# *Combining Random Forest and Copula Function: a heuristic approach for selecting assets in a financial crisis perspective*

Giovanni De Luca[a], Giorgia Rivieccio[a] and Paola Zuccolotto[b*]

[a] *Università degli Studi di Napoli Parthenope*

*Dipartimento di Statistica e Matematica per la Ricerca Economica*

[b] *Università degli Studi di Brescia*

*Dipartimento Metodi Quantitativi*

**Abstract**

In this paper we propose a heuristic strategy aimed at selecting and analyzing a set of financial assets, focusing attention on their multivariate tail dependence structure. The selection, obtained through an algorithmic procedure based on data mining tools, assumes the existence of a reference asset we are specifically interested to. The procedure allows to opt for two alternatives: to prefer those assets exhibiting either a minimum lower tail dependence or a maximum upper tail dependence. The former could be a recommendable opportunity in a financial crisis period. For the selected assets, the tail dependence coefficients are estimated by means of a proper multivariate copula function.

**Keywords**: Copula functions, Archimedean copula, tail dependence, Random Forest.

[*] Correspondence to: Paola Zuccolotto, Dipartimento Metodi Quantitativi, C.da S. Chiara, 50 - 25122 Brescia, Italy. Email: zuk@eco.unibs.it

# 1 Introduction

In the last decades financial markets have been characterized by an increasing globalization and a complex set of relationships among asset returns has been established. In spite of these close connections, cross market correlation coefficients rarely assume significantly high values. In this context, Engle (2002) has demonstrated that we should preferably analyse dynamic correlation, showing that asset returns are positively associated conditionally on market volatility. The presence of a stronger dependence when markets are more volatile (and especially during crises) suggests to investigate the presence of an appreciably higher association in the tails of the joint distribution. In the literature this phenomenon is known as tail dependence. The main feature of joint distributions characterized by tail dependence is the presence of heavy and possibly asymmetric tails, thus the traditional hypothesis of (multivariate) Gaussianity is completely inadequate. In absence of a reasonable alternative distributional assumption, a copula approach can be particularly interesting.

Copula functions are effective quantitative tools for modelling the joint dependence of random variables (see for example Joe, 1997, Cherubini *et al.*, 2004 and Nelsen, 2006). The use of copula functions in finance is recent and the history of its rapid growth can be read in Genest *et al.* (2009). Applications of copula functions to bivariate financial time series have been carried out for capturing the dynamics of the dependence structure (Jondeau and Rockinger, 2006, Patton, 2006 and Bouyé and Salmon, 2009), for estimating the Value-at-Risk (Palaro and Hotta, 2006) or for measuring the tail dependence (Fortin and Kuzmics, 2002).

The main advantage of copula functions is that they allow us to preliminarily and separately model the marginal distributions, which are then joined into a multivariate distribution. A second desirable property is that some copula functions imply very flexible joint distributions, able to fulfil an in-depth analysis of the tail dependence structure. Unfortunately, this is not the case of the most common copula family, the Elliptical family, including the Gaussian and Student-$t$, which suffer from an absent or symmetric lower and upper tail dependence, respectively. On the other hand, the Archimedean family allows for different lower and upper tail dependence.

Due to the complex structure of financial markets, a high-dimensional multivariate approach to tail dependence analysis is surely more insightful. However, a growing number of jointly modeled variables hugely and rapidly increases formal and computational complexity. In an analysis of geographical indices, for example, a joint study of all the markets in the world is impossible. In general, even a drastic restriction of the study to the so called *developed markets* is not sufficient to provide a manageable number of variables.

To cope with the dimensionality problem, a number of strategies based on the reduction of a multivariate copula to a cascade of bivariate copulae can be found in the literature (see Aas and Berg, 2009 for a detailed comparison of the proposed techniques). Alternatively, a selection procedure of the most suitable (according to some definite rule) assets is necessary. However, in high-dimensional contexts, the selection procedure itself can be computationally burdensome.

In this paper we propose to realize the selection using data mining tools. We face the problem with a heuristic reasoning and we propose an algorithmic procedure, based on the recent technique of Random Forest (see Breiman, 2001), in order to opportunely select the assets we want to introduce in an analysis aimed at investigating tail dependence. The selection is built with a hierarchical structure *around* an asset we are interested to. In other words, we firstly choose a reference asset we want to necessarily include in the set of analysed assets, as frequently happens in investment strategies, and then, after filtering the data from autocorrelation and heteroskedasticity, we select step-by-step the other assets, by adding an asset at each step until a termination criterion is satisfied. For the selected assets, we propose to use a copula approach in order to estimate the tail dependence coefficients.[1]

Thus, the aim of this paper is to propose a structured procedure for selecting and analyzing a set of financial assets, focusing attention on their joint tail dependence.

Given a large set of financial assets, the proposed strategy is organized in three steps:

1. application of univariate models to the financial returns, in order to filter the data from autocorrelation and heteroskedasticity;
2. selection of $k$ financial assets (including the reference one) whose returns exhibit a low (high) level of lower (upper) tail dependence. In detail, in an investment perspective, we could aim at including in a portfolio assets with low association in case of negative shocks, or assets with high association in case of positive shocks. The choice between the two strategies depends upon the expectation about the future trends of the financial markets and upon the desired risk degree of the investors. In this paper we mimic a financial crisis perspective, hence we will focus on the selection of assets with low association in the lower tail. The extension to the upper tail is straightforward;
3. estimation of the multivariate tail dependence coefficients.

For each step we propose a specific statistical tool:

1. an AR-GARCH model for univariate returns in order to obtain standardized residuals;
2. an algorithmic heuristic selection procedure with a hierarchical structure, based on the analysis of joint association in the extreme values by means of a data mining technique called Random Forest;
3. a copula function for the estimation of multivariate tail dependence coefficients.

[1] An alternative idea in this context is to model the tails of the multivariate distributions using extreme value theory (see McNeil, 1999).

The paper is organized as follows. In Section 2 the theory of copula functions is briefly recalled and the tail dependence coefficients are defined. Section 3 describes an empirical problem showing that a selection procedure using copula functions is computationally unfeasible. Section 4, after briefly recalling the Random Forest algorithm, illustrates the functioning of the proposed asset selection procedure, also presenting the results of two simulation studies. An application to real data is shown in (Section 5). Section 6 concludes.

## 2 Copula functions and tail dependence

In the multivariate analysis of returns, the assumption about their distribution is a critical issue. In the past the hypothesis of Gaussianity has been largely exploited but it has seldom provided satisfactory results in terms of density forecasts which are very useful in risk management (e.g. to compute the Value-at-Risk or the Expected Shortfall). In the recent years, a great deal of interest in non-normal probability laws has contributed to overcome the traditional Gaussian distribution. The multivariate $t$ and Skew-$t$ distributions are remarkable examples. However, a high degree of flexibility can be reached using a copula function.

A copula function is a multivariate distribution function with standard uniform marginal distributions. According to the theorem proposed by Sklar (1959), each distribution function $H(x_1, x_2, \dots, x_n)$ can be expressed by a copula function whose arguments are the univariate distribution functions, that is

$$H(x_1, x_2, \dots, x_n) = C(F_1(x_1), F_2(x_2), \dots, F_n(x_n)). \tag{1}$$

If the distribution function $H$ is continuous, then the copula $C$ is unique. Conversely, if $C$ is a copula and $F_1(x_1)$, $F_2(x_2)$,. . .,$F_n(x_n)$ are the marginal distributions, then $H(x_1, x_2, \dots, x_n)$ is a joint distribution function with margins $F_i(\cdot)$.

The main advantage of using a copula function is that the specification of the marginal distributions can be separated from the definition of the dependence structure.

It is common to denote $u_i = F_i(x_i)$, so that (1) is usually presented as

$$H(x_1, x_2, \ldots, x_n) = C(u_1, u_2, \ldots, u_n).$$

The most popular families of copula functions are the Elliptical and the Archimedean. Among the Elliptical copulae a prominent role is assigned to the Gaussian and the Student's $t$ copulae.

The Archimedean copulae are defined through a generator function, $\Phi : I \to R^+$, continuous, decreasing and convex, such that $\Phi(1) = 0$. A bivariate Archimedean copula is expressed as

$$C(u_1, u_2) = \Phi^{-1}(\Phi(u_1) + \Phi(u_2)).$$

In the $n$-dimensional case, an Archimedean copula function is defined as

$$C(u_1, \ldots, u_n) = \Phi^{-1}(\Phi(u_1) + \Phi(u_2) + \ldots + \Phi(u_n)).$$

A remarkable example, widely used in financial time-series analyses, is the Clayton copula, given by

$$C(u_1, \ldots, u_n) = \left( \sum_{i=1}^{n} u_i^{-\theta} - (n-1) \right)^{-1/\theta} \tag{2}$$

characterized by the parameter $\theta > 0$. The presence of a unique parameter, which captures the co-movements only between extremely low values, that is in the lower tail of the distribution, gives to this copula function a limited capacity of explaining very complex relationships among $n$ variables. A more satisfying description of these relationships can be achieved by considering a copula function able to capture the movements in both the tails, such as the Joe-Clayton copula (also known as BB7 copula in the bivariate case according to classification in Joe, 1997), which is a generalization of the Clayton copula. It is formalized as

$$C(u_1, \ldots, u_n) = 1 - \left\{ 1 - \left[ \sum_{i=1}^{n} (1 - (1 - u_i)^{\kappa})^{-\theta} - (n-1) \right]^{-1/\theta} \right\}^{1/\kappa}, \tag{3}$$

and is then characterized by two parameters, $\theta > 0$ and $\kappa \geq 1$. The generator function and its inverse are given, respectively, by $\Phi(t) = [1-(1-t)^{\kappa}]^{-\theta} - 1$ and $\Phi^{-1}(t) = 1-[1-(1+t)^{-1/\theta}]^{1/\kappa}$. When $\kappa = 1$ we turn back to the Clayton copula.

In general, for a $n$-variate copula, the density is obtained by computing the $n$-th derivative,

$$\frac{\partial C(u_1, u_2, \ldots, u_n)}{\partial u_1 \partial u_2 \ldots \partial u_n}. \tag{4}$$

The parameters are usually estimated through the maximum likelihood method.

## 2.1 Tail dependence

Given two random variables, $X_i$ and $X_j$, several measures of association can be considered. The most popular measures are the linear correlation and the concordance. However, there are many other measures. The tail dependence is a key measure when we are interested to risk management. It captures the concordance between extreme values of the variables. More specifically, we distinguish the lower tail dependence and the upper tail dependence.

The lower tail dependence coefficient between two variables $X_i$ and $X_j$, denoted as $\lambda_L^{i|j}$, is given by

$$\begin{aligned} \lambda_L^{i|j} &= \lim_{v \to 0^+} P(F_i(X_i) \leq v | F_j(X_j) \leq v) \\ &= \lim_{v \to 0^+} P(U_i \leq v | U_j \leq v). \end{aligned}$$

It measures the concordance between extremely low values of $X_i$ and $X_j$.

Similarly, the upper tail dependence coefficient, denoted as $\lambda_U^{i|j}$, is defined as

$$\begin{aligned} \lambda_U^{i|j} &= \lim_{v \to 1^-} P(F_i(X_i) > v | F_j(X_j) > v) \\ &= \lim_{v \to 1^-} P(U_i > v | U_j > v) \end{aligned}$$

and measures the concordance between extremely high values of $X_i$ and $X_j$.

The choice of the family and of the specific copula can be driven by the observed dependence between extreme values. For example, the Gaussian copula does not allow for any tail dependence while the $t$-copula models dependence in the two tails in the same way. These copulae should be chosen when the hypothesis of absence or equality of upper and lower tail dependence, respectively, is reasonable.

On the other hand, Archimedean copulae are more manageable from this point of view. They admit lower or upper tail dependence, or both, in a non-symmetric way.

In the financial context, the lower tail dependence assumes a very important role. In fact, the dependence between extremely low values of returns is a measure of the risk related to a set of assets. Thus a significant statistical tool for risk management can be a copula function able to properly model at least the lower tail dependence.

The Joe-Clayton copula (3) admits both lower and upper tail dependence whose coefficients are given by

$$\lambda_L^{i|j} = 2^{-1/\theta}$$

and

$$\lambda_U^{i|j} = 2 - 2^{1/\kappa}$$

for $i, j = 1, \ldots, n$ and $i \neq j$.

## 2.2 Multivariate tail dependence

Tail dependence is a bivariate concept. An extension of its definition to a multivariate context has been provided in De Luca and Rivieccio (2010). Given $n$ variables, it is possible to compute a lot of conditional probabilities. For instance, we could be interested in the probability of an extremely low value of asset $i$, given that extremely low values have occurred for $m$ assets,

$$\lambda_L^{i|j_1 \ldots j_m} = \lim_{v \to 0^+} P(U_i \leq v | U_{j_1} \leq v, \ldots, U_{j_m} \leq v)$$

or we could be interested in the probability of extremely low values of $m$ assets, given an extremely low value of the $j$-th asset,

$$\lambda_L^{i_1 \ldots i_m | j} = \lim_{v \to 0^+} P(U_{i_1} \leq v, \ldots, U_{i_m} \leq v | U_j \leq v),$$

an so on. These conditional probabilities are easily evaluated from the equation of the copula function. Multivariate upper tail dependence coefficients can be defined in an analogous way.

When we have to select a certain number of assets starting from a reference asset (referred to as asset 1) to which other assets (referred to as assets $2, 3, \ldots$) are added in turn, we can compute a sequence of lower tail dependence coefficients, following the order of the added assets, enlarging in turn the information set, that is the conditioning event. So, after estimating

$$\lambda_L^{1|2} = \lim_{v \to 0^+} P(U_1 \leq v | U_2 \leq v),$$

the lower tail dependence coefficient between asset 1 and asset 2, we can estimate

$$\lambda_L^{1|23} = \lim_{v \to 0^+} P(U_1 \leq v | U_2 \leq v, U_3 \leq v),$$

until

$$\lambda_L^{1|2 \ldots n} = \lim_{v \to 0^+} P(U_1 \leq v | U_2 \leq v, \ldots, U_n \leq v). \tag{5}$$

This sequence of conditional probabilities can be seen as a measure of the tail dependence of a reference asset (asset 1) on the other assets. It offers a clear view of the riskiness of asset 1 in a crisis period describing the possible contagion in terms of probabilities.

At the same time, we could be interested in

$$\lambda_L^{12|3} = \lim_{v \to 0^+} P(U_1 \leq v, U_2 \leq v | U_3 \leq v),$$

until

$$\lambda_L^{1 \ldots n-1|n} = \lim_{v \to 0^+} P(U_1 \leq v, \ldots, U_{n-1} \leq v | U_n \leq v). \tag{6}$$

It is worth noting that the following relation holds between the coefficient (6) and some marginal coefficients of the form of (5):

$$\lambda_L^{1 \ldots n-1|n} = \lambda_L^{1|2 \ldots n} \lambda_L^{2|3 \ldots n} \lambda_L^{3|4 \ldots n} \ldots \lambda_L^{n-1|n}. \tag{7}$$

Thus the tail dependence coefficient (6) is able to measure what we call *chain effect risk* in an $n$-set of assets, that is the probability of a crisis for the entire $n$-set, given that a default has occurred for asset $n$.

Applying these definitions, De Luca and Rivieccio (2010) have shown that for the $n$-dimensional Joe-Clayton copula (3), the above mentioned lower tail dependence coefficients are given by

$$\lambda_L^{1|2\ldots n} = \left(\frac{n}{n-1}\right)^{-1/\theta}$$

and

$$\lambda_L^{1\ldots n-1|n} = (n)^{-1/\theta} .$$

# 3 The problem of selecting assets in a financial crisis perspective

The estimation of tail dependence coefficients with a copula function can be used in order to select, out of a great set of $p$ possible investment alternatives, a subset of $k$ assets with minimum or maximum multivariate tail dependence, which hereafter will be called $k$-set. This idea is quite general and allows to opt for two different risk management strategies: defensive, when we choose assets with minimum lower tail dependence, or aggressive, when we choose assets with maximum upper tail dependence. The former is recommendable in a financial crisis perspective, as the latter can be used when high returns are expected in the market. As stated before, in this paper we limit our empirical analysis to the first case. In addition, we suppose to have a reference asset we want to necessarily include in the $k$-set.

The first problem is to define what we mean when we speak of *minimum tail dependence*. As a matter of fact, while in a bivariate context tail dependence is described by a unique coefficient, in a multivariate analysis there are a lot of coefficients describing the tail dependence structure of a $k$-set. For example there are coefficients like (5) or like (6) and, for each of these two types, all the marginal coefficients measuring tail dependence in subsets of the $k$-set. Hence we have to decide which is our main goal, that is which is the multivariate tail dependence coefficient we want to minimize. From a financial point of view, recalling the *chain effect* described by (7), we think that the coefficient (6) provides a good description of the tail dependence structure of the $k$-set, thus its minimization will be the goal of our assets selection.

The second problem is to formulate a procedure to select the $k$ assets with the minimum (6). Given a reference asset, a first way consists in using copula functions in order to estimate (6) relatively to all the $\binom{p-1}{k-1}$ $k$-sets containing the reference asset and then choosing the $k$-set exhibiting the minimum estimated value. Although the approach with copula functions has been recognized to offer good estimates of the tail dependence coefficients, it rapidly becomes cumbersome to apply in a high-dimensional context, because it requires the optimization of a very complex likelihood function. Hence this procedure can be applied only for very small values of $k$, as will be clear in the real data example presented in the next subsection.

The third problem is to set the value $k$. Since the coefficient (6) tends to decrease as $k$ increases, a straightforward solution is to run the procedure for $k = 2$, $k = 3$, $\ldots$, until it reaches a defined size. In other words, for each value of $k$ we select the set minimizing (6), and we stop when this minimum is sufficiently low, then choosing the smallest set of indices with a *chain effect risk* lower than a given threshold.

## 3.1 Empirical analysis of MSCI geographical indices

The analysed dataset is composed of 23 MSCI market capitalization weighted indices, addressed to measure the equity market performance of developed markets[2], recorded daily from June 3, 2002 to June 10, 2010 (2095 observations; Source: MSCI Barra).

Let us suppose to designate MSCI-Italy as the reference asset. We want to select, out of the $p = 23$ indices, the $k$-set ($k < p$) containing MSCI-Italy, with

[2] The MSCI World Index consists of the following developed market country indices: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Sweden, Switzerland, the United Kingdom, and the United States.

the minimum (6) and we desire that this minimum is lower than a threshold that we decide to set at the value 0.005. The log-returns are preliminary filtered by means of univariate Student-$t$ AR-GARCH models and the tail dependence is evaluated on the corresponding standardized residuals. For $k = 2$ we estimate all the 22 possible bivariate Joe-Clayton copulae (3) between MSCI-Italy and each of the candidate assets. We find that the minimum lower tail dependence coefficient (6) is obtained in correspondence of the couple MSCI-Italy and MSCI-Japan, with $\lambda_L^{It|Ja} = 0.0201$. Since the value is over the defined threshold, we have to increase $k$.

For $k = 3$ we estimate all the possible trivariate Joe-Clayton copulae. We have to estimate 231 copulae. We find that the minimum lower tail dependence coefficient (6) is obtained in correspondence of the triplet MSCI-Italy, MSCI-Japan, MSCI-USA, with $\lambda_L^{It,Ja|US} = 0.0061$. The value is again over the threshold, so we should continue increasing $k$, but the computational burden is now very high. In fact, for $k = 4$ we have to estimate 1540 copulae, for $k = 5$, 7315 copulae, and so on. It is clear that when the dimension of the problem increases, the described method rapidly becomes dramatically time-consuming, both for the increasing number of estimates and for the increasing complexity of the likelihood function. An alternative procedure is then recommended.

## 4 A heuristic procedure for asset selection

As shown above, when a large set of investment alternatives is available, we need a manageable selection procedure in terms of the computational burden. In this section we propose a heuristic approach (subsection 4.2), based on the variable importance measurement by means of the algorithmic technique of Random Forest, well-known in the field of data mining, which will be preliminarily recalled in subsection 4.1. The procedure is aimed at selecting a $k$-set exhibiting a low (high) joint association in lower (upper) extreme values out of a great set of investment opportunities. This is made building a hierarchical structure around a reference asset we are specifically interested to. The performance of the proposed procedure is inspected with some simulation studies (subsection 4.3).

## 4.1 Variable importance measurement with Random Forest

Given a response variable and a set of covariates, variable importance measurement allows to identify the most important predictors for the response variable within the set of covariates. The prediction problem is called classification if the response variable is categorical and regression if it is numerical. Some powerful data mining tools have been recently proposed in the framework of learning ensembles, algorithmic techniques able to face both the problems of prediction and of variable importance measurement, even in presence of many redundant predictors and of complex relationships among the variables. Each ensemble member is given by a different function of the input covariates and predictions are obtained by a linear combination of the prediction of each member (see Breiman, 1996, Friedman and Popescu, 2005). Learning ensembles can be built using different prediction methods, that is different base learners as ensemble members. The most interesting proposals use decision trees (more specifically CART, Classification And Regression Trees, Breiman *et al.*, 1984) as base learners and are called tree-based learning ensembles. Popular examples are the Random Forest technique (RF, Breiman, 2001) or the tree-based Gradient Boosting Machine (GBM, Friedman, 2001). Both these algorithmic techniques identify the most important predictors within the set of covariates, by means of the computation of some variable importance measures.

RF with randomly selected inputs are sequences of trees grown by selecting at random *at each node* a small group of $F$ input variables to split on. This procedure is often used in tandem with *bagging* (Breiman, 1996), that is with a random selection of a subsample of the original training set at each tree. This simple and effective idea is founded on a complete theoretical apparatus analytically described by Breiman (2001) in his seminal work. The RF prediction is computed as an average of the single trees predictions. This successfully neutralizes the well-known instability of decisions trees. In addition, four measures of variable importance $M1$, $M2$, $M3$, $M4$ are available in order to identify informative predictors (Breiman, 2002). In the recent literature $M1$ and $M4$ are addressed as the two main RF variable importance measures, as $M2$ and $M3$ have revealed somehow redundant with $M1$:

- **Measure 1 - Mean Decrease in Accuracy**: at each tree of the RF all the values of the $h$-th covariate are randomly permuted. New predictions are obtained with this dataset, where the role of the $h$-th covariate is completely destroyed. The prediction error provided by this new dataset is compared with the prediction error of the original one and the $M1$ measure for $h$-th variable is given by the difference of these two errors.
- **Measure 4 - Total Decrease in Node Impurities**: at each node $z$ in every tree only a small number of variables is randomly chosen to split on, relying on some splitting criterion given by a variability/heterogeneity index such as the MSE for regression and the Gini index or the Shannon entropy for classification. Let $d(h,z)$ be the maximum decrease (over all the possible cutpoints) in the index allowed by variable $X_h$ at node $z$. $X_h$ is used to split at node $z$ if $d(h,z) > d(w,z)$ for all variables $X_w$ randomly chosen at node $z$. The $M4$ measure is calculated as the sum of all decreases in the RF due to $h$-th variable, divided by the number of trees.

## 4.2 The algorithm

In order to select a $k$-set of assets with a low (high) mutual extreme values association we set up a heuristic algorithmic procedure based on the analysis of the association among extreme values within a set of financial assets.

The proposed technique is based on RF variable importance measures, used for the special purpose to identify alternatively the most or the least influential predictors of a given outcome.

The procedure can be summarized as follows. Let $X_{1t}, X_{2t}, \cdots, X_{pt},\ t = 1, 2, \cdots, T$ be the log-returns time series of $p$ assets. A Student-$t$ AR-GARCH model is fitted to each series. Let $Z_{1t}, Z_{2t}, \cdots, Z_{pt}$ be the standardized residuals, and $\hat{\nu}_1, \hat{\nu}_2, \cdots, \hat{\nu}_p$ their estimated degrees of freedom. At this point we have to choose a "reference asset", say the $h$-th, whose extreme values association with the others we want to analyse.

### (A) Lower extreme values association

1. Set $\alpha = \alpha_0 = L_t/100$ where $L_t \in \mathbb{N}$ and $0 < L_t < 50$;

2. using Student-$t$ distributions with the estimated degrees of freedom, $T_\nu$, compute the quantiles $q_{1,\alpha}, q_{2,\alpha}, \cdots, q_{p,\alpha}$ of the $p$ standardized residuals, where
$$q_{j,\alpha} : Pr(Z_{jt} \leq q_{j,\alpha}) = Pr(T_{\nu_j} \leq q_{j,\alpha}) = \alpha;$$

3. create $p$ binary 0/1 series $Y_{1t}, Y_{2t}, \cdots, Y_{pt}$ according to the following rule:
$$y_{jt} = \begin{cases} 1 & \text{if} \quad z_{jt} \leq q_{j,\alpha} \\ 0 & \text{otherwise} \end{cases} ; \tag{8}$$

4. perform a RF classification with response variable $Y_{ht}$ and the others as predictors;

5. compute variable importance measures and let $\mathbf{I}^{(\alpha)} = \{I_1^{(\alpha)}, \cdots, I_{h-1}^{(\alpha)}, I_{h+1}^{(\alpha)}, \cdots, I_p^{(\alpha)}\}'$ be the vector containing the relative deviations of the $p-1$ measures and their average (relative importances);

6. repeat steps (2) through (5) for $\alpha = \alpha_0 - 0.01, \alpha_0 - 0.02, \cdots, 0.01$;

7. collect the $(p-1) \times L_t$ relative importances in the matrix $\mathbf{I}_L = \{\mathbf{I}^{(\alpha_0)}, \mathbf{I}^{(\alpha_0 - 0.01)}, \cdots, \mathbf{I}^{(0.01)}\}$;

8. compute the vector of the average relative importances of the $p-1$ predictors $\bar{\mathbf{I}}_L = \dfrac{\mathbf{I}_L \cdot \mathbf{1}}{L_t}$, where $\mathbf{1}$ is the $L_t \times 1$ vector of ones.

### (B) Upper extreme values association

1. Set $\alpha = \alpha_0 = U_t/100$ where[3] $U_t \in \mathbb{N}$ and $50 < L_t < 100$;

2. using Student-$t$ distributions with the estimated degrees of freedom, $T_\nu$, compute the quantiles $q_{1,1-\alpha}, q_{2,1-\alpha}, \cdots, q_{p,1-\alpha}$ of the $p$ standardized residuals, where
$$q_{j,1-\alpha} : Pr(Z_{jt} \leq q_{j,\alpha}) = Pr(T_{\nu_j} \leq q_{j,\alpha}) = 1 - \alpha;$$

3. create $p$ binary 0/1 series $Y_{1t}, Y_{2t}, \cdots, Y_{pt}$ according to the following rule:
$$y_{jt} = \begin{cases} 0 & \text{if} \quad z_{jt} \leq q_{j,1-\alpha} \\ 1 & \text{otherwise} \end{cases} ; \tag{9}$$

4. perform a RF classification with response variable $Y_{ht}$ and the others as predictors;

5. compute variable importance measures and let $\mathbf{I}^{(\alpha)} = \{I_1^{(\alpha)}, \cdots, I_{h-1}^{(\alpha)}, I_{h+1}^{(\alpha)}, \cdots, I_p^{(\alpha)}\}'$ be the vector containing the relative deviations of the $p-1$ measures and their average (relative importances);

6. repeat steps (2) through (5) for $\alpha = \alpha_0 + 0.01, \alpha_0 + 0.02, \cdots, 0.99$.

7. collect the $(p-1) \times (100 - U_t)$ relative importances in the matrix $\mathbf{I}_U = \{\mathbf{I}^{(\alpha_0)}, \mathbf{I}^{(\alpha_0+0.01)}, \cdots, \mathbf{I}^{(0.99)}\}$;

8. compute the vector of the average relative importances of the $p-1$ predictors $\bar{\mathbf{I}}_U = \dfrac{\mathbf{I} \cdot \mathbf{1}}{100 - U_t}$, where $\mathbf{1}$ is the $(100 - U_t) \times 1$ vector of ones.

The generic $j$-th elements of the vectors $\bar{\mathbf{I}}_L$ and $\bar{\mathbf{I}}_U$ are measures of the average importance of $j$-th asset extreme values in the prediction of the extreme values of the "reference asset", conditionally on the extreme values of the other

[3] Relying on heuristic reasoning, we recommend to set $10 \leq L_t \leq 20$ and $80 \leq U_t \leq 90$. Simulation studies show that within these ranges the procedure is very robust and the choice of $L_t$ and $U_t$ does not affect results.

assets. Even so, the selection of a $k$-set of assets cannot be based only on the vectors $\bar{\mathbf{I}}_L$ and $\bar{\mathbf{I}}_U$. If we want, for example, $k$ assets with a mutually low association in lower extreme values, it is not correct to run procedure (A) and select the $k-1$ assets having the lowest importance in predicting the "reference asset" extreme values. In fact this could lead to select assets with low association with the "reference asset", but highly associated each other. In order to avoid this misconstruction, procedures (A) and (B) have to be iterated $k-1$ times, as described in the following procedure (C).

In general we desire a set of assets having a low association in lower extreme values, in order to counterbalance negative shocks, or a high association in upper extreme values, in order to accentuate positive shocks. Hereafter we will assume that procedures (A) and (B) are iterated in (C) in order to select assets respectively with low and high association.

### (C) Selection of $k-1$ assets

1. Choose the "main reference asset", the asset which has to be necessarily included in the $k$-set;

2. set $w = 1$;

3. apply procedure (A) and let $\bar{\mathbf{I}}_L^{(w)}$ be the vector of the $p-w$ average relative importances contained in $\bar{\mathbf{I}}_L$.

4. select the asset, say the $g$-th, which satisfies the following rule:

$$\sum_{i=1}^{w} \bar{I}_{gL}^{(i)} = \min_{s=1,2,\cdots,p-w} \sum_{i=1}^{w} \bar{I}_{sL}^{(i)} \tag{10}$$

   where $\bar{I}_{gL}^{(i)}$ is the average relative importance of $g$-th asset when $w = i$;

5. prepare a new iteration of procedure (A) by removing the actual reference asset and setting the $g$-th asset as new reference asset;

6. repeat steps (2) through (5) for $w = 2, 3, \cdots, k-2$.

7. repeat steps (2) through (4) for $w = k - 1$.

8. repeat steps (1) through (7) using procedure (B) and replacing (10) with

$$\sum_{i=1}^{w} \bar{I}_{gU}^{(i)} = \max_{s=1,2,\cdots,p-w} \sum_{i=1}^{w} \bar{I}_{sU}^{(i)}. \tag{11}$$

At the end of procedure (C), two $k$-sets of assets are selected out of the $p$ assets in the dataset, having respectively a mutual low and high extreme values association. This selection is obtained with an algorithmic approach which does not need any prior assumption on the multivariate joint distribution.

After the selection, we are ready to perform a more complete and accurate analysis through the estimation of tail dependence coefficients with copula functions.

### 4.3 Simulation studies in the lower tail

As stated before, in this paper we limit our analysis to a financial crisis perspective. Hence we employ the procedure (A) described in the previous paragraph, thus iterating only step (1) through (7) of procedure (C). In this context we carry out the following two simulation studies in order to check the performance of the algorithm.

**Simulation 1.** In the first simulation study $N = 1000$ observations are randomly drawn from a multivariate 25-dimensional Student-$t$ distribution $\mathbf{X} = (X_1, \cdots, X_{25})$ with zero mean vector and correlation matrix

$$\mathbf{P_X} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_5 \end{bmatrix}$$

where $\mathbf{0}$ is a $(5 \times 5)$ matrix of zeros and $\mathbf{P}_1 = \cdots = \mathbf{P}_5 = (1-\rho)\mathbf{I}_5 + \rho$, with $\mathbf{I}_5$ denoting the 5-dimensional identity matrix. In detail, the 25 variables are divided into 5 blocks. Within each block the variables are mutually correlated with correlation coefficient $\rho$. Variables belonging to different blocks are uncorrelated. This data generating process tries to emulate the existence of 5 different markets composed by assets associated each other, but not with assets belonging to other markets. Since with a multivariate Student-$t$ distribution with $\nu$ degrees of freedom, correlation between two marginal distributions directly reflects on their lower tail dependence, in this situation, a good portfolio should contain assets of different markets. We suppose to desire a $k$-set composed by $k = 5$ variables. We decide to set $X_1$ as main reference variable, thus selecting $k - 1 = 4$ further variables by means of the proposed heuristic procedure with $L_t = 10$. Fixed $\nu = 5$, the procedure is repeated $r = 50$ times for different values of $\rho$, ranging from 0.05 to 0.9. We consider "correct" a $k$-set composed by one variable per block. The rate of correct $k$-sets rapidly increases when the association among variables becomes stronger (Figure 1, left). The simulation has been carried out also fixing different degrees of freedom ($\nu = 15$ and $\nu = 30$), with the same results.

Figure 1 about here

**Simulation 2.** In the second simulation study $N = 1000$ observations are randomly drawn from a multivariate 15-dimensional Student-$t$ distribution $\mathbf{X} = (X_1, \cdots, X_{15})$ with zero mean vector and correlation matrix

$$\mathbf{P_X} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{P}_{12} & \mathbf{P}_{13} \\ \mathbf{P}_{21} & \mathbf{P}_2 & \mathbf{P}_{23} \\ \mathbf{P}_{31} & \mathbf{P}_{32} & \mathbf{P}_3 \end{bmatrix}$$

where $\mathbf{P}_1 = \mathbf{P}_2 = \mathbf{P}_3 = (1 - 0.6)\mathbf{I}_5 + 0.6$ and $\mathbf{P}_{12} = \mathbf{P}_{13} = \mathbf{P}_{23} = \mathbf{P}'_{21} = \mathbf{P}'_{31} = \mathbf{P}'_{32}$ are $(5 \times 5)$ matrices obtained by multiplication of the column vector $(0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4)'$ and the 5-dimensional row vector of ones. Thus the 15 variables are divided into 3 blocks. Each variable has a moderately high correlation ($\rho = 0.6$) with variables belonging to the same block, but has also an increasing correlation with the variables belonging to different blocks ($\rho = 0$ with the first variable of each block, $\rho = 0.1$ with the second, $\rho = 0.2$ with the third, $\rho = 0.3$ with the fourth, $\rho = 0.4$ with the last). In this situation, the selection of the $k$-set is more challenging. If we set $X_1$ as main reference variable, the algorithm should proceed selecting firstly one of the first variables of each block, then one of the second ones, then one of the third ones, and so on, until the remaining $k - 1$ variables are selected (Table 1).

Table 1 about here

The procedure is repeated $r = 50$ times with $L_t = 10$, for different values of $k$, from 3 to 12. For each repetition, let $ncv$ be the number of correct variables in the selected $k$-set, the rate of correctly selected variables

$$CSV = av_r \left\{ \frac{ncv - 1}{k - 1} \right\}$$

ranges from 0.644 to 0.89 (Figure 1, right).

# 5 Application of the heuristic procedure to MSCI indices

In this section we apply the proposed algorithmic procedure to the dataset described in subection 3.1, in order to select a $k$-set with minimum lower tail dependence, choosing the smallest value $k$ ensuring that coefficient (6) is lower than 0.005. As pointed out in subsection 3.1, we preliminarily depurate data from autocorrelation and heteroskedasticity, by means of univariate Student-$t$ AR-GARCH models applied to the log-returns. Then the procedure is carried out using the standardized residuals, setting $L_t = 10$ and with MSCI-Italy as *main reference asset.*

We run the procedure for $k = 2$, $k = 3$, ..., each time computing the coefficient (6) and stopping when it is lower than the chosen threshold[4]. The step-by-step results are the following:

[4] The computations of the asset selection algorithm are carried out using the library `randomForest` of the R package (R Development Core Team, 2006). The R script is available on email request to the Corresponding Author.

- $k = 2$

  Procedure (A) is applied with MSCI-Italy as *reference asset* and the vector $\bar{\mathbf{I}}_L$, which we will call $\bar{\mathbf{I}}_L^{(1)}$ because it is computed at the first iteration, is obtained (Fig. 2, top left). MSCI-Japan is selected. The first $k$-set ($k = 2$) is then $\{It, Ja\}$ (consistently with the results obtained in subsection 3.1), with $\lambda_L^{It|Ja} = 0.0201$. This value is greater than the given threshold of 0.005, so we have to increase $k$.

- $k = 3$

  Since the algorithmic procedure is hierarchical, the first step is the same as for $k = 2$. After the selection of MSCI-Japan, MSCI-Italy is removed from the dataset, procedure (A) is applied with MSCI-Japan as *reference asset* and the vector $\bar{\mathbf{I}}_L^{(2)}$ is obtained (Fig. 2, top right). For each index in the dataset the values in $\bar{\mathbf{I}}_L^{(1)}$ and $\bar{\mathbf{I}}_L^{(2)}$ are summed (Fig. 3, top right). MSCI-USA (exhibiting the minimum summed value) is selected. The second $k$-set ($k = 3$) is then $\{It, Ja, US\}$ (again, consistently with the results obtained in subsection 3.1), with $\lambda_L^{It,Ja|US} = 0.0061$. We have to increase $k$.

- $k = 4$

  Again, the first two steps are the same as for $k = 3$. After the selection of MSCI-USA, MSCI-Japan is removed from the dataset, procedure (A) is applied with MSCI-USA as *reference asset* and the vector $\bar{\mathbf{I}}_L^{(3)}$ is obtained (Fig. 2, bottom left). For each index in the dataset the values in $\bar{\mathbf{I}}_L^{(1)}$, $\bar{\mathbf{I}}_L^{(2)}$ and $\bar{\mathbf{I}}_L^{(3)}$ are summed (Fig. 3, bottom left). MSCI-Greece (exhibiting the minimum summed value) is selected. The third $k$-set ($k = 4$) is then $\{It, Ja, US, Gr\}$, with $\lambda_L^{It,Ja,US|Gr} = 0.0055$. We have to increase $k$.

- $k = 5$

  The first three steps are the same as for $k = 4$. After the selection of MSCI-Greece, MSCI-USA is removed from the dataset, procedure (A) is applied with MSCI-Greece as *reference asset* and the vector $\bar{\mathbf{I}}_L^{(4)}$ is obtained (Fig. 2, bottom right). For each index in the dataset the values in $\bar{\mathbf{I}}_L^{(1)}$, $\bar{\mathbf{I}}_L^{(2)}$, $\bar{\mathbf{I}}_L^{(3)}$ and $\bar{\mathbf{I}}_L^{(4)}$ are summed (Fig. 3, bottom right). MSCI-New Zealand (exhibiting the minimum summed value) is selected, with $\lambda_L^{It,Ja,US,Gr|NZ} = 0.0015$. The desired threshold has been reached, so the final $k$-set ($k = 5$) is given by $\{It, Ja, US, Gr, NZ\}$.

Figure 2 about here

Figure 3 about here

Table 2 about here

Table 3 about here

Table 4 about here

Table 5 about here

Tables 2 through 5 show details about the estimation of the parameters of the Joe-Clayton copulae. In order to check the goodness-of-fit of the estimated copulae, extending Dobric and Schmid (2006), we have considered the general null hypothesis that a multivariate data set can be described by a specified copula

$$H_0 : (X_1, X_2 \dots, X_n) \text{ has copula } C$$

through the auxiliary hypothesis

$$H_0^* : S(X_1, X_2, \dots, X_n) \sim \chi_n^2,$$

where

$$\begin{aligned} S(X_1, X_2, \dots, X_n) &= [\Phi^{-1}(F_1(X_1))]^2 + [\Phi^{-1}(C(F_2(X_2)|F_1(X_1)))]^2 + \dots \\ &+ [\Phi^{-1}(C(F_n(X_n)|F_1(X_1), F_2(X_2), \dots, F_{n-1}(X_{n-1})))]^2. \end{aligned}$$

This hypothesis can be tested using one of the goodness-of-fit tests popularized in the literature, such as the Cramer-von Mises (CvM) test. In almost all the cases, the hypothesis is accepted at the 1% significance level (critical value = 0.743). In the last case ($n = 5$), the test statistic is slightly below the critical value.

## 5.1 An example of portfolio selection

In this subsection we show an example of how the proposed asset selection procedure can be employed in a simple portfolio selection problem. We use MSCI data focusing attention on the financial crisis period occurred in the second semester of 2008. The asset selection procedure described in Section 5 is then carried out on daily data from June 3, 2002 to May 30, 2008 (1564 observations) and the selected $k$-set ($k$=5) is $\{It, Ja, US, Gr, NZ\}$, just the same as that obtained using the full dataset (until June 10, 2010).

Using the popular Markowitz portfolio selection procedure we compute the efficient frontier of this set of indices and compare it with those of 100 randomly selected $k$-sets containing the *main reference asset* MSCI-Italy. It is interesting to note that the portfolios reaching the lowest risk levels are obtained using the selected $k$-set (Fig. 4, left). Thus in this example the asset selection procedure aimed at minimizing the lower tail dependence provides a $k$-set able to minimize the traditional risk measure given by variance, too.

In a financial crisis perspective, we choose the portfolio characterized by the minimum variance and compare its returns in the crisis period from June 2, 2008 to December 31, 2008 to the corresponding returns obtained with the minimum variance portfolios of the 100 above mentioned randomly selected $k$-sets (Fig. 4, right). The selected portfolio outperforms the main part of the competitors during the crisis period.

Figure 4 about here

# 6 Concluding remarks

This paper deals with the problem of selecting a subset of financial assets out of a great set of investment alternatives. The aim driving the selection is to choose assets having either low association in the lower tail, or high association in the upper tail, according to the investment strategy. We mainly inspected the former case, assuming a financial crisis perspective for the empirical analysis.

Copula functions are powerful tools for estimating tail dependence coefficients. Nevertheless, we have found that their use for the above described selection problem is unfeasible due to an excessive computational burden. For this reason we have explored the possibility to build a heuristic procedure making use of algorithmic tools widely used in the field of data mining. The proposed selection procedure has been preliminarily checked with two simulation studies, then applied to real data of MSCI indices, in order to select a subset of developed markets to invest on. In the lower tail, we focused our attention on the *chain effect risk*, that is the probability of a default of the whole set of indices, given the default of one of them.

The proposed procedure is heuristic and merely descriptive, but can be applied to many empirical contexts. With large datasets it can provide a good preliminary inspection of the relationships among variables in the tails of their joint distributions. This makes possible a sort of dimensionality reduction which allows to employ more efficient estimation tools, like copula function.

A number of possible developments in this direction are possible.

Firstly, the procedure can be generalized avoiding the preliminarily definition of a reference asset. Actually, the idea of choosing a reference asset can be justified in many empirical analyses, when there effectively exists an asset we want to necessarily include in the portfolio, or when we want to add new investments to an existing one. Nonetheless, in some cases it can be limiting. The problem can be easily overcome, simply running the procedure as many times as the number $p$ of variables in the dataset, each time choosing a different main reference asset. At the end, $p$ different $k$-sets are selected. Each one can be deeply analysed by means of copula functions and finally the best one can be chosen, according to a defined criterion.

Secondly, in the presented analysis, the *chain effect* is evaluated at the same time $t$. In other words, the coefficient (6) measures the probability of a simultaneous default of all the indices. This is not so limiting, as it is well-known that a default in a developed market rapidly propagates to the others. However the inclusion of lagged variables in the dataset could probably improve the results, taking account of some delay in the default and also of effects due to different time zones.

Thirdly, other types of copula functions can be used to estimate the tail dependence coefficients, then using goodness-of-fit statistics in order to decide which one best reproduces the joint distribution of data.

Finally, some reasoning could be done about the distribution of tail dependence coefficient estimates.

# References

Aas K, Berg D. 2009. Models for construction of multivariate dependence - a comparison study. *The European Journal of Finance* **15**: 639–659.

Bouyé E, Salmon M. 2009. Dynamic copula quantile regressions and tail area dynamic dependence in Forex markets, *The European Journal of Finance* **15**: 721–750.

Breiman L. 1996. Bagging Predictions. *Machine Learning* **24**: 123–140.

Breiman L. 2001. Random Forests. *Machine Learning* **45**: 5–32.

Breiman L. 2002. Manual on setting up, using, and understanding random forests v3.1. *http://oz.berkeley.edu/users/breiman.*

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984 *Classification and Regression Trees*, Chapman & Hall, New York, 1984.

Cherubini U, Luciano E, Vecchiato W. *Copula Methods in Finance*, Wiley, New York, 2004.

De Luca G, Rivieccio G. 2010. Multivariate tail dependence coefficients for Archimedean copulae, submitted.

Dobric J, Schmid F. 2006. A goodness of fit test for copulas based on Rosenblatts transformation, *Computational Statistics and Data Analysis* **51**: 4633-4642.

Engle RF. 2002. Dynamical conditional correlation: a simple class of multivariate generalized autoregressive heteroscedasticity models. *Journal of Business and Economic Statistics* **20**: 339–350.

Fortin I, Kuzmics C. 2002. Tail-dependence in stock-returns pairs, *International Journal of Intelligent Systems in Accounting, Finance & Management* **11**: 89–107.

Friedma JH. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**: 1189–1232.

Friedman JH, Popescu BE. 2005. Predictive Learning via Rule Ensembles, *Technical report*, Stanford University.

Genest G, Gendronb M, Bourdeau-Brienb M. 2009. The Advent of Copulas in Finance, *The European Journal of Finance* **15**: 609-618.

Joe H. 1997. *Multivariate models and dependence concept*, Chapman & Hall, New York.

Jondeau E, Rockinger M. 2006. The Copula-GARCH model of conditional dependencies: an international stock market application. *Journal of International Money and Finance* **25**: 827-853.

McNeil AJ. 1999. Extreme value theory for risk managers. In *Internal Modelling and CAD II*, RISK Books: 93-113.

Nelsen RB. 2006. *An introduction to copulas*, Springer-Verlag, New York.

Palaro HP, Hotta LK. 2006. Using conditional copula to estimate Value at Risk. *Journal of Data Science* **4**: 93-115.

Patton A. 2006. Modelling asymmetric exchange rate dependence, *International Economic Review* **47**: 527-556.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Wien.

Sklar A. 1959.Fonctions de repartition ‘a n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris* **8**: 229-231.

Table 1: Correct selections in Simulation 2 for different values of $k$

| $k$ | Correct selections | | |
|---|---|---|---|
| 3 | $s_3 = (X_1, X_6, X_{11})$ | | |
| 4 | $s_{41} = (s_3, X_2)$ | $s_{42} = (s_3, X_7)$ | $s_{43} = (s_3, X_{12})$ |
| 5 | $s_{51} = (s_{41}, X_7)$ | $s_{52} = (s_{42}, X_{12})$ | $s_{53} = (s_{43}, X_2)$ |
| 6 | $s_6 = (s_3, X_2, X_7, X_{12})$ | | |
| 7 | $s_{71} = (s_6, X_3)$ | $s_{72} = (s_6, X_8)$ | $s_{73} = (s_6, X_{13})$ |
| 8 | $s_{81} = (s_{71}, X_8)$ | $s_{82} = (s_{72}, X_{13})$ | $s_{83} = (s_{73}, X_3)$ |
| 9 | $s_9 = (s_6, X_3, X_8, X_{13})$ | | |
| 10 | $s_{101} = (s_9, X_4)$ | $s_{102} = (s_9, X_9)$ | $s_{103} = (s_9, X_{14})$ |
| 11 | $s_{111} = (s_{101}, X_9)$ | $s_{112} = (s_{102}, X_{14})$ | $s_{113} = (s_{103}, X_4)$ |
| 12 | $s_{12} = (s_9, X_4, X_9, X_{14})$ | | |

Table 2: Estimates of the parameters of the Joe-Clayton copula in the bivariate case (Italy, Japan).

| Parameter | Estimate | St. error |
|---|---|---|
| $\theta$ | 0.1773 | 0.0289 |
| $\kappa$ | 1.0757 | 0.0261 |
| CvM | 0.1126 | |
| $\lambda_L^{It\|Ja}$ | 0.0201 | |

Table 3: Estimates of the parameters of the Joe-Clayton copula in the trivariate case (Italy, Japan, USA).

| Parameter | Estimate | St. error |
|---|---|---|
| $\theta$ | 0.2155 | 0.0191 |
| $\kappa$ | 1.0857 | 0.0192 |
| CvM | 0.4222 | |
| $\lambda_L^{It,Ja\|US}$ | 0.0061 | |

Table 4: Estimates of the parameters of the Joe-Clayton copula in the quadrivariate case (Italy, Japan, USA, Greece).

| Parameter | Estimate | St. error |
|---|---|---|
| $\theta$ | 0.2664 | 0.0159 |
| $\kappa$ | 1.1007 | 0.0168 |
| CvM | 0.7393 | |
| $\lambda_L^{It,Ja,US\vert Gr}$ | 0.0055 | |

Table 5: Estimates of the parameters of the Joe-Clayton copula in the 5-variate case (Italy, Japan, USA, Greece, New Zealand).

| Parameter | Estimate | St. error |
|---|---|---|
| $\theta$ | 0.2486 | 0.0130 |
| $\kappa$ | 1.0912 | 0.0145 |
| CvM | 0.8593 | |
| $\lambda_L^{It,Ja,US,Gr\|NZ}$ | 0.0015 | |

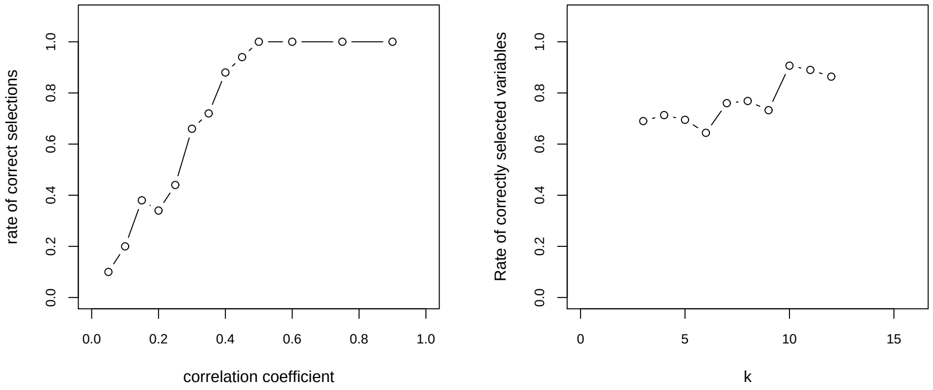

Figure 1: left: Rate of correct $k$-sets vs correlation coefficient (Simulation 1); right: Rate of correctly selected variables vs $k$ (Simulation 2)

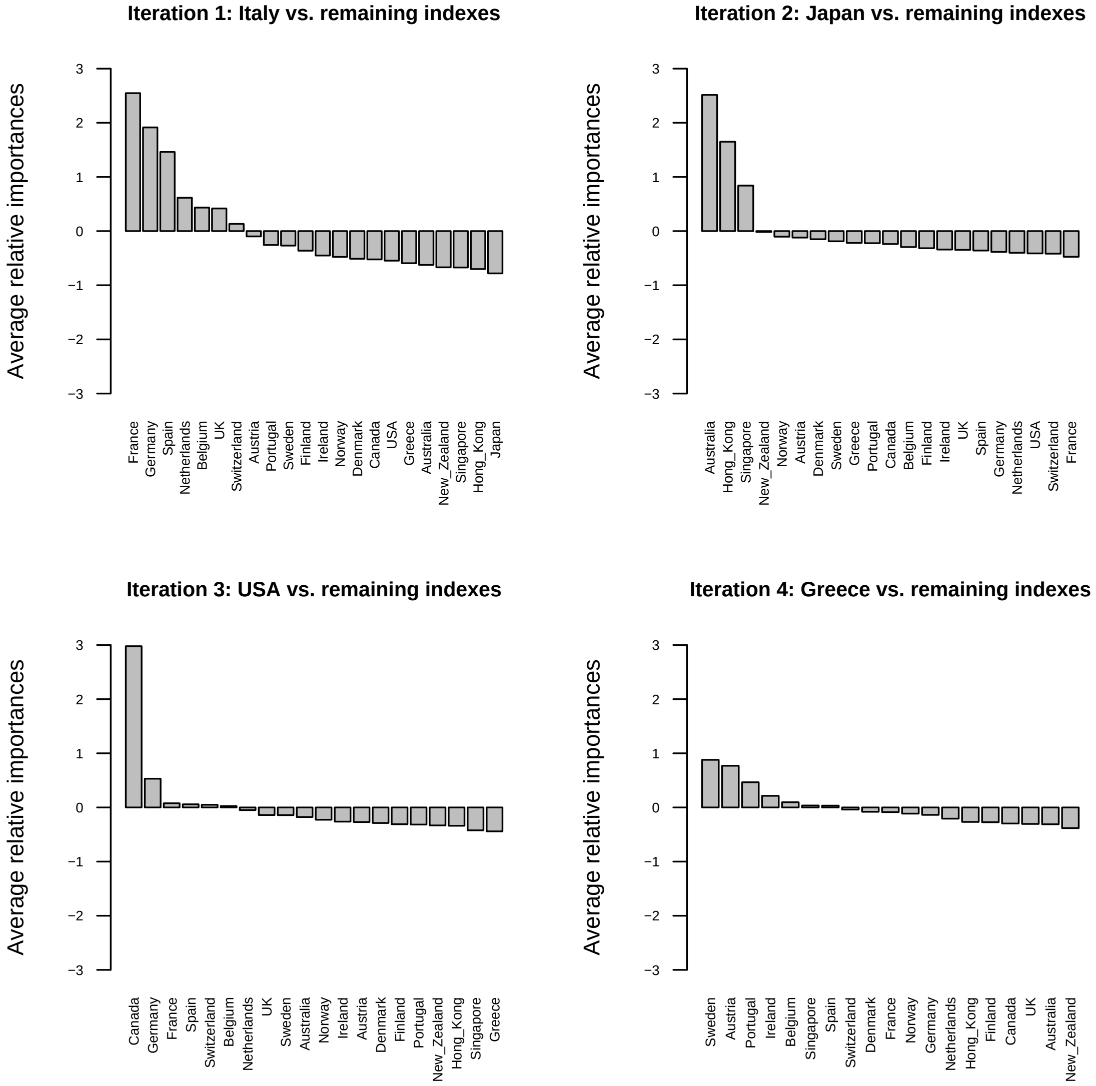

Figure 2: Bar graphs of the standardized VIMs $\bar{I}_{sL}^{(w)}$ for $w = 1, 2, 3, 4$ and $s = 1, 2, \cdots, 23 - w$

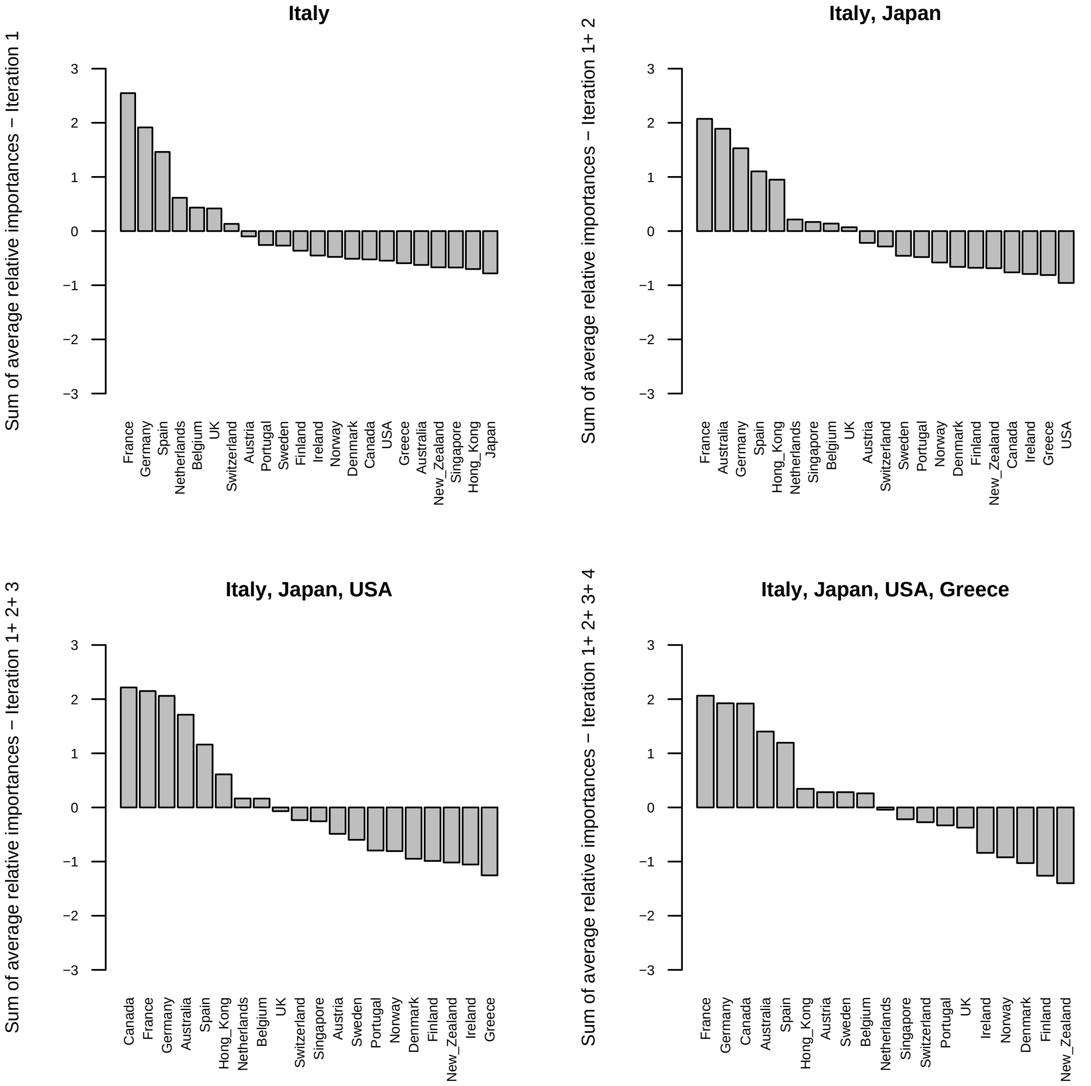

Figure 3: Bar graphs of the summed standardized VIMs $\sum_{i=1}^{w} \bar{I}_{sL}^{(i)}$ for $w = 1, 2, 3, 4$ and $s = 1, 2, \cdots, 23 - w$

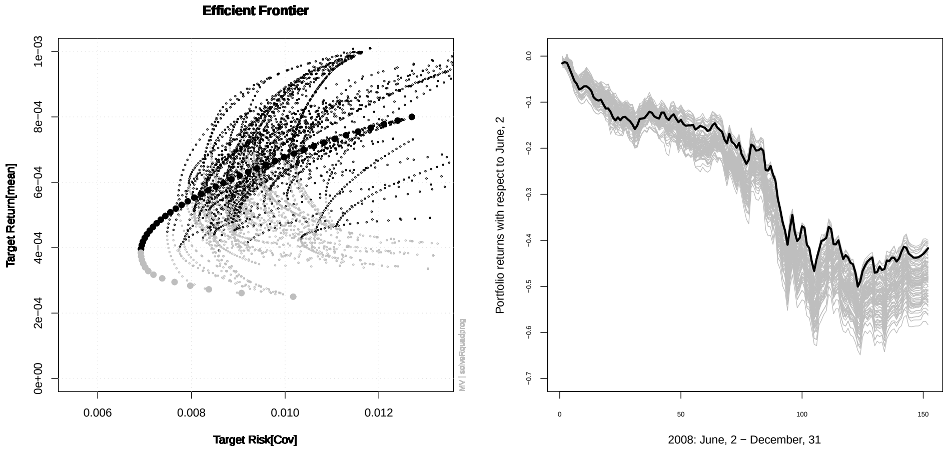

Figure 4: Left: efficient frontier of the selected set of assets together with the efficient frontiers of 100 randomly selected 5-dimensional set of assets including the reference asset. Right: returns of the minimum variance portfolio for the selected assets against returns of the 100 minimum variance competitor portfolios in the second semester of 2008.