990. A distance measure for classifying ARMA models. *Journal of Time Series analysis*, **11**, 153–164.

TANIGUCHI, M., & KAKIZAWA, Y. 2000. *Asymptotic theory of statistical inference for time series*. New York: Springer.

VILAR, J.A., ALONSO A.M., & VILAR, J.M. 2010. Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics and Data Analysis*, **54**, 2850–2865.

WENG, X., & SHEN, J. 2008. Classification of multivariate time series using two-dimensional singular value decomposition. *Knowledge-Based Systems*, **21**, 535–539.