Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER Journal of Economic Dynamics & Control 32 (2008) 235–258
JOURNAL OF Economic Dynamics & Control
www.elsevier.com/locate/jedc

Cluster analysis for portfolio optimization

Vincenzo Tola[a,b,1], Fabrizio Lillo[c,d,e], Mauro Gallegati[a], Rosario N. Mantegna[c,d,*]

[a]*Dipartimento di Economia, Università Politecnica delle Marche, Piazza Martelli 8, I-60121 Ancona, Italy*
[b]*Banca d'Italia, Servizio Vigilanza sugli Enti Creditizi, Via Piacenza 6, 00184 Rome, Italy*
[c]*Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, viale delle Scienze, I-90128 Palermo, Italy*
[d]*INFN, Sezione di Catania, Catania, Italy*
[e]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Received 1 May 2005; received in revised form 1 December 2006; accepted 26 January 2007
Available online 2 April 2007

Abstract

We consider the problem of the statistical uncertainty of the correlation matrix in the optimization of a financial portfolio. By assuming idealized conditions of perfect forecast ability for the future return and volatility of stocks and short selling, we show that the use of clustering algorithms can improve the reliability of the portfolio in terms of the ratio between predicted and realized risk. Bootstrap analysis indicates that this improvement is obtained in a wide range of the parameters N (number of assets) and T (investment horizon). The predicted and realized risk level and the relative portfolio composition of the selected portfolio for a given value of the portfolio return are also investigated for each considered filtering method. We also show that several of the

*Corresponding author. Dipartimento di Fisica e Tecnologie Relative, Universita degli studi di Palermo, Viale delle Scienze, 90128 Palermo, Italy. Tel.: +39 91 661 5074; fax: +39 91 661 5063.
E-mail address: mantegna@unipa.it (R.N. Mantegna).

[1]Opinions expressed in this paper are exclusively by the authors and do not necessarily reflect those of Banca d'Italia.

doi:10.1016/j.jedc.2007.01.034

results obtained by assuming idealized conditions are still observed under the more realistic assumptions of no short selling and mean return and volatility forecasting based on historical data.

JEL classification: G11; C30; C15

Keywords: Portfolio optimization; Clustering methods; Correlation matrices; Random matrix theory

1. Introduction

The problem of portfolio optimization is one of the most important issues in asset management (Elton and Gruber, 1995). Since the seminal work of Markowitz (1959), which solved the problem under a certain number of simplifying assumptions (see also Section 2), many other studies have been devoted to consider several aspects of portfolio optimization both from a theoretical and from an applied point of view. A huge number of studies considering key aspects of portfolio optimization theory are present in the finance literature. Here we refer to a few studies considering, for example, the real performance of portfolios constructed using sample moments (Jorion, 1985), the realized Sharpe ratio of the global minimum variance portfolio (Jagannathan and Ma, 2003), the role of constraints in portfolio optimization (Eichhorn et al., 1998; Jagannathan and Ma, 2003), the shrinkage estimator of large dimensional covariance matrices (Ledoit and Wolf, 2004a,b). The aim of the present study is to focus on the role of the correlation coefficient matrix in portfolio optimization. The estimation of the correlation matrix is unavoidably associated with a statistical uncertainty, which is due to the finite length of the asset return time series. Recently, there have been several contributions in the econophysics literature devoted to quantify the degree of statistical uncertainty present in a correlation matrix. The results of these investigations have been obtained by using concepts and tools of random matrix theory (RMT) (Metha, 1990). The RMT quantification of the statistical uncertainty associated with the estimation of the correlation coefficient matrix of a finite multivariate time series has been recently used to devise a procedure to filter the information present in the correlation coefficient matrix which is robust with respect to the unavoidable statistical uncertainty (in the econophysics literature the term of noise dressing has been used) (Galluccio et al., 1998; Laloux et al., 1999, 2000; Plerou et al., 1999, 2002; Gopikrishnan et al., 2001; Drozdz et al., 2001; Rosenow et al., 2002; Pafka and Kondor, 2003, 2004; Rosenow et al., 2003; Guhr and Kalber, 2003; Malevergne and Sornette, 2004; Sharifi et al., 2004; Burda and Jurkiewicz, 2004). The correlation matrices obtained by this filtering procedure has been used in portfolio optimization in some studies (Laloux et al., 2000; Rosenow et al., 2002), which have shown that under the assumption of perfect forecasting of future returns and volatilities the distance

between the predicted optimal portfolio and the realized one is smaller for the filtered correlation matrix than for the original one at a given level of the portfolio return.

In recent years, other filtering procedures of the correlation coefficient matrix performed using correlation based clustering procedures have also been proposed in the econophysics literature (Mantegna, 1999; Kullmann et al., 2000, 2002; Bonanno et al., 2000, 2001, 2003, 2004; Giada and Marsili, 2001; Maslov, 2001; Bernaschi et al., 2002; Onnela et al., 2002; Mendes et al., 2003; Micciche et al., 2003; Maskawa, 2003; Di Matteo et al., 2004; Basalto et al., 2005; Tumminello et al., 2005). These methods also select information of the correlation coefficient matrix which is representative of the entire matrix and it is often less affected by the statistical uncertainty and therefore more stable than the entire matrix during the time evolution of the system.

In this paper we investigate how the portfolio optimization procedure is sensitive to different filtering procedures applied to the correlation coefficient matrix. Specifically, we consider filtering procedures based on RMT and on correlation based clustering procedures. We proceed in two steps. In the first step, similarly as in Laloux et al. (2000) and Rosenow et al. (2002), we assume perfect forecasting ability of future returns and volatilities. We also assume that short selling is allowed in the portfolio optimization procedure. These are quite idealized conditions. In the second step, we assume more realistic conditions. Specifically, we consider no short selling constraint and we assume imperfect forecasting of mean returns and volatilities. The forecast is obtained just by using estimation based on historical data. We limit the number of potential control parameters by investigating the global minimum variance portfolio. This is not a severe limitation because it has been shown that under weight constraints the ex-post Sharpe ratio of the global minimum variance portfolio is often no smaller than other efficient portfolios (Jagannathan and Ma, 2003). We verify that several of the results obtained under idealized conditions are still observed for the global minimum variance portfolio under much more realistic conditions.

The paper is organized as follows. In Section 2 we describe briefly the mean variance optimization problem, we define the notation and we summarize the problem of the estimation of the correlation matrix. In Section 3 we review the approach recently introduced (Laloux et al., 2000; Rosenow et al., 2002) which makes use of the RMT to improve the portfolio optimization in the presence of estimation errors due to the finiteness of sample data. In Section 4 we describe the clustering algorithms used to perform the portfolio optimization. These algorithms are the average linkage and the single linkage. In Section 5 we describe the portfolio optimization procedure performed with clustering algorithms and we compare the obtained results with the one of the RMT under the idealized assumptions of perfect forecasting and short selling allowed. The more realistic conditions of forecasting based on historical data and no short selling for the global minimum variance portfolios are investigated in Section 6. Finally in Section 7 we summarize our results and indicate future work extending and possibly improving our method.

2. Portfolio optimization

2.1. Markowitz's solution

In this section we briefly discuss the basic aspects of Markowitz portfolio optimization. This is also useful to set the notation and to state the assumptions made and the methods used. Given N risky assets the portfolio composition is determined by the weights w_i ($i = 1, \ldots, N$) giving the fraction of wealth invested in asset i. The weights are normalized as $\sum_{i=1}^{N} w_i = 1$. The average return and the variance of the portfolio are

$$r_p \equiv \sum_{i=1}^{N} w_i m_i, \tag{1}$$

$$\sigma_p^2 \equiv \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij}, \tag{2}$$

where m_i is the mean return of asset i and σ_{ij} is the covariance between returns of asset i and j. The optimization problem consists in finding the vector $\mathbf{w}$ which minimizes σ_p for a given value of r_p. In Section 5, we assume that short selling is allowed, i.e. w_i can assume negative values whereas in Section 6 we relax this assumption. As known the solution of this optimization problem has been found by Markowitz (1959) and it is

$$\mathbf{w}^* = \lambda \mathbf{\Sigma}^{-1}\mathbf{1} + \gamma \mathbf{\Sigma}^{-1}\mathbf{m}, \tag{3}$$

where $\mathbf{\Sigma}$ is the covariance matrix, $\mathbf{1}^T = (1, \ldots, 1)$ and $\mathbf{m}$ is the vector of the mean returns of the N assets. The other parameters are

$$\lambda = \frac{C - r_p B}{\Delta}, \quad \gamma = \frac{r_p A - B}{\Delta},$$
$$A = \mathbf{1}^T \mathbf{\Sigma}^{-1}\mathbf{1}, \quad B = \mathbf{1}^T \mathbf{\Sigma}^{-1}\mathbf{m},$$
$$C = \mathbf{m}^T \mathbf{\Sigma}^{-1}\mathbf{m}, \quad \Delta = AC - B^2.$$

When $\gamma = 0$, the optimal portfolio is the global minimum variance portfolio. We will focus on this portfolio for the investigations of portfolio optimization performed under realistic condition described in Section 6.

2.2. Curse of dimensionality and adopted method

Markowitz's solution to the optimization problem relies upon a series of assumptions that are rarely observed in practice. First of all the asset returns are assumed to be Gaussian variables whereas fat tails in price return distribution are observed. Second the parameters used in the optimization, i.e. the mean values $\mathbf{m}$ and the covariance matrix $\mathbf{\Sigma}$, are assumed constant. Finally even if these quantities are really constant in the time horizon relevant for the problem, their statistical estimation over finite time intervals T leads to the problem known as *curse of*

dimensionality. Since the covariance matrix has $N(N-1)/2 \sim N^2/2$ distinct entries whereas the number of records used in the estimation is NT, one needs time series of length $T \gg N$ in order to have small error on the covariance. But for long T non-stationarity becomes more and more important. For these reasons it is important to develop methods able to filter the part of the covariance matrix which is less likely to be affected by statistical uncertainty, and use (when possible) the filtered information to build portfolios.

In this paper we are mainly concerned with the problems in the portfolio optimization due to the estimation of the correlation matrix, i.e the matrix whose entries are the correlation coefficients between returns of different assets. The correlation coefficient is defined as $\rho_{ij} \equiv \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$. In the first part of the paper, we will use the following procedure (Laloux et al., 2000; Rosenow et al., 2002) to assess the effectiveness of the filtering procedure of the correlation coefficient matrix based on RMT. Given N assets, a portfolio horizon of T trading days and a time t_0 when the optimization is supposed to take place, we compute the correlation matrix in the T days preceding t_0 but we compute the mean returns m_i and the volatilities $\sigma_i \equiv \sqrt{\sigma_{ii}}$ in the T days following t_0. We use this data to compute the covariance matrix and the predicted optimal portfolio at time t_0. We then compare the predicted risk-return curve with the realized risk-return curve obtained by computing

$$\hat{\sigma}_{\mathrm{p}}^2 \equiv \sum_{i=1}^{N}\sum_{j=1}^{N} w_i^* w_j^* \hat{\sigma}_{ij}, \tag{4}$$

where w_i^* are the weights obtained in the optimization and $\hat{\sigma}_{ij}$ is the covariance matrix observed between t_0 and $t_0 + T$. By using this procedure we are able to decouple the problem of estimating cross-correlations from the problem of estimating mean returns and volatilities. In other words, in the first part of the paper, we will assume that the investor has a perfect forecast of m_i and σ_i and all her uncertainty is in the estimation of the cross-correlation matrix. These assumptions will be relaxed in Section 6 where we consider the optimization procedure under more realistic conditions.

In order to quantify and compare the goodness of different filtering methods we make use of three measures. The first quantity measuring the reliability of the portfolio is obtained by comparing, for a given value of expected return, the risk σ_{p} predicted by using the past correlation matrix with the realized risk $\hat{\sigma}_{\mathrm{p}}$ of Eq. (4). A portfolio is more reliable when

$$\mathscr{R} \equiv \frac{|\hat{\sigma}_{\mathrm{p}} - \sigma_{\mathrm{p}}|}{\sigma_{\mathrm{p}}} \tag{5}$$

is small. We have also used different measures of the reliability, such as $|\hat{\sigma}_{\mathrm{p}} - \sigma_{\mathrm{p}}|$, obtaining similar results.

The second quantity used to compare different methods is simply the realized risk $\hat{\sigma}_{\mathrm{p}}$. Clearly a portfolio is less risky than another one when its realized risk is smaller. Note that in general a portfolio with a small *predicted* risk is not necessarily better than a (*predicted*) more risky portfolio. In fact if the uncertainty on the *realized* risk

of the safe portfolio is very large, an investor could face large fluctuations and therefore an ex-post larger loss for the ex-ante less risky portfolio than for the one ex-ante considered more risky.

The third characteristic for evaluating portfolio optimization methods is the degree of reduction in the effective dimension of the portfolio. Dealing with a large portfolio can be very costly because of the transaction costs that the investor has to face any time she wants to rebalance the weights. Even if we do not consider here the problem of portfolio rebalancing and benchmarking, we wish to quantify the 'effective' number of stocks with a significant amount of money invested in. By following Bouchaud and Potters (2003), we quantify this number as

$$\mathcal{N}^{(\mathrm{eff})} = \frac{1}{\sum_{i=1}^{N} w_i^2}. \tag{6}$$

This quantity is equal to 1 when all the wealth is invested in only one asset, whereas it is equal to N when the wealth is divided equally among the N assets, i.e. $w_i = 1/N$. It may be worth noting that the quantity $\mathcal{N}^{(\mathrm{eff})}$ does not give the number of assets where a non-vanishing amount of wealth is invested in. It simply gives a rough estimate of the number of assets that could effectively be used to build a smaller portfolio with risk-return properties not too far from the original N asset portfolio.

In the next sections, the different filtering procedures considered in this paper are investigated by using the set of data of 1071 stocks continuously traded at New York Stock Exchange (NYSE) during the period 1988–1998. This overall period is divided into shorter time intervals to test both the role of a different number of time records and the behavior of different filtering methods in distinct market phases characterized by different levels of market volatility. In several cases we consider a two-year time interval. In all this study, we consider daily returns.

3. Random matrix theory approach

Recently (Laloux et al., 1999; Plerou et al., 1999) it has been shown that the RMT can be useful to investigate the properties of return correlation matrices of financial assets. The simplest random matrix is a matrix of given type and size whose entries consist of random numbers from some specified distribution (Metha, 1990). RMT was developed originally in nuclear physics and then applied to many different fields. In the context of asset portfolios RMT is useful because it allows one to compute the effect of statistical uncertainty in the estimation of the correlation matrix. Suppose that the N assets are described by N time series of length T and that the returns are independent Gaussian random variables with zero mean and variance σ^2. The correlation matrix of this set of variables in the limit $T \to \infty$ is simply the identity matrix. When T is finite the correlation matrix will in general be different from the identity matrix. RMT proves that in the limit $T, N \to \infty$, with a fixed ratio $Q = T/N \geqslant 1$, the eigenvalue spectral density of the

covariance matrix is given by

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2\lambda}\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}, \tag{7}$$

where $\lambda_{\min}^{\max} = \sigma^2(1 + 1/Q \pm 2\sqrt{1/Q})$. The spectral density is different from zero in the interval $]\lambda_{\min}, \lambda_{\max}[$. When one considers a correlation coefficient matrix σ^2 can be set equal to one. The spectrum described by Eq. (7) is different from $\delta(\lambda - 1)$ which is expected from an identity correlation matrix. In other words RMT quantifies the role of the finiteness of the length of the time series on the spectral properties of the correlation matrix.

RMT has been applied to correlation matrices of returns of financial assets (Laloux et al., 1999; Plerou et al., 1999) and it has been shown that the spectrum of a typical portfolio can be divided into three classes of eigenvalues. The largest eigenvalue is totally incompatible with Eq. (7) and describes the common behavior of the stocks composing the portfolio. A fraction of the order of 5% of the eigenvalues are also incompatible with the RMT because they fall outside the interval $]\lambda_{\min}, \lambda_{\max}[$. These eigenvalues probably describe economic information stored in the correlation matrix. The remaining large part of the eigenvalues is between $\lambda_{\min}$ and $\lambda_{\max}$ and thus one cannot say whether any information is contained in the corresponding eigenspace.

The fact that by using RMT it is possible, under certain assumptions, to identify the noisy part of the correlation matrix suggested to several authors (Laloux et al., 2000; Rosenow et al., 2002) to use RMT in the optimization of financial portfolios. Specifically the suggested method (Rosenow et al., 2002) is the following.

One computes the correlation matrix and finds the spectrum ranking the eigenvalues such that $\lambda_k < \lambda_{k+1}$. In spite of the fact that we are considering the spectrum of the correlation coefficient matrix, it is worth noting that σ^2 needs to be redefined in the estimation of $\lambda_{\min}$ and $\lambda_{\max}$ because one needs a procedure to separate the overall behavior of the market described by the largest eigenvalue from the bulk of eigenvalues affected by statistical uncertainty. This is done by computing the variance of the part not explained by the highest eigenvalue as $\sigma^2 = 1 - \lambda_1/N$ and by using this value in Eq. (7) to obtain $\lambda_{\min}$ and $\lambda_{\max}$. One then constructs a filtered diagonal matrix obtained by setting to zero all the eigenvalues smaller than $\lambda_{\max}$ and leaving unaltered the remaining ones. Finally, one obtains the filtered correlation matrix by transforming the filtered diagonal matrix into the original basis. In order to obtain a meaningful correlation matrix we set to one the diagonal elements of the filtered correlation matrix. This matrix preserves only the information of the original correlation matrix that the RMT recognizes as signal. In Laloux et al. (2000) and Rosenow et al. (2002), it has been shown that the portfolio obtained by using the filtered correlation matrix has a smaller value of $\mathscr{R}$ than a portfolio with weights obtained with Markowitz's procedure and by using the whole correlation matrix. As an example we show in Fig. 1 the predicted and realized risk for a portfolio of 150 highly capitalized stocks traded at NYSE in a period of $T = 500$ trading days (approximately corresponding to two calendar years). In this and in the other figures the risk and return are annualized. In the shown example, we

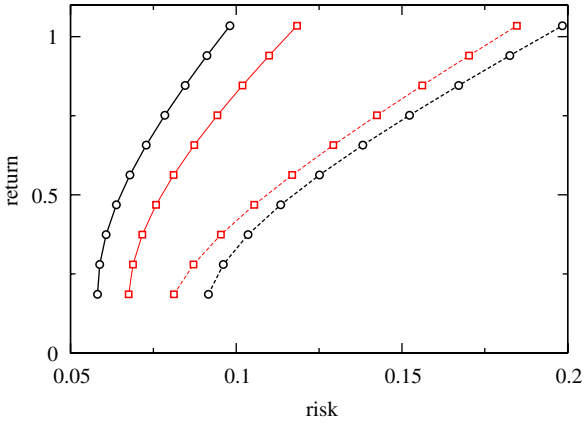

Fig. 1. The continuous lines are the predicted risk and the dashed lines are the realized risk. The circles refer to the Markowitz portfolio optimization, whereas the squares are the predicted and realized risk curves obtained by filtering the correlation matrix with the random matrix theory approach. We assume that the only uncertainty of the investor is on the correlation matrix. The dataset is composed by 150 highly capitalized stocks traded at NYSE in the period 1989–1992. The first two years are used for the estimation of the correlation matrix and the other two years are the investment period.

have estimated the correlation matrix in the two-year period 1989–1990 and the realized risk is computed in the two-year period 1991–1992. The figure shows the risk-return curve for the Markowitz portfolio and for a portfolio obtained by filtering the correlation matrix with the RMT method outlined above. Before to continue, for the sake of completeness and contextualization we wish to provide some summary statistics of the data used in the optimization procedure shown in Fig. 1 and also in the followingFigs. 2 and 5. Specifically, the ex-ante and ex-post mean stock return, volatility and correlation coefficient of the portfolio in the considered time period and also the weights of the minimum variance portfolio (MVP) for the Markowitz and RMT optimization. In Table 1 we also report the weights of the average linkage (AVG) and single linkage (SIN) optimization that are discussed below and shown in Figs. 2 and 5. It is also worth noting that equally weighted and value-weighted portfolios are characterized by a predicted and realized annual portfolio return equals to 0.256 and 0.232, respectively. The annual predicted risk is equal to 0.147 and 0.155, respectively, whereas the annual realized risk is equal to 0.126 and 0.131, respectively. We note that for all the values of the expected return r_p the parameter $\mathscr{R}$ for the RMT portfolio is significantly smaller than for the Markowitz portfolio. For this portfolio the realized risk of the RMT portfolio is smaller than the realized risk of the Markowitz portfolio. This is in agreement with what, for example, Rosenow et al. (2002) find for another set of data. We wish to point out that the behavior of Fig. 1 is rather common in different portfolios composed by a varying number of stocks selected from our database of 1071 continuously traded stocks when the portfolios have been evaluated in several distinct time periods. However, we have also found that for some portfolios and time

Table 1
Summary statistics of the data used in the optimization procedure summarized in Figs. 1, 2 and 5

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Std.
Mean return ex-ante	−0.524	0.0267	0.141	0.131	0.255	0.573	0.198
Mean return ex-post	−0.292	0.133	0.239	0.257	0.359	1.03	0.203
Volatility ex-ante	0.141	0.225	0.252	0.273	0.312	0.677	0.0763
Volatility ex-post	0.129	0.221	0.257	0.270	0.295	0.626	0.0815
Correlation ex-ante	−0.232	0.229	0.309	0.309	0.381	0.747	0.124
Correlation ex-post	−0.114	0.162	0.215	0.222	0.269	0.724	0.106
MVP-Markowitz weights	−0.0739	−0.0132	−0.00122	0.00666	0.0240	0.130	0.0332
MVP-AVG weights	−0.0335	−0.00935	0.00216	0.00666	0.0171	0.135	0.0255
MVP-SIN weights	−0.057	−0.0113	−0.00188	0.00666	0.0158	0.181	0.0309
MVP-RMT weights	−0.0685	−0.00923	0.002684	0.00666	0.0177	0.146	0.0272

Mean returns and volatility are annualized. AVG and SIN indicate average and single linkage optimization, respectively (see Section 5). RMT indicates random matrix theory optimization. The summary statistics of portfolio weights refers to the minimum variance portfolio (MVP).

periods the realized risk profile obtained with the RMT filtering is larger than the one obtained with the Markowitz approach.

4. Clustering algorithms

In this paper, we introduce a new portfolio optimization technique which is based on clustering algorithms. Clustering is a common practice in multivariate data analysis (Mardia et al., 1979). The purpose of clustering analysis is to obtain a meaningful partition of a set of N variables in groups according to their characteristics. For example in correlation based clustering algorithms (adopted here) the correlation coefficient between two time series is assumed to be a measure of the similarity between the two time series. Correlation based clustering has been recently used to infer the hierarchical structure of a portfolio of stocks from its correlation coefficient matrix (Mantegna, 1999; Bonanno et al., 2001, 2003). Correlation based clustering may be seen as a filtering procedure, i.e. a matrix transformation retaining a smaller number of distinct elements. After its application one usually retains a subset of the distinct elements composing the correlation coefficient matrix. For example, in the clustering algorithm of the single linkage (Gower and Ross, 1969) the number of distinct elements present in the filtered matrix is $n-1$ whereas the number of distinct elements present in the original matrix is $n(n-1)/2$. The selection of these $n-1$ elements is done according to some widespread algorithm (Papadimitriou and Steiglitz, 1982). A possible conceptual description of the algorithm is the following. Let us assume that a similarity measure S of pairs of elements is defined, e.g. the correlation coefficient between pairs of elements of the system. An ordered list S_{ord} of pairs of elements can be constructed

by arranging them in a descending order accordingly with the value of the similarity s_{ij} between element i and element j. Different elements are iteratively included in clusters starting from the first two elements of the similarity measure ordered list. At each step, when two elements or one element and a cluster or two clusters p and q merge in a wider single cluster t, the similarity or distance between the new cluster t and cluster r is determined as follows: if s_{ij} is a correlation-like measure (for example, $s_{ij} = \rho_{ij}$)

$$s_{tr} = \max\{s_{pr}, s_{qr}\}, \tag{8}$$

indicating that the similarity between any element of cluster t and any element of cluster r is the similarity between the two most similar entities in clusters t and r. Conversely, if s_{ij} is a distance-like measure (for example, $s_{ij} = d_{ij} = \sqrt{2(1-\rho_{ij})}$ Gower, 1966),

$$s_{tr} = \min\{s_{pr}, s_{qr}\}. \tag{9}$$

By applying iteratively this procedure $n-1$ of the $n(n-1)/2$ distinct elements of the correlation coefficient matrix are selected. When a distance-like measure d_{ij} is used, the distance matrix obtained by applying the single linkage procedure is an ultrametric matrix comprising the $n-1$ distinct selected elements. Ultrametric distances $d_{ij}^{<}$ are distances satisfying an inequality $d_{ac}^{<} \leqslant \max\{d_{ab}^{<}, d_{bc}^{<}\}$ stronger than the customary triangular inequality $d_{ac} \leqslant d_{ab} + d_{bc}$ (Rammal et al., 1986). In particular, the single linkage clustering procedure is associated with an ultrametric correlation coefficient matrix which is the subdominant ultrametric matrix of the original correlation coefficient matrix. For a didactic description of the method used to obtain the ultrametric matrix one can consult Mantegna and Stanley (2000).

In Tumminello et al. (2007) it is proved that the ultrametric correlation matrix obtained by the single linkage clustering procedure of the correlation coefficient matrix is always positive definite when all the elements of the obtained ultrametric correlation matrix are positive. This condition is rather common in financial data of stock portfolio and it has always been observed for all the investigations we have performed so far. The effectiveness of the single linkage clustering procedure in pointing out the hierarchical structure of the investigated portfolio has been shown by several studies (Mantegna, 1999; Bonanno et al., 2000, 2001, 2003, 2004; Kullmann et al., 2002; Onnela et al., 2002; Micciche et al., 2003; Di Matteo et al., 2004). However, the single linkage is just one correlation based filtering method. Other methods have also been applied to financial portfolios (Kullmann et al., 2000, 2002; Giada and Marsili, 2001; Maslov, 2001; Bernaschi et al., 2002; Onnela et al., 2002; Mendes et al., 2003; Maskawa, 2003; Basalto et al., 2005; Tumminello et al., 2005). Each method puts a specific emphasis on some aspects of the original matrix and is usually able to point out a series of aspects that might not be elucidated by a different filtering procedure. The choice of the filtering method must therefore be guided by the specific goals that one pursues. In the present study, we decide to consider the average linkage procedure in addition to the single linkage procedure. The average linkage is another widespread clustering algorithm (Anderberg, 1973).

The difference to the single linkage algorithm is that the similarity measure between an element and the closest cluster is given by the mean similarity measure between the considered element and each element of the closest cluster. In other words, if s_{ij} is a similarity-like measure, at each stage one obtains s_{tr} between clusters t and r defined as above, as the average distance $s_{tr} = \text{avg}\{s_{pr}, s_{qr}\}$ between all pairs of links of the elements belonging to the two clusters. For a detailed discussion of the average linkage cluster algorithm see, for example Anderberg (1973). Also in the case of the average linkage the filtered correlation coefficient matrix is an ultrametric distance. In Tumminello et al. (2007) it is proved that the ultrametric correlation coefficient matrix associated with a given average linkage clustering procedure is positive definite under the same general conditions valid for the case of the single linkage. However, this property is not generic to all clustering procedures. We have verified that it does not apply for the cases of the complete linkage and for the Ward clustering method.

5. Portfolio optimization with clustering algorithms

The portfolio optimization method we propose here is based on the use of the ultrametric matrix associated with a given clustering method as a meaningful and robust filtration of the original correlation matrix. In other words we construct the portfolio by solving the Markowitz optimization problem by using the ultrametric matrix rather than the original correlation matrix or the RMT filtered matrix. The reasons for this choice are: (i) it is known that clustering algorithms are able to filter the relevant information in a multivariate set of data (ii) other studies indicate that clustering algorithms are quite robust with respect to measurement noise due to the finiteness of sample size. This is particularly true for a set of variables hierarchically organized (Tumminello et al., 2007).

From the filtered (ultrametric) correlation matrix we build the portfolio by using the Markowitz result (Eq. (3)) and for each value of the portfolio return r_{p} we find the predicted risk σ_{p}. We note that in order to consider the ultrametric matrix as a meaningful correlation matrix it is important that the matrix is positive definite (or semidefinite). To assess the generality of our results obtained under idealized conditions, we have performed a very large number of portfolio optimizations using real data and we have not found a single case in which the ultrametric matrix is not positive definite. We used the weights obtained from the optimization to compute the realized risk by using Eq. (4) where $\hat{\sigma}_{ij}$ is the *original* covariance matrix. In other words we use the filtered matrix only for obtaining the weights w_i, whereas the realized risk is clearly determined by the whole correlation matrix.

5.1. Average linkage

We consider first portfolios built by using correlation matrices filtered with the average linkage cluster algorithm.

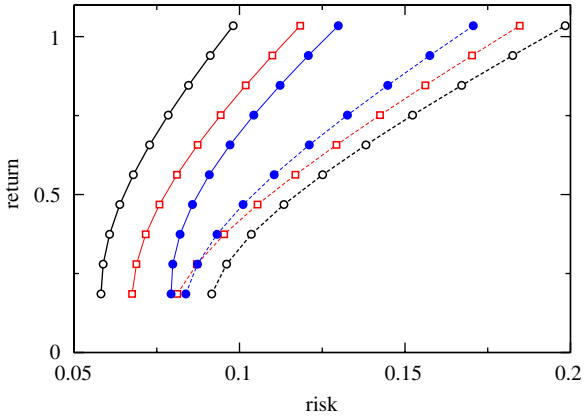

Fig. 2. The continuous lines are the predicted risk and the dashed lines are the realized risk. The filled blue circles refer to the portfolio obtained with the average linkage method. The empty circles refer to the Markowitz portfolio optimization, whereas the squares are the predicted and realized risk curves obtained by filtering the correlation matrix with the random matrix theory approach (same data as in Fig. 1). We assume that the only uncertainty of the investor is on the correlation matrix. The dataset is composed of 150 highly capitalized stocks traded at NYSE in the period 1989–1992. The first two years are used to estimate the correlation matrix and the other two years are the investment period.

5.1.1. *Reliability*

Fig. 2 shows the predicted and realized risk for the portfolio obtained with the average linkage considering the same set of stocks and the same time period as in Fig. 1. The distance between predicted and realized risk for the portfolio obtained with average linkage is significantly smaller than the distance for the portfolio obtained with the RMT. This result indicates clearly that, under idealized conditions, the use of clustering methods to build financial portfolios is able to provide more reliable portfolios (in terms of the error in the forecasted risk) than the ones obtained with RMT and with Markowitz optimization. We also note that for this set of data the realized risk of the portfolio obtained with the clustering method is almost always smaller than the realized risk of the RMT portfolio.

In order to verify the robustness of these results we have performed an extensive bootstrap experiment. We have considered many different values of the portfolio size N and of the investment horizon T and for each couple (N, T) we have randomly sampled 50 portfolios each composed of N stocks randomly selected from the 1071 stocks available in our database and we have randomly selected 50 initial times t_0. In our calculations, we have varied N from 10 to 100 in steps of 10 and from 100 to 500 in steps of 50. The parameter T has been varied from $N + 50$ to $N + 900$ in steps of 100. For each portfolio we have considered 10 values of the expected portfolio return r_p. Specifically we have taken 10 equispaced values of r_p between the value of r_p associated with the global minimum variance portfolio and the highest expected return among the N stocks of the portfolio. For each expected return and for each portfolio we have computed the parameter $\mathscr{R}$ and we have counted the fraction of

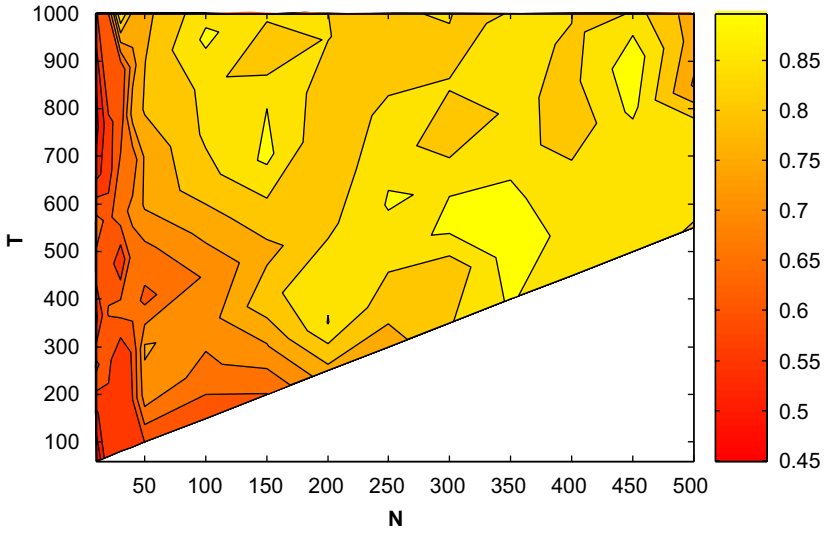

Fig. 3. Summary of a bootstrap experiment measuring the reliability of the average linkage portfolio optimization with respect to the RMT optimization method. The contour plot is showing the percentage of success of the average linkage portfolio optimization technique over the random matrix theory approach as a function of the number of assets N and of the investment period T. For each value of N and T, we perform $2 \times 10 \times 50$ different portfolio optimization procedures by randomly selecting N stocks from our database and by varying T from 100 to 1000. The white area corresponds to cases where $T < N + 50$, which are not considered in our investigation.

times that $\mathscr{R}_{\text{av.link}} < \mathscr{R}_{\text{RMT}}$, i.e. the percentage of cases p_s in which the portfolio obtained with average linkage is more reliable than the portfolio obtained with random matrix theory. The result of this analysis is shown in Fig. 3. The figure indicates that the average linkage portfolio outperforms the RMT portfolio almost for any value of N and T. For $N \simeq 350$ and $T \simeq 500$ the average linkage portfolio is 85% more reliable than the RMT portfolio. The standard error σ_e of the mean percentage of success p_s for 50 realizations is given by the relation $\sigma_e = \sqrt{p_s(1-p_s)/500}$. This implies that the values of p_s are characterized by an uncertainty ranging from $\sigma_e = 0.022$ when $p_s = 0.45$ to $\sigma_e = 0.013$ when $p_s = 0.90$ in Fig. 3. The reliability of the average linkage portfolio is higher when the number of stocks is large. For small size ($N < 50$) portfolios the two methods are statistically equally reliable.

5.1.2. *Riskiness*

We now compare the realized risk of the portfolios obtained with the two methods, i.e. RMT and average linkage. The realized risk is a measure of the riskiness of the portfolio. We observe that small size portfolios tend to be less risky when obtained with the average linkage, whereas as N increases the RMT portfolios

become less risky. The boundary between the two regions is approximately for $N \sim 75$. By comparing this result with Fig. 3 we see that when the average linkage is more reliable it is also riskier and vice versa. There is a small region around $N \simeq 50$ where it is possible to build portfolios with average linkage which are reliable and not too risky.

It is important to stress that, as for Fig. 3, the above result on riskiness is obtained by putting together all the values of the portfolio expected return r_p. On the other hand we find that the riskiness of the average linkage portfolio compared to RMT portfolio strongly depends on r_p especially for large portfolios. Specifically when we consider large portfolios ($50 < N < 500$) we find that for small r_p only in $\sim 25\%$ of the cases the average linkage portfolio is less risky than the RMT portfolio. When r_p is large this fraction is of the order of $\sim 45\%$. In other words for portfolios with large r_p average linkage portfolios are approximately as risky as the RMT portfolios.

5.1.3. *Effective size*

Finally, we consider the effective size $\mathcal{N}^{(\mathrm{eff})}$ of the portfolio as quantified by Eq. (6). We consider three portfolio sizes, i.e. $N = 50, 300$, and 500, and we select two values of the portfolio expected return r_p, i.e. the minimum value (corresponding to the global minimum variance portfolio) and an intermediate value between the minimum and the maximum. Fig. 4 shows $\mathcal{N}^{(\mathrm{eff})}$ as a function of the investment horizon T. Similar results are observed for high values of expected return, but in this case the dimensionality of the portfolio becomes smaller and smaller and the wealth is more and more concentrated in the asset with highest return. We note that for small portfolios ($N = 50$) the effective size of RMT portfolios is slightly smaller than the effective size of average linkage portfolios. On the other hand for larger portfolios the effective size of average linkage portfolio is significantly smaller than the effective size of RMT portfolio and the average linkage portfolio is, in general, more risky. This result shows that portfolios built with average linkage have a smaller effective dimensionality, i.e. the maintenance cost of these portfolios is smaller than for the one of RMT portfolios.

5.2. *Single linkage*

We performed the same analysis by using a different clustering algorithm, specifically the single linkage cluster analysis.

5.2.1. *Reliability*

By using the same data as in Figs. 1 and 2 we compute the curves for predicted and realized risk for a portfolio built by using the ultrametric matrix associated with the single linkage algorithm. The result is in Fig. 5. Also in this case the predicted and realized risk for single linkage portfolios are significantly closer than the corresponding quantities for Markowitz and for RMT portfolios. In this case it is more evident that the realized risk of the single linkage portfolio is larger than the

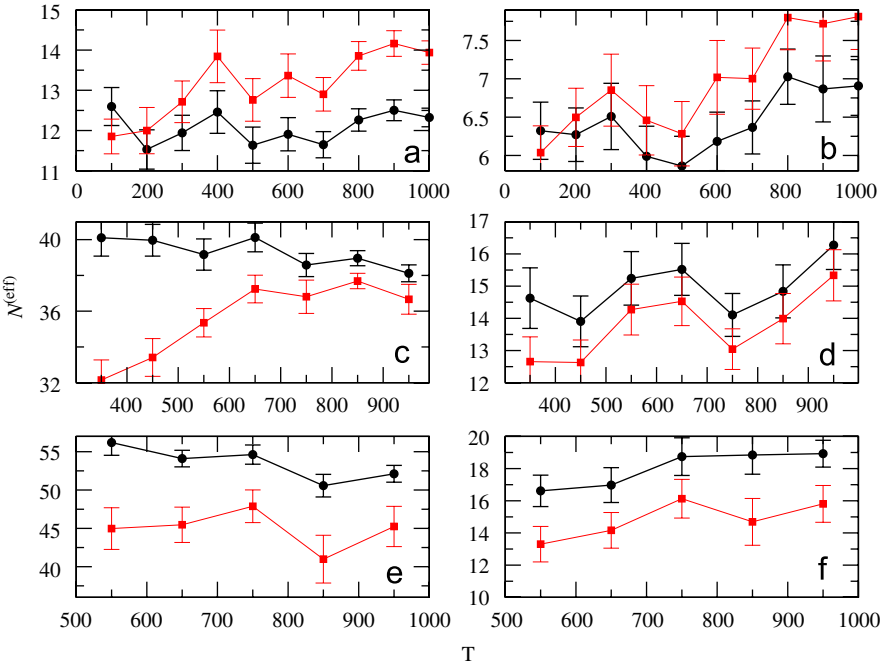

Fig. 4. Effective size $\mathcal{N}^{(\mathrm{eff})}$ of the porfolio as defined in Eq. (6) as a function of investment horizon T. The black circles refer to the RMT portfolio and the red squares to the average linkage portfolio. The portfolio size is $N = 50$ (panels (a) and (b)), $N = 300$ (panels (c) and (d)) and $N = 500$ (panels (e) and (f)). The left panels (a,c, and e) refer to the MVP and the right panels (b,d, and f) refer to a portfolio with an intermediate value of r_{p}. Every point is the average over 50 realizations obtained by bootstrapping and the error bars are standard errors.

other two realized risks. Thus the single linkage portfolio is riskier but more reliable when compared with the RMT portfolio.

Also for the single linkage method we perform a bootstrap analysis similar to the one described above for the average linkage method in order to compare the reliability of the single linkage method as compared to the RMT method. The result is the density plot shown in Fig. 6. We find that the single linkage method is able to provide more reliable portfolios in wide ranges of the parameters N and T. This is more and more evident for portfolios with $T \simeq N$, i.e. portfolios for which the investment horizon (in trading days) is comparable with the portfolio size N. It is interesting to note that for these portfolios the effect of the measurement noise ('noise dressing') is particularly high. It is worth noting that two effects might be simultaneously relevant in this respect. Specifically, the investment horizon and/or the estimation period of optimization parameters. In the investigations presented here these two parameters are assuming the same value in each optimization procedure and therefore we cannot discriminate about their specific role. A future study will consider this aspect by investigating different values of investment horizon and estimation period.

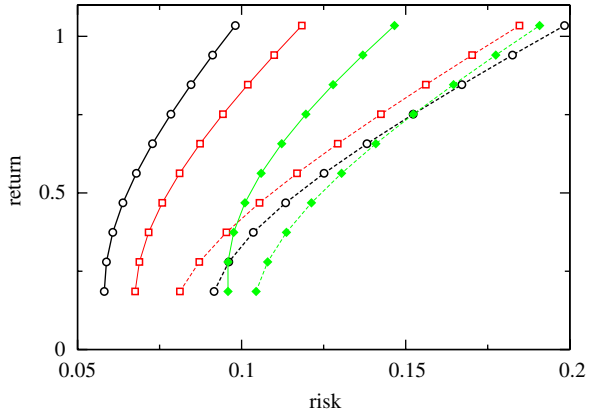

Fig. 5. The continuous lines are the predicted risk and the dashed lines are the realized risk. The filled green diamonds refer to the portfolio obtained with the single linkage method. The empty circles refer to the Markowitz portfolio optimization, whereas the squares are the predicted and realized risk curves obtained by filtering the correlation matrix with the random matrix theory approach (same data as in Fig. 1). We assume that the only uncertainty of the investor is on the correlation matrix. The dataset is composed of 150 highly capitalized stocks traded at NYSE in the period 1989–1992. The first two years are used to estimate the correlation matrix and the other two years are the investment period.

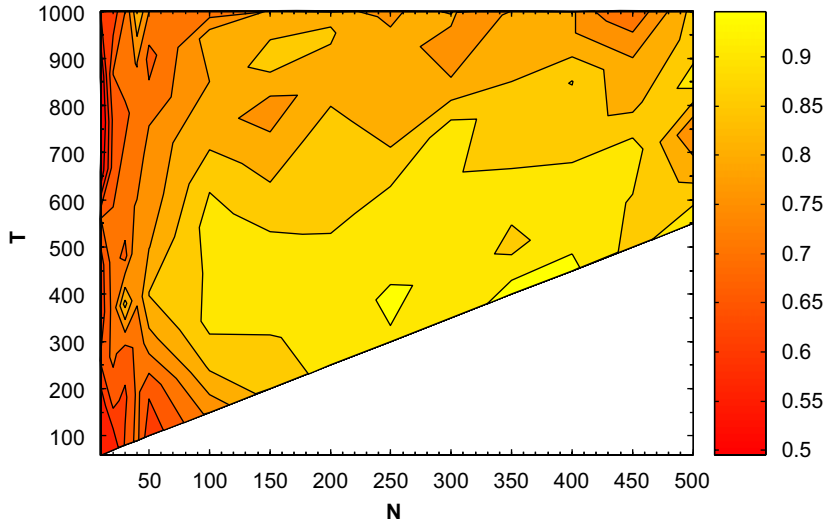

Fig. 6. Summary of a bootstrap experiment measuring the reliability of the single linkage portfolio optimization with respect to the RMT optimization method. The contour plot is showing the percentage of success of the single linkage portfolio optimization technique over the random matrix theory approach as a function of the number of assets N and of the investment period T. For each value of N and T, we perform $2 \times 10 \times 50$ different portfolio optimization procedures by randomly selecting N stocks form our database and by varying T from 100 to 1000. The white area corresponding to cases where $T < N + 50$, which are not considered in our investigation.

5.2.2. Riskiness

The analysis of the riskiness, i.e. the realized risk, of single linkage portfolios shows that these portfolios are systematically riskier than RMT portfolios. This is also seen in the example shown in Fig. 5. Only for very small sizes ($N<15$) are the single linkage portfolios less risky than the RMT portfolios. As for the average linkage portfolio this effect strongly depends on the portfolio expected return. When we consider large portfolios ($50<N<500$) we find that for small r_p only in ~0.3% of the cases the single linkage portfolio is less risky than RMT portfolio. When r_p is large this fraction is of the order of ~10%. In any case these figures indicate that single linkage portfolios are risky, even if they can be quite reliable.

5.2.3. Effective size

The effective size $\mathcal{N}^{(\mathrm{eff})}$ of the portfolio as quantified by Eq. (6) for the single linkage portfolio shows interesting properties. Fig. 7 shows the effective size for different portfolio conditions and should be compared with Fig. 4. We see that for

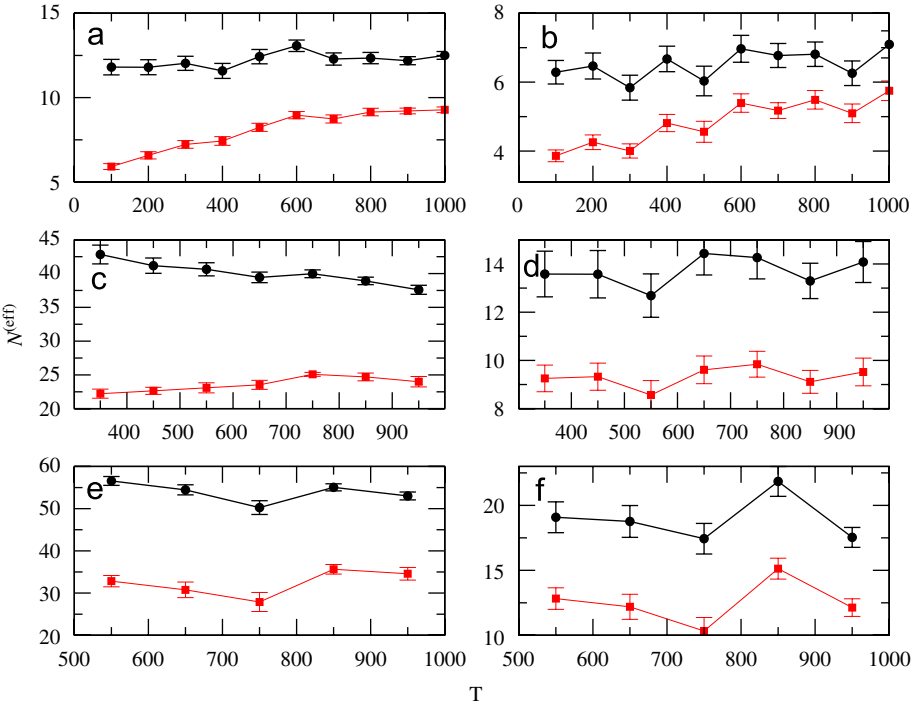

Fig. 7. Effective size $\mathcal{N}^{(\mathrm{eff})}$ of the porfolio as defined in Eq. (6) as a function of investment horizon T. The black circles refer to RMT portfolio and the red squares to single linkage portfolio. The portfolio size is $N = 50$ (panels (a) and (b)), $N = 300$ (panels (c) and (d)) and $N = 500$ (panels (e) and (f)). The left panels (a,c, and e) refer to the MVP and the right panels (b,d, and f) refer to a portfolio with an intermediate value of r_p. Every point is the average over 50 realizations obtained by bootstrapping and the error bars are standard errors.

any value of the portfolio size N the effective size of the single linkage portfolio is significantly smaller than the one of the RMT portfolio and the single linkage portfolio is, in general, more risky. The risk increase is probably related to the diminished amount of diversification unavoidably associated with the effective reduction of the portfolio size (Campbell et al., 2001). The size reduction of portfolio is observed also for small size portfolios, in contrast with what we observe for average linkage portfolios. Even more important, for large portfolios (where the size reduction is an important issue) the single linkage portfolios have an effective size which is roughly half the effective size of the RMT portfolio. This result suggests that single linkage portfolios could be used to detect a small subset of stocks which is representative of the whole portfolio, and thus to replace the original portfolio with another one of significantly smaller size with reduced maintenance costs. This possibility will be explored in a future work.

6. Portfolio optimization with clustering algorithms under more realistic conditions

In the previous section, we have investigated the use of clustering techniques for portfolio optimization under conditions which are quite unrealistic. First of all, in order to focus on the problem of cross-correlation matrix estimation, we have assumed that the investor has perfect knowledge of the future value of the return and volatility of all the stocks composing the portfolio of interest. Second, we have placed no constraints on the weights of the portfolio, except the normalization $\sum_i w_i = 1$. Thus in many portfolios a large fraction of weights are negative indicating that the investor has to perform short selling operations.

In this section we consider portfolio optimization under more realistic conditions. For the sake of simplicity, in the present study we only consider the global minimum variance portfolio, leaving the investigation of the whole efficient frontier to a future paper. Specifically, first we constrain the weights to be non-negative. In this way the investor cannot do short selling. Second, we consider that volatilities σ_i of the future are not known. The investor uses the historical data to compute σ_i (as well as the whole covariance matrix). As in the first part of the paper, the sampling period used to compute the parameters is the same as the investment period. Our results are obtained considering the portfolio of $N = 150$ highly capitalized stocks traded at NYSE in several two-year investigated periods. For the time periods 1989–1990 (for estimation) and 1991–1992 (realized portfolio), this is the same portfolio as in Figs. 1, 2, and 5. In order to check the robustness of the results we consider nine different two-year ($T = 500$) time periods. Specifically, we use 1987–1988 for estimation and 1988–1989 as the investment period, then 1988–1989 for estimation and 1990–1991 for investment, and so on. We distinguish four types of portfolio optimization. The first is the same as in the rest of the paper, i.e. we assume perfect knowledge of the return and volatility of all the assets and we allow the investor to do short selling (Case I). In the second optimization scheme (Case II) the investor use the historical data to compute mean returns and volatilities and she is allowed to do short selling. In Case III short selling is not allowed but the investor has a perfect forecast of mean

returns and volatilities. Finally, in Case IV short selling is forbidden and the investor uses historical data to compute mean returns and volatilities.

For each investigated portfolio we measure the reliability $\mathscr{R}$ of Eq. (5), the value of the realized (i.e. ex-post) risk, the ex-post Sharpe ratio and the portfolio effective size. For the computation of the Sharpe ratio we use the two-year US Treasury bonds rate for the risk free rate. Specifically we use the value of the bonds at the date when the investment is supposed to take place. The results are summarized in Tables 2–5.

Table 2 shows that also in more realistic conditions portfolio optimization techniques based on clustering algorithms are often very reliable in terms of the difference between the predicted and realized risk. In the more realistic case (Case IV) seven over nine portfolios are more reliable when clustering algorithms are used to optimize the portfolio.

Table 3 indicates that in general RMT optimization technique selects less risky portfolios under idealized conditions (Case I). Interestingly this is not true for Case III, i.e. when short selling is not allowed although the investor has still a perfect forecast of mean returns and volatilities. In this case average linkage based portfolios are typically less risky. For the most realistic conditions (Case IV), the algorithm providing the least risky portfolio is different in different time periods suggesting that the best outcome might depend on the specific market phase of mean return and volatility.

Table 4 indicates the role of a bad specification of mean returns and volatilities on the Sharpe ratio of a portfolio. In fact in Cases II and IV, where the mean returns and volatilities are estimated from the historical data, the Sharpe ratio is quite small and often Markowitz optimization gives the highest Sharpe ratio. In Cases I and III where we assume a perfect forecast of mean returns and volatilities, RMT and average linkage optimization give a larger Sharpe ratio. This is again a confirmation that the noisiness of sample means and volatilities plays an important role in

Table 2
Reliability $\mathscr{R}$ of Eq. (5) for the four types of optimization conditions (Cases I–IV) and the four optimization algorithms (MAR = Markowitz, RMT = random matrix theory, AVG = average linkage, and SIN = single linkage)

Investment period	Case I				Case II				Case III				Case IV			
	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN
1989–1990	1.4	0.71	0.67	**0.54**	0.47	0.079	0.071	**0.050**	0.11	**0.089**	0.15	0.21	0.18	0.19	**0.17**	0.23
1990–1991	0.83	0.40	**0.39**	0.44	1.1	0.56	0.40	**0.34**	0.11	0.086	**0.0069**	0.15	0.51	0.44	0.41	**0.25**
1991–1992	0.58	0.21	**0.060**	0.093	0.57	0.24	**0.0027**	0.036	**0.023**	0.035	0.10	0.21	0.0032	**0.0012**	0.069	0.19
1992–1993	0.66	0.39	**0.24**	0.33	0.33	**0.067**	0.099	0.11	0.048	**0.044**	0.14	0.17	**0.089**	0.11	0.17	0.25
1993–1994	0.42	0.18	**0.10**	0.18	0.62	0.34	**0.16**	0.18	**0.020**	0.039	0.058	0.22	0.13	0.15	**0.087**	0.13
1994–1995	0.55	0.23	**0.13**	0.15	0.62	0.31	0.12	**0.12**	0.15	0.12	**0.074**	0.20	0.22	0.19	0.14	**0.026**
1995–1996	0.61	0.30	0.25	**0.19**	0.56	0.27	0.24	**0.13**	0.22	0.20	**0.14**	0.15	0.25	0.23	0.21	**0.079**
1996–1997	0.77	0.43	0.36	**0.19**	1.5	1.2	1.1	**0.71**	0.48	0.46	0.38	**0.0021**	1.07	1.0	1.09	**0.60**
1997–1998	0.60	0.30	0.21	**0.15**	1.7	1.1	0.93	**0.65**	0.34	0.34	0.18	**0.058**	1.2	1.1	1.0	**0.73**

The description of the four cases is in the text. For each investment period and each case we indicate with boldface the most reliable algorithm.

Table 3
Realized risk (annualized in percent) for the four types of optimization conditions (Cases I–IV) and the four optimization algorithms (MAR = Markowitz, RMT = random matrix theory, AVG = average linkage, and SIN = single linkage)

Investment period	Case I				Case II				Case III				Case IV			
	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN
1989–1990	11.2	**9.9**	11.0	11.9	10.9	**9.7**	11.2	11.6	9.5	9.6	**9.5**	10.0	11.2	**10.9**	11.6	12.0
1990–1991	10.3	**9.3**	10.4	12.2	11.5	**10.3**	10.3	11.4	10.3	10.2	**9.8**	10.0	12.1	11.8	**11.7**	12.0
1991–1992	9.1	**8.1**	8.4	10.4	9.7	8.8	**8.6**	10.1	8.3	8.3	**8.1**	8.5	8.8	8.9	**8.7**	9.2
1992–1993	9.3	**8.7**	9.5	12.0	8.8	**8.2**	8.5	10.5	7.7	7.8	**7.3**	8.3	8.2	8.2	**8.0**	8.6
1993–1994	7.8	**7.5**	7.9	9.9	9.2	8.8	**8.8**	11.4	7.5	**7.4**	7.6	7.8	**8.7**	9.0	8.9	11.0
1994–1995	7.2	**6.6**	7.0	9.0	8.2	**7.8**	8.0	10.5	6.9	6.9	**6.8**	7.0	**7.8**	7.8	7.8	9.4
1995–1996	7.7	**7.2**	7.9	10.2	8.37	**7.7**	8.8	10.0	7.7	**7.6**	7.7	7.7	8.4	**8.3**	8.4	8.6
1996–1997	10.1	**9.6**	10.3	11.3	12.3	**11.8**	13.5	14.4	11.0	10.8	11.1	**10.8**	12.7	**12.7**	13.5	13.8
1997–1998	11.0	**10.4**	11.0	13.3	14.5	**12.6**	14.1	15.1	12.1	12.0	**11.6**	11.7	14.9	**14.2**	14.9	16.0

The description of the four cases is in the text. For each investment period and each case we indicate with boldface the least risky algorithm.

Table 4
Sharpe ratio for the four types of optimization conditions (Cases I–IV) and the four optimization algorithms (MAR = Markowitz, RMT = random matrix theory, AVG = average linkage, and SIN = single linkage)

Investment period	Case I				Case II				Case III				Case IV			
	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN
1989–1990	1.7	**1.9**	1.8	1.6	**1.2**	0.64	−0.0079	0.088	0.40	0.40	**0.41**	0.38	−0.20	−0.25	**−0.12**	−0.26
1990–1991	0.58	**0.64**	0.58	0.491	**0.24**	0.13	−0.38	−0.70	0.48	0.48	**0.51**	0.50	**0.55**	0.53	0.48	0.39
1991–1992	1.2	**1.4**	1.4	1.1	0.24	**0.60**	−0.53	−1.0	1.6	1.6	**1.7**	1.6	**1.4**	1.3	1.1	0.79
1992–1993	1.2	**1.3**	1.2	0.93	**0.73**	0.48	−0.18	−0.59	1.3	1.3	**1.4**	1.3	1.6	1.5	**1.7**	1.5
1993–1994	0.065	**0.068**	0.065	0.051	−0.065	**0.024**	−0.779	−1.140	0.26	**0.26**	0.26	0.25	**0.27**	0.16	0.22	−0.040
1994–1995	2.1	**2.2**	2.1	1.657	**1.4**	1.2	0.68	−0.073	1.8	1.8	**1.9**	1.8	**1.4**	1.4	1.0	0.5
1995–1996	2.3	**2.4**	2.2	1.7	2.4	**2.6**	2.2	1.7	2.5	**2.6**	2.5	2.5	2.6	**2.7**	2.5	2.3
1996–1997	1.3	**1.4**	1.3	1.2	**1.8**	1.7	1.4	1.1	1.5	**1.6**	1.5	1.5	**1.9**	1.9	1.7	1.7
1997–1998	1.2	**1.3**	1.2	1.022	1.1	**1.2**	1.0	0.71	1.0	1.0	**1.1**	1.1	1.1	**1.2**	1.1	1.2

The description of the four cases is in the text. For each investment period and each case we indicate with boldface the algorithm with the highest Sharpe ratio.

decreasing the effectiveness of portfolio optimization. Investors typically use adjusted estimates of mean returns and volatilities for example by computing weighted averages (J.P. Morgan 1996) and/or by incorporating economic information in the estimates. We interpret the results of Cases I and III as idealized targets which are approachable by improving mean return and volatility forecasting ability. In these idealized limits the RMT and the average linkage clustering portfolio optimization procedures are highly effective. In more realistic conditions (Case IV), Sharpe ratios of all methods are quite close the one with the other. There is no

Table 5
Effective size $\mathcal{N}^{(\mathrm{eff})}$ of the porfolio as defined in Eq. (6) for the four types of optimization conditions (Cases I–IV) and the four optimization algorithms (MAR = Markowitz, RMT = random matrix theory, AVG = average linkage, and SIN = single linkage)

Investment period	Case I				Case II				Case III				Case IV			
	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN	MAR	RMT	AVG	SIN
1989–1990	**4.3**	6.2	5.9	5.2	**4.4**	6.0	6.4	5.3	9.7	9.7	8.4	**5.2**	6.8	6.8	5.6	**4.0**
1990–1991	**5.8**	10.0	9.2	7.3	**5.4**	7.8	8.8	5.7	12.1	12.4	12.8	**9.1**	13.1	13.4	11.4	**8.5**
1991–1992	**5.8**	8.5	9.6	6.7	**5.3**	7.1	9.0	6.4	9.5	9.8	11.5	**6.6**	10.9	9.6	11.4	**6.9**
1992–1993	**5.8**	8.3	9.8	6.6	**5.8**	8.4	10.2	7.3	11.6	10.9	13.2	**7.7**	10.3	10.4	12.8	**7.0**
1993–1994	**7.5**	11.8	12.6	8.5	**6.8**	10.7	12.2	8.5	19.6	19.4	19.9	**12.4**	13.4	13.1	13.7	**7.3**
1994–1995	8.7	15.2	14.6	**8.6**	**8.2**	14.5	19.4	10.6	21.9	24.7	25.1	**14.1**	20.1	21.2	25.0	**12.1**
1995–1996	**9.4**	15.0	16.0	9.9	10.3	12.0	16.4	**10.1**	23.4	24.2	24.3	**15.0**	27.0	27.7	26.3	**15.7**
1996–1997	9.7	14.1	18.0	**9.5**	10.2	14.7	15.3	**9.0**	24.0	22.7	26.4	**13.2**	24.1	26.5	25.3	**14.9**
1997–1998	10.1	12.0	13.3	**8.3**	**7.8**	12.8	12.9	8.4	18.3	18.9	18.1	**10.0**	18.6	19.9	19.7	**10.7**

The description of the four cases is in the text. For each investment period and each case we indicate with boldface the algorithm with the smallest effective size.

method systematically outperforming the others and no simple rule is deducted about the reason of success of a specific method indicating that the success of a given method is probably related to the specific market phase characterizing the financial market. However, different optimal portfolios might be quite different with respect to stock composition. In fact, Table 5 shows that the effective size of the global minimum variance portfolio can be significantly different in Case IV when one uses the single linkage clustering portfolio optimization. Specifically, in Case IV for the single linkage procedure the values of $\mathcal{N}^{(\mathrm{eff})}$ are approximately half of the values of all other optimization methods whereas the Sharpe ratios show just a small percent change. A similar pattern is also observed in Case III clearly indicating a role of the no short selling constraint.

In summary, the investigation of hierarchical clustering portfolio optimization under more realistic conditions shows that these methods might contribute in selecting portfolios characterized by high reliability, level of realized risk and Sharpe ratios comparable to other methods and significantly reduced effective size. The selection of the preferred clustering or filtering method depending on the importance the investor is associating to the properties of reliability, realized risk, Sharpe ratio and effective size of the portfolio of interest.

7. Conclusions

In this paper we have performed portfolio optimization by using filtered correlation coefficient matrices. These matrices have been obtained by applying different filtering methods to the original correlation coefficient matrix. We have proposed two filtering methods based on the average linkage and single linkage

clustering procedures. The optimal portfolios obtained with these two new methods have been compared with the one based on RMT recently proposed in Laloux et al. (2000) and Rosenow et al. (2002) both under idealized conditions and under more realistic conditions.

In the idealized case, a large set of simulations have shown that clustering methods are outperforming RMT filtering when we consider the reliability of the estimation of the realized portfolio with respect to the predicted one for portfolios with a number of assets $\approx 50 < N < \approx 500$. Hence, for relatively large portfolios the clustering filtering methods provide a more reliable estimation of the predicted risk-return profile both with respect to the Markowitz basic estimation and with respect to the determination of the correlation coefficient done with the RMT filtering

The portfolios obtained with the average linkage show a predicted and realized risk-return profile which are characterized by high values of reliability and are often located within the corresponding profiles obtained both with the Markowitz basic estimation and after the RMT filtering. In the case of the single linkage clustering method the risk-return profile shows risk levels which are systematically higher than the ones obtained both with the Markowitz basic estimation and after the RMT filtering. Therefore with respect to the aspect of the level of risk associated with the selected portfolios the most successful methods are the average linkage and the RMT filtering.

Another aspect investigated in our study refers to the composition of the portfolios selected. We have quantified the degree of homogeneity of the distribution of the wealth across the stocks of the portfolio through the 'effective size' of the portfolio. A small number of this parameter indicates an uneven distribution of the portfolio wealth suggesting that during portfolio re-balancing only a subset of stocks will be significantly involved. The investigation of the 'effective size' of the portfolio has shown that the average linkage and the RMT are characterized by not too different values of the 'effective size'. In fact for small portfolios (e.g. $N = 50$) the RMT has for most values of T a smaller value of the 'effective size' whereas the pattern is reversed for medium ($N = 300$) and large portfolios ($N = 500$) both for the minimum and intermediate value of r_p. The pattern is clearly different for the case of the single linkage filtering. In this case the 'effective size' is always significantly less than the one observed in the cases of RMT filtering.

The above discussion of the idealized case shows that the different filtering procedures provide different portfolio optimization results that are characterized by specific strengths or weaknesses. In other words, for each value of N and T, the most useful filtered correlation coefficient matrix can be different depending on the strongest constraint the investor has among the risk level of the portfolio, the reliability of the estimation and the portfolio 'effective size'. We believe that the two clustering methods we have proposed here and the RMT are not exhaustive with respect to all potential aspects of portfolio optimization and, probably, other filtering methods could also provide very interesting results in specific regimes of the different control parameters.

These findings obtained in the idealized case have been compared with the results obtained for the global minimum variance portfolio selected under more realistic

conditions. We have verified that the reliability results of clustering portfolio optimization methods are largely still present under the realistic assumptions of mean return and volatility forecasts and no short selling. Results about portfolio realized risk and Sharpe ratio indicate that the average linkage clustering portfolio optimization method often outperforms Markowitz or RMT methods when one assumes perfect forecasting of returns and volatilities but no short selling conditions (Case III of Tables 3 and 4). Always considering the realized risk and Sharpe ratio, in the absence of perfect forecasting ability and no short selling (Case IV of Tables 3 and 4), we do not observe any method systematically outperforming the other ones. The last observation concerns the effective size of portfolios. It is worth noting that, under realistic conditions (Case IV of Table 5), different portfolios characterized by similar values of the Sharpe ratio can present a much lower value of $\mathcal{N}^{(\mathrm{eff})}$ when one uses the single linkage clustering portfolio optimization procedure. This result can be achieved at the cost of accepting a slightly higher level of the realized risk.

The different results of the different filtering methods raise the scientific question of which is the reason for the difference between the various filtering procedures. A precise quantification of the information retained by the different filtered matrices would be very useful. This goal is left for future research.

Acknowledgments

Authors acknowledge support from the research project MIUR 449/97 'High frequency dynamics in financial markets'. F.L. and R.N.M. acknowledge support also from the research project MIUR-FIRB RBNE01CW3M 'Cellular self-organizing nets and chaotic nonlinear dynamics to model and control complex system' and from the European Union STREP project n. 012911 'Human behavior through dynamics of complex social networks: an interdisciplinary approach.'

References

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.
Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., Pascazio, S., 2005. Physica A 345, 196.
Bernaschi, M., Grilli, L., Vergni, D., 2002. Physica A 308, 381.
Bonanno, G., Vandewalle, N., Mantegna, R.N., 2000. Physical Review E 62, R7615.
Bonanno, G., Lillo, F., Mantegna, R.N., 2001. Quantitative Finance 1, 96.
Bonanno, G., Caldarelli, G., Lillo, F., Mantegna, R.N., 2003. Physical Review E 68, 046130.
Bonanno, G., Caldarelli, G., Lillo, F., Micciche, S., Vandewalle, N., Mantegna, R.N., 2004. European Physical Journal B 38, 363.
Bouchaud, J.-P., Potters, M., 2003. Theory of Financial Risk and Derivative Pricing, second ed. Cambridge University Press, Cambridge, New York.
Burda, Z., Jurkiewicz, J., 2004. Physica A 344, 67.
Campbell, J.Y., Lettau, M., Malkiel, B.G., Xu, Y.X., 2001. Journal of Finance 56, 1.
Di Matteo, T., Aste, T., Mantegna, R.N., 2004. Physica A 339, 181.
Drozdz, S., Kwapien, J., Grummer, F., Ruf, F., Speth, J., 2001. Physica A 299, 144.
Eichhorn, D., Gupta, F., Stubbs, E., 1998. Journal of Portfolio Management 24, 41.
Elton, E.J., Gruber, M.J., 1995. Modern Portfolio Theory and Investment Analysis. Wiley, New York.

Galluccio, S., Bouchaud, J.-P., Potters, M., 1998. Physica A 259, 449.
Giada, L., Marsili, M., 2001. Physical Review E 63, 061101.
Gopikrishnan, P., Rosenow, B., Plerou, V., Stanley, H.E., 2001. Physical Review E 64, 035106.
Gower, J.C., 1966. Biometrika 53, 325.
Gower, J.C., Ross, G.J.S., 1969. Applied Statistics 18, 54.
Guhr, T., Kalber, B., 2003. Journal of Physics A 36, 3009.
Jagannathan, R., Ma, T.S., 2003. Journal of Finance 58, 1651.
Jorion, P., 1985. Journal of Business 58, 259.
J.P. Morgan 1996. RiskMetrics Technical Document, New York.
Kullmann, L., Kertesz, J., Mantegna, R.N., 2000. Physica A 287, 412.
Kullmann, L., Kertesz, J., Kaski, K., 2002. Physical Review E 66, 026125.
Laloux, L., Cizeau, P., Bouchaud, J.-P., Potters, M., 1999. Physical Review Letters 83, 1468.
Laloux, L., Cizeau, P., Potters, M., Bouchaud, J.-P., 2000. International Journal of Theoretical Applied Finance 3, 391.
Ledoit, O., Wolf, M., 2004a. Journal of Portfolio Management 30, 110.
Ledoit, O., Wolf, M., 2004b. Journal of Multivariate Analysis 88, 365.
Malevergne, Y., Sornette, D., 2004. Physica A 331, 660.
Mantegna, R.N., 1999. European Physical Journal B 11, 193.
Mantegna, R.N., Stanley, H.E., 2000. An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge, pp. 120–128.
Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, San Diego.
Markowitz, H., 1959. Portfolio Selection: Efficient Diversification of Investment. Wiley, New York.
Maskawa, J., 2003. Physica A 324, 317.
Maslov, S., 2001. Physica A 301, 397.
Mendes, R.V., Araujo, T., Louca, F., 2003. Physica A 323, 635.
Metha, M.L., 1990. Random Matrices. Academic Press, New York.
Micciche, S., Bonanno, G., Lillo, F., Mantegna, R.N., 2003. Physica A 324, 66.
Onnela, J.P., Chakraborti, A., Kaski, K., Kertesz, J., 2002. European Physical Journal B 30, 285.
Pafka, S., Kondor, I., 2003. Physica A 319, 487.
Pafka, S., Kondor, I., 2004. Physica A 343, 623.
Papadimitriou, C.H., Steiglitz, K., 1982. Combinatorial Optimization. Prentice-Hall, Englewood Cliffs.
Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E., 1999. Physical Review Letters 83, 1471.
Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., Stanley, H.E., 2002. Physical Review E 65, 066126.
Rammal, R., Toulouse, G., Virasoro, M.A., 1986. Reviews of Modern Physics 58, 765.
Rosenow, B., Plerou, V., Gopikrishnan, P., Stanley, H.E., 2002. Europhysics Letters 59, 500.
Rosenow, B., Gopikrishnan, P., Plerou, V., Stanley, H.E., 2003. Physica A 324, 241.
Sharifi, S., Crane, M., Shamaie, A., Ruskin, H., 2004. Physica A 335, 629.
Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R.N., 2005. Proceedings of the National Academy Science of the United States of America 102, 10421.
Tumminello, M., Lillo, F., Mantegna, R.N. 2007. Hierarchically nested time series models from dendrograms, Europhysics Letters, in press.