Maria Brigida Ferraro · Paolo Giordani
Barbara Vantaggi · Marek Gagolewski
María Ángeles Gil · Przemysław Grzegorzewski
Olgierd Hryniewicz
Editors

Soft Methods for Data Science

Editors
Maria Brigida Ferraro
Department of Statistical Sciences
Sapienza University of Rome
Rome
Italy

Paolo Giordani
Department of Statistical Sciences
Sapienza University of Rome
Rome
Italy

Barbara Vantaggi
Department of Basic and Applied Sciences
for Engineering
Sapienza University of Rome
Rome
Italy

Marek Gagolewski
Department of Stochastic Methods, Systems
Research Institute
Polish Academy of Sciences
Warsaw
Poland

María Ángeles Gil
Department of Statistics and Operational
Research and Mathematics Didactics
University of Oviedo
Oviedo
Spain

Przemysław Grzegorzewski
Department of Stochastic Methods, Systems
Research Institute
Polish Academy of Sciences
Warsaw
Poland

Olgierd Hryniewicz
Department of Stochastic Methods, Systems
Research Institute
Polish Academy of Sciences
Warsaw
Poland

A Portfolio Diversification Strategy via Tail Dependence Clustering

Hao Wang, Roberta Pappadà, Fabrizio Durante and Enrico Foscolo

Abstract We provide a two-stage portfolio selection procedure in order to increase the diversification benefits in a bear market. By exploiting tail dependence-based risky measures, a cluster analysis is carried out for discerning between assets with the same performance in risky scenarios. Then, the portfolio composition is determined by fixing a number of assets and by selecting only one item from each cluster. Empirical calculations on the EURO STOXX 50 prove that investing on selected assets in trouble periods may improve the performance of risk-averse investors.

1 Introduction

In recent years, financial markets have been characterized by an increasing globalization and a complex set of relationships among asset returns. Moreover, it has been recognized that the linkages among different assets vary across time and that their strength tends to increase especially during crisis periods. The presence of a stronger dependence when markets are experiencing losses, is of utmost interest from a risk manager perspective. In fact, it has been recognized that investors can reduce the risk of their portfolios through diversification, i.e. allocating their investments in various classes and/or categories that would move in different ways in response to the same event.

H. Wang
School of Economics, Jilin University, Changchun 130012, China
e-mail: haowang@jlu.edu.cn

R. Pappadà
Department of Economics, Business, Mathematics and Statistics "Bruno De Finetti", University of Trieste, 34127 Trieste, Italy
e-mail: rpappada@units.it

F. Durante · E. Foscolo (✉)
Faculty of Economics and Management, Free University of Bozen-Bolzano, Bolzano, Italy
e-mail: enrico.foscolo@unibz.it

F. Durante
e-mail: fabrizio.durante@unibz.it

In order to provide a suitable diversification of a portfolio that takes into account the occurrence of extreme scenarios, various clustering techniques for multivariate time series have been proposed in the literature, mainly based on measures of association like Pearson correlation coefficient (see, e.g., [13]). Recently, such techniques have also been applied in order to group financial time series that are similar in extreme scenarios by using tail dependence coefficients (see, e.g., [2, 3] and [7]), or conditional measures of association, like Spearman's correlation, as done in [6]. For an alternative approach, see also [9, 10].

The aim of this contribution is to exploit recent tail-dependence clustering methods in order to select a weighted portfolio in a group of assets. In particular, it will be shown how the adoption of fuzzy clustering methodology (see, e.g., [8] and references therein) may provide some advantages in terms of both performance and computational tractability of the model.

2 The Clustering Procedure

Several clustering procedures are based on the choice of a suitable dissimilarity measure that expresses the relations among the financial time series of the asset returns under consideration. Following previous approaches, we present here a procedure to group time series based on their tail behaviour, as done in [6]. This methodology is summarized below.

Consider a matrix of d financial time series $(x_{it})_{t=1,\dots,T}$ $(i = 1, 2, \dots, d)$ representing the log–returns of different financial assets. We assume that each time series $(x_{it})_{t=1,\dots,T}$ is generated by the stochastic process $(\mathbf{X}_t, \mathcal{F}_t)$ such that, for $i = 1, \dots, d$,

$$X_{it} = \mu_i(\mathbf{Z}_{t-1}) + \sigma_i(\mathbf{Z}_{t-1})\varepsilon_{it}, \tag{1}$$

where $\mathbf{Z}_{t-1}$ depends on $\mathcal{F}_{t-1}$, the available information up to time $t-1$, and the innovations ε_{it} are distributed according to a distribution function F_i for each t. Moreover, the innovations ε_{it} are assumed to have a constant conditional distribution F_i (with mean zero and variance one, for identification) such that for every t the joint distribution function of $(\varepsilon_{1t}, \dots, \varepsilon_{dt})$ can be expressed in the form $C(F_1, \dots, F_d)$ for some copula C. Such a general model includes many multivariate time series models presented in the literature (see, for instance, [14]).

Then the following steps can be performed in order to divide the time series into sub-groups such that elements in each sub-group have strong tail dependence between each other.

1. Choose a copula-based time series model in order to describe separately the marginal behavior of each time series and the link between them.
2. Estimate a (pairwise) tail dependence measure among all the time series.
3. Define a dissimilarity matrix by using the information contained in the tail dependence matrix.

4. Apply a suitable cluster algorithm for grouping time series according to the tail behavior.

Steps 1–3 described above have been discussed in details in [6]. Here (and in the following illustration), these steps are specified in the following way:

1. We fit an appropriate ARMA-GARCH model to each univariate time series and, using the estimated parameters, we construct the standardized residuals that are helpful in determining the joint distribution of the innovations.
2. As a measure of tail dependence, we use the conditional Spearman's correlation ρ_α that expressed the Spearman's correlation between two random variables X and Y given that they are both under their α–quantile (here, $\alpha = 0.10$). The estimation is based on the procedure described in [4, 5].
3. Once the conditional Spearman's correlation has been computed for all pairs extracted from the time series, we transform it through a monotonic function f in such a way that the obtained dissimilarity between two time series is small when their tail dependence is high, and monotonically increases when their tail dependence decreases. Thus, for $i, j = 1, \ldots, d$, we define $\Delta = \left(\Delta_{ij}\right)$ whose elements are given by

$$\Delta_{ij} = \sqrt{2(1 - \hat{\rho}_\alpha^{ij})}, \tag{2}$$

where $\hat{\rho}_\alpha^{ij}$ is the conditional Spearman's correlation between time series i and j.

Starting from the dissimilarity matrix defined in (2), we can perform a cluster analysis by different techniques. Here we focus on a fuzzy clustering algorithm, i.e. the *fanny algorithm* by [12], since it allows to quantify the degree of membership of an object to the different clusters by means of a coefficient, which ranges from 0 to 1. In order to determine the optimal number k of clusters, we use the average silhouette index [11], which reflects the within-cluster compactness and between-cluster separation of a clustering.

Fanny algorithm aims to minimize the objective function

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} m(i,v)^r m(j,v)^r \Delta_{ij}}{2\sum_{j=1}^{n} m(j,v)^r}$$

where n is the number of involved time series, k is the number of clusters, $r > 1$ is the membership exponent (usually, $r = 2$), $m(i, v)$ the membership of time series i to cluster v, and Δ_{ij} is the dissimilarity between the time series i and j. The algorithm returns the membership degree of each time series i to any cluster. Obviously, if a crisp assignment of each time series to only one cluster is necessary, then one could proceed according to the highest membership degree.

3 The Portfolio Selection

Once the cluster analysis is carried out for identifying assets with the same performance during risky scenarios, a portfolio selection procedure can be implemented by fixing the number of assets per portfolio equal to the number of clusters, and by selecting only one item from each cluster. The rationale is that, since assets in different clusters are weakly associated with each other (in risky periods), then they form a well-diversified portfolio. This idea has been used, for instance, in [2, 3] and is slightly modified here by exploiting the advantages of fuzzy clustering.

Specifically, suppose that n time series have been classified by means of the procedures described in Sect. 2 into $k \geq 2$ groups. Let $m(i, v)$ be the membership degree of time series i to cluster C_v. The selection algorithm goes as follows:

The portfolio selection algorithm

1. Fix $T \in [0, 1]$, which represents a cut-off value for the degree of membership to a cluster.
2. For $i = 1, 2, \ldots, n$, assign the time series i to the cluster C_v if it holds that $m(i, v) = \max_{v'=1,\ldots k} m(i, v')$.
3. For each cluster C_v $(v = 1, \ldots k)$, remove the element j in C_v provided that $m(j, v) < T$. The resulting clusters are denoted by D_v $(v = 1, \ldots, k)$. Notice that some D_v can be empty.
4. Determine all possible portfolios composed by (at most) k assets obtained by selecting exactly one asset from each element of $\{D_1, \ldots, D_k\}$.
5. For these portfolios, calculate the optimal weights assigned to each of its assets by Minimum Conditional-Value-at-Risk (CVaR) strategy.
6. Select the Minimum CVaR portfolio with the lowest CVaR value.

Some comments are needed here.

Step 3 guarantees that we only focus on those assets that can be assigned to a given cluster with a membership degree larger than T. It avoids the selection of assets that are likely to be associated with more than one cluster (and, hence, tend to downgrade the effects of diversification).

Step 4 is usually computationally expensive; however, the computational burden can be limited by a careful selection of the cut-off value T. In particular, this aspect highlights the main difference between the proposed algorithm and the methodology discussed in [2].

Step 5 suggests a portfolio selection procedure that focuses on extreme events and, hence, is coherent with the tail dependence approach developed here (see also [3]). Specifically, the procedure optimizes the CVaR, defined as the expected loss exceeding VaR_β (for more details, see [15]). Below, we set $\beta = 0.10$.

For the illustration of the algorithm, we consider time series related to EURO STOXX 50 stock index and its components in the period from January 2, 2003 to July 31, 2011. Moreover, as out-of-sample period, we will show the performance of our procedure in the period from August 1, 2011 to September 9, 2011. The period

Table 1 Cluster composition of the EURO STOXX 50 constituents by using conditional Speaman's correlation ρ_α with $\alpha = 0.1$ and fanny algorithm. The assets whose maximal membership degree is smaller than 0.90 are denoted in bold

Cluster							
1	D.DTEX	E.IND	D.SAPX	F.EI	D.BAYX	F.UBL	D.BMWX
	F.CRFR	**D.RWEX**					
2	F.SQ.F	F.FTEL	F.OR.F	H.ASML	F.EX.F	F.BSN	
3	E.BBVA	D.ALVX	H.ING	D.MU2X	E.SCH	F.TAL	F.BNP
	D.BASX	E.REP	I.ENEL	I.ENI	**D.DBKX**	**F.SGE**	
4	E.TEF	F.GSZ	H.MT	M.NOK1	CRGI	D.EONX	F.AIR
	B.ABI	**E.IBE**					
5	F.QT.F	F.GOB	H.PHIL	D.SIEX	I.ISP	H.UNIL	I.UCG
	I.G	**F.MIDI**	**D.DAIX**	**F.LVMH**	**F.DG.F**	**D.VO3X**	

has been selected due to the fact that EURO STOXX 50 was experiencing severe losses (see Fig. 2).

We preliminary apply a univariate Student-t ARMA(1,1)-GARCH(1,1) model to each time series of log–returns of the constituents of the index to remove autocorrelation and heteroscedasticity from the data and compute the standardized residuals.

Then, we compute the conditional Spearman's ρ_α (here we select $\alpha = 0.10$) for all pairs of times series. By means of the procedures illustrated in Sect. 2, we determine a dissimilarity matrix and apply the fanny algorithm. According to the average silhouette index, the optimal number of cluster, k, is set equal to 5 (we run different algorithms with $k = 2, 3, \ldots, 8$).

Table 1 presents the cluster composition of the portfolio, when each asset is assigned to a cluster in a crisp way. Moreover, we highlighted in bold all the assets whose maximal membership degree is smaller than $T = 0.90$.

Thus, we run the portfolio selection algorithm by considering all the assets (i.e. by setting $T = 0$) or by considering the assets whose maximal membership degree is larger than $T = 0.90$). All the possible 82134 portfolios composed by 5 assets, such that each asset belongs to a different cluster, are calculated and visualized in Fig. 1, where the 25872 possible portfolios obtained by adopting the threshold $T = 0.90$ are colored in grey. As can be seen, the minimal CVaR portfolios generated by the algorithm with $T = 0$ and $T = 0.90$ coincide; however, the latter is obtained under a smaller computational effort.

In order to verify the performance of the methodology in an out-of-sample comparison, we consider the period from August 1, 2011 to September 9, 2011 as out-of-sample period, and compare the performance of the minimum CVaR portfolios obtained from our algorithm (with $T = 0$ and $T = 0.90$) with, respectively, the minimum variance portfolio and the minimum CVaR portfolio built from the whole set of assets, the equally weighted portfolio (obtained by assigning the same weight to

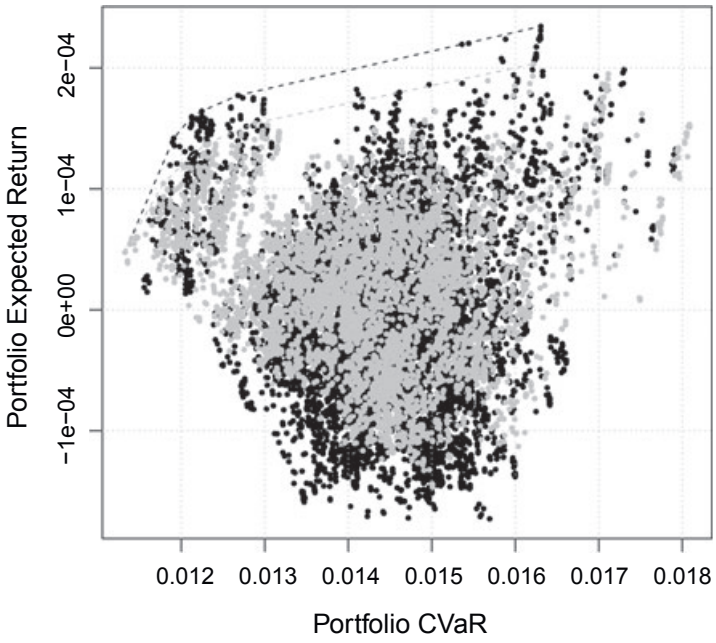

Fig. 1 Portfolio CVaR–Portfolio Expected Return plot of 5-asset portfolios generated at Step 4 of the portfolio selection algorithm. i highlights the portfolio frontier obtained from our algorithm with $T = 0$ (*black*) and $T = 0.90$ (*gray*)

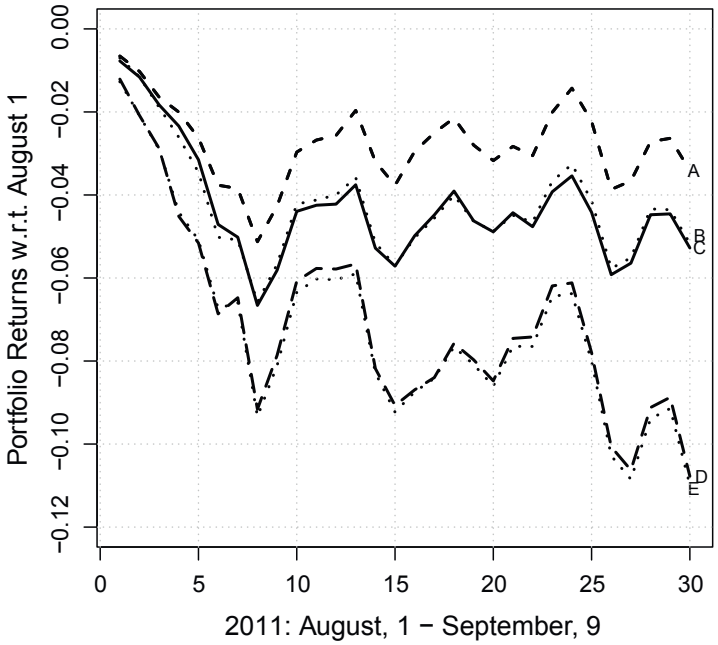

Fig. 2 Out-of-sample performance of the following portfolios: *A* minimum CVaR portfolio produced by our algorithm, *B* minimum variance portfolio from all 50 assets, *C* minimum CVaR portfolio from all 50 assets, *D* equally weighted portfolio, *E* EURO STOXX 50 index

each asset) and the benchmark index EURO STOXX 50. As it can be seen in Fig. 2, the performance of the portfolios selected from the proposed algorithm is better than the benchmark and outperforms the global minimum variance portfolio. This seems to confirm the idea that, when markets are experiencing a period of losses, a diversification strategy could be beneficial.

4 Conclusions

We have introduced a procedure aiming at selecting a portfolio from a group of assets in such a way that the assets are diversified in their tail behavior. The procedure exploits some features of fuzzy clustering algorithms. It is intended to be used by an investor to have more insights into the relationships among different assets in crisis periods.

Although these preliminary findings are promising, further analysis is necessary to assess the validity of the procedures. First, more benchmark datasets should be analyzed to assess the real usefulness of the proposed algorithm. Second, different tail dependence measures and/or clustering procedures (in particular, fuzzy c–medoids algorithms [1]) should be considered. Finally, as kindly suggested by one of the reviewers, in order to mitigate the computational burden, it could be also convenient to rank all the possible portfolios according to the sum of the membership degrees of their components and, hence, select the top p portfolios (p should be decided by the user) for further analysis. All these aspects will be the object of future investigations.

Acknowledgments The first author acknowledges the support of the Major Program of the National Social Science Foundation of China (No. 15ZDA017), and the support of Jilin University via the "Fundamental Research Funds for the Central Universities" (No. 450060522110) and via "Young Academic Leaders Training Program" (No. 2015FRLX07). The second author acknowledges the support of the Department of Economics, Business, Mathematics and Statistics "Bruno De Finetti" (University of Trieste, Italy), via the project "FRA". The third and fourth author acknowledge the support of the Faculty of Economics and Management (Free University of Bozen-Bolzano, Italy), via the project "COCCO".

References

1. Coppi R, D'Urso P, Giordani P (2006) Fuzzy C-Medoids clustering models for time-varying data. In: Bouchon-Meunier B, Coletti G, Yager R (eds) Modern information processing: from theory to applications. Elsevier Science, Amsterdam, pp 195–206
2. De Luca G, Zuccolotto P (2011) A tail dependence-based dissimilarity measure for financial time series clustering. Adv Data Anal Classif 5(4):323–340
3. De Luca G, Zuccolotto P (2015) Dynamic tail dependence clustering of financial time series. Stat Pap. doi:10.1007/s00362-015-0718-7
4. Dobrić J, Frahm G, Schmid F (2013) Dependence of stock returns in bull and bear markets. Depend Model 1:94–110
5. Durante F, Jaworski P (2010) Spatial contagion between financial markets: a copula-based approach. Appl Stoch Models Bus Ind 26(5):551–564
6. Durante F, Pappadà R, Torelli N (2014) Clustering of financial time series in risky scenarios. Adv Data Anal Classif 8:359–376
7. Durante F, Pappadà R, Torelli N (2015) Clustering of time series via non-parametric tail dependence estimation. Stat Pap 56(3):701–721
8. D'Urso P (2015) Fuzzy clustering. In: Meila M, Murtagh F, Rocci R (eds) Handbook of Cluster Analysis. Hennig C. Chapman & Hall,
9. Haerdle W, Nasekin S, Chuen D, Fai P (2014) TEDAS—Tail Event Driven ASset Allocation. Sfb 649 discussion papers, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2014-032.pdf

10. Haerdle W, Chuen D, Nasekin S, Ni X, Petukhina A (2015) Tail event driven ASset allocation: evidence from equity and mutual funds' markets. Sfb 649 discussion papers, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2015-045.pdf
11. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, data mining, inference, and prediction, Springer Series in Statistics, 2nd edn. Springer, New York
12. Kaufman L, Rousseeuw P (1990) Finding groups in data. Applied probability and statistics. Wiley Series in probability and mathematical statistics. John Wiley & Sons Inc., New York
13. Mantegna R (1999) Hierarchical structure in financial markets. Euro Phys J B 11(1):193–197
14. Patton AJ (2013) Copula methods for forecasting multivariate time series. In: Elliott G, Timmermann A (eds) Handbook of economic forecasting, vol 2. Elsevier, Oxford, pp 899–960
15. Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. J Risk 2:21–41